Publishing in the networked world: transforming the nature of communication

14th International Conference on Electronic Publishing 16 - 18 June 2010, Helsinki, Finland

http://www.elpub.net

Edited by

Turid Hedlund
Hanken School of Economics, Helsinki, Finland

Yasar Tonta
Hacettepe University, Ankara, Turkey

Helsinki 2010
Hanken School of Economics

Publishing in the networked world: transforming the nature of communication
14th International Conference on Electronic Publishing 16-18 June 2010

Keywords: Electronic publishing, scholarly communication, social networks

Disclaimer
Any views or opinions expressed in any of the papers in this collection are those of their respective authors. They do not represent the view or opinion of the Hanken School of Economics, the Hacettepe University, the editors and members of the Programme Committee, nor of the publisher or conference sponsors.

Distributor:

Library
Hanken School of Economics
P.O.Box 479
00101 Helsinki, Finland

Telephone: +358-40-3521376, +358-40-3521265
Fax: +358-40-3521425
E-mail:publ@hanken.fi
http://www.hanken.fi

ISBN: 978-952-232-086-5

Edita Prima Ltd, Helsinki 2010

Members of the 2010 Programme Committee

Baptista, Ana Alice *University of Minho* (Portugal)

Björk Bo-Christer *Hanken School of Economics* (Finland)

Borbinha, José *INESC-ID / IST – Lisbon Technical University* (Portugal)

Chan, Leslie *University of Toronto Scarborough* (Canada)

Costa, Sely M.S. *University of Brasilia* (Brazil)

Delgado, Jaime *Universitat Politècnica de Catalunya* (Spain)

Dobreva, Milena *University of Strathclyde* (Scotland) *&IMI-BAS* (Bulgaria)

Engelen, Jan *Katholieke Universiteit Leuven* (Belgium)

Gargiulo, Paola *CASPUR* (Italy)

Gradmann, Stefan *University of Hamburg* (Germany)

Güntner, Georg *Salzburg Research* (Austria)

Halttunen Kai *University of Tampere* (Finland)

Hedlund, Turid *Hanken School of Economics* (Finland)

Hindersson-Söderholm Tua *Hanken School of Economics* (Finland)

Horstmann, Wolfram *University of Bielefeld* (Germany)

Iyengar, Arun *IBM Research* (USA)

Jezek, Karel *University of West Bohemia in Pilsen* (Czech Republic)

Kurbanoglu Serap *Hacettepe University* (Turkey)

Linde, Peter *Blekinge Institute of Technology* (Sweden)

Lioma,Christina *Konstanz University*, (Germany)

Mac An Airchinnigh Micheal *Trinity College Dublin* (Ireland)

Martens, Bob *Vienna University of Technology* (Austria)

Mendéz, Eva *Universidad Carlos III, Madrid* (Spain)

Mornati, Susanna *CILEA* (Italy)

Morrison, Heather *British Columbia Electronic Library Network* (Canada)

Nisheva-Pavlova, Maria *Sofia University* (Bulgaria)

Opas-Hänninen, Lisa Lena *University of Oulu* (Finland)

Roos Annikki *University of Helsinki,* (Finland)

Smith, John *University of Kent at Canterbury* (UK)

Tonta, Yasar *Hacettepe University* (Turkey)

Preface

The title of the 14th International Conference on Electronic Publishing (ELPUB), “Publishing in the networked world: Transforming the nature of communication”, is a timely one. Scholarly communication and scientific publishing has recently been undergoing subtle changes. Published papers are no longer fixed physical objects, as they once were. The “convergence” of information, communication, publishing and web technologies along with the emergence of Web 2.0 and social networks has completely transformed scholarly communication and scientific papers turned to living and changing entities in the online world. The themes (electronic publishing and social networks; scholarly publishing models; and technological convergence) selected for the conference are meant to address the issues involved in this transformation process. We are pleased to present the proceedings book with more than 30 papers and short communications addressing these issues.

What you hold in your hands is a by-product and the culmination of almost a Year long work of many people including conference organizers, authors, reviewers, editors and print and online publishers. The ELPUB 2010 conference was organized and hosted by the Hanken School of Economics in Helsinki, Finland. Professors Turid Hedlund of Hanken School of Economics and Yaşar Tonta of Hacettepe University Department of Information Management (Ankara, Turkey) served as General Chair and Program Chair, respectively. We received more than 50 submissions from several countries. All submissions were peer-reviewed by members of an international Program Committee whose contributions proved most valuable and appreciated.

The 14th ELPUB conference carries on the tradition of previous conferences held in the United Kingdom (1997 and 2001), Hungary (1998), Sweden (1999), Russia (2000), the Czech Republic (2002), Portugal (2003), Brazil (2004), Belgium (2005), Bulgaria (2006), Austria (2007), Canada (2008) and Italy (2009). The ELPUB Digital Library, http://elpub.scix.net serves as archive for the papers presented at the ELPUB conferences through the years. The 15th ELPUB conference will be organized by the Department of Information Management of Hacettepe University and will take place in Ankara, Turkey, from 14-16 June 2011. (Details can be found at the ELPUB web site as the conference date nears by.)

We thank Marcus Sandberg and Hannu Sääskilahti for copyediting, Library Director Tua Hindersson – Söderholm for accepting to publish the online as well as the print version of the proceedings. Thanks also to Patrik Welling for maintaining the conference web site and Tanja Dahlgren for administrative support. We warmly acknowledge the support in organizing the conference to colleagues at Hanken School of Economics and our sponsors.

Turid Hedlund
General Chair

Yaşar Tonta
Program Chair

Contents

Sessions: Thursday 17.6.2010

Sessions: Friday 18.6.2010

Short papers

BUSINESS MODELS FOR ELECTRONIC OPEN ACCESS JOURNALS AND DISCIPLINARY DIFFERENCES: A PROPOSAL

Katiúcia Araujo Gumieiro[1]; Sely Maria de Souza Costa[2]
[1] Deputies Chamber
Brazil
e-mail: kathygumieiro@gmail.com;
[2] University of Brasilia
Brazil
e-mail: selmar@unb.br

Abstract

Reports results of a research that aimed at studying the use of business models in the context of open access electronic scholarly journals publishing. Additionally, the work approaches disciplinary differences, particularly in terms of three issues, namely required publication speed, funding and features that involve the edition of a scholarly journal. In this context, the study aimed at proposing a model that allows identifying required elements to design business models appropriated to open access scholarly journals publishing. Along with identifying the elements, the study looked at the relationships between these elements and differences found between knowledge fields. Based on a bibliographic survey, the research adopted a qualitative approach that consisted of analysing the content of the literature reviewed. As a result, a business model for the activity of open access electronic journal publishing has been proposed. Based on Stähler's approach, the model entails a set of four components, namely value proposition, products and/or services, value architeture and source of resources. Derived from this basic model, three other models are presented, each one representing particularities of the three major divisions of knowledge, Sciences, Social & Human Sciences and Arts & Humanities. As conclusion, features of business models for Sciences are considerably different from the other two divisions. On the other hand, there are important similarities between business models for the Social & Human Sciences and for Arts & Humanities.

Keywords: Business models; Open access to scientific information; Scholarly communication; Disciplinary differences.

1. Introduction

Science advancement occurs when knowledge is shared amongst members of the scientific world. Researchers discussions both promote and improve science constructs, although barriers are constantly found within the scholarly communication system. High prices of scholarly journals subscription, for instance, have made access to science findings unfeasible. Moreover, there is a high preoccupation amongst scholarly journal publishers regarding the protection of their rights.

Due to this fact, the movement of open access to scientific information is brought to light as a major initiative in favour of the wide and unrestricted dissemination of research results in electronic media. Both the green road (institutional repositories) and the gold road (open access journals) have become the two main ways of providing open access to scientific information. The present study focus on the later, taking into account that it consists of a feasible alternative to the traditional scholarly journal publication model.

It seems natural to ask how to maintain the publication of an open access scholarly journal without having resources from subscription or access charges. The answer comes from the use of business models in a creative way, as they constitute a method through which each publisher can build and use its own resources in order to offer a better value than its competitors and, then, achieve a long-term sustainability [1]. Such method allows an entrepreneur to better understand his/her own business when outlining it in a simplified way. From the resulting models, it is feasible to organise businesses, besides increasing value appropriateness to a given business.

Taking account of the present time, in which economic environment is highly uncertain, competitive and changing, business decisions become difficult and complex. In this sense, the use of such models is strategic to any kind of organisation, including open access scholarly journal publishers. This is because using these models facilitates analysing, understanding and explaining empirical relationships found in this kind of businesses [2].

Van Der Beek et al. [3] emphasise that studies about business models can be grouped in two categories. The first one describes specific business models. They consist of model taxonomies in which business models pertaining to the same category share common features such as price policies and clients relationship. The second one comprises studies that define and analyse business models components. Within this later, Linder & Cantrell [4] explain that business models components are simply bits of a model, each of them representing a specific feature of a business. The present work adopted this later approach and it is justified by Mahadevan [5], who reports that studying only the models without looking at their components leads to focusing on very specific features of how a sector makes business.

It is important to notice that apparently, there is no consensus on which components should comprise a business model. Hence, this research objective is, from the perspective of open access electronic scholarly journal publishing, identify a set of components that better correspond to such reality.

In the elaboration of a business model it is fundamental for a journal publisher to consider, before any other thing, particularities concerning the knowledge field with which his/her journal is concerned. It is even more important when these particularities involve disciplinary communication patterns. Meadows [6] explains that the nature and features of each filed of knowledge lead to the adoption of different ways of carrying out research. Consequently, the way of communicating results is different, too. Therefore, publishers as intermediates in the scholarly communication process need to focus on these patterns in order to produce and offer outputs that better attend the needs of their clients. Because of being fairly recent as compared to the existence of scholarly journals as a whole, the suitability of business models for open access journals from different fields of knowledge becomes a relevant factor to the success of these journals.

2. Research methodology

The purpose of this study is both exploratory and descriptive. Exploratory, because in the literature reviewed no studies were found having the same focus of this research, that is, to study the main components of business models not limiting to that concerned with profits. Descriptive, to the extent that there are, already, data respecting disciplinary differences in the literature pertaining to this topic.

Additionally, the study adopted a methodology essentially qualitative, building itself on the interpretation of the literature. It is important to notice that the present research makes use, during the analysis, of the inductive reasoning, assuming that the model generated has the potential to reflect itself on a broader reality. Conjointly, it availed itself of another kind of reasoning: the deductive. By studying business models in the electronic environment, the researchers inferred deductively that this knowledge is applicable to the activity of publishing scientific periodicals of open access, since it is produced in the electronic environment.

Bibliographic research was the technical procedure of choice. In analysing the texts, two approaches were used. The first one is the codification and categorization method, proposed by. Kvale & Brinkman [7], who explain that this method attributes to one or more keywords the capability of identifying a communication appearing subsequently. The other method used was that of interpretation, whose key feature is to allow the interpreter to move beyond what is actually said, bringing out structures and relationships not apparent in the text.

3. Discussion

Based on the literature analysis, the present study discusses the use of business models in the context of open access scholarly journals. The study sought for knowledge on the business models theme in order to apply it to the scientific publication activity. Therefore, business models components that are feasible to open access electronic scholarly journal publishing have been looked at.

After a careful analysis of the literature, it has been decided to adopt Stähler's [8] approach, because it allows the analysis of key aspects involving journal publication. The author describe four components of a business model:

- Value proposition. It is concerned with the offer of differential values for users, in view of the intense market competitiveness. Within the context of journal publishing, these values can be offered to business clients (readers, libraries), internal partners (reviewers, authors) and external partners (sponsors, publicity teams.
- Services and/or products. It consists of the description of services and products offered, taking careful account of their feasibility to user needs. In the present research, it was necessary to characterise journals in relation to writing style, presentation (text proportion, graphs, figures and tables), average number of pages per article, periodicity, minimum number of articles per year and average number of refused submissions.
- Value architeture. This component is strongly associated with intrinsic aspects of a specific enterprise, as it is the description of how it is organised in order to offer values to its clients and partners. The present research took into account specific aspects of a publisher in terms of market design (target audience), as well as internal and external architeture.
- Source of resources. It describes the way a business obtain resources needed to is sustainability. These resources can come from three sources. The first concerns additional services (in the context of this research they can consist of selling print copies, convenient forms of licenses, specific charges for different types of distribution and so on). The second is related to external partners (sponsorship, publicity, expositions and conference co-work). Finally, there are contributions and funds from foundations,

institutional subsidies, government agencies, voluntary contributions and so on [9].

These components are hereafter adopted in the proposition of business models for open access scholarly journals publishing. The first model is generic and from this three more models have been proposed for the three major divisions of knowledge.

3.1 Generic business model for open access scholarly journals publishing

The relationship between these four components allowed the proposition of a generic business model (Fig. 1) for open access scholarly journals. This model shows how sources of revenue serve as input to the component 'value architeture', which, in turn, drive other characteristics of the editorial business, making it cyclical.

As can be observed, value architeture better organises the publisher business, helping him/her to offer the correspondent value proposition to its clients and partners. Clients are then attracted to have the journal, bringing about a greater demand, which, in turn, calls the attention of sponsors and advertisers, who financially invest in the business. The same happens to authors and reviewers as partners. When a publisher offers services that correspond to their yearnings, there is a tendency of getting a greater offer of their work, as well as an increase of better offerers' work. This, in turn, attracts sponsors and advertisers.

In the context of disciplinary differences, particularities of the three major divisions of knowledge have been associated to each component of the generic model. Such association has allowed the proposition of three additional, specific models. The model for the Sciences (Fig 2) shows a distinct configuration from those for Social & Human Sciences (Fig. 3) and Arts & Humanities (Fig. 4). An additional observation is the inference that the Sciences business model should attract a greater number of clients and partners than the other two divisions, because their authors make more use of journals than those from the others.

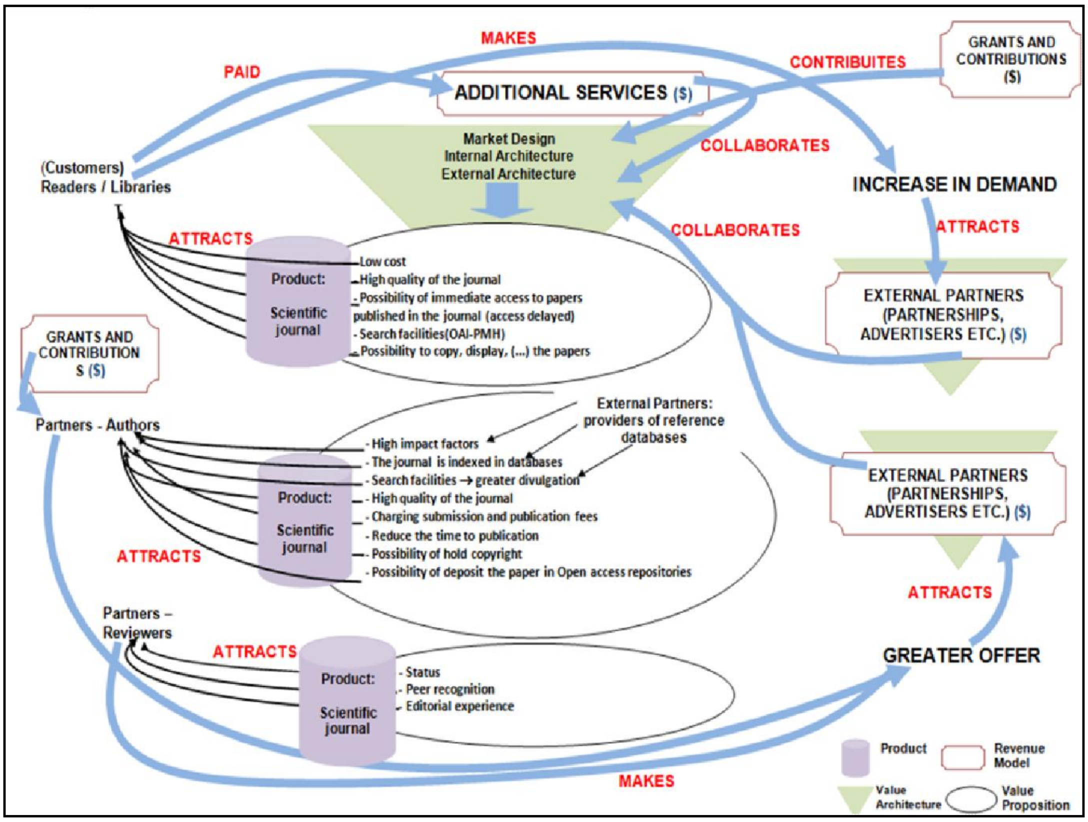

MAKES
PAID
CONTRIBUITES
GRANTS AND CONTRIBUTIONS ($)
ADDITIONAL SERVICES ($)
Market Design
Internal Architecture
External Architecture
COLLABORATES
(Customers) Readers / Libraries
INCREASE IN DEMAND
ATTRACTS
COLLABORATES
ATTRACTS
Product: Scientific journal
Low cost
High quality of the journal
Possibility of immediate access to papers published in the journal (access delayed)
Search facilities(OAI-PMH)
Possibility to copy, display, (...) the papers
EXTERNAL PARTNERS (PARTNERSHIPS, ADVERTISERS ETC.) ($)
GRANTS AND CONTRIBUTION S ($)
Partners - Authors
External Partners: providers of reference databases
- High impact factors
- The journal is indexed in databases
- Search facilities → greater divulgation
- High quality of the journal
- Charging submission and publication fees
- Reduce the time to publication
- Possibility of hold copyright
- Possibility of deposit the paper in Open access repositories
Product: Scientific journal
ATTRACTS
EXTERNAL PARTNERS (PARTNERSHIPS, ADVERTISERS ETC.) ($)
ATTRACTS
Partners – Reviewers
ATTRACTS
Product: Scientific journal
- Status
- Peer recognition
- Editorial experience
GREATER OFFER
Product
Revenue Model
MAKES
Value Architecture
Value Proposition

Figure 1 – Business model for the activity of open access electronic scholarly journals publishing

3.2 Business model for open access scholarly journals in the Sciences

Each particularity of the Sciences, as compared to the other two divisions of knowledge (Social & Human Sciences and Arts & Humanities) is reflected on components of the business model, as shown below and depicted in figure 2.

Value proposition	– Immediate access to readers is more applicable to Sciences than to the other two divisions. Publication speed is higher [10] and citations achieve the top faster [11]. – Shorter time between submission and publication because of its dynamic aspect, making time an important value. – The possibility authors have to deposit a preprint correspond to the needs of researchers from the Sciences [12]. There is actually a tendency of researchers from this division to use less formal methods of disseminating their results [13].
	– Authors from the Sciences write shorter sentences, therefore, easier to be read [14]. – Literature review found mostly as footnotes [15].

Products and/or services	– Articles with more figures and equations [16], which may lead to higher editorial costs. – Average number of pages is lower [16]. – Higher amount of articles [13], perhaps justifying more options of titles available to publish in. – Higher proportion of articles co-authored [10]. – Lower refusal rates [10].
Sources of resources	– Research in the Sciences requires greater support, making contributions and funding higher [10]; – Because of that, the "author pays" model is more attractive, leading to a likely greater impact factor.

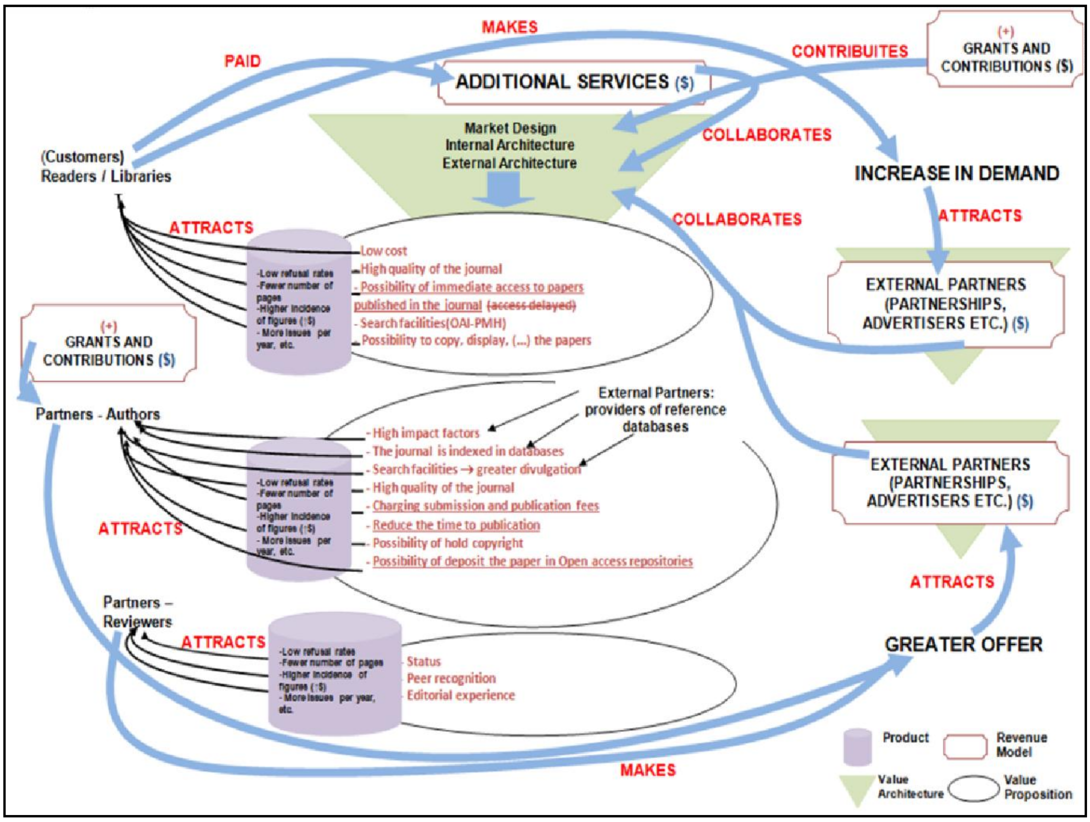

MAKES
PAID
CONTRIBUITES
(+)
GRANTS AND CONTRIBUTIONS ($)
ADDITIONAL SERVICES ($)
Market Design
Internal Architecture
External Architecture
COLLABORATES
(Customers)
Readers / Libraries
INCREASE IN DEMAND
ATTRACTS
COLLABORATES
ATTRACTS
-Low refusal rates
-Fewer number of pages
-Higher incidence of figures (↑$)
- More issues per year, etc.
Low cost
High quality of the journal
- Possibility of immediate access to papers published in the journal ~~(access delayed)~~
- Search facilities(OAI-PMH)
Possibility to copy, display, (...) the papers
EXTERNAL PARTNERS (PARTNERSHIPS, ADVERTISERS ETC.) ($)
(+)
GRANTS AND CONTRIBUTIONS ($)
Partners - Authors
External Partners: providers of reference databases
- High impact factors
- The journal is indexed in databases
- Search facilities → greater divulgation
- High quality of the journal
- Charging submission and publication fees
- Reduce the time to publication
- Possibility of hold copyright
- Possibility of deposit the paper in Open access repositories
ATTRACTS
-Low refusal rates
-Fewer number of pages
-Higher incidence of figures (↑$)
- More issues per year, etc.
EXTERNAL PARTNERS (PARTNERSHIPS, ADVERTISERS ETC.) ($)
ATTRACTS
Partners – Reviewers
ATTRACTS
-Low refusal rates
-Fewer number of pages
-Higher incidence of figures (↑$)
- More issues per year, etc.
Status
Peer recognition
Editorial experience
GREATER OFFER
Product
Revenue Model
Value Architecture
Value Proposition
MAKES

Figure 2 – Business model for the activity of open access electronic scholarly journals publishing in the Sciences

3.3 Business model for open access scholarly journals in the Social & Human Sciences

With reference to Social & Human Sciences, because this division encompasses a variety of disciplines, there are also a variety of communication patterns, ranging from the Humanities to the Sciences. So, grouping them in a unique set is a limitation of this study. However, according to what has been found in the literature, it was possible to obtain a list of interesting particularities for the proposition of a business model, as shown below and in figure 3.

Value proposition	– Publishing slowness [10] makes the possibility of immediate access to results non-attractive. However, an exception is found concerning disciplines with communication patterns close to the Sciences. – A smaller period of time between submission and publication is not an attractive issue, because of the slowness cited above [10]. For the same reason, the delayed open access model becomes attractive. – Depositing in preprint repositories is not a well-accepted praxis [12] and does not constitute a differential value. Although researchers from more flexible disciplines can informally communicate their work in progress, they do prefer to publish results in more formal channels [13]. – Offering of low access cost journals does constitute a differential value because research funding is smaller [10] as also is the number of researchers with access.
Products and/or services	– Sentences are longer and more difficult of being read [14]. – Amongst empirical disciplines, literature review and methodology are sections appearing in the beginning of the text and references at the end[15]. – Literature is purely in textual form with occasional occurrence of tables and illustrations [16]. – The average number of pages is greater [16]. – The amount of articles is higher[13]. – Co-authored articles are lower than in the Sciences and higher than in the Humanities [10].
Sources of resources	– Research funding is smaller as is the number of researchers with access to it [10]. The author-pay model is, therefore, not attractive either

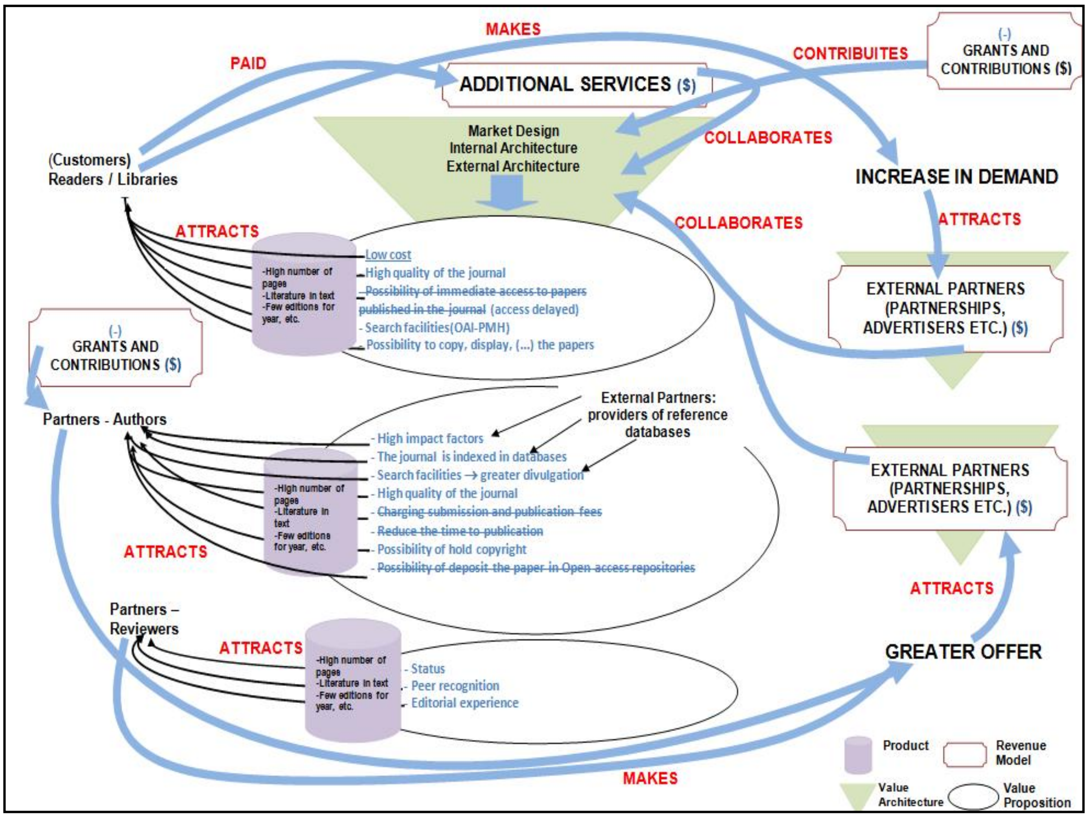

MAKES
PAID
CONTRIBUITES
(-) GRANTS AND CONTRIBUTIONS ($)
ADDITIONAL SERVICES ($)
Market Design
Internal Architecture
External Architecture
COLLABORATES
(Customers) Readers / Libraries
INCREASE IN DEMAND
ATTRACTS
COLLABORATES
ATTRACTS
-High number of pages
-Literature in text
-Few editions for year, etc.
Low cost
High quality of the journal
~~Possibility of immediate access to papers published in the journal~~ (access delayed)
- Search facilities(OAI-PMH)
Possibility to copy, display, (...) the papers
EXTERNAL PARTNERS (PARTNERSHIPS, ADVERTISERS ETC.) ($)
(-) GRANTS AND CONTRIBUTIONS ($)
Partners - Authors
External Partners: providers of reference databases
- High impact factors
- The journal is indexed in databases
- Search facilities → greater divulgation
- High quality of the journal
- ~~Charging submission and publication fees~~
- ~~Reduce the time to publication~~
- Possibility of hold copyright
- ~~Possibility of deposit the paper in Open access repositories~~
-High number of pages
-Literature in text
-Few editions for year, etc.
ATTRACTS
EXTERNAL PARTNERS (PARTNERSHIPS, ADVERTISERS ETC.) ($)
ATTRACTS
Partners – Reviewers
ATTRACTS
-High number of pages
-Literature in text
-Few editions for year, etc.
- Status
- Peer recognition
- Editorial experience
GREATER OFFER
MAKES
Product
Revenue Model
Value Architecture
Value Proposition

Figure 3 – Business model for the activity of open access electronic scholarly journals publishing in the Social & Human Sciences

3.4 Business model for open access scholarly journals in Arts & Humanities

It is well known within the scholarly community that researchers from Arts and Humanities make more use of books than of journals [17]. However, journals have their proper importance in the division. Therefore, the proposition of a business model for the activity of open access scholarly journal in Arts & Humanities should take into account particularities shown below. Some peculiarities are presented in comparison with Sciences and Social & Human Sciences.

Value proposition	– Immediate access to published work does not constitute a differential; neither does the smaller period of time between submission and publication. This is because speed of publication is low [10]. Delayed access model might be feasible to the peculiarities of the area. – Allowing researchers to deposit results in a digital repository is not a well-accepted praxis. Researchers from more flexible disciplines may informally communicate their work in progress but do prefer formal channels to their final results [13].

	– Offering of low access cost journals does constitute a differential value because research funding is smaller [10] as also is the number of researchers with access.
Products and/or services	– Sentences are longer and more difficult of being read [14]. – Amongst some specialties, literature review and methodology are sections appearing in the beginning of the text and references on footnotes [15]. – In some disciplines articles have less informative titles than the common praxis in other areas [10]. – Abstracts, though very usual in most areas, are rare [10]. – Literature is purely in textual form with occasional occurrence of tables and illustrations [16]. – o número médio de páginas de um artigo é maior nas Humanidades do que nas Ciências Naturais [16]; – The average number of pages is higher [13]. Researches count on less journal alternatives to publish. – Co-authored articles are lower than in the Sciences and higher than in the Humanities [10]. – Refusal rates are much higher [10].
Sources of resources	– Research funding is smaller as is the number of researchers with access to it [10]. The author-pay model is, therefore, not attractive either submissão de trabalhos não é um diferencial nessa área.

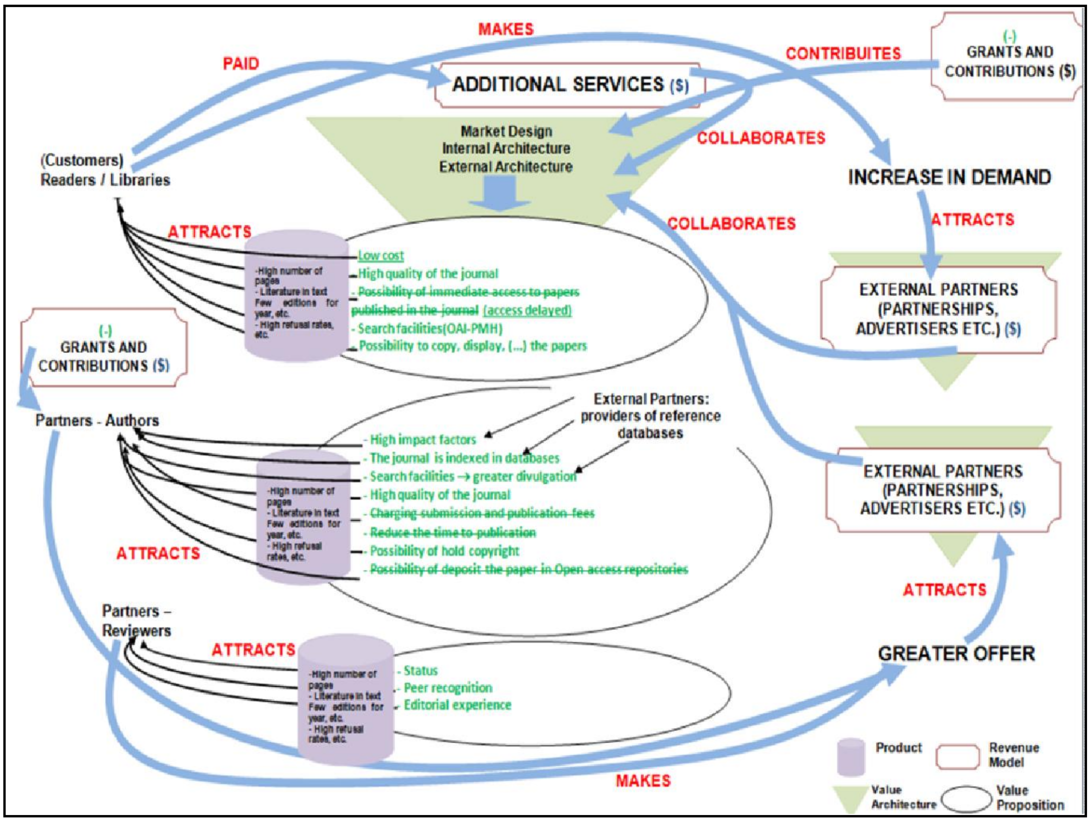

MAKES
PAID
CONTRIBUITES
(-) GRANTS AND CONTRIBUTIONS ($)
ADDITIONAL SERVICES ($)
Market Design
Internal Architecture
External Architecture
COLLABORATES
(Customers) Readers / Libraries
INCREASE IN DEMAND
ATTRACTS
COLLABORATES
ATTRACTS
-High number of pages
- Literature in text Few editions for year, etc.
- High refusal rates, etc.
Low cost
High quality of the journal
~~- Possibility of immediate access to papers published in the journal~~ (access delayed)
- Search facilities(OAI-PMH)
Possibility to copy, display, (...) the papers
EXTERNAL PARTNERS (PARTNERSHIPS, ADVERTISERS ETC.) ($)
(-) GRANTS AND CONTRIBUTIONS ($)
Partners - Authors
External Partners: providers of reference databases
- High impact factors
- The journal is indexed in databases
- Search facilities → greater divulgation
- High quality of the journal
~~- Charging submission and publication fees~~
~~- Reduce the time to publication~~
- Possibility of hold copyright
~~- Possibility of deposit the paper in Open access repositories~~
EXTERNAL PARTNERS (PARTNERSHIPS, ADVERTISERS ETC.) ($)
ATTRACTS
ATTRACTS
Partners – Reviewers
ATTRACTS
GREATER OFFER
-High number of pages
- Literature in text Few editions for year, etc.
- High refusal rates, etc.
- Status
- Peer recognition
- Editorial experience
Product
Revenue Model
MAKES
Value Architecture
Value Proposition

Figure 4 – Business model for the activity of open access electronic scholarly journals publishing in Arts & Humanities

4. Conclusion

The results obtained and discussed in this research enable to conclude that the conception of a business model for the editorial milieu is strongly associated with two important conditions. On a macro level, it is associated to the peculiarities of the different disciplinary areas. On a micro level, it is concerned with the context of a given publisher. Specifically, regarding to the disciplinary differences, the study showed that the configuration of business models for the Sciences distinguishes itself markedly from the other areas. On the other hand, the business models for the Social Sciences and Humanities and the Arts and Humanities are similar.

Perhaps the most critical issue in planning is the process of choosing and integrating the different overtones of a business setting and to integrate them into a model. The manner a publisher selects, implements and combines sundry components will reflect its idiosyncratic context—philosophical, cultural, technical and disciplinary. The business models proposed herein are just some amongst many resulting from the analysis of the publication context of open access scholarly journals. Therefore, it is beyond the intent to consider the present model as a standard for the publication of scholarly journals; on the contrary, it intends to serve as a spawning ground for new and more perfected ideas.

References

[1] AFUAH, Allan; TUCCI, Christopher. *Internet business models and strategies*. New York: McGraw-Hill, 2001.

[2] YUE, Gin Kwan. *Modelo de negócio:* uma proposta de visão integrada de processos logísticos em redes de restaurantes fast food. 2007. Thesis (PhD). University of São Paulo. Available at http://www.teses.usp.br/teses/disponiveis/3/3136/tde-31032008-145820/ (May 2009).

[3] VAN DER BEEK, Kornelia; KRÜGER, Cornelia C.; SWATMAN, Paula M.C. Business model formation within the on-line news market: the core + complement business Model Framework. In: BLED ELECTRONIC COMMERCE CONFERENCE, 16., 2003. Slovenia. *Proceedings*... Slovenia: IJEC, 9-11 June, 2003.

[4] LINDER, Jane; CANTRELL, Susan. *Changing business models:* surveying the landscape. Carlsbad, U.S.A: Institute for Strategic Change, 2000. Available at http://www.riccistreet.net/dwares/lane/mba600/linder.pdf (January 2009).

[5] MAHADEVAN, B. Business models for internet-based e-commerce: an anatomy. *California Management Review*, v. 42, n. 4, p. 55-69, Summer 2000.

[6] MEADOWS, A. J. *Communicating research.* San Diego: Academic Press, 1998.

[7] KVALE, Steinar; BRINKMANN, Svend. *Interwiews:* learning the craft of qualitative research interviewing. 2. Ed. Los Angeles: Sage, 2009.

[8] STÄHLER, Patrick. *Business models as an unit of analysis for strategizing*. 2002. Available at http://www.geschaeftsmodellinnovation.de/english/definitions.htm (May 2009).

[9] CROW, R.; GOLDSTEIN, H. *Guide to Business Planning for Launching a New Open Access Journal.* 2. Ed. Open Society Institute, 2003. Available at http://www.soros.org/openaccess/oajguides/business_planning.pdf (March 2009)

[10] MEADOWS, A. J. *A comunicação científica.* Brasília: Briquet de Lemos, 1999.

[11] TESTA, James. A base de dados ISI e seu processo de seleção de revistas. *Ciência da Informação*, Brasília, v. 27, n. 2, p. 233-235, maio/ago. 1998. Available at http://www.scielo.br/pdf/ci/v27n2/testa.pdf (April 2009)

[12] CRONIN, B. Scholarly Communication and Epistemic Cultures. *New Review of Academic Librarianship*, v.9, n. 1, p.1-24, Dec. 2003.

[13] SPARKS, Sue. *JISC Disciplinary Differences Report*. 2005. Available at http://www.jisc.ac.uk/media/documents/themes/infoenvironment/disciplinarydifferencesneeds.pdf (November 2008).
[14] HARTLEY J.; SOTTO, E.; FOX, C. Clarity across the disciplines: an analysis of texts in the Sciences, Social Sciences, and Arts and Humanities. *Science Communication*, v.26, n. 2, p. 188-210, Dec. 2004.
[15] THODY, Angela. *Writing and Presenting Research*. London: Sage Publications, 2006.
[16] HAYASH, Takayuk; FUJIGAKI, Yuko. Differences in knowledge production between disciplines based on analysis of paper styles and citation patterns. *Scientometrics*, v. 46, n. 1, p. 73-86, 1999.
[17] MOREIRA, A. C. S.; COSTA, S. M. S. Um modelo de comunicação eletrônica para os cientistas sociais e humanistas. In: SIMPOSIO INTERNACIONAL DE BIBLIOTECAS DIGITAIS, 3, 2005, São Paulo. *Proceedings*... São Paulo: University of São Paulo: Universidade Estadual Paulista, 2005. 29 p. Available at http://bibliotecas-cruesp.usp.br/3sibd/docs/moreira165.pdf (March 2009).

The Impact Factor of Open Access journals: data and trends

Elena Giglia [1]

1 Sistema Bibliotecario di Ateneo,
University of Turin,
via Verdi, 8
e-mail: elena.giglia@unito.it

Abstract

In recent years, a large debate has arisen about the citation advantage of Open Access (OA). Many studies have been conducted on different datasets and according to different perspectives, which led to different and somehow contradictory results depending on the considered disciplinary field, the researchers' attitude and citational behaviour, and the applied methodology. One of the bibliometric indicators most used worldwide to measure citations is Impact Factor – not free from criticisms and reservations – but it has only been tested on Open Access journals once, in 2004.

The aim of this preliminary work, focused on "Gold" Open Access, is to test the performance of Open Access journals with the most traditional bibliometric indicator – Impact Factor, to verify the hypothesis that unrestricted access might turn into more citations and therefore also good Impact Factor indices. Other indicators, such as Immediacy Index and 5-year Impact Factor, will be tested too.

The preliminary step of the work was fixing the list of Open Access journals tracked by Thomson Reuters in «Journal Citation Reports» (JCR). JCR was compared to the Directory of Open Access Journals (DOAJ) as of 31 December of the corresponding year.

As to coverage, Open Access journals in «Journal Citation Reports» are still a small percentage, even though there has been a large increase since 2003 in the *Science* edition (from 1.47% to 5.38%), less visible in the *Social Science* edition (from 1.05% to 1.52%, with a slight decrease from the 2007 1.71%).

In order to obtain comparable data, absolute Impact Factor or Immediacy Index values were not considered, but rather converted into percentiles for each category. The rank of the Open Access journals was analyzed in each single category. The titles were then clustered in disciplinary macro-areas, and data were aggregated.

Open Access journals in JCR 2008 *Social Sciences* edition rank in the top fifty percentiles (0-50) with a 54.5% share.

With substantial differences between macro-areas, in JCR 2008 *Science* edition Open Access journals rank in the top fifty percentiles (0-50) with a 38.62% share when considering Impact Factor, and with a 37.68% share referring to Immediacy Index. When considering 5-year Impact Factor, the share is 40.45%.

Open Access journals are relatively new actors in the publishing market, and gaining reputation and visibility is a complex challenge. Some of them show impressive Impact Factor trends since their first year of tracking. The collected data show that the performance of Open Access journals, also tested with the most traditional bibliometric indicator, is quite good in terms of citations.

Keywords: Open Access journals, Impact Factor, impact, scholarly communication, citations.

1. Impact, citations, Open Access, and Impact Factor

“Impact” in scientific communication is hard to define and moreover harder to measure. If we agree that «Science is a gift-based economy; value is defined as the degree to which one’s ideas have contributed to knowledge and impacted the thinking of others» [1], we should also admit that citation count is only one of the possible impact indicators, a proxy measure referring only to the academic context. This concept is even more true in the digital era, where a great variety of new impact measures – based on social network analysis and usage log data – are under development or already in use [2]. The notion of impact as a «multi-dimensional construct» and the suggestion that usage measures actually better describe in their connections and correlations the complexity of “impact” in the scientific process [3, 4] cannot be ignored, and we expect in a future further, new functional implications of this approach [5]. The new “article level metrics” suggested by PLoS One goes straight on this pathway [6].

However, “impact” has traditionally been expressed in terms of quantitative indicators, among which Impact Factor can be considered a standard *de facto*: or, at least, it is in the Italian academic context. Impact Factor has also gained a privileged position in the research evaluation system, with all its implications. But Impact Factor is only a proxy measure, and it should be used with caution in evaluating a single article and a single researcher [7]; reasonable critics and reservations on Impact Factor have been widely discussed by different actors involved in scientific publishing, such as recently summarized by Cope and Kalantzis and by Young et al [8]. Yet, focus of this work is to test an indicator and to present raw data; therefore it will not address the question and the related debate on the value of Impact Factor in itself.

The author is interested in matching the most traditional quantitative impact indicator, Impact Factor, and «one of the most exciting and radical events in publishing in recent years» [9], i.e. Open Access. One of the most debated arguments between Open Access advocates and detractors is its alleged citation advantage, which would stem by the « free, irrevocable, worldwide, right of access» stated by the Berlin Declaration [10]. Many studies have been carried out to determine if there is an actual Open Access advantage in citations [11] and, once established, to measure its value and understand its causes. Alma Swan edited a sort of systematic review of these studies and discussed methodological and interpretive issues, starting from the point that «citability rests upon the quality, relevance, originality and influence of a piece of work» and stating that «that OA would produce an automatic citation boost for every article was never the expectation» [12]. Different selected datasets and control-cases, different measures, e.g. citations or downloads, different time-spans led to different and somehow contradictory results, depending on the considered disciplinary field, the researchers’ attitude and citational behaviour, and the applied methodology [13]. Except for the two reports of Marie E. McVeigh of former ISI Thomson [14], since 2004 no more investigations have been conducted on the Impact Factor value trends of Open Access journals. The author thought it could be interesting to test again, after some years, the performance of Open Access journals in terms of citations, by applying the most commonly used quantitative indicator, Impact Factor. The author does not intend to deal with the debate about Impact Factor appropriateness or exhaustiveness, as just stated.

2. Do Open Access journals have good Impact Factor indices?

The 2009 RIN survey on *Communicating knowledge: how and why researchers publish and disseminate their findings*, shows, in addition to other fundamental findings about researchers' citing behaviour, that availability and easy access are one of the key criteria in citing an article [15]. The hypothesis the author intends to verify is that the "open" access, by raising the level of readership, might easily turn into more citations and therefore also good Impact Factor indices. Dealing with Impact Factor, this study forcedly addresses only Open Access journals – referred to as the "Gold Road" to Open Access. All the pre-prints and post-prints self-archived by authors in institutional or subject-based repositories have not been considered. They are referred to as the "Green Road", a preferential channel in early and free dissemination of research outputs, and they have been the object of recent bibliometric studies [16].

Sources of the work were:

- Thomson Reuters «Journal Citation Reports» (JCR), published every year in June, for the data about Journal Impact Factor, Immediacy Index and 5-year Impact Factor. It has a *Science* and a *Social Sciences* edition. No coverage is provided for Humanities;
- Directory of Open Access Journals (DOAJ) edited by Lund University, as the most accredited list of Open Access journals [17].

In order to define the method and in setting the research criteria, the author would have tried when possible to follow the choices of McVeigh's 2004 analysis, but it wasn't so easy partly because McVeigh, inside the former ISI, had had access to a great amount of complementary data, partly because McVeigh's sources at that time were different. In 2004 DOAJ was at the beginning, so McVeigh had to consider also SCiELO, whose titles now appear in DOAJ, and J-Stage, which also includes journals that are free on the Web, but not strictly Open Access [18].

Although the same framework has been maintained (4 disciplinary macro areas, reduction in percentiles and so on), it is hard to make a direct comparison because of the different list of titles examined and the adopted principle of inclusion [19]. In the present work, only DOAJ has been considered as a source, because with its 4,833 titles (as of March, 21st 2010) and its rigorous selection it is now supposed to be somehow an official register of Open Access journals.

3. Open Access journals coverage in Journal Citation Reports

Fixing the list of Open Access journals included in Journal Citation Reports was the first step of the work. There is no automatic filter to extract them, so the author has to achieve them by comparison.

The Impact Factor of a journal is «the average number of times articles from the journal published in the past two years have been cited in the JCR year » and it is calculated «by dividing the number of citations in the JCR year by the total number of articles published in the two previous years» [20]. JCR 2008 edition, published in June 2009, contains data about 2007 and 2006 articles' citations in 2008 journals. The author then decided to compare the titles present in DOAJ as of December, 31st of the corresponding JCR year, i.e. those on which Impact Factor has been calculated.

A query run by ISSN number gave a first automatic extraction. Then, a manual comparison drove to the inclusion of titles which for whatsoever reason had different ISSN numbers in the two sources.

The same method has been applied both within the JCR *Sciences* and *Social Sciences* editions, considering the online original version as of June, 2009. Further inclusions in the 2009 Fall revision of JCR have not been considered, in order to set a definite edition for future comparisons.

In JCR 2008 *Social Science* edition resulted a list of 30 Open Access titles out of 3,801 (1.52%); in JCR 2008 *Sciences* edition resulted a list of 355 Open Access titles out of 6,598 (5.38%). The coverage in 2003-2008 is presented in Table 1 (JCR *Social Sciences* edition) and 2 (JCR *Sciences* edition).

Year	Titles in JCR	Titles in DOAJ 31-12	OA titles with IF	OA titles with IF (%)
2003	1714	602	18	1.05%
2004	1712	1194	19	1.11%

2005	1747	1811	22	1.26%
2006	1768	2357	24	1.36%
2007	1866	2954	32	1.71%
2008	1980	3801	30	1.52%

Tab. 1: Open Access titles in JCR – *Social Sciences* edition.

Year	Titles in JCR	Titles in DOAJ 31-12	OA titles with IF	OA titles with IF (%)
2003	5907	602	87	1.47%
2004	5968	1194	168	2.82%
2005	6088	1811	218	3.58%
2006	6164	2357	259	4.20%
2007	6417	2954	315	4.91%
2008	6598	3801	355	5.38%

Tab. 2: Open Access titles in JCR – *Science* edition

It is to be noticed that the lists of titles are not homogeneous. In JCR 2008 *Science* edition 110 titles were excluded compared to the 2007 edition, including 6 Open Access titles; in JCR 2008 *Social Sciences* 23 titles were excluded, including 3 Open Access titles. In DOAJ, too, there have been variations, and 8 former Open Access titles listed in 2007 were not included as of December 2008.

In JCR 2008 *Science* edition 355 titles have been counted instead of 356 because of the changing title of *Acta Phytotaxonomica Sinica* in *Journal of Systematics and Evolution*. The journal maintained the same ISSN but has no 2008 data. There are also two titles which were assigned to a different category compared to 2007 (*Interciencia* and *Journal of Research of the National Institute of Standards and Technology*).

These tables show the coverage of Open Access journals within Journal Citation Reports. While in the *Science* edition they are represented in a still small but growing percentage, the small number and percentage of titles included in the Social Sciences edition, 1.52%, representing a decrease from 2007, has not been investigated in depth, as the numbers are not sufficient to draw any conclusions. In DOAJ as of December, 31st 2008, at least 533 titles

(14%) can be referred to the Social Sciences area. So we have to wait for their inclusion in JCR in the future.

Some more comparisons can be added, in order to clarify the size of the sample: in Ulrichsweb, we find 26,710 active refereed academic/scholarly journals as of March 21st, 2010. Compared to this, the 4,833 Open Access titles listed in DOAJ the same day represent a 18.09%.

4. Open Access journals in Journal Citation Reports: where do they come from?

Focusing on the *Science* edition, the author looked for the geographical distribution of the list of 355 Open Access journals, taking the publisher's country as the point of origin. The results are shown in Table 3.

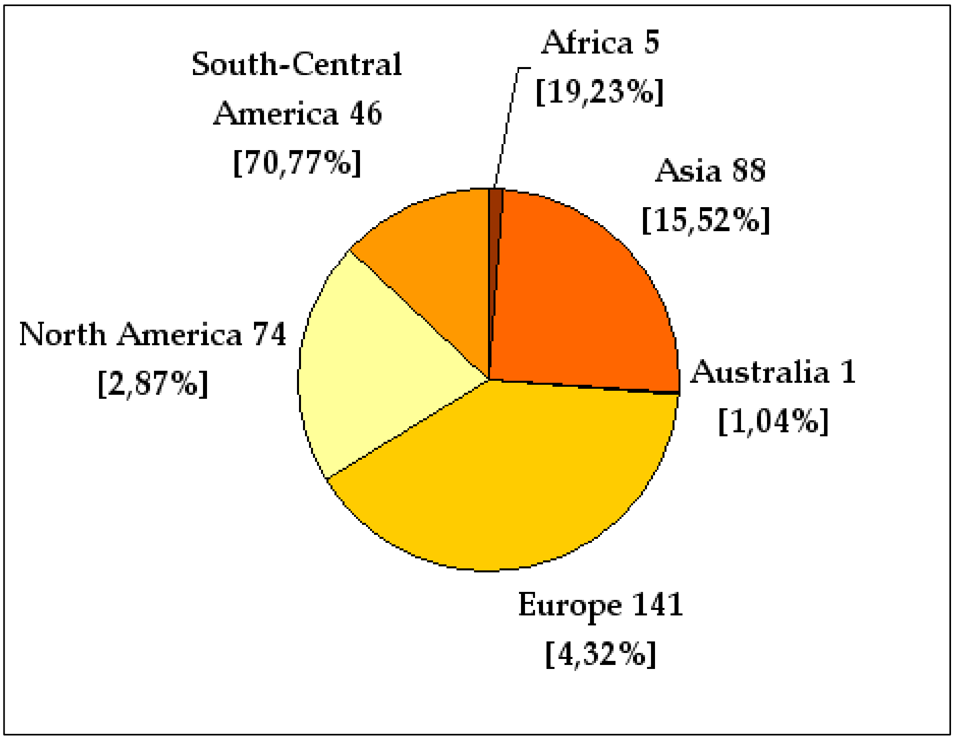


Table 3: Geographical distribution of OA journals in JCR 2008 - *Science* ed.

Ratios generated in the comparison with the geographical distribution of all 6,598 titles in JCR 2008 *Science* edition are shown in Table 4, in association with 2007 data (6,417 titles):

Area	JCR *Science* Titles		OA titles		%		
	2007	2008	2007	2008	2007	2008	Variat.
Africa	24	26	4	5	16.67%	19.23%	+2.56%
Asia	547	567	74	88	13.53%	15.52%	+1.99%
Australia – New Zealand	89	96	1	1	1.12%	1.04%	-0.08%
Europe	3177	3264	118	141	3.71%	4.32%	+0.61%
North America	2529	2580	80	74	3.16%	2.87%	-0.29%
South-Central America	51	65	38	46	74.51%	70.77%	-3.74%
Tot.	6,417	6,598	315	355			

Table 4: Percentages of OA titles by geographical distribution – JCR *Science* ed.

It's important to notice that 70.77% (74.51% in 2007) of covered titles from South-Central America are available as Open Access: this could be a demonstration of the international quality, visibility and reputation of the cited SCiELO platform. The Africa and Asia ratios are also interesting, with a good presence of Open Access journals and a growing trend, while North America, Europe and Australia show lower percentages rates.

5. Open Access journals in Journal Citation Reports: what do they talk about?

Following Mc Veigh's method, the 355 Open Access titles of JCR 2008 *Science* edition have been clustered in 4 disciplinary macro-areas, Chemistry [CH], Mathematics-Physics-Engineering [M-P-E], Life Sciences [LS], Medicine [MED], relating to the category assigned in JCR, as shown in Table 5. Titles referring to two or more categories have been duplicated, so the total amount counted 479 items. In 2007, 315 titles had originated 422 items. The table shows also the growing trend in inclusion of Open Access titles in each macro-area, with the caution, as we said above, that not all the 2007 Open Access titles are still represented in the 2008 edition.

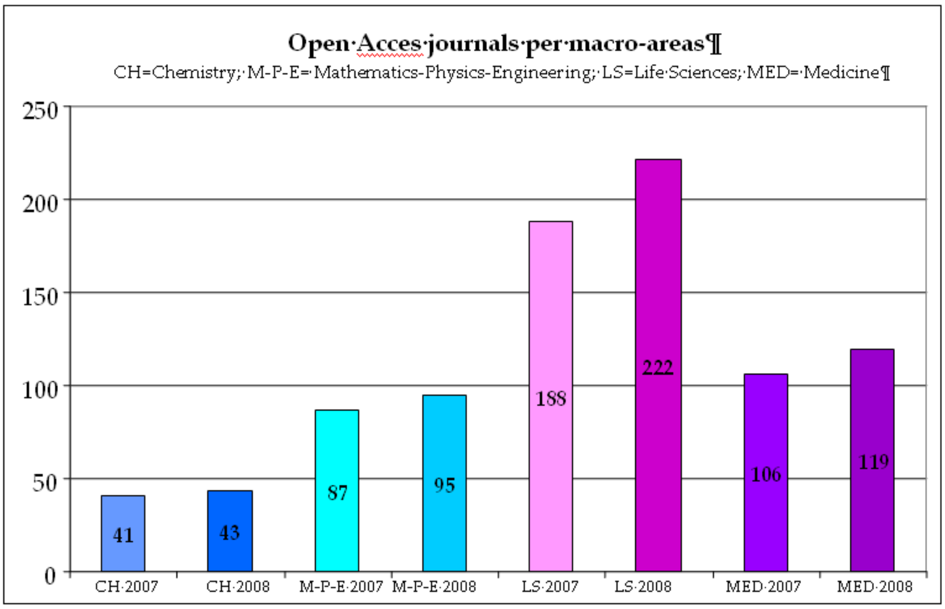


Table 5: OA journals by macro disciplinary areas in JCR *Science* ed.

6. Open Access journals ranking in Journal Citation Reports by Impact Factor

The author then ranked the Open Access titles by Impact Factor.

Impact Factor's values range is widely distributed among the categories: *CA - A cancer journal for clinicians*, an Open Access journal which runs first in its category (Oncology) and which runs also first among all the 6,598 titles, has a 74.575 index value as Impact Factor. *Communications on pure and applied mathematics*, which runs as well first in its category (Mathematics), has a 3.806 index value.

Therefore, in order to obtain comparable data, absolute Impact Factor was not considered. Impact Factor was converted to percentile rank as follows

$$p_n = \frac{100}{N}\left(n - \frac{1}{2}\right)$$

where p is the percentile, *N* the number of items in a category and *n* the rank value of the title.

Percentiles 0-10 include the highest Impact Factor values, 91-100 the lower ones.

This is the only analysis carried out on JCR 2008 *Social Science* edition, to have a preliminary benchmark result for future comparisons. There are 30 Open Access titles which, once duplicated because of the pertaining category, generated 37 items. Due to the small size of the sample, no subdivision in categories was performed. Results are shown in synopsis in Table 6. Open Access titles rank in the top fifty percentiles (0-50) with a 54.05% share (20 out of 37).

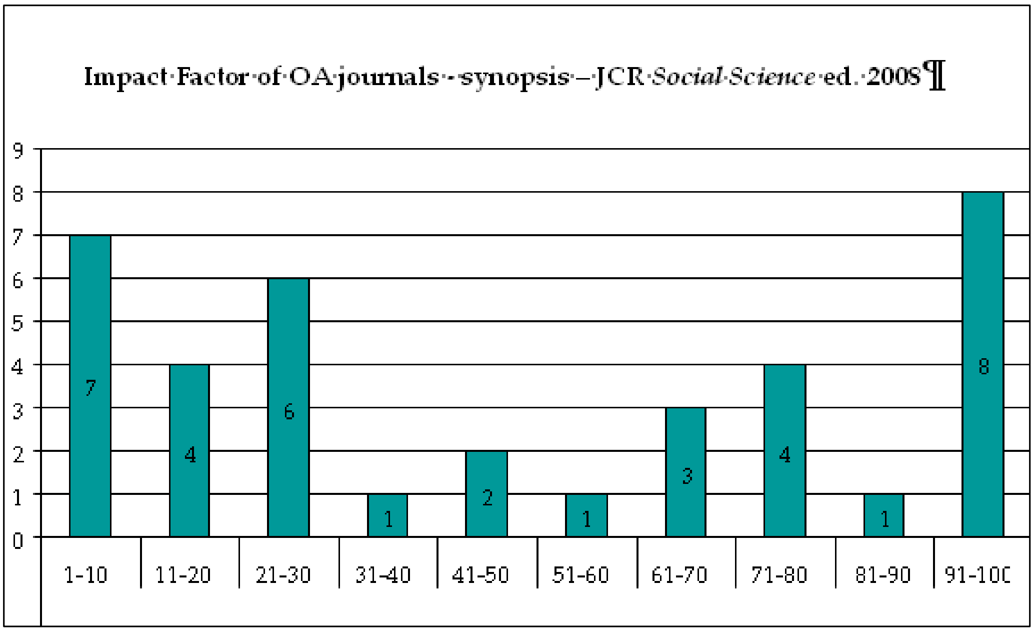


Table 6: OA journals in JCR 2008 *Social Sciences* ed. ranking by Impact Factor (synopsis).

Referring to JCR 2008 *Science* edition, the author then analyzed the 479 Open Access titles, duplicates included.

Percentile rank was first analyzed for each title in its assigned category within JCR: Chemistry [CH]: 43 titles in 15 categories, Mathematics-Physics-Engineering [M-P-E]: 95 titles in 32 categories, Life Sciences [LS]: 222 titles in 46 categories, Medicine [MED]: 119 titles in 31 categories.

Results were then aggregated by disciplinary macro-area, as shown in Tables 7-10, in comparison with 2007 data.

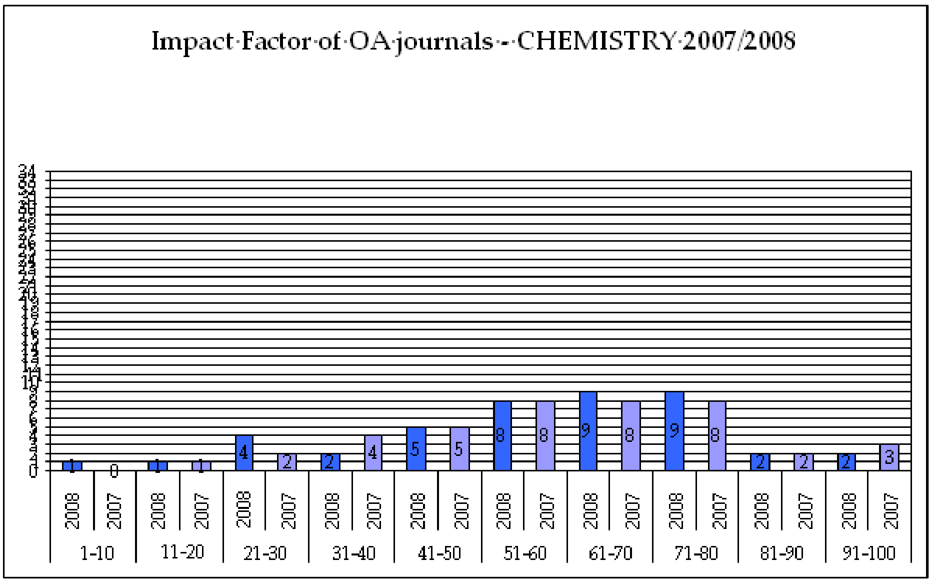


Table 7 Impact Factor of OA journals Chemistry 2007/2008

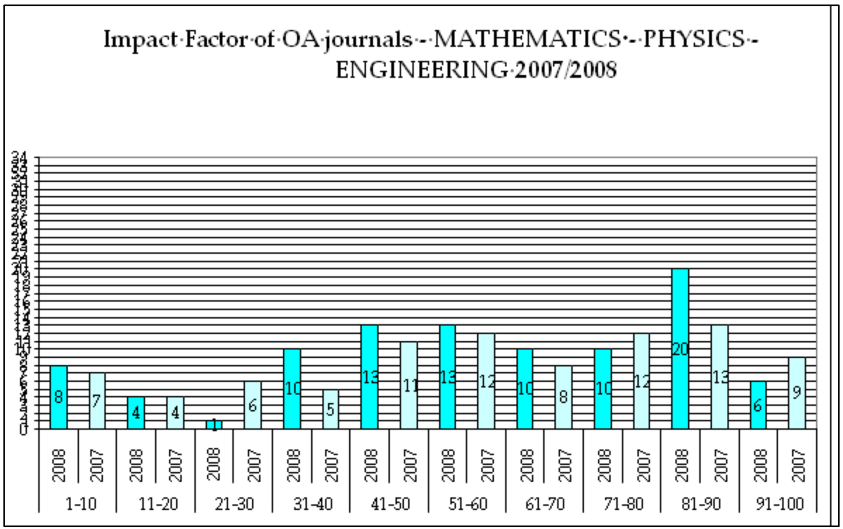


Table 8 Impact Factor of OA journals Mathematics, Physics, Engineering 2007/2008

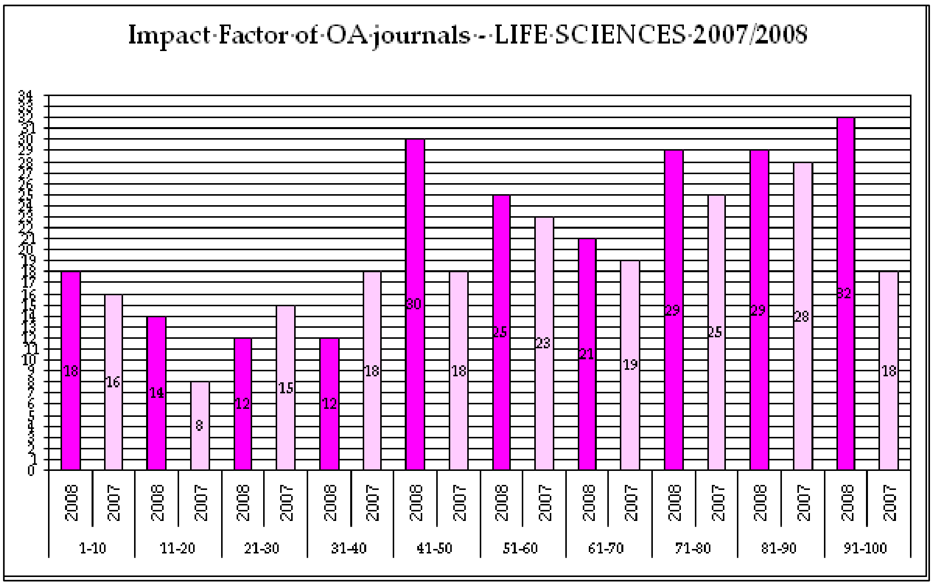


Table 9 Impact Factor of OA journals Life Sciences 2007/2008

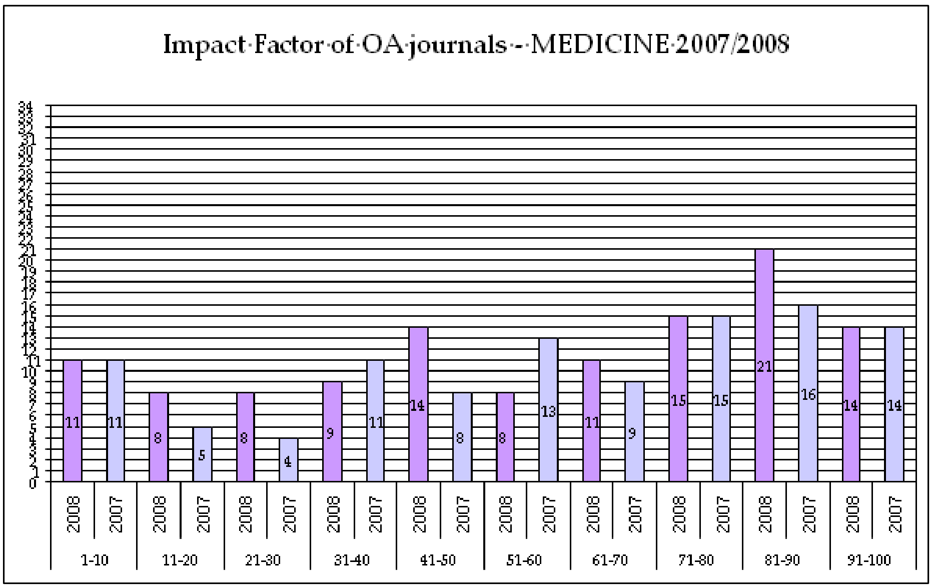


Table 10 Impact Factor of OA journals Medicine 2007/2008

There are as expected strong differences among disciplinary areas. When considering the best performances, in Medicine there is a strong presence in the top twenty (0-20) percentiles (15.96%); slightly lower in Life Sciences and in Mathematics-Physics-Engineering (respectively 14.42% and 12.63%), absolutely lower in Chemistry (4.66%). Data in synopsis are shown in Table 11.

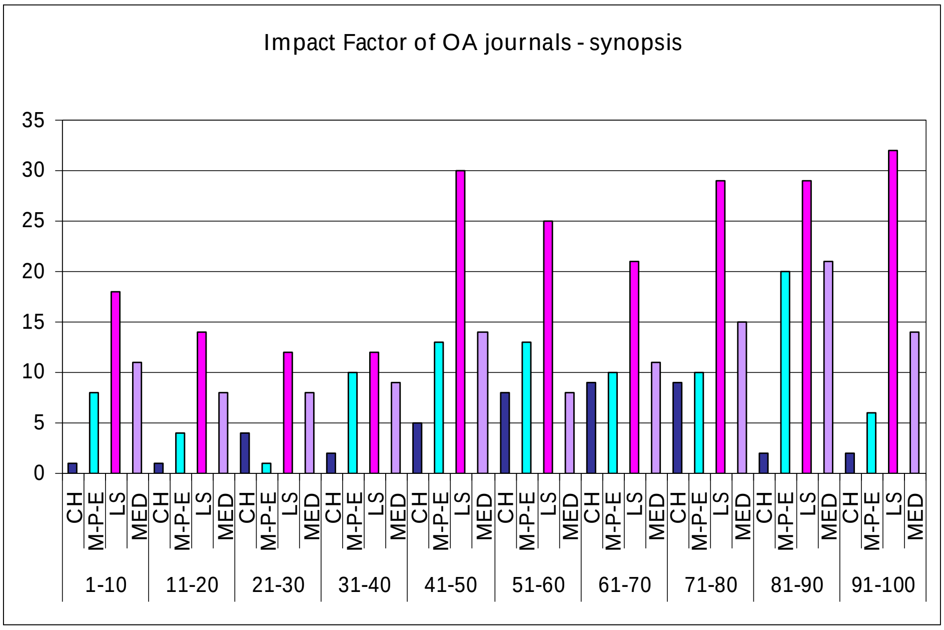


Table 11: OA journals in JCR 2008 *Science* ed. ranking by Impact Factor (synopsis).

In a global outlook, Open Access journals rank in the top fifty percentiles (0-50) with a 38.62% share (185 titles out of 479) when considering Impact Factor, as shown in Table 12. The table also outlines the distribution in each disciplinary macro area: in Medicine, 42.02% titles rank in the top fifty percentiles. 2007 values are included in the table in red.

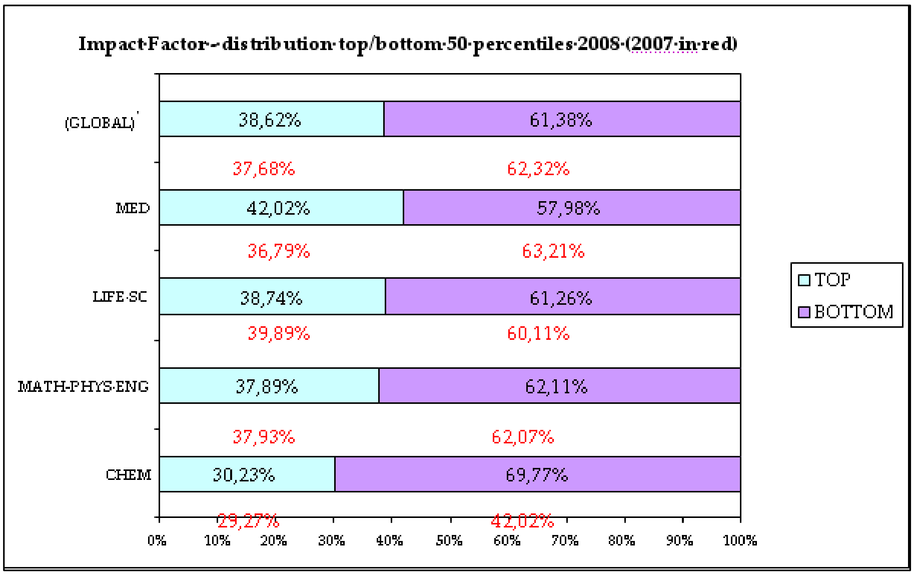


Table 12: distribution top/bottom percentiles in JCR 2008 *Science* ed. (in red 2007 data)

In Fall, 2009, Thomson Reuters released a revised version of JCR 2008. In the *Science* edition, titles became 6,620 (+22). 10 titles out of these 22 are Open Access. Open Access titles moved from 355 to 365, and from 479 to 492 duplicates included. In some cases, wrong assigned Impact Factor values have been rectified. Global data then moved from a 38.62% to a 39.43% share ranking in the top fifty (0-50) percentiles (194 titles out of 492), with a shift from 30.23% to 31.11% in Chemistry, from 37.89% to 39.58% in Mathematic-Physics-Engineering, from 38.74% to 39.04% in Life Sciences, and from 42.02% to 43.09% in Medicine. However, according to the purpose of this study, aimed at future assessments, only the official June 2009 edition has to be considered.

Even though a direct comparison with McVeigh's 2004 data is not possible, as we said above, we can try at least to relate the final results. McVeigh's global data showed in JCR 2002 edition a 34% share in the top fifty (which are 51-100, because she used a different formula) percentiles and a 66% share in the bottom ones [21]. Six years later (according to JCR date of publication), the ratio is 38% [39% in Fall revised edition] against 62% [61%]. It seems to be a little change. But it is to be noticed that the list of 355 titles in JCR 2008 *Science* edition is the whole sample of strictly Open Access journals with Impact Factor, obtained by matching DOAJ and JCR. DOAJ has rigorous selection criteria in defining what an "Open Access journal" is. In 2004, Mc Veigh considered as a source also J-Stage, a Japan gateway which includes simply "free on web" journals [22]. So, McVeigh's sample seems to have been built on wider inclusion criteria: therefore results might be overrated and the resulting gap with JCR 2008 data underestimated. A new study with the same methodology and criteria of the analysis presented in these pages is going to be carried on next JCR 2010 edition, in order to obtain comparable data to set up a trend.

5. Open Access journals ranking in Journal Citation Reports by Immediacy Index

In order to test a potential early advantage, the author then ranked Open Access journals in JCR 2008 *Science* edition by Immediacy Index. Immediacy Index is calculated by dividing the number of citations to articles published in a given year by the number of articles published in the same year. Possible biases within this measure are that frequently issued journals, with articles published early in the year, had more chances of being cited and that large journals have advantage over small ones: these are cautions notified in JCR itself [23].

Among the 355 Open Access titles, 33% are quarterly, 21% bimonthly, and 17% monthly. 13% have no issues per year declared in JCR, comprising both irregular and e-only titles. Only 3% have 20 or more issues per year.

To obtain comparable data, also Immediacy Index was converted to percentile rank with the same formula: $p_n = \frac{100}{N}(n - \frac{1}{2})$ where p is the percentile, *N* the number of items in a category and *n* the rank value of the title.

According to the same methodology applied to Impact Factor values, percentile rank was first analyzed for each title in its assigned category within JCR. Results were then aggregated by disciplinary macro-area.

Global results are shown in Table 13 in comparison with Impact Factor data.

Immediacy Index seems to be higher in the top thirty (0-30) percentiles. In a global outlook, in JCR 2008 *Science* edition Open Access journals rank in the top fifty (0-50) percentiles by Immediacy Index with a 37.16% share (178 titles out of 479), slightly lower than the same year's Impact Factor (-1.46%).

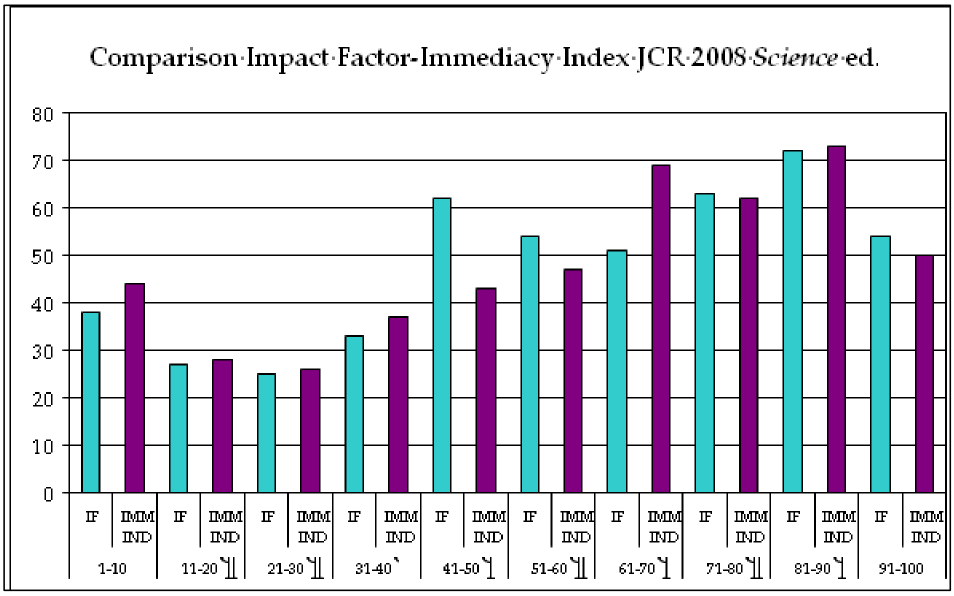
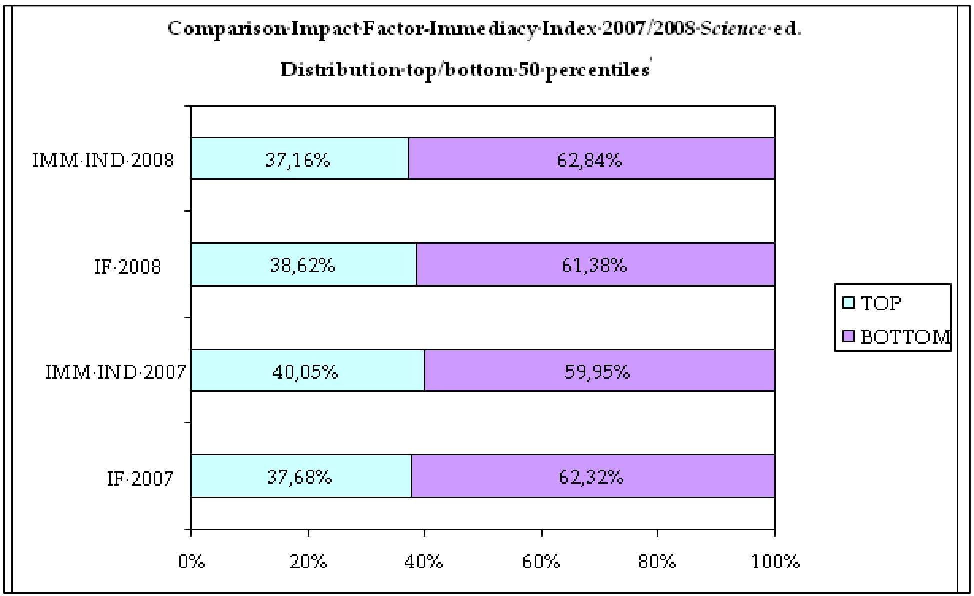


Table 13: Impact Factor compared to Immediacy Index – JCR 2008 *Science* ed.



In 2007, the trend was

the opposite: they ranked in the top fifty (0-50) percentiles with a 40.05% share (169 titles out of 422) when considering Immediacy Index, a 2.37 % higher than Impact Factor (159 titles, 37.68%). Data are collected in Table 14.

It is interesting to notice some cases of many titles which rank low by Impact Factor but high by Immediacy Index. 225 titles out of 479 (47%) show a best performance in Immediacy Index than in Impact Factor (56% in Chemistry 56% in Mathematics-Physics-Engineering, 41% in Life Sciences and 49% in Medicine)

Table 14: Impact Factor to Immediacy Index – global data JCR *Science* ed. 2007/2008

The median value of the difference between the two values is 8, with 104 titles under the median and 121 above. The peaks are represented by *Kyushu Journal of Mathematics* (184th by Impact Factor and 36th by Immediacy Index), *Abstract and Applied Analysis* (116th and 9th), *Boundary value problems* (118th and 14th), *Revista Chilena de Historia Natural* (96th and 8th).

6. A further analysis: 5-year Impact Factor

Considering that one of the most diffused criticisms against Impact Factor is its time span – two years is often a too narrow period to test the impact of a research article, especially in certain disciplines – a new indicator has been provided in JCR starting with the 2007 edition, 5-year Impact Factor. It is calculated by dividing the number of citations in the JCR year by the total number of articles published in the five previous years.

As with Impact Factor and with Immediacy Index, absolute values of 5-year Impact Factor were converted to percentile rank with the same formula: $p_n = \frac{100}{N}(n - \frac{1}{2})$ where p is the percentile, *N* the number of items in a category and *n* the rank value of the title.

According to the same methodology applied to Impact Factor and Immediacy Index values, percentile rank was first analyzed for each title in its assigned category within JCR. Results were then aggregated by disciplinary macro-area.

In 2007 JCR *Science* edition 315 titles out of 422 (75% of the total) have a 5-year Impact Factor. They rank in the top fifty percentiles (0-50) with a 40% share (126 titles out of 315).

In 2008 JCR *Science* edition 356 titles out of 479 have a 5-year Impact Factor (74% of the total). They rank in the top fifty percentiles (0-50) with a 40.45% share (144 titles out of 356). Results are shown in Table 15.

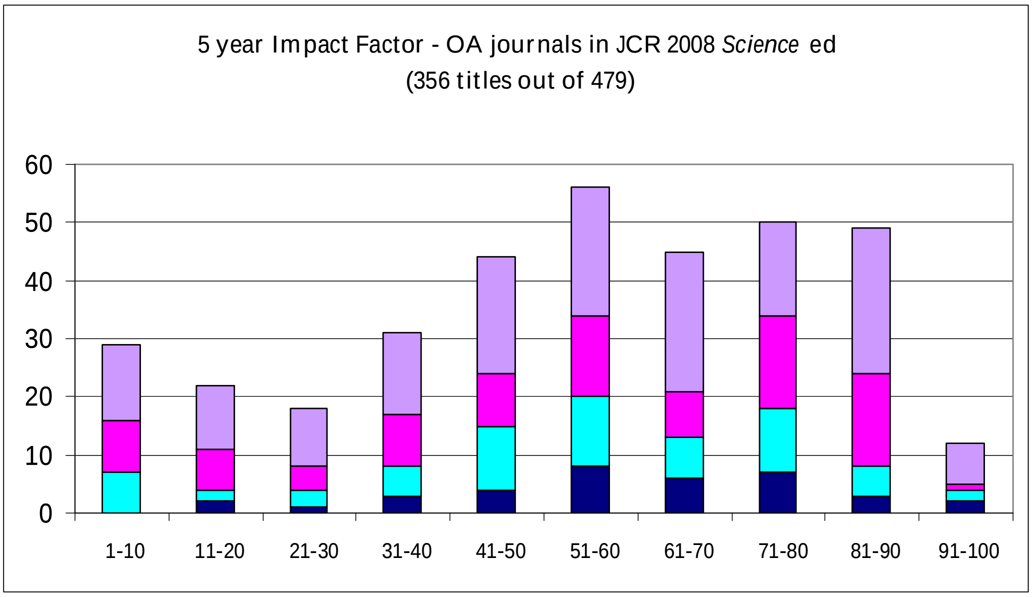


Table 15: 5-year Impact Factor for OA journals JCR 2008 *Science* ed. (only for 356 titles).

7. Open Access journals in Journal Citation Reports: how old are they?

In the asymmetry of the inelastic scholarly communication market, there are prestigious titles with reputations acquired over a period of many years.

Therefore the journal age has been analyzed, in order to find if there might be any correlation between age and performance. Once obtained the splitting into categories and percentiles for JCR 2008 *Science* edition titles, the author tried to collect data in Table 16. Only the first year of publication could have been considered; as known, some journals are Open Access-natives, other are Open Access-converted, so these data are just relative. Although you can access a list of converted titles in Open Access Directory [24], information dates back only to 2006, and the list is not exhaustive; in most cases, it is impossible to establish the year of conversion. However, the author considered the median starting year of publication for journals within their own percentile by Impact Factor rank. At the left and right side of the median year is the number of older and younger/equal titles respectively. Younger/equal titles are in majority.

Percentile	CHEMISTRY	MATH-PYS-ENG	LIFE SCIENCES	MEDICINE
0-10	2001 1	4 1994 4	7 2003 11	5 1999 6
11-20	2001 1	2 1999 2	7 2001 7	3 2001 5
21-30	2 2000 2	1997 1	2 2001 10	3 2000 5
31-40	1 2003 1	5 1997 5	4 2000 8	4 2003 5
41-50	2 1990 3	6 1999 7	13 2000 17	6 2001 8
51-60	3 2000 5	5 1998 8	10 2000 15	4 2002 4
61-70	4 2000 5	5 1997 5	8 1999 13	5 2000 6
71-80	4 2002 5	4 1997 6	10 2000 19	7 1999 8
81-90	1 1998 1	9 1999 11	12 2000 17	6 2000 15
91-100	1 2004 1	3 2003 3	9 2001 23	6 2001 8

Table 16: Open Access journals in JCR 2008 *Science* ed.: median first year of publication.

Distribution is uneven, so that a direct causal relationship between age and visibility and prestige in terms of citations cannot be straightforwardly inferred.

At a glance, lower median years can be found in the top fifty (0-50) percentiles only in Mathematics-Physics-Engineering, where the lowest percentile corresponds to the most recent median year. In Life Sciences, in the top ten (0-10) percentiles, the median year is 2003, but seven titles were born in 2005 (out of 18). In Medicine, in the top ten (0-10) percentiles there are a 2003, a 2004 and a 2007 title. The last one is *PLoS Neglected tropical diseases*, which ranks first in its first year of tracking.

Thus, there seems to be no strong correlation between the age of a journal and its performance according to Impact Factor. There are some striking examples, such as the cited young PLoS journals which since their first tracking year ranked in the first percentiles – *PLoS Biology* ranked first in its category in its first year, with an Impact Factor quite double over the second in ranking – or such as BioMedCentral *BMC Bioinformatics*, or *Atmospheric*

Chemistry and Physics, with its innovative concept of peer-review, always in the first positions of its category [25]. They could be a proof that the pre-reputation period – i.e. the time span requested for a journal to establish in the scholarly publications market – could result shortened in an Open Access environment [26]. Otherwise, the great number of young Open Access journals ranking in the bottom fifty percentiles (51-100) could be a sign of the difficulty of competing with traditional and established titles. More detailed analyses and comparisons with non-Open Access titles trends are due to address the question.

8. Conclusions and further researches

Open Access journals presence in JCR 2008 *Social Sciences* edition (1.52%) is so low that claims, as to now, no more investigations than the simple trend in Impact factor value. These few Open Access journals rank in the top fifty (0-50) percentiles with a 54.05% share.

Open Access journals in JCR 2008 *Science* edition are still represented in a small percentage, even though the large increase since 2003 (from 1.47% to 5.38%).

As for Impact Factor performance, a 38.62% share [39.43% in Fall edition] in the top fifty (0-50) percentiles is a good although not striking result, such as a 37.16% share as for Immediacy Index and a 40.45% as for 5-year Impact Factor (the latter only for 356 titles out of 479).

These results are not outstanding, but they represent only the first step of an ongoing work. A fair discussion should require a comparison with JCR 2010 data, to set a trend which is expected to be highly positive.

The preliminary data reported in this contribution might be useful to further comparisons, more elaborated reflections and in-depth analysis. Further researches might concern the Impact Factor values trend of Open Access journals over several years, in comparison with that of traditional journals, and the performance in terms of Impact Factor of Open Access and traditional titles of the same age.

Open Access journals are relatively new actors in the scholarly publishing market; and gaining reputation and visibility is a complex challenge among established titles. Our collected data, nevertheless, show that the performance of Open Access journals, as tested with the most traditional bibliometric indicator, Impact Factor, is quite good in terms of citations. They can compete with older actors; in other words, as Peter Suber puts it, quality can keep pace with prestige and reputation [27].

Acknowledgements

I am deeply indebted to Marta Plebani (Thomson Reuters) who provided me with 2007 raw data to work on; to Professor Giuseppe Migliaretti (University of Turin) for his statistic expertise; to Susanna Mornati (CILEA) for her initial methodological suggestions; to Heather Morrison (Simon Fraser University) and Paola Gargiulo (CASPUR) for their comments.

Notes and References

[1] BOLLEN J; et al. A principal component analysis of 39 scientific impact measures. PLoS ONE 4 (6), 2009, e6022. Available at http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0006022 (March 2010).

[2] BOLLEN J; et al. A principal component analysis of 39 scientific impact measures. PLoS ONE 4 (6), 2009, e6022. Available at http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0006022 (March 2010). See also FRANCESCHET, M. Journal influence factors. Journal of Informetrics, 2010, in press. Available at doi:10.1016/j.joi.2009.12.002 (March 2010).

[3] BOLLEN J; et al. A principal component analysis of 39 scientific impact measures. PLoS ONE 4 (6), 2009, e6022. Available at http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0006022 (March 2010).

[4] BOLLEN J; et al. Clickstream Data Yields High-Resolution Maps of Science. PLoS ONE 4 (3), 2009, e4803. Available at http://www.plosone.org/article/info:doi/10.1371/journal.pone.0004803 (March 2010).

[5] BOLLEN J. Studying scientific activity from large-scale usage data. Presentation. *CERN workshop on innovations in scholarly communication - OAI 6*, Geneva 17-19 June 2009. Available at http://indico.cern.ch/contributionDisplay.py?contribId=22&sessionId=8&confId=48321 (March 2010). see also MESUR, Metrics from Scholarly Usage of Resources. Available at http://www.mesur.org/MESUR.html (March 2010).

[6] BINFIELD P. PLoS One: background, future development, and article-level metrics. In MORNATI S; HEDLUND T. editors. *Rethinking electronic publishing*, ELPUB 2009 proceedings. Milan: Nuova Cultura, 2009, pp. 69-86. Available at http://conferences.aepic.it/index.php/elpub/elpub2009/paper/view/114/51 (March 2010). See also PLoS One (2009) Article-level metrics. Available at http://article-level-metrics.plos.org/ (March 2010).

[7] CAMPBELL P. Escape from the impact factor. ESEP 8, 2008, p. 5-7. Available at http://www.int-res.com/articles/esep2008/8/e008p005.pdf (March 2010). GARFIELD E. The Impact Factor and using it correctly. Der Unfallchirurg, 101(6), June 1998 p.413. English translation. Available at http://garfield.library.upenn.edu/papers/derunfallchirurg_v101(6)p413y1998english.html (March 2010).

[8] COPE B; et al. Signs of epistemic disruption: Transformations in the knowledge system of the academic journal. First Monday, 14 (4) 6 April 2009, Available at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2309/2163 (March 2010).

YOUNG NS; et al. Why Current Publication Practices May Distort Science. PLoS Med 5 (10), 2008, e201. Available at http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0050201 (March 2010).

[9] MC VEIGH ME. Open Access journals in the ISI citation databases: analysis of Impact Factors and citation patterns. A citation study from Thomson Scientific, 2004. Available at http://scientific.thomsonreuters.com/m/pdfs/openaccesscitations2.pdf (March 2010).

[10] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003). Available at http://oa.mpg.de/openaccess-berlin/berlindeclaration.html (March 2010).

[11] The effect of open access and downloads ('hits') on citation impact: a bibliography of studies (Op-Cit project). Available at http://opcit.eprints.org/oacitation-biblio.html (March 2010). See also WAGNER, AB. Open Access Citation Advantage: An Annotated Bibliography. Issues in Science and Technology Librarianship, 60. Winter

2010. Available at http://www.istl.org/10-winter/article2.html (March 2010).

[12] SWAN, A. The Open Access citation advantage: Studies and results to date. Technical Report, 2010. Available at http://eprints.ecs.soton.ac.uk/18516/ (March 2010).

[13] Stevan Harnad is as usual very sharp and meticulous in commenting and pointing out methodological questions about OA advantage. A thread of his exchanges with the authors of many studies could be followed on his blog, Open Access Archivangelism (http://openaccess.eprints.org/) searching "OA advantage". A collection of these contributions can also be found on Connotea, tag "OA advantage Harnad".

[14] MC VEIGH ME. Open Access journals in the ISI citation databases: analysis of Impact Factors and citation patterns. A citation study from Thomson Scientific, 2004 Available at http://scientific.thomsonreuters.com/m/pdfs/openaccesscitations2.pdf (March 2010). See also MC VEIGH ME. The Impact of Open Access journals. Report, 2004. Available at http://thomsonscientific.jp/event/oal/impact-oa-journals.pdf (March 2010).

[15] Research Information Network – RIN. Communicating knowledge: how and why researchers publish and disseminate their findings. Report, 2009, p. 31. Available at http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/communicating-knowledge-how-and-why-researchers-pu (March 2010).

[16] GARGOURI, Y; et al. Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. PLOS ONE, 2010. (Submitted). Available at http://eprints.ecs.soton.ac.uk/18493/ (March 2010).

GENTIL-BECCOT, A; et al. Citing and Reading Behaviours in High-Energy Physics. How a Community Stopped Worrying about Journals and Learned to Love Repositories. arXiv.org, 30 Jun 2009. Available at http://arxiv.org/ftp/arxiv/papers/0906/0906.5418.pdf (March 2010).

GARGOURI, Y; et al. Logistic regression of potential explanatory variables on citation counts. Preprint 11 Apr. 2009. Available at http://www.crsc.uqam.ca/yassine/SelfArchiving/LogisticRegression.htm (March 2010).

[17] DOAJ - Directory of Open Access Journals. Available at http://www.doaj.org/ (March 2010). Journal Citation Reports is available only for subscribers.

[18] SciELO – Scientific Electronic Library Online. Available at http://www.scielo.br/ (March 2010); J-Stage – Japan Science and Technology Information Aggregator Electronic. Available at http://www.jstage.jst.go.jp/browse/ (March 2010).

[19] MC VEIGH ME. Open Access journals in the ISI citation databases: analysis of Impact Factors and citation patterns. A citation study from Thomson Scientific, 2004. Available at http://scientific.thomsonreuters.com/m/pdfs/openaccesscitations2.pdf (March 2010).

[20] JCR Help. Available for subscribers: http://admin-apps.isiknowledge.com/JCR/help/h_impfact.htm#impact_factor (March 2010).

[21] MC VEIGH ME. The Impact of Open Access journals. Report, 2004. Available at http://thomsonscientific.jp/event/oal/impact-oa-journals.pdf (March 2010).

[22] J-Stage - Japan Science and Technology Information Aggregator, Electronic. Available at http://www.jstage.jst.go.jp/browse/ (March 2010).

[23] JCR Help. Available for subscribers: http://admin-apps.isiknowledge.com/JCR/help/h_impfact.htm#impact_factor (March 2010).

[24] OAD - Open Access Directory. Available at http://oad.simmons.edu/oadwiki/Journals_that_converted_from_TA_to_OA (March 2010).

[25] POESCHL, U. Open peer review and interactive open access publishing: the effectiveness of transparency and self-regulation in scientific quality assurance. Presentation. *CERN workshop on innovations in scholarly communication - OAI 6*, Geneva 17-19 June 2009. Available at http://indico.cern.ch/contributionDisplay.py?contribId=23&sessionId=8&confId=48321 (March 2010).

[26] WILLINSKI J. Open Access and academic reputation. Slaw.Ca, 16 Jan 2009. Blog post. Available at http://www.slaw.ca/2009/01/16/open-access-and-academic-reputation/ (March 2010).

[27] SUBER, P. Thinking about prestige, quality and Open Access. SPARC Open Access Newsletter, Sept. 2008. Available at

http://www.earlham.edu/~peters/fos/newsletter/09-02-08.htm (March 2010).

Predictive validity of editorial decisions at an open access journal: A case study on *Atmospheric Chemistry and Physics*

Lutz Bornmann[1]*; Hans-Dieter Daniel*[1,2]

1 ETH Zurich, Professorship for Social Psychology and Research on Higher Education, Zähringerstr. 24,
CH-8092 Zurich
bornmann@gess.ethz.ch;
2 University of Zurich, Evaluation Office
Mühlegasse 21, CH-8001 Zurich
daniel@evaluation.uzh.ch

Abstract

In this study we investigate the quality of the selection process of an open access (OA) journal, taking as an example the journal *Atmospheric Chemistry and Physics* (ACP). ACP is working with a new system of public peer review. We examined the predictive validity of the ACP peer review system – namely, whether the process selects the best of the manuscripts submitted. We have data for 1111 manuscripts that went through the complete ACP selection process in the years 2001 to 2006. The predictive validity was investigated on the basis of citation counts for the later published manuscripts. The results of the citation analysis confirm the predictive validity of the editorial decisions at ACP: They covary with citation counts for the published manuscripts.

Keywords: public peer review, open access, predictive validity

1. Introduction

More than 4500 open access (OA) journals have now become established in science that either still use the traditional peer review system or have introduced the 'new' system of public peer review (see http://www.doaj.org/). "The difference compared to traditional. . .journals is that OA journals let authors retain the copyright, and that they have a different business strategy: they are free of charge at the point of use. . .Many – but not all – of the OA publishers adopt the 'author/institution pays' policy, that is, paying once and in advance and grant free access for everyone, worldwide" [1]. The greatest reservation about OA journals is whether they achieve adequate quality control [2]. "In the open-access business model, it is widely accepted that authors (or their funding agencies or universities) pay. This means that the journals' revenues depend directly on the number of articles published. Only fools would believe that editors wouldn't then tend to accept a manuscript in the many borderline cases" [3].

Taking as an example the journal *Atmospheric Chemistry and Physics* (ACP), we present the – according to our literature search – first results of an evaluation study on the quality of the selection process of an electronic OA journal. The study examines whether the ACP peer review system actually does select the 'best' manuscripts among those submitted. For that, the citation impact of papers is compared which, after a positive evaluation either in ACP or if rejected after a negative evaluation, were submitted and published elsewhere. As the number of citations of a publication reflects the international impact of the reported research and in the absence of other operationalizable indicators, it is a common approach in peer review research to evaluate the success of a peer review process on the basis of the citation count of the reviewed manuscripts [4]. According to Jennings [5] "the most important question is how accurately the peer review system predicts the longer-term judgements of the scientific community." Scientific judgements on manuscripts are said to show predictive validity in peer review research if the citation counts of manuscripts receiving different decisions differ to a statistically significant degree.

2. Methodology

ACP was launched in September 2001. It is produced and published by the European Geosciences Union (EGU) (www.copernicus.org/EGU/EGU.html) and the Copernicus Society (www.copernicus.org). ACP is freely accessible via the Internet (www.atmos-chem-phys.org). ACP has a two-stage publication process, with a 'new' peer review process consisting of a public peer review and interactive discussion [6, 7]. In the first stage, manuscripts that pass a rapid pre-screening process (access review) are immediately published as 'discussion papers' on the journal's Web site (as a result, they are published in Atmospheric Chemistry and Physics Discussions, ACPD). After the end of the discussion phase, based on the revised manuscript and in the light of the access peer review and interactive public discussion, the editor accepts or rejects the revised manuscript for publication in ACP.

For the investigation of peer review at ACP we had data for 1111 manuscripts that went through the complete ACP selection process in the years 2001 to 2006. These manuscripts reached one of the following final statuses: 958 (86%) were published in ACPD and ACP, 74 (7%) were published in ACPD but not in ACP (here, the editor rejected the revised manuscript), and 79 (7%) were not published in either ACPD or ACP (these manuscripts were rejected during the access review). Some of the manuscripts submitted to ACP but not published there (because they were rejected during the access review, for example) were submitted by the authors, as described in the following, to another journal and published there. According to Schultz [8], there are two reasons for the high publication rate of submissions to ACP [see also 9]: By using the public peer review and interactive discussion, (1) ACP can expect a high average quality of submitted manuscripts, and (2) ACP works harder than journals working with the traditional peer review to keep and improve the submissions.

For manuscripts published in ACP, ACPD or elsewhere, we determined the number of citations for a fixed time window of three years including the publication year. The citation analyses were based on the Science Citation Index (SCI, Thomson Reuters, Philadelphia, PA, USA), Chemical Abstracts (CA, Chemical Abstracts Services, Columbus, Ohio, USA) and Scopus (Elsevier, Amsterdam, The Netherlands).

3. Results

The search for the fate of the manuscripts that were not published in ACP (n=153) was conducted using two research literature databases, Web of Science (WoS, Thomson Reuters) and CA. Two Ph.D. environmental research scientists carried out the search. The results of the investigation revealed that of the 153 manuscripts, 38 (25%) were published in other journals. No publication information was found for 115 (75%) manuscripts, whereby 70 of the 115 manuscripts (61%) were published in ACPD. Other studies on the fate of manuscripts that were rejected by a journal reported percentages ranging from 28% to nearly 85% for manuscripts later published elsewhere [10], whereby the journals examined do not work with a two-stage publication process as does ACP. For manuscripts rejected by AC-IE at the beginning of the year 2000, Bornmann and Daniel [11] determined a percentage of 95%.

The 38 manuscripts that were published as contributions in other journals were published in 25 different journals within a time period of five years (that is, between 2005 and 2009). Six manuscripts were published in the *Journal of Geophysical Research*; three manuscripts were published in *Geophysical Research Letters*. The other 23 journals each published one or two manuscripts.

Table 1 shows the mean number of citations found in CA, SCI and Scopus for manuscripts published in ACP and ACPD (group 1), published in ACPD only or in ACPD and elsewhere (group 2), or published neither in ACP nor in ACPD, but elsewhere (group 3). The medians are printed in bold in the table since the median – unlike the arithmetic mean – is not affected by outliers. The high standard deviations indicate that the distributions of the citation counts are characterized by a multitude of outliers.

Table 1: Descriptive statistics about citation counts for manuscripts published in ACP and ACPD (group 1), published in ACPD only or in ACPD and elsewhere (group 2), or published neither in ACP nor in ACPD, but elsewhere (group 3).

Group	Statistic	CA	SCI	Scopus
Group 1	n	958.00	958.00	951.00
	mean	8.49	9.72	11.87
	sd	11.32	12.99	15.68
	median	6.00*	6.00$^{\$}$	7.00§
Group 2	n	74.00	74.00	51.00
	mean	1.76	2.04	3.82
	sd	2.69	2.83	4.47
	median	1.00*	1.00$^{\$}$	2.00§
Group 3	n	17.00	17.00	15.00
	mean	1.29	1.71	2.73
	sd	2.20	2.37	2.74
	median	0.00*	1.00$^{\$}$	2.00§

Notes: The citation counts were searched in the databases Chemical Abstracts (CA), Science Citation Index (SCI) and Scopus for a fixed three-year citation window. Since citation counts could not be searched for all manuscripts in the databases, the number of manuscripts in the table differs from the number of manuscripts stated in the methodology section.*

* $\chi^2 = 99.6$, $P < .001$; $^{\$}$ $\chi^2 = 108.2$, $P < .001$; § $\chi^2 = 56.7$, $P < .001$.

As the results in Table 1 show, independently of the literature database in which the citation search was conducted, manuscripts in group 1 are cited more frequently on average than those in group 3. For example, manuscripts that were published in ACP and ACPD (group 1) were cited, according to the SCI, on average 6 times (median); manuscripts that were published neither in ACP nor in ACPD, but elsewhere (group 3) were cited on average once (median). It is also evident that manuscripts in group 1 are cited much more frequently than those published only in ACPD or in ACPD and elsewhere (group 2). In contrast, hardly any differences are detectable between the median citation counts of group 2 and group 3 manuscripts. Regardless of the citation database, the differences between the three groups in Table 1 are statistically significant.

4. Discussion

Many OA journals come into being in recent years. It is hoped that unrestricted access to scientific publications will have a positive effect on scientific progress: According to Borgman [12], “scholarship is a cumulative process, and its success depends on wide and rapid dissemination of new knowledge so that findings can be discarded if they are unreliable or built on if they are confirmed. Society overall benefits from the open exchange of ideas within the scholarly community” (p. 35). Some of the OA journals are using public or open peer review, for one, in the interest of higher quality submissions: “Open review has the advantage of speeding and democratizing reviewing, and could result in better manuscripts being submitted” [13]. Furthermore, “reviewers would be more tactful and constructive” [14]. And for another, “there is a widely held suspicion (certainly amongst commercial publishers and to a lesser extent amongst authors) that articles in … OA journals are less well peer-reviewed than their counterparts in toll-access journals. This perception has two roots; firstly, as … OA journals are new, they have not yet had a chance to attain high status, and secondly, there is a feeling that because income depends on the number of accepted articles, the editors will be under pressure to accept poor quality manuscripts to keep the income stream up” [15].

Contrary to those fears, the results of this study show – in agreement with the results on various closed peer review systems of traditional journals [see an overview in 4] – that in the journal examined here, public peer review is able to assess the quality of manuscripts ‘validly’ and to select the ‘best’ manuscripts among the manuscripts submitted. The results of the citation analysis confirm the predictive validity of the editorial decisions: They correlate statistically significantly with citation counts. When interpreting these results, however, it should be taken into consideration that the ACP peer review system, through the high acceptance rate among submissions, in many cases exercises a different function that the peer review system at many traditional journals, such as at AC-IE: It is more about improving manuscripts prior to publication than about selecting among submissions. In the words of Shashock [16], journals like *Science*, *Nature*, or the AC-IE skim off the cream and discard everything else among the submissions. ACP, in contrast, in the first review step screens out unsuitable manuscripts only and eliminates them from the further selection process. Through the use of public peer review in the second review step, a large part of the manuscripts that in the access review were deemed potentially suitable for publication in ACP are published after varying degrees of revision.

5. Conclusions

For Anderson [17], open and closed peer review systems are each suitable for different publication environments: “Closed peer review works best in scarce environments, where many papers fight for a few coveted journal slots. Open peer review works best in an abundant environment of online journals with unlimited space. In the scarce world of limited pages in top journals, prestige is earned through those journals’ high standard and exclusivity. That comes, in part, from the process, which involves impressing the very discriminating combination of an editor and a few respected researchers.” Since the number of OA journals can be expected to increase in coming years, future studies on predictive validity should examine in particular their peer review systems. Here, studies are needed that investigate not only the selection function, as in this study, but also the improvement function of peer review.

Acknowledgements

We would like to thank Dr. Hanna Joos (at the Institute for Atmospheric and Climate Science of the ETH Zurich, Switzerland) and Dr. Hanna Herich (at EMPA – a research institution within the ETH Domain) for the investigation of the manuscripts rejected by *Atmospheric Chemistry and Physics* and published elsewhere. We also thank Dr. Werner Marx and Dr. Hermann Schier of the Central Information Service for the institutes of the Chemical Physical Technical (CPT) Section of the Max Planck Society (located at the Max Planck Institute for Solid State Research in Stuttgart, Germany) for conducting the citation search for citations of the accepted and rejected (but published elsewhere) manuscripts in the literature database Chemical Abstracts (CA) and Science Citation Index (SCI).

The research project, which is investigating quality assurance of interactive open access journals, is supported by a grant from the Max Planck Society (Munich, Germany). We thank Dr. Ulrich Pöschl, Chief Executive Editor of *Atmospheric Chemistry and Physics*, the Editorial Board of *Atmospheric Chemistry and Physics*, and Copernicus Publications (Katlenburg-Lindau, Germany) for permission to conduct the evaluation of the selection process of the journal, and thank the members of Copernicus Systems + Technology (Berlin, Germany) for their generous technical support during the carrying out of the study.

References

[1] GIGLIA, E. Open Access in the biomedical field: a unique opportunity for researchers (and research itself). Europa Medicophyisica, 2007, vol. 43, no. 2, p. 203-213, p. 208.

[2] JOINT INFORMATION SYSTEMS COMMITTEE *Journal authors survey report*. Truro, UK: Key Perspectives Ltd., 2004.

[3] GÖLITZ, P. Twitter, Facebook, and Open Access ... Angewandte Chemie International Edition, 2010, vol. 49, no. 1, p. 4-6.

[4] BORNMANN, L. Scientific peer review. Annual Review of Information Science and Technology, in press.

[5] JENNINGS, C.G. Quality and value: the true purpose of peer review. What you can't measure, you can't manage: the need for quantitative indicators in peer review. 2006. Retrieved July 6, 2006, from http://www.nature.com/nature/peerreview/debate/nature05032.html.

[6] KOOP, T. AND PÖSCHL, U. Systems: an open, two-stage peer-review journal. The editors of *Atmospheric Chemistry and Physics* explain their journal's approach. 2006. Retrieved 26 June 2006, from http://www.nature.com/nature/peerreview/debate/nature04988.html.

[7] PÖSCHL, U. Interactive journal concept for improved scientific publishing and quality assurance. Learned Publishing, 2004, vol. 17, no. 2, p. 105-113.

[8] SCHULTZ, D.M. Rejection rates for journals publishing atmospheric science. Bulletin of the American Meteorological Society, 2010, vol. 91, no. 2, p. 231-243.

[9] PÖSCHL, U. Interactive Open Access publishing and peer review: the effectiveness and perspectives of transparency and self-regulation in scientific communication and evaluation. LIBER Quarterly, submitted.

[10] WELLER, A.C. *Editorial peer review: its strengths and weaknesses*. Medford, NJ, USA: Information Today, Inc., 2002.

[11] BORNMANN, L. AND DANIEL, H.-D. The effectiveness of the peer review process: inter-referee agreement and predictive validity of manuscript refereeing at *Angewandte Chemie*. Angewandte Chemie International Edition, 2008, vol. 47, no. 38, p. 7173-7178.

[12] BORGMAN, C.L. *Scholarship in the digital age. Information, infrastructure, and the Internet*. Cambridge, MA, USA: MIT Press, 2007.

[13] BORGMAN, p. 61.

[14] DECOURSEY, T. Perspective: The pros and cons of open peer review. Should authors be told who their reviewers are? 2006. Retrieved June 20, 2006, from http://www.nature.com/nature/peerreview/debate/nature04991.html.

[15] OPPENHEIM, C. Electronic scholarly publishing and open access. Journal of Information Science, 2008, vol. 34, no. 4, p. 577-590, p. 582.,

[16] SHASHOK, K. Standardization vs diversity: how can we push peer review research forward? Medscape General Medicine, 2005, vol. 7, no. 1, p. 11.

[17] ANDERSON, C. Technical solutions: wisdom of the crowds. Scientific publishers should let their online readers become reviewers. 2006. Retrieved June 15, 2006, from http://www.nature.com/nature/peerreview/debate/nature04992.html. para. 14, 15.

Search engine in a class of academic digital libraries

Maria Nisheva-Pavlova, Pavel Pavlov

Faculty of Mathematics and Informatics, Sofia University
5 James Bourchier Blvd., Sofia 1164, Bulgaria
{marian, pavlovp}@fmi.uni-sofia.bg

Abstract

The paper discusses some aspects of an ongoing project aimed at the development of a methodology and proper software tools for building and usage of academic digital libraries. A particular functional model of academic digital library has been proposed and analyzed. The emphasis falls on some solutions of the large set of problems concerning the development of adequate mechanisms for semantics-oriented search in multilingual digital libraries. An ontology-based approach is suggested in order to standardize the semantic annotation of the library resources and to facilitate the implementation of the functionality of the search engine. The main features of a prototype of knowledge-based search engine for a multilingual academic digital library with research and learning materials are discussed. This search engine uses proper ontologies describing the conceptual knowledge considerable for the chosen domains and in this way it is capable of retrieving and filtering documents by their semantic properties.

Keywords: Digital Library; Metadata; Semantic Annotation; Ontology; Search Engine

1. Introduction

Research and practical activities in the field of Digital Libraries during the last two decades lead to significant results in the development and management of digital collections, in the innovation in scholarly publishing and the long-term preservation of digital information. Many institutions are actively involved in building suitable repositories of the institution's books, papers, theses, and other works which can be digitized or were “born digital”. In particular, universities and other academic institutions participate successfully in lots of projects directed to the development of different types of *academic digital libraries*. Academic digital libraries are committed to maintaining valuable collections of scholarly information. To this end, essential information resources should remain available and accessible into the future – a real challenge in the cases of digital resources that are increasingly transient and at risk.

The paper is aimed at the presentation of an ongoing project which is directed to the development of a methodology and corresponding software tools for building academic digital libraries. A special attention has been paid to the elaboration of means for semantics-oriented search in multilingual digital libraries. The study and the practical experiments are oriented to the development of DigLib-CI – a digital library with research and learning materials (articles, dissertations, monographs, lecture notes, textbooks, presentations, example program sources, data sets, quizzes, manuals etc.) created at the Department of Computer Informatics of the Faculty of Mathematics and Informatics (FMI), Sofia University, or especially selected from among the scholarly materials freely available on the Web.

2. Related Work

Digital Libraries can mainly be characterized as a converging point where disparate communities have been meeting to address common issues related with the creation, management and usage of digital information [1]. The goal of a digital library and especially of an academic library is to provide access to selected intellectual works. Moreover, academic digital libraries are usually aimed at some specific challenges like digital preservation of valuable scientific heritage collections and investigation of innovative methods for automatic indexing, metadata extraction, document search and retrieval etc. In this sense, academic digital libraries are the front-rankers in the discussed area.

The digital libraries of Cornell University [2], the University of Michigan [3] and Carnegie Mellon University [4] are considered as leaders in the field of academic digital library creation and management.

The Cornell University Library is the eleventh largest academic library in the United States, ranked by number of volumes held. In 2005 it held 7.5 million printed volumes in open stacks, 8.2 million microfilms and microfiches, and a total of 440,000 maps, motion pictures, DVDs, sound recordings, and computer files in its collections, in addition to extensive digital resources and the University Archives.

The Cornell Library Digital Collections Project integrates online collections of historical documents. Featured collections include the Database of African-American Poetry, the Historic Math Book Collection, the Samuel May Anti-Slavery Collection, the Witchcraft Collection, and the Donovan Nuremberg Trials Collection.

The University of Michigan Digital Library Project (UMDL) is based on the traditional values of service, organization, and access that have made libraries powerful intellectual institutions in combination with open, evolving, decentralized advantages of the web. The content of UMDL will emphasize a diverse collection, focused on earth and space sciences, which can satisfy the needs of many different types of users. The content will be supplied by publishers, although the project will eventually allow all users to publish their work.

The implementation of the current prototype of UMDL requires the integration of numerous agent technologies for knowledge exchange, commerce, learning, and modelling. Recently, the efforts have been concentrated on developing technologies that, for example, manipulate ontological descriptions of the elements of a digital library to help agents find services and auctions for exchanging goods and services under various conditions. These technologies allow flexibility in the UMDL configuration policies, extensibility and scalability by using demand as incentive for replicating services.

Carnegie Mellon University Libraries became very popular with the Million Book (or the Universal Library) project which was aimed to digitize a million books by 2007. The activities within the project include scanning books in many languages, using OCR to enable full text searching, and providing free-to-read access to the books on the web. As of 2007, they have completed the scanning of the planned number of books and have made accessible the corresponding database.

The research within the Million Book project includes developments in machine translation, automatic summarization, image processing, large-scale database management, user interface design, and strategies for acquiring copyright permission at an affordable cost.

Compared to these well-known large scale initiatives, our project is of a significantly smaller scale, but in contrast to all of them, it investigates the use of a set of subject ontologies to provide flexible, semantics-oriented access to the library resources for users with different profiles and language skills.

3. Architecture of DigLib-CI

DigLib-CI is designed as a typical academic digital library. It has been under development at FMI in order to provide open access to various kinds of scholarly and instructional content, mainly in a wide range of subfields of Computer Science and Information Systems. The functional structure of DigLib-CI is shown in Figure 1.

The content repositories include research and learning materials of different types (books, dissertations, periodicals and single articles, manuals, lecture notes, presentations, source code of computer programs, data sets, tests, quizzes etc.) in the areas of Computer Science and Information Systems. These library resources are available in various digital formats: pdf, html, plain text, doc, ppt, jpeg etc. Most of them are developed by faculty members, the others are especially selected from among the scholarly materials freely available on the Web. The content repositories are stored in a small number of locations. The materials in them are written in Bulgarian or in English language.

The metadata catalogues are destined to facilitate the identification of the needed research or learning materials by the search engine. They contain descriptive metadata stored in XML format and support the reusability of all library resources and facilitate their interoperability.

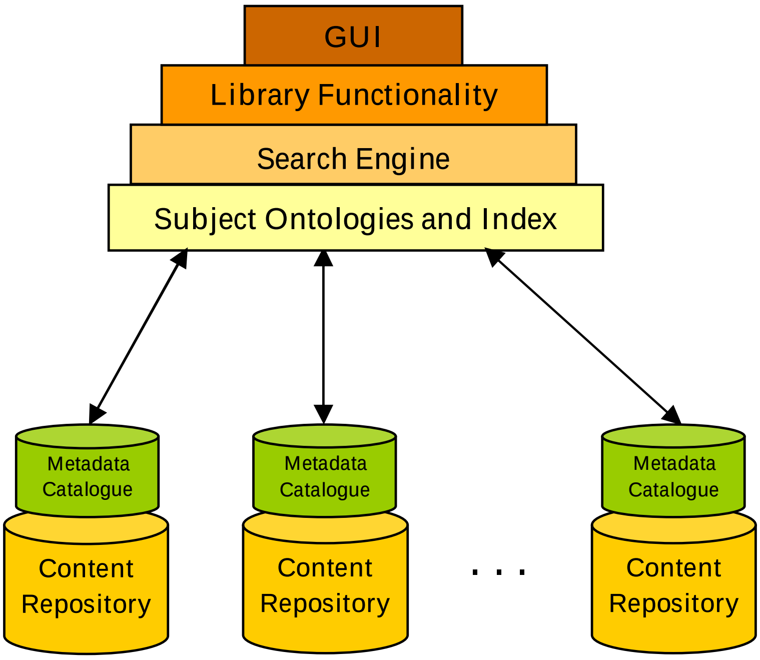


Figure 1: Functional Model of DigLib-CI

The subject ontologies include large sets of concepts of the areas of Computer Science and Information Systems, with description of their properties and different kinds of relationships among them. They play a significant role in the implementation of the full functionality of the search engine.

The purpose of the search engine is to provide adequate access to the complete palette of resources stored in DigLib-CI.

The library functionality and the user interface of DigLib-CI are designed in accordance with the expected needs and requirements of the basic types of users of the library. The interface module provides adequate online access to the corresponding library resources and supporting software tools.

4. Catalogue Metadata

The library catalogues contain metadata which support the identification of the requested resources by the search engine. These metadata are stored in XML format and comply with the IEEE Standard for Learning Object Metadata [5].

Typical examples of relevant attributes of most kinds of research and learning materials are: type of the material; author; title of the material; language(s) (human and/or programming one(s)); digital format; location; version; date of creation; completion status; restrictions on use; semantic annotation – list of concepts from a proper subject ontology describing the Computer Science or Information Systems subfields and/or concepts covered or treated by the material. Learning materials have been characterized also by their educational level and the principal types of users for which the corresponding material was designed; officially published research materials and textbooks are supplied with the corresponding bibliographic metadata.

Each catalogue entry (i.e., each resource description) consists of two equivalent parts in which the element values are texts in Bulgarian or English language, respectively. The search engine examines the corresponding parts of the descriptions according to the language of the user query.

The elements <ontologyRefs> and <keywords> of the resource descriptions play the role of semantic annotations of the corresponding library materials. The values of the child elements of <ontologyRefs> are concepts of the suitable subject ontologies (names of classes in these subject ontologies) which present most precisely the content of the corresponding document.

The concepts of the subject ontologies are too general from the point of view of the expectations of the typical users of DigLib-CI. For that reason one can include in the resource descriptions additional lists of keywords which describe the content of the corresponding documents at the necessary level of abstraction. These keywords are set as values of the child elements of the <keywords> resource description elements.

The names of the respective subject areas and names of the files containing the suitable subject ontologies have been assigned as values of the child elements of the catalogue description elements <subjects> and <ontologies> respectively.

5. Subject Ontologies

The subject ontologies include a large set of concepts in the fields of Computer Science and Information Systems, with description of their properties and the different kinds of relationships among them. Two subject ontologies are included in the current version of DigLib-CI. The Computer Science ontology is based on the Computer Science Curriculum 2008 of ACM and IEEE/CS [6]. Using the curriculum as a guideline, this ontology defines the atomic knowledge units for the University courses and available research materials in the field of Computer Science and makes them sharable and reusable. Its current version includes approximately 300 concepts with their relationships.

The Information Systems ontology has been under development using the Model Curriculum and Guidelines for Undergraduate Degree Programs in Information Systems of ACM, AIS and AITP [7].

The subject ontologies are designed in order to play the role of information sources describing the hierarchy and the other relationships between the main concepts in the discussed domains. A dictionary of synonyms has also been under development with the purpose of providing the search engine with other viewpoints to the conceptual structure of the areas of Computer Science and Information Systems.

The body of knowledge in the areas of Computer Science and Information Systems is formulated in the terms of a considerable number of common concepts, therefore the two subject ontologies discussed above contain many common classes (with equal or similar names and intersecting properties and restrictions on them). Because of that our further plans include the development of an approach to the integration of domain ontologies relevant to the contents of multilingual academic digital libraries which will be based on some of our former results [8].

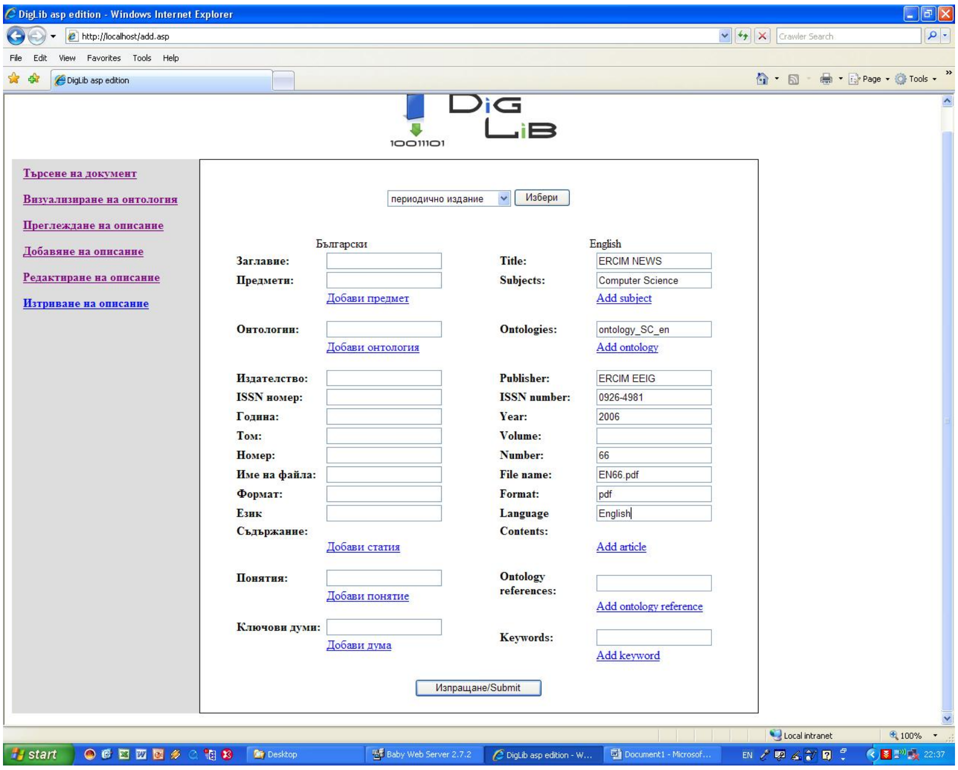
6. User Interface

The library functionality and the user interface of DigLib-CI are designed in accordance with the expected requirements of the basic types of users of the library. The interface module provides adequate online access to the corresponding library resources and supporting software tools.

The current version of the user interface allows one to formulate queries in Bulgarian or English language. It is intended for four types of users:

- FMI students – they may read/download textbooks, open lecture notes and presentations from all public sections of the library as well as all manner of other kinds of materials (monographs, dissertations, articles, periodicals, degree theses, lecture notes, presentations, exercises, programs, data sets, quizzes, tests etc.) from fixed public library sections;
- FMI lecturers and researchers – in addition to the students' access rights, they may upload materials to fixed public sections as well as create and update private sections and use materials in some of them;
- librarians (library administrators) – they have full access to all public resources of the library (may download and upload materials destined for all public sections of the library);
- general citizen – they may read and download public materials of fixed types (e.g., dissertations, textbooks, open lecture notes and presentations).

All types of users of DigLib-CI may use the standard input interface which provides convenient means for entering, editing and submitting queries for various kinds of document search and retrieval. FMI lecturers and researchers as well as the library administrators may play the role of authors of library resources and have an access to the author's part of the user interface. This part of the user interface places at the authorized persons' disposal appropriate forms enabling one to enter and edit catalogue descriptions of all types of library resources (Figure 2). More precisely, the user may enter the values of some of the elements or pick out the values of others from previously drawn lists. In particular, the available subject ontologies can be properly visualized and the necessary concepts in them can be picked out as values of the child elements of the element <ontologyRefs>.



	Български		English
Заглавие:		Title:	ERCIM NEWS
Предмети:		Subjects:	Computer Science
	Добави предмет		Add subject
Онтологии:		Ontologies:	ontology_SC_en
	Добави онтология		Add ontology
Издателство:		Publisher:	ERCIM EEIG
ISSN номер:		ISSN number:	0926-4981
Година:		Year:	2006
Том:		Volume:	
Номер:		Number:	66
Име на файла:		File name:	EN66.pdf
Формат:		Format:	pdf
Език		Language	English
Съдържание:		Contents:	
	Добави статия		Add article
Понятия:		Ontology references:	
	Добави понятие		Add ontology reference
Ключови думи:		Keywords:	
	Добави дума		Add keyword

Figure 2: User interface of DigLib-CI (author's view – form for entering catalogue metadata of periodicals)

7. Working Principles of the Search Engine

The purpose of the search engine is to provide adequate access to the complete palette of resources stored in DigLib-CI.

The search engine maintains several types of search and document retrieval within DigLib-CI. The user queries define restrictions on the values of certain metadata attributes of the required research or learning materials. Generally, the search mechanism may be formulated as follows: the document descriptions included in all permissible user sections of the library are examined one by one and these descriptions which have a specific element (determined by the type of the user query) with a value matching the user query, are marked in order to form the search result. The matching process is successful if the value of the element or the value of one of its child elements is equal to the user query. The documents pointed by the marked descriptions are retrieved and the user is given an access to these documents and their catalogue descriptions.

The current implementation of the search engine supports four types of search and document retrieval:

- full search – search and retrieval of all available library resources, ordered by title, by author, by category, by date of creation or by date of inserting in the library;
- author search (search and retrieval of the documents created by a given author) – the search is performed in the value of the element <authors>;
- ontological search – the search is performed in the value of the element <ontologyRefs>;
- keyword search – the search is performed in the value of the element <keywords>.

During the ontological search the user query is augmented with regard to the concepts searched out in the semantic annotations of the required research or learning materials. The more specific concepts from each of the subject ontologies indicated by the user are added to the original one in the resulting query. Then the search engine retrieves all documents in the library containing in their descriptions at least one component of the augmented query as the value of a child element of <ontologyRefs>. In this way the ontological search enables one to find documents described by ontology concepts which are semantically related to the concept defining the user query.

Till now, we have no disposal of an accomplished proper dictionary of synonyms of the concepts in the areas of Computer Science and Information Systems neither in Bulgarian, nor in English, but our idea is to provide a possibility for two-stage augmentation of the user query. At the first stage the request for ontological search will be extended with the more specific concepts (its successors) from the indicated subject ontologies. At the second stage the synonyms found in the dictionary will be added to the main (given by the user) concept and its successors.

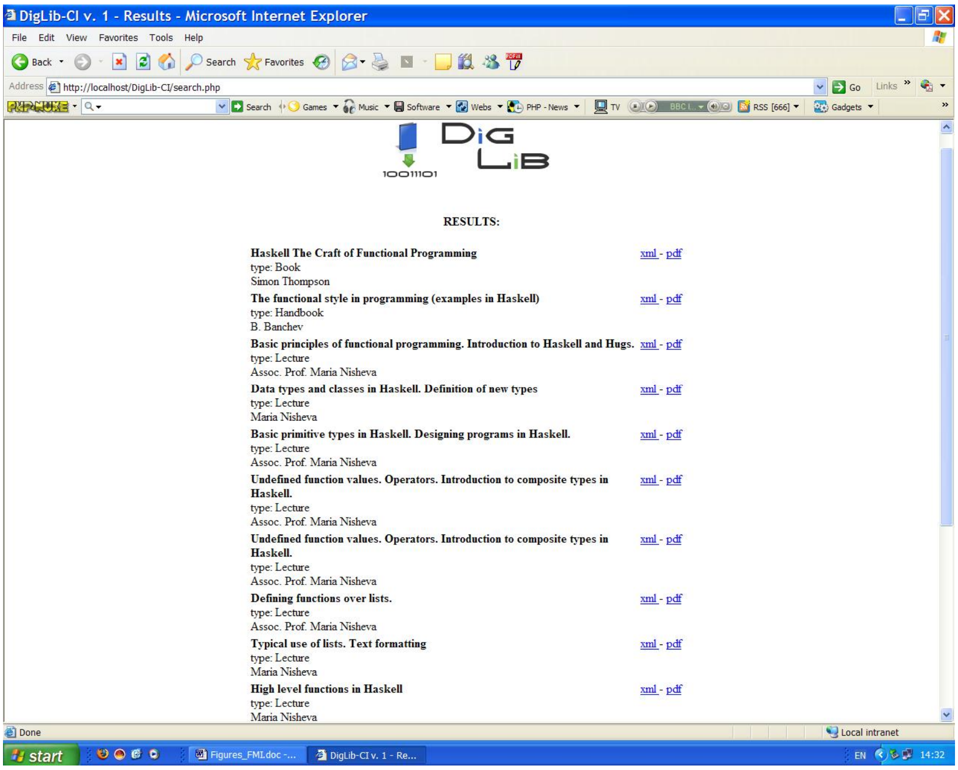
We allow in the current version of the implementation of the search engine only "atomic" user queries that do not contain conjunctions or disjunctions of words or phrases. The next step will be to elaborate a sophisticated version of the search engine which will be capable to analyze and execute queries in the form of conjunctions or disjunctions of phrases of interest for the user. Some of our former ideas suggested in [9] will be used for the purpose.

The discussed working principles of the search engine of DigLib-CI are designed in order to support flexibility, interoperability and reusability. These principles could be applied in the implementation of the search engines of a whole class of academic digital libraries that provide semantics oriented access to their resources.

8. An Example of Ontological Search

Let us suppose for example that the user defines a request (a query) for ontological search concerning the concept “fundamental constructs”. First an extension of this request will be generated. It will include all ontological concepts which are special cases of the concept given by the user (with respect to the ontologies indicated by the user). For this purpose, breadth-first search in the graphs that represent the ontologies will be performed, starting in each one from the concept chosen by the user.

Assume that the Computer Science ontology is chosen by the user. In this case the extended request (the augmented query) will include the concepts “fundamental constructs”, “basic syntax and semantics”, “binding and scope”, “conditional structures”, “declarations”, “expressions”, “functions and procedures”, … , “variables”, “bindings”, “blocks”, … , “simple variables”.


DigLib-CI v. 1 - Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost/DigLib-CI/search.php
Go Links
DiG LiB
10011101
Done
Local intranet
start
Figures_FMI.doc -...
DigLib-CI v. 1 - Re...
EN 14:32

RESULTS:

Haskell The Craft of Functional Programming xml - pdf
type: Book
Simon Thompson

The functional style in programming (examples in Haskell) xml - pdf
type: Handbook
B. Banchev

Basic principles of functional programming. Introduction to Haskell and Hugs. xml - pdf
type: Lecture
Assoc. Prof. Maria Nisheva

Data types and classes in Haskell. Definition of new types xml - pdf
type: Lecture
Maria Nisheva

Basic primitive types in Haskell. Designing programs in Haskell. xml - pdf
type: Lecture
Assoc. Prof. Maria Nisheva

Undefined function values. Operators. Introduction to composite types in Haskell. xml - pdf
type: Lecture
Assoc. Prof. Maria Nisheva

Undefined function values. Operators. Introduction to composite types in Haskell. xml - pdf
type: Lecture
Assoc. Prof. Maria Nisheva

Defining functions over lists. xml - pdf
type: Lecture
Assoc. Prof. Maria Nisheva

Typical use of lists. Text formatting xml - pdf
type: Lecture
Maria Nisheva

High level functions in Haskell xml - pdf
type: Lecture
Maria Nisheva

Figure 3: Some search results for the query “fundamental constructs”

After that, a consecutive search in the catalogue descriptions follows. In this search all documents with descriptions that are juxtaposed with at least one element of the extended request are extracted. In the current implementation each document appears as many times in the result list, as many elements of the augmented query are juxtaposed with its description (which means that the element <ontologyRefs> of the description includes a sub-element that has value, coincident with an element of the augmented query).

Figure 3 shows a screenshot displaying part of the ontological search results for the query “fundamental constructs”.

If the user indicates more than one subject ontology (e.g., the Computer Science ontology and the Information Systems ontology), the procedure described above is repeated consecutively for each of these ontologies.

Our current activities are directed to the selection of a proper set of relationships between the ontology concepts that should be taken into account in the process of ontological search along with the hierarchical ones. We envisage for the near future the development of a more flexible and user-friendly mechanism for ontological search which will not expect the user to indicate explicitly the subject ontologies appropriate for every particular case.

9. Conclusions

The most considerable results of the discussed project obtained so far may be summarized as follows:

- A functional model of an academic digital library was proposed. This model provides tools for semantics oriented access to learning and research materials in various digital formats written in different languages;
- A prototype of DigLib-CI – an academic digital library with research and learning materials in the areas of Computer Science and Information Systems, was developed.

The main advantage of the suggested approach to building academic digital libraries consists in the provided facilities for flexible and adequate semantics-oriented access to the library resources for users with various professional profiles and language skills.

The complete implementation of the project will help to enhance the research activities and the exchange of teaching innovation and thus will improve the overall scholarly and teaching quality in Computer Science and Information Systems at FMI. It will also contribute to the methodology of development of innovative software systems maintaining the entire lifecycle of academic digital content.

Acknowledgements

This work has been partly funded by the Sofia University SRF within a project titled "Methods and Tools Supporting the Lifecycle of Rich Digital Content".

References

[1] BORBINHA, J. The Age of the Digital Library. In: D. Castelli, E. Fox (Eds.), Pre-proceedings of the first International Workshop on Foundations of Digital Libraries, Vancouver, Canada, 2007, pp. 31-36.

[2] Cornell University Library Digital Collections. Available at http://cdl.library.cornell.edu/ (March 2010).

[3] University of Michigan Digital Library. Available at http://www.si.umich.edu/UMDL/ (March 2010).

[4] Carnegie Mellon University Libraries: Digital Collections. Available at http://diva.library.cmu.edu/ (March 2010).

[5] IEEE Standard for Learning Object Metadata. Available at http://ltsc.ieee.org/wg12/20020612-Final-LOM-Draft.html (March 2010).

[6] Association for Computing Machinery; IEEE Computer Society. Computer Science Curriculum 2008: An Interim Revision of CS 2001. Available at http://www.acm.org/education/curricula/ (March 2010).

[7] ACM; AIS; AITP. IS 2002: Model Curriculum and Guidelines for Undergraduate Degree Programs in Information Systems. Available at http://www.acm.org/education/ (March 2010).

[8] ZLATAREVA, N; NISHEVA, M. Alignment of Heterogeneous Ontologies: A Practical Approach to Testing for Similarities and Discrepancies. In: D. Wilson, H. Chad Lane (Eds.), Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference, AAAI Press, Menlo Park, CA, 2008, pp. 365-370.

[9] PAVLOV, P; NISHEVA-PAVLOVA, M. Knowledge-based Search in Collections of Digitized Manuscripts: First Results. In: Proceedings of the 10th ICCC International Conference on Electronic Publishing, FOI-Commerce, Sofia, 2006, pp. 27-35.

Reliable scholarly objects search and interchange framework

Victor Torres[1]; Ruben Tous[2]; Jaime Delgado[2]

1 Departament de Tecnologies de la Informació i les Comunicacions,
Universitat Pompeu Fabra (UPF)
Carrer Roc Boronat 128, 08018 Barcelona, Spain
victor.torres@upf.edu;
2 Departament d'Arquitectura de Computadors,
Universitat Politècnica de Catalunya (UPC)
Campus Nord. Carrer Jordi Girona 1-3, 08034 Barcelona, Spain
{rtous, jaime.delgado}@ac.upc.edu

Abstract

Authors of scholarly objects might fear that there is a potential risk that the original material they publish in online sites or that they submit for evaluation to scientific journals or conferences is used by others as their own material. In such cases, it would not be easy for the original authors to prove authorship of the original contribution. In similar circumstances, it is very difficult to prove the authorship or origin of some materials that are being distributed amongst social networks, private or institutional websites or any other means through the Internet, namely documents, papers, images, data, etc. Those materials can be easily plagiarised (e.g. partially or totally translated) and redistributed without any control and with no means to prove authorship. In this context, we propose an online framework for the registration, search, interchange and trade of scholarly objects, which helps to overcome the potential drawbacks of online distribution and publishing. This framework acts as an intellectual property repository and sales point, where people is able to register content and determine the way they want to trade it, while providing innovative search capabilities based on the MPEG Query Format standard [1]. Creative Commons (CC) [2] limitations are identified and overcome by means of a licensing approach that combines Rights Expression Languages and the MPEG-21 Media Value Chain Ontology [3].

Keywords: intellectual property rights; scholarly objects; creative commons; multimedia information retrieval.

1. Introduction

In general, it is very difficult to prove the authorship or origin of some materials that are being distributed amongst social networks, private or institutional websites or any other means through the Internet, namely documents, papers, images, data, etc.

Although there are some initiatives focused to detect plagiarism [4] regarding well-known contributions to literature, it is very difficult to prove authorship for other minor or recent works that are not yet consolidated or present in global databases. Those materials can be easily plagiarised and redistributed without any control and even partially or totally translated.

In this paper, we analyse current approaches and initiatives that deal with intellectual property (IP) rights, determining up to which point they can be considered a secure means for protecting IP from the authors' perspective. After this analysis, we describe the desirable features that an ideal system would have. This framework would act as an intellectual property repository and sales point, where people would be able to register content and determine the way they want to trade it, while providing innovative search capabilities based on the MPEG Query Format standard [1]. Creative Commons (CC) [2] limitations will be identified in section 1.3 and overcome by means of a licensing approach that combines the flexibility of rights expression languages and the MPEG-21 Media Value Chain Ontology [3].

2. Intellectual property, services and initiatives

Intellectual property rights is the set of rights that correspond to authors and other entities (artists, producers, broadcasters, companies, etc.) with respect to works and other types of creations and inventions [5].

Copyright rights apply to literary and artistic works (e.g. written compositions, musical works, photographs, paintings, etc.) and they involve economic rights regarding the work reproduction, distribution, public performance, adaptation and translation and moral rights regarding the right to claim authorship and the right of integrity [6].

2.1. Copyright protection

In general, in most countries, any document, work, or creative project is protected by copyright by virtue of its creation from the date it is created. The inclusion of the author's name, date of creation and a copyright statement or the symbol "©" within or accompanying the work is a valid means for declaring copyright. However, the presence of this statement does not fully protect the author in case of litigation. Other types of qualified proofs such as written or documentary evidence of the date and time of registration are needed to be sure that a work is safely protected.

2.2. Intellectual Property registry offices

Intellectual Property (IP) registry offices, which usually depend on national governments, provide a mechanism for registering and proving content authorship in both the analogue and digital world.

Although the inscription of content in such registries is not compulsory, they are useful to provide qualified proofs stating that copyright exists for a work and it belongs to someone. Some intellectual property registries already offer online registration facilities [7] [8], easing authors the tedious process of the traditional manual and on-site registration. However, those registries lack other functionalities than the mere registration, such as the interaction with other applications that build upon them via APIs, the possibility for authors to determine other licensing schemes than the de facto “all rights reserved”, powerful searching facilities and even trading options.

2.3. Creative Commons licensing

Creative Commons (CC) [2] is a non-profit organisation that provides a set of reference licensing models that can be used by authors which hold IP rights to enable people to easily change their copyright terms from the default of “all rights reserved” to “some rights reserved”, while being consistent with the rules of copyright.

Following the CC approach, authors can mark their content with some specific licenses that grant some permissions regarding copyright rights to anyone that accesses the content. It is relevant to remark the importance for customers of being capable to prove that they own the appropriate rights for using a specific content.

We could imagine an editor that is used to work with images subject to any of the CC licensing models, which are used for illustrating their online newspaper or blog with, let’s say, photographs obtained from the Flickr [9] site. It may happen that in a certain moment in time an image is licensed under the CC Attribution (BY) model [10], which lets others copy, distribute, display, and perform the copyrighted work, and derivative works based upon it, but only if they give credit the way the author requests. Later in time, the author of the image may decide to change the licensing model to the CC Attribution Non-Commercial (BY-NC) model [10], which lets others copy, distribute, display, and perform the work, and derivative works based upon it, but for non-commercial purposes only. What happens then? Which is the licensing model that applies to the image? According to the CC model, it would depend on the moment the content is accessed. That is, if the editor accessed the content after the change in the licensing model, the Attribution model would apply. Moreover, as stated in the CC BY-NC license [10], this license would be royalty-free and perpetual (for the duration of the applicable copyright). However, licensing in CC does not provide any proof on how or when content is accessed, since there is no link or association between content and its licensing model, so it would be the editor's responsibility to prove that the licensing model being offered in that moment was the appropriate one in case a legal dispute occurs. The main problems that arise from the usage of Creative Commons licenses are, thus:

1) The lack of protection for content authors or rights holders regarding the content commercial use. Enabling a commercial usage of content does not mean they resign a part of the income perceived by the party that exploits it. However, the possibilities for authors of perceiving any income are reduced, since CC licenses do not contemplate the possibility for stating such compensations. In general, content consumers do not have the initiative to reach an agreement with content authors. Thus, authors need to start a legal dispute, which is often a long, expensive and non-fruitful process.

2) The lack of protection for content authors or rights holders regarding the license duration. CC licenses grant perpetual rights for those that can prove the content was accessed under that specific licensing model. Any change in the licensing model being used for the content will not apply to the users that accessed the content under the previous model.

3) The lack of protection for content consumers, since they need to prove which is (or was) the licensing model applicable to the content they use in case of litigation.

Regarding the first problem, CC has defined the CCPlus (CC+) model [11], which enables authors to express where consumers can get rights beyond those granted by the CC license, which is a non-commercial license (e.g. CC BY-NC). The CCPlus license can include a link to an external site or service, which can be a specialized commercial license broker as e.g. gettyimages [12], or even an email address to be contacted by the consumer. The CCPlus approach is still not available at many sites.

Regarding the third problem, some initiatives already tackle it for specific fields. ImageStamper [13] is a free online tool that generates and keeps a timestamp that includes the image, the license that applies and the date. However, it is only useful for images.

2.4. Safe Creative

SafeCreative [14] is a global, free, open and independent intellectual property registry that allows creators and rights holders to register their works and obtain a valid proof suitable to be used on court hearings.

One of the main differences between SafeCreative and other registries is the possibility to state the rights that apply to works by means of predefined or customized licenses. CC and GNU [15] licenses are included between the templates offered to users, whereas, for other customized models, users need to provide their own specific text. The licensing model being applied can be changed any time by the content author or rights holder.

SafeCreative also provides proofs for content consumers that can be used to certify the licensing model being applied to the content when accessed. In that sense, SafeCreative solves the third issue identified previously, since it provides proofs for content consumers, while keeping track of licensing changes. In order to have reliable proofs, SafeCreative uses officially recognised timestamping services and accepts some X.509 digital certificates issued by trusted issuers. API interfaces are also provided to enable the integration of their services in other web applications.

2.5. Copyright associations

The Writers' Copyright Association (WCA), Webmaster's Copyright Association (WMCA) and Musician's Copyright Association (MCA) [16] [17] [18] are different associations that provide registration services for the type of content they deal with. WCA accepts literary work for film and television, books, poems, artwork, lyrics, teleplays, game shows, storyboards, animations and cartoons, web pages etc. WMCA deals with websites, e.g. zipped entire sites, flash movies, custom java scripts, etc. Finally, MCA accepts music files and documents containing scores.

They all function in the same manner. The user uploads a file, pays a fee and receives a registration number that should be applied to the front page of the author's work. If necessary, a Registry employee may produce registration information or material as evidence if legal or official guild action is initiated. The processing fees, which are common for the three associations are available at their sites. Although they provide a simple interface for authors, they act as an unofficial intellectual property registry, while providing some basic search functionalities to browse amongst their registry entries.

Some possible problems when relying on these 3 registries are related to its terms and conditions of use, available at their web sites: 1) any of the three associations provides a formal copyright; 2) they do not “verify the originality or authenticity of the material, make comparisons of registration deposits, provide any statutory protections, nor give legal advice”; 3) “In the unlikely event that said file is lost, corrupted, damaged or destroyed due to the WCA's failure to maintain reasonable care or by any other cause whatsoever, it is agreed between both parties that the liquidated damages for the loss of the manuscript shall be £1.00”.

3. Proposed System

In order to deal with the potential drawbacks of online distribution and publishing and respecting intellectual property rights, we propose an online framework for the registration, search and trade of scholarly objects.

A means to prove authorship is the first functionality needed for such framework. Thus, it would act as an intellectual property repository where people would be able to post their content prior to any other action they may want to do with them, such as submitting a paper for evaluation or publishing it elsewhere. Digital signatures applied to the XML [19] representation of works will be a reliable proof for authorship. Two approaches can be followed here. A digital signature from the framework will be trustable as long as we trust in the framework management. A digital signature from the user will require the usage of a recognised X.509 certificate and private key from the user’s side and will be a more reliable proof of registration. The combination of both approaches would be the optimum solution. We must say that even adopting these mechanisms, still some legal disputes may arise regarding content ownership. However, the chances of happening so will be clearly reduced and limited to some active thefts.

Another desirable feature is the possibility to trade or share the registered content. In this way, the proposed framework acts as a sales point, easing the distribution and commercialisation of content, and always giving digital evidence of all the transactions being executed in the system, not only for authors but also for customers. The rights to be considered for being traded are, on one hand, those involved in content creation and distribution, which are defined in the MPEG-21 Media Value Chain Ontology (MVCO) [3]: make adaptation, make instance and make copy (useful for determining the type of works that can be derived), distribute, produce, public communication and synchronization. On the other hand, we should consider those rights related to content consumption and fruition such as render/play, embed, extract, enlarge, diminish, enhance, etc., which are defined in the MPEG-21 Rights Expression Language (REL) [20]. With these two sets of rights we can refer to any action that can be exercised over the content both during creation and distribution and its consumption by final users. Apart from rights, we need to consider conditions, which restrict how rights can be exercised. MPEG-21 REL defines different types of conditions, amongst which we find the following: temporal (e.g. from/to or time interval), payment (e.g. flat or per use fees to be cleared), territorial (e.g. country or region) and the number of times the right can be exercised. Additionally, in order to cover a wider range of agreements, such as those offered in CC and others, some conditions should be added: attribution, exclusivity, non-territory. Finally, a new condition should be considered: the possibility to keep a percentage of the income generated by derived works.

Once authors can determine the rights they want to trade and the applicable conditions, the same framework can act as a sales point where other users (business or end users), to whom we call consumers, could get and clear licenses that grant them some rights or permissions under certain conditions. A license would, thus, formalise the ownership of rights by a consumer and the related conditions. In that context, both authors and consumers would need to be registered so as to be able to identify them and generate the corresponding licenses that act as proofs, since they are expressed as digitally signed XML documents.

Once we have authors and consumers identified in the system, we can also provide an advanced functionality that enables authors to trade content with everyone or just with a limited and selected set of trusted users. We could even decide to offer different conditions for different sets of selected users, depending on our needs or will. In fact, this can be seen as a social network functionality that empowers social relationships. It is important to remark that the author does not need for the permission of target users, since these targeted offers are not public and will be only accessible by target users when accessing that specific content in the framework or otherwise notified by the author, depending on the implementation. The only requirement is that the author is able to identify the target user by means of a nick-like identifier.

As we have seen with other initiatives, this framework can be managed by anyone, as long as security mechanisms are deployed so as to have a trustable system. In that sense, any of the following aspects will help: security audits, use of external timestamping services, use of digital signatures and recognised X.509 certificates. The business model relying under the framework may be diverse, varying from the payment model to the free model. It is worth noting that the free model may be free for the general public but not free for massive registration through applications that make use of a specific API. Other approaches may include the payment for some value-added services such as the usage of pseudonyms, advanced statistics, registration with more than one author, use of advanced and/or customised licensing models, preview images, etc., for which authors may appreciate a real and useful value so that they may be willing to pay.

In order to deal with the problems identified in section 2.3 for the Creative Commons (CC) approach, the proposed framework separates licensing into two parts. First, authors decide how content is to be traded, by editing the rights they offer and determining the conditions that apply to them. This edition is done through an intuitive and simple interface which hides the complexity of legal texts. Pre-defined templates are also available for common licenses. After selecting rights and conditions, an equivalent legal text is automatically produced. Finally, any rights acquisition is formalized by means of a digitally signed license, expressed in a Rights Expression Language (REL) (e.g. MPEG-21 REL [20]), which links together consumer identity, consumer rights and conditions and the content identification. Thus, a license acts as a proof for both the author and customer. Whenever the license does not state any temporal conditions, it will be forever. In general, a license will apply as long as conditions are fulfilled. In this way, licenses equivalent to the CC models can be generated, but with the possibility of adding new conditions such as those previously mentioned. On the other hand, authors will still be able to modify the rights they offer from a given moment, but without affecting any licenses that might have been acquired prior to the change. Another important feature proposed for the framework is the possibility to search for content. Authors and consumers should be able to express complex conditions to filter the potentially huge amount of documents. Current information retrieval technologies allow extending the traditional search functionalities beyond the traditional keywords-based or metadata-based querying. New approaches allow, for instance, searching for research papers containing potential image copyright infringements (through content based image retrieval techniques). It would be desirable that both the traditional and the advanced search functionalities would be provided through and open query interface for search, providing high expressive power to allow users formulate sophisticated conditions over the scholarly objects' metadata and contents (textual or audiovisual). We envisage that this interface is based on the MPEG Query Format (MPQF) [1]. MPQF is a recent standard of the MPEG standardization committee (i.e. ISO/IEC JTC1 SC29/WG11), which provides a standardized interface to multimedia document repositories, including but not limited to multimedia databases, documental databases, digital libraries, spatio-temporal databases and geographical information systems.

The MPEG Query Format offers a new and powerful alternative to the traditional scholarly communication model. MPQF provides scholarly repositories with the ability to extend access to their metadata and contents via a standard query interface, in the same way as Z39.50 [21], but making use of the newest XML querying tools (based in XPath 2.0 [22] and XQuery 1.0 [23]) in combination with a set of advanced multimedia information retrieval capabilities defined within MPEG. This would allow, for example, querying for journal papers by specifying constraints over their related XML metadata (which is not restricted to a particular format) in combination with similarity search, relevance feedback, query-by-keywords, query-by-example media (using an example image for retrieving papers with similar ones), etc. MPQF has been designed to unify the way digital material is searched and retrieved. This has important implications in the near future, when scholarly users' information needs will become more complex and will involve searches combining (in the input and the output) documents from different nature (e-prints, still images, audio transcripts, video files, etc.).

4. Results

In this section we present the system we have developed, which tackles some of the problems identified in section 2.3, and which can be used not only for dealing with scholarly objects, but also for musical compositions, audiovisual works and many other types of creations.

The Intellectual Property Operations System – Digital Shadow (IPOS-DS) [24] is a service-oriented architecture that consists of a main web application, accessible through a web browser, which interacts with different web services. It also includes a user desktop application which deals with the rendering of protected content. Figure 1 depicts the overall architecture. Further details can be found at [25] and [26].

IPOS-DS main features include: 1) Content registration and certification. The IPOS-DS system digitally signs an XML representation of any work registered in the system including the identification of the work and author. Content ownership is ensured and content lineage can be traced thanks to the presence of a reference to its ancestor (e.g. adaptation to work) in the representation. 2) Content licensing according to MVCO and MPEG-21 REL capabilities. Authors decide how they want to trade content and they can modify it any time without affecting previous purchases. Conditions include not only standard temporal limitations, territory restrictions, fees to be cleared and limited number of executions of the right, but also specific IPOS-DS conditions such as keeping percentage of the income generated by derived content, and determining for whom rights will be available to be acquired. Customers formalise the rights' acquisition through personal user-specific licenses. 3) Content access and monitoring. Content is encrypted and can only be accessed by those who have purchased a license. For those users that are entitled to access content, it can be stored in clear so that they can use it without Digital Rights Management (DRM) restrictions. IPOS-DS keeps track of the licenses being purchased and when content is accessed so that the authors can have detailed usage information. 4) IPOS-DS provides search interfaces based on main content metadata fields.

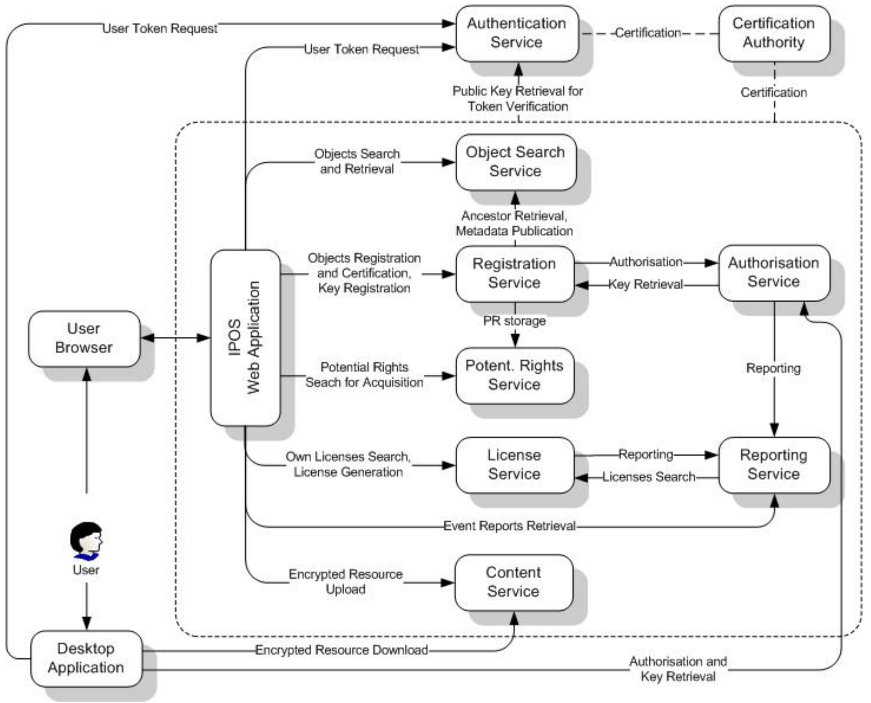


Figure 1: The IPOS-DS System

IPOS-DS is still being improved in some aspects in order to fulfil all the features proposed in section 3: 1) New conditions need to be added to generate Creative Commons-equivalent licenses: exclusivity, attribution. 2) A thorough usability analysis is needed (e.g. use of license templates). 3) Provide better searching capabilities by adopting the MPQF approach. 4)

Interface with official IP registries or recognised timestamping services in order to improve trust.

The IPOS-DS system was commissioned for development to the DMAG (Distributed Multimedia Applications Group) of the UPC (Universitat Politècnica de Catalunya) [27] by the company NetPortedItems S.L. (NPI) [24], which is responsible for its exploitation. It has been made accessible [24] for the public in a pre-exploitation phase.

Regarding the business model, IPOS-DS can be exploited independently by a private company, or even adopted by collecting societies, as it provides much added value by offering their constituents and other users that later may become members the benefit of their collective management services.

5. Conclusions

In this paper we have identified several initiatives that deal with the protection and management of intellectual property rights, which can be applied to scholarly objects.

After describing their main features and analysing their operation, we have identified the drawbacks of current systems and proposed a set of desirable functionalities that an intellectual property registry should have. Our proposal has been made with the aim not only to give protection to authors in terms of copyright but also to give them the freedom to trade their content and provide powerful and innovative searching capabilities in a standardised and automated manner.

Finally, we have also presented the IPOS-DS [24] system, which partially implements the proposed features and which will be extended to fulfil them.

Acknowledgements

Part of this work has been supported by the Spanish administration: Multimedia Content Management Life Cycle project, TEC2008-06692-C02-01.

References

[1] ISO/IEC 15938-12:2008 “Information Technology -- Multimedia Content Description Interface -- Part 12: Query Format”.

[2] Creative Commons, http://creativecommons.org/.

[3] ISO/IEC FDIS 21000-19. Information technology - Multimedia framework (MPEG-21) - Part 19: Media Value Chain Ontology.

[4] Turnitin, http://turnitin.com/

[5] World Intellectual Property Organization (WIPO). About Intellectual Property, http://www.wipo.int/about-ip/en/

[6] World Intellectual Property Organization (WIPO). Understanding Copyright and Related Rights, http://www.wipo.int/freepublications/en/intproperty/909/wipo_pub_909.html

[7] Spanish Intellectual Property Registry, https://wwws.mcu.es/RPIntelectual

[8] US Intellectual Property Registry, http://www.copyright.gov/eco/

[9] Flickr, http://www.flickr.com/

[10] Creative Commons licenses, http://creativecommons.org/licenses/

[11] CCPlus, http://wiki.creativecommons.org/CCPlus

[12] gettyimages, http://www.gettyimages.com/

[13] ImageStamper, http://www.imagestamper.com/

[14] SafeCreative, http://www.safecreative.org/

[15] GNU licenses, http://www.gnu.org/licenses/

[16] Writers' Copyright Association, http://www.wcauk.com/

[17] Webmasters Copyright Association, http://www.wmcaglobal.org/

[18] Musicians' Copyright Association, http://www.mcaglobal.org/

[19] XML-Signature Syntax and Processing W3C recommendation, http://www.w3.org/TR/xmldsig-core/.

[20] ISO/IEC, Information Technology – Multimedia framework (MPEG-21) – Part 5: Rights Expression Language, ISO/IEC 21000-5:2004.

[21] ISO 23950. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification.

[22] XML Path Language (XPath) 2.0. W3C Recommendation. 23 January 2007, http://www.w3.org/TR/xpath20/.

[23] XQuery 1.0: An XML Query Language. W3C Recommendation. 23 January 2007, http://www.w3.org/TR/xquery/.

[24] IPOS-DS (Intellectual Property Operations System – Digital Shadow), http://www.digitalmediavalues.com/.

[25] TORRES, V.; DELGADO, J. et al. A web-based rights management system for developing trusted value networks. Proc. of the 18th International World Wide Web Conference Developer's Track, p. 57-59.

[26] TORRES, V.; DELGADO, J. et al. Enhancing rights management systems through the development of trusted value networks. Proc. of the 7th International Workshop on Security in Information Systems, pp. 26-35.

[27] Distributed Multimedia Applications Group (DMAG), http://dmag.ac.upc.edu/.

A collaborative faceted categorization system – User interactions

Kurt Maly; Harris Wu; Mohammad Zubair

Old Dominion University, Norfolk, VA, USA

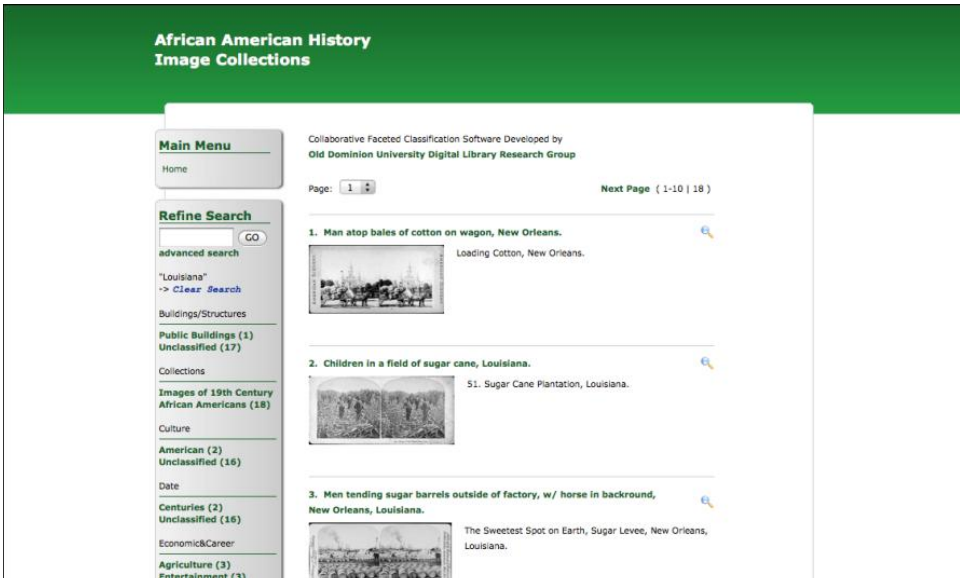
Abstract

We are building a system that improves browsing and searching access to a large, growing collection by supporting users to build a faceted (multi-perspective) classification schema collaboratively. The system is targeted in particular to collections of photographs and images that, in general, have few textual metadata. Our system allows users to build and maintain a faceted classification schema collaboratively and have the system help to classify documents into the evolving facet schema automatically. This paper focuses on the evolution of faceted classification schema for a large growing collection.

Keywords: collaborative faceted classification; schema enrichment; anomaly detection; user feedback; category ordering.

1. Introduction

We are building a system that improves browsing and searching access to a large, growing collection by supporting users to build a faceted (multi-perspective) classification schema collaboratively [1,3]. The system is targeted in particular to collections of photographs and images that, in general, have few textual metadata. A facet is an attribute (dimension) of an item in a collection that gives one perspective of that item. For example, in a collection of wine, “color” could be one facet. Other facets could be “origin”, “price”, etc. for the wine collection. This allows different users to navigate the collection using the facet of most interest to them. What is a good set of facets for a given collection is very much dependent on the given collection and the target users. Some example commercial sites that use facet-based classification are Amazon and eBay. The facets can evolve with time because of change in target users or change in interest of existing users in how they want to navigate the collection. For example, after a given facet schema has stabilized there may be a need to add another facet, for example, "healthy ingredients" for the wine collection. Some example categories in this facet are resveratrol, flavonoids, and non-flavonoids. For collections that grow in both volume and variety, a major challenge is to evolve the facet schema, and to reclassify existing objects into the modified facet schema. Centrally managed classification systems often find it difficult to adapt to evolving collections. It is hoped that through users' collective efforts the faceted classification schema will evolve along with the user interests and thus help them navigate through the collection quickly and intuitively. Our system (a) allows users to build and maintain a faceted classification collaboratively, (b) enriches the user-created facet schema systematically, and (c) classifies documents into an evolving, user-managed facet schema automatically. Readers can explore the current system by browsing the African History Image Collection on our website (http://facet.cs.odu.edu/), shown in Figure 1. In order for the faceted categorization to be effective in our system: (a) there needs to be a sufficient set of categories; (b) improper categories need to be removed, and (c) the schema's size needs to be regulated.



Collaborative Faceted Classification Software Developed by
Old Dominion University Digital Library Research Group

1. Man atop bales of cotton on wagon, New Orleans.

Loading Cotton, New Orleans.

2. Children in a field of sugar cane, Louisiana.

51. Sugar Cane Plantation, Louisiana.

3. Men tending sugar barrels outside of factory, w/ horse in backround, New Orleans, Louisiana.

The Sweetest Spot on Earth, Sugar Levee, New Orleans, Louisiana.

Figure 1: Screenshot of the system front page

In this paper, we focus on the enrichment and evolution of facet based classification for a large growing collection. First, we review the background of tagging and classification. Then we present an approach on schema enrichment that utilizes a statistical co-occurrence algorithm to produce possible new categories based on the existing metadata. In addition, we present an algorithm that harnesses the lexical power of WordNet in order to detect possible anomalies in the evolving category schema. We also present a statistical algorithm to visually rearrange the schema in the navigation panel of the user interface in order to minimize the time spent finding relevant items. Finally, we discuss an approach to capture user feedback on classifications produced by an automated process in order to control the quality of the overall classification.

2. Background

In this section we summarize previously reported related research [1]. Categories represent a way content is organized into a structure that is both meaningful and traversable and which allows for objects to be easily retrieved for later usage. Images, in particular, need such organization because an image itself is not easily searchable for any specific information that is useful to the requestor. A commonly used approach is "tagging" images with keywords which can later be searched for. However, tags do not fully allow for browsing a collection by selecting and narrowing down collective criteria. It is categories that allow for multiple images that share common traits to be arranged together and, consequently, found together. Faceted categorization is an extension to the common category structure. Facets allow for an image to belong to more than one collective criterion (the facet). Within each facet, a regular, multi-tier category structure is developed. By allowing an image to posses several descriptive categorizations, browsing for specific needs becomes much easier.

Traditionally, tagging and categorization in image classification systems have been the tasks of two dissimilar human groups. Tagging an image with keywords is generally the task of the users of the system. It represents their ability to associate what they are seeing with an idea or an object which they can easily recall later and search for. Very little input is needed by an administrative entity to collect and support such metadata. Faceted categorization systems, on the other hand, are typically created and maintained by a central entity. Facets and categories are created by the administrator and, with the exception of occasional changes, remain very much the same. As a result, many users' ideas of new classifications are not included in the schema which can potentially reduce the intuitiveness of browsing the collection.

One major obstacle for a user-created facet schema is its initial quality when compared to a centrally-compiled facet schema. Collaborative facet schema building depends on users' continuous improvement over time. Initially, a new facet may not contain all pertinent categories it should contain. The categories may have mixed-level or misplacement problems. For example, a user may create a "Location" facet with US states and cities as two levels of categories. Many states may be missing from the user-created facet initially. Some cities may be listed at the same level as states are listed. Some cities may be under the wrong states. Such data quality problems in the facet schema will burden users' classification efforts, and potentially lead to misclassification problems.

Several researchers have attempted to build topical ontology using metadata such as tags from social tagging. Reference [4] built a hierarchy of Flickr tags. Similarly, [5] built a concept hierarchy on the image collection provided by the ImageCLEF 2005 conference. Both studies adapted the subsumption model [2], a simple statistical co-occurrence model that identifies parent-child relationships: X subsumes Y if: P(x|y >= 0.8) and P(y|x < 1). For example, suppose X = "glass", and Y = "stained glass". If most documents tagged with "stained glass" are also tagged with "glass", then "glass" subsumes "stained glass". In [7] researchers built a hierarchy of del.icio.us tags using graph centrality analysis. In [8] a faceted classification of concepts was built using WordNet, a large lexical database of English [9]. References [9] and [10] have induced ontology using statistical NLP (natural language processing) techniques for textual documents.

Several research projects have also attempted to categorize items into a pre-existing ontology utilizing tags from social tagging. In [11] researchers were able to map about 75% of the tags from social tagging applications to the Dublin Core metadata standard elements (subject, date, title, description, format, publisher, etc.), so that the tags could be used for semantic web applications. The researchers also attempted to augment the Dublin Core standard with several proposed new elements, such as Action, Utility, Category, Depth, Notes and User name, so that another 20% of the tags can be mapped to these new metadata elements. In [12], the researchers built an image repository based on the hierarchical structure provided by Wordnet, utilizing search engine and large-scale human labelling efforts. The project used the service of Amazon Mechanical Turk, an online platform on which one can put up tasks for users to complete and to get paid. As of 2009, the repository collected 3.2 million human-labelled images and consisted of 12 subtrees (a small subset of Wordnet): mammal, bird, fish, reptile, amphibian, vehicle, furniture, musical instrument, geological formation, tool, flower, and fruit. Quintarelli et al's Facetag project [13] first manually creates a faceted classification schema. Then, their system guides users' tagging by presenting the facets to users. The system encourages users to use facets instead of free-form tags through a thoughtfully designed interface. If a user enters a tag, the completion tool suggests similar categories from the pertaining facet. Our research is similar to these projects in that we try to utilize large-scale human efforts. In contrast to these projects, the multi-faceted category schema in our approach is created and evolved by users.

3. Schema Evolution

Facet and Category Enrichment

As a collection grows new categories (or even facets) need to be created to improve the user browsing experience.

We designed an algorithm that adds categories utilizing the metadata pool and a statistical co-occurrence model. The model identifies parent-child relationship between x and y if all documents tagged with y are also tagged with x (so-called subsumption) [2]. For example, if all images labelled with “liberty bell” are also labelled with “independence”, where “independence” is an existing category, the algorithm will suggest “liberty bell” as a subcategory under the “independence” category. For an existing tagword t in the metadata pool, the algorithm identifies all documents with tag t. If these documents have a common category c, the rule of subsumption implies that t is a possible subcategory of c. We adapted the tool to our system and ran it on an African American History image collection and found the results encouraging enough to include this in future version. For example, the following suggestions were made:

Category	Suggested sub-category
American Civil War	military life
China	boxer rebellion

The suggestion feature will be added to the user interface in the future so that any user who is in the process of modifying the schema can receive instant recommendations for the particular facet or category she is modifying.

Schema Cleansing

In a collaborative classification system, it is likely that categories are created under the wrong facet, or child categories might represent a broader concept than the parent category. –To detect such anomalies we utilize WordNet [4].

WordNet is a semantic lexicon, which stores hierarchical relationships among words. The example below shows a hierarchy in WordNet starting with the search term “dog”.

```
dog, domestic dog, Canis familiaris
    => canine, canid
         => carnivore
             => placental, placental mammal, eutherian, eutherian mammal
                  => mammal
                      => vertebrate, craniate
                           => chordate
                                => animal, animate being, beast, brute, creature, fauna
                                     => ...
```

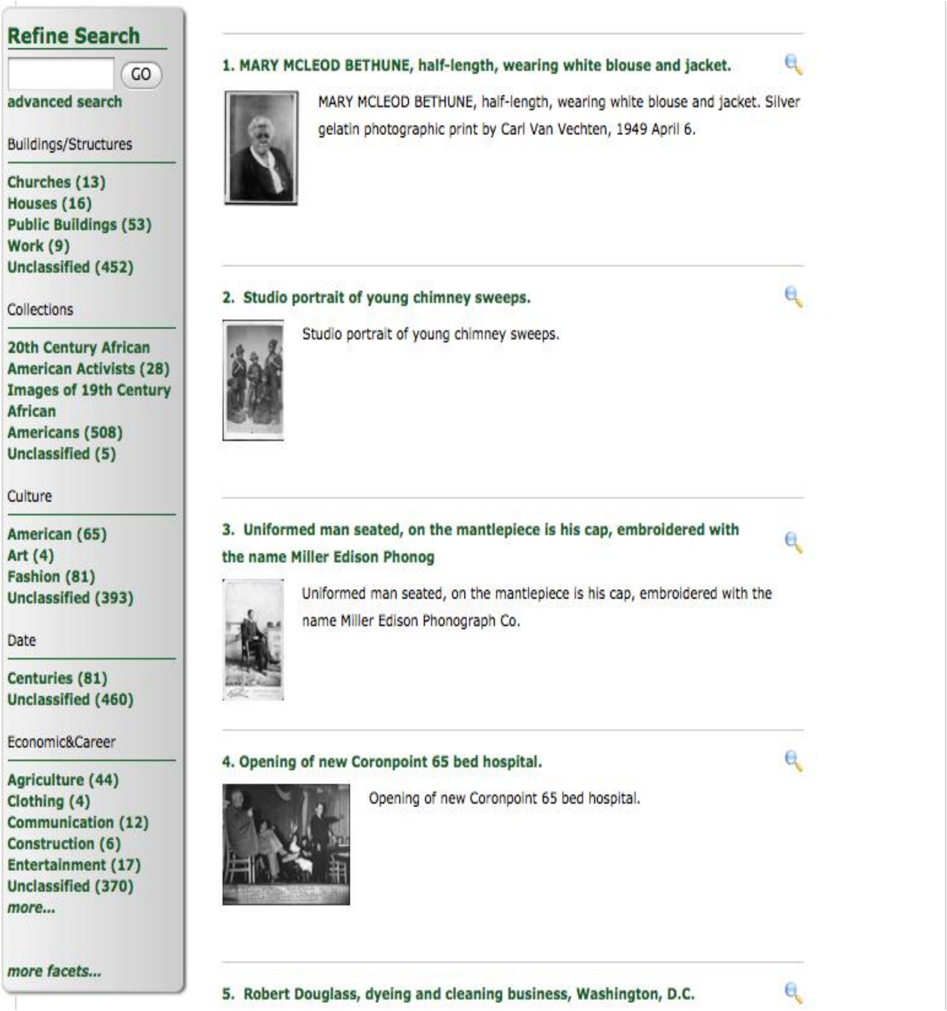
This example describes the relation hyponymy. It is commonly known as “is a” relationship (“dog is a mammal”). An anomaly detection algorithm detects issues by running a number of administrator-defined rules against the facet schema. An example of real anomalies detected is:

Category	Parent Cat	Grandparent	Category Problem
President	Holiday	Politics	more closely related to grandparent than to parent

4. Ordering of Schema Display

In a collaborative classification system, it is possible that a significant number of categories are created under a given facet (or another category), or large number of facets are created.

For usability the system displays only a limited number of child categories under a parent. The number of visible children, v, can be configured through the administrative user interface. While the initial display (see Figure 2) is limited to the first v categories, a “More” link (Figure 3) expands all categories under a node. A proper display order is critical to the usability of the system. The simplest display order would be an alphabetical one.



1. MARY MCLEOD BETHUNE, half-length, wearing white blouse and jacket.

MARY MCLEOD BETHUNE, half-length, wearing white blouse and jacket. Silver gelatin photographic print by Carl Van Vechten, 1949 April 6.

2. Studio portrait of young chimney sweeps.

Studio portrait of young chimney sweeps.

3. Uniformed man seated, on the mantlepiece is his cap, embroidered with the name Miller Edison Phonog

Uniformed man seated, on the mantlepiece is his cap, embroidered with the name Miller Edison Phonograph Co.

4. Opening of new Coronpoint 65 bed hospital.

Opening of new Coronpoint 65 bed hospital.

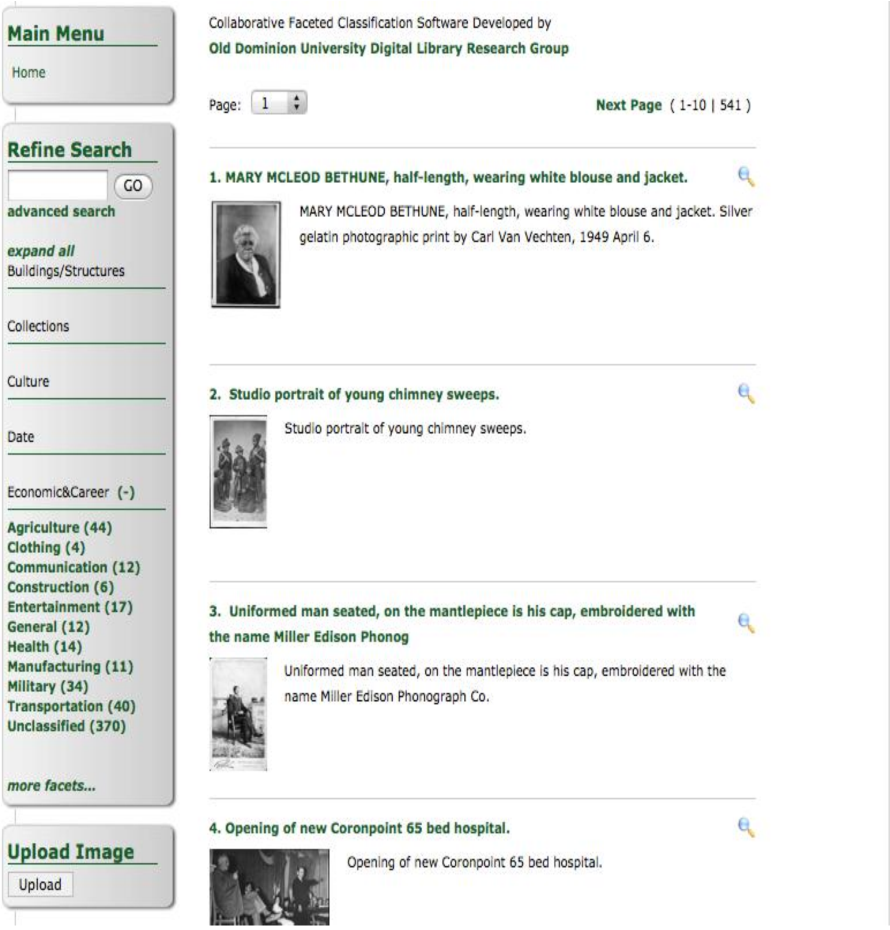
5. Robert Douglass, dyeing and cleaning business, Washington, D.C.

Figure 2: Limiting category display using the “more…” link

Our system orders the display by a popularity (P) measure, which favors the biggest, most used, and fastest growing facets and categories. The total number of items in a category (including subcategories) PN determines the biggest category. The growth rate of a category is produced by the number of new (recent) items for a unit of time (PR). Finally, a popular category will see a larger number of browsing hits (number of clicks on the category link in the browsing menu) which are measured over a period of time (PC). We combine the measure by adjustable weights and a normalizing function f:

$$P = 0.5*f(PN*PC)+ 0.5*PR \qquad (1)$$

The retrieval effort against a category is –in proportion to their size and the number of times they were accessed. Therefore, the first part of the popularity measure is a product of the atomic popularity values for size and browsing clicks. The latter, representing the number of new items in a category, symbolizes the categorizing effort. We believe that retrieval and categorization are two equally important user actions that consume roughly equal user efforts. Therefore, they are assigned equal weights after the normalizing function f is applied. The weights can be adjusted for different environments where browsing and categorizing consume unequal amount of user efforts.



Collaborative Faceted Classification Software Developed by
Old Dominion University Digital Library Research Group

1. MARY MCLEOD BETHUNE, half-length, wearing white blouse and jacket.

MARY MCLEOD BETHUNE, half-length, wearing white blouse and jacket. Silver gelatin photographic print by Carl Van Vechten, 1949 April 6.

2. Studio portrait of young chimney sweeps.

Studio portrait of young chimney sweeps.

3. Uniformed man seated, on the mantlepiece is his cap, embroidered with the name Miller Edison Phonog

Uniformed man seated, on the mantlepiece is his cap, embroidered with the name Miller Edison Phonograph Co.

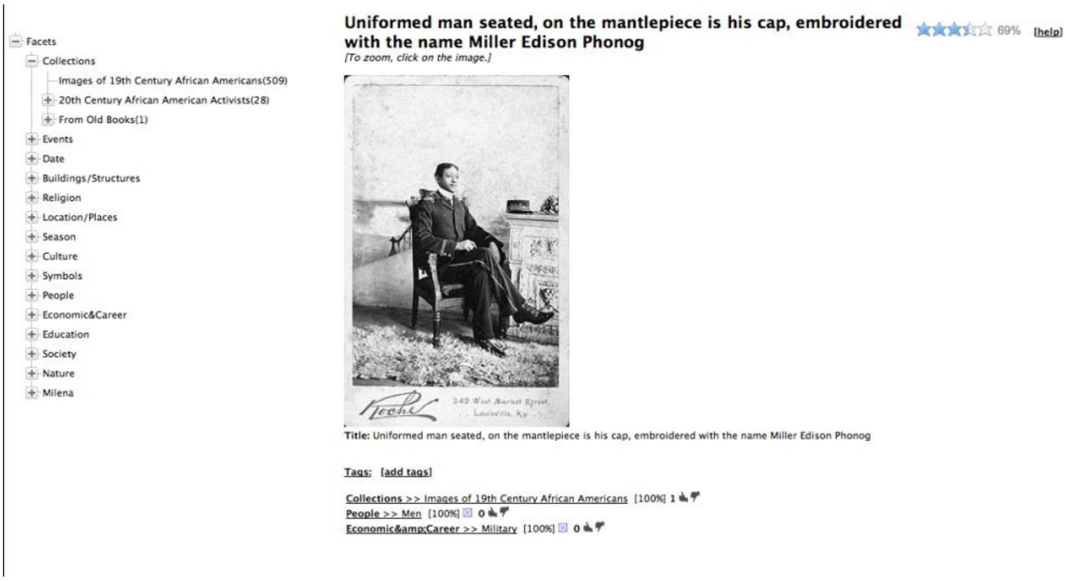
4. Opening of new Coronpoint 65 bed hospital.

Opening of new Coronpoint 65 bed hospital.

Figure 3: Expanding category display using the “more…” link

5. Quality Assessment through User Feedback

Our Faceted classification system uses user feedback to assess the quality of a classification and to remove it if needed. It is collected in the form of “thumb-up” and “thumb-down” buttons available for every association (see Figure 4). Users can vote up or down for the association between an image and a category on the basis of how relevant and accurate they think it is. The value of this explicit feedback determines when a classification can be deleted or, conversely, when it becomes “hard”, i.e., it is confirmed. Each feedback action will update the confidence value of an association by increasing or decreasing it by 0.05 based on whether a user believes it is a correct classification or not. If the confidence value reaches 1.00, it is hardened and essentially becomes a user-created classification. On the other hand, if users vote an association down below the threshold value, the system will allow them to delete it.



Uniformed man seated, on the mantlepiece is his cap, embroidered with the name Miller Edison Phonog

[To zoom, click on the image.]

Title: Uniformed man seated, on the mantlepiece is his cap, embroidered with the name Miller Edison Phonog

Tags: [add tags]

Collections >> Images of 19th Century African Americans [100%] 1
People >> Men [100%] ☒ 0
Economic&amp;Career >> Military [100%] ☒ 0

Figure 4: Feedback on category associations

6. Conclusions and Future Work

In an open community, management is needed to avoid problems created by the multitude of diverse users. This paper has presented several algorithms in the areas of schema enrichment, cleansing and ordering. With these automated algorithms, the burden on the administrator is reduced to responding to alarms and suggestions rather than laborious manual efforts. Future improvements include recording actual administrator actions for training purposes.

Notes and References

[1] K. Maly, H. Wu, and M. Zubair, “Automated Support for a Collaborative System to Organize a Collection using Facets”, proceedings, CD-ROM, ELPUB, Milan, June 2009

[2] M. Sanderson and B. Croft, "Deriving concept hierarchies from text," in Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval, 1999, pp. 206-213. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.3667

[3] H. Wu, M. Zubair, and K. Maly, "Collaborative classification of growing collections with evolving facets," in HT '07: Proceedings of the 18th conference on Hypertext and hypermedia. New York, NY, USA: ACM Press, 2007, pp. 167-170. Available at: http://dx.doi.org/10.1145/1286240.1286289

[4] Schmitz and Patrick, Inducing Ontology from Flickr Tags. Workshop in Collaborative Web Tagging, 2006.

[5] Clough, P., H. Joho, and M. Sanderson, Automatically Organising Images using Concept Hierarchies. Proceedings of Multimedia Information Retrieval 2005.

[7] Heymann, P. and Garcia-Molina, H., Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Stanford Technical Report InfoLab 2006-10, 2006.

[8] Yee, K.P., et al., Faceted Metadata for image searching and browsing. Proceeding of CHI 2003, 2003.

[9] Hearst, M., Automatic Acquisition of Hyponyms from Large Text Corpora. Proc. of COLING 92", Nantes, 1992.

[10] Mani, I., et al., Automatically Inducing Ontologies from Corpora. Proceedings of CompuTerm2004: 3rd International Workshop on Computational Terminology, 2004.

[11] Maria Elisabete Catarino and Ana Alice Baptista. "Relating folksonomies with Dublin Core", Proceedings of the 2008 International Conference on Dublin Core and Metadata Application, Berlin, Germany, Pages: 14-22, 2008.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. To Appear in IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

[13] Quintarelli, E., L. Rosati, and Resmini, A. Facetag: Integrating Bottom-up and Top-down Classification in a Social Tagging System. EuroIA 2006, Berlin.

[14] Harris Wu, Kurt Maly and Mohammad Zubair, Harvesting social knowledge from folksonomies. ACM Hypertext 2006, Odense, Denmark.

[15] Wu, Zubair, and Maly, 2007.

PROBADO3D – Towards an automatic multimedia indexing workflow for architectural 3D models

René Berndt[1]; *Ina Blümel*[2]; *Raoul Wessel*[3]

1 Computer Science, Graz University of Technology
Inffeldgasse 16c, A-8010 Graz, Austria
r.berndt@cgv.tugraz.at;
2 German National Library of Science and Technology TIB
D-30167 Hannover, Germany
ina.bluemel@tib.uni-hannover.de
3 Institute of Computer Science II Computer Graphics, University of Bonn
D-53117 Bonn, Germany
wesselr@cs.uni-bonn.de

Abstract

In this paper, we describe a repository for architectural 3D-CAD models which is currently set up at the German National Library of Science and Technology (TIB), Hannover, as part of the larger German PROBADO digital library initiative: The proposed PROBADO-framework is integrating different types of multimedia content-repositories and adding features available in text-based digital libraries. A workflow for automated content-based data analysis and indexing is proposed.

Keywords: Digital libraries, multimedia indexing, content-based retrieval

1. Motivation

The amount of newly generated multimedia content increases year by year and the use of this complex, non-textual data are becoming more and more important. However, this data is not analyzed and indexed sufficiently within the workflow of today's digital libraries, which are focusing on textual documents. Even though a lot of research has been done on how to manage, search, retrieve and present multimedia documents, there is still the need for integrating multimedia documents in existing library workflows. User-friendly tools must be developed so that both the management of multimedia documents for librarians and the user access to these documents (both content-based and in the conventional way of searching metadata) become possible.

Manual indexing multimedia content with keywords often results in a loss of information. To give an example: if a complex 3D model of a roof is just marked with the keyword "roof", it is not identifiable as a flat roof, a pitched roof or a cupola. In current document interpreting processes two different persons have to use the same keywords to make the document detectable: the person interpreting the document and, to detect the document, the person who is searching. Another issue is that authors are not motivated to extend their documents with metadata, even in the presence of suitable tools. Furthermore, it is also nearly impossible to interpret all existing data manually. The result is that in most cases multimedia documents are “black boxes” whose content could not be made accessible individually.

2. Introduction

As a concrete step into this direction, the ongoing cooperative German digital library project PROBADO [1] aims for setting up a framework for integrating repositories containing multimedia documents. The project targets to develop an integrated workflow for both document handling and cataloguing according to the classical library workflow and content-based document processing, i.e., making the collection accessible through content-based retrieval, the latter involving automatic content-based document analysis and indexing.

To achieve these goals, project partners from the University of Bonn and Graz University of Technology, each having expertise in distinct areas of multimedia document analysis and retrieval, are cooperating with partners of two large German libraries, the German National Library of Science and Technology in Hannover (TIB) and the Bavarian State Library in Munich (BSB).

Rather than being a pure research project, it is a special focus of PROBADO to achieve long-term usage of the developed systems and workflows at the cooperating libraries.

Two multimedia repositories are currently set up: one for architectural 3D models at the TIB and one for music at the BSB. As key contributions, we

- describe methods to support automatic processing of general documents in the library processing chain of document acquisition, annotation, search, delivery, and storage,

- develop and implement a common PROBADO platform serving as a web-based access point for searching and accessing general document types stored in the connected repositories. A service-oriented framework allows easy integration of new multimedia type repositories,
- develop and implement PROBADO-enabled multimedia repositories which are located at particular libraries and that are suitable for both conventional and content-based access.

3. Related Work

Within this field there are related scientific initiatives for 3D search engines. There are the Princeton Shape Retrieval [2] group with content-based search engines and Aim@Shape [3] with content-based and metadata based search engines.

And there exists the former EU-project MACE[1], that aims to connect various repositories of architectural knowledge and enrich their contents with metadata. Searching and browsing are very much based on architects needs, e.g. by conceptual connection, geography, language, competence. The search engine is only able to process metadata. There are no 3D models integrated.

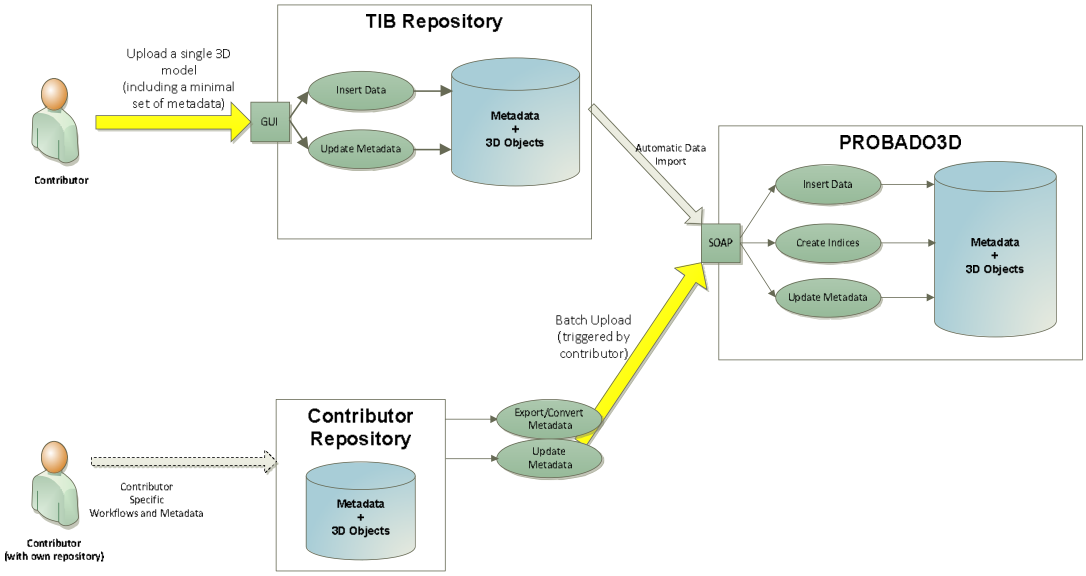


Figure 1. Use-case for integrating 3D models into PROBADO3D

[1] http:// http://portal.mace-project.eu/

4. Use-case: Import of 3D models into PROBADO3D

A repository for architectural 3D models is currently set up at the TIB in Hannover. As being the German National Library of Science and Technology, the TIB references relevant scientific material for superregional supply of technical literature and data.

Momentarily, the repository contains about 7,000 building, construction unit and object models which are converted, indexed and described with metadata as well as approximately 13,000 models that are to be analyzed and indexed.

The source 3D model can be categorized into two major groups:

- Models which are hosted by the contributor itself (Type-A Model) – usually provided by architectural practices
- Models which cannot be accessed through the internet (Type-B Model) – e.g. a submission from a student's master thesis.

Especially Type-A models can be subject to access restriction by the contributor, e.g. pay-per-view or IP-based access for certain groups. Usually the contributor has already implemented the technical details for access or payment.

In the paper we will present the integration workflow (see Figure 1) for models of Type-A and Type-B in detail.

5. Processing Pipeline

One major goal of the PROBADO project is to minimize the manual cataloguing work and to automatically generate the appropriate metadata wherever possible. As a 3D model normally does not bring along any describing data (if it is not catalogued in a database beforehand), the main source for first metadata is the automatic deduction.

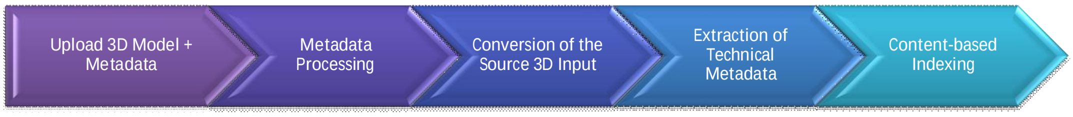

Upload 3D Model + Metadata
Metadata Processing
Conversion of the Source 3D Input
Extraction of Technical Metadata
Content-based Indexing

Figure 2. Schematic overview of the PROBAD3D processing pipeline

The following steps are a short excerpt of the different stages of the processing pipeline for the input data (see Figure 2):

Upload 3D Model and Metadata

The first step in the processing pipeline is the upload of the original 3D model and optional metadata by the contributor. This functionality is provided by a SOAP-based web service. The data is submitted as a single archive (see Figure 3), containing one XML file (METADATA.XML) and the 3D model file (plus additional files e.g. materials, etc. belonging to the model).

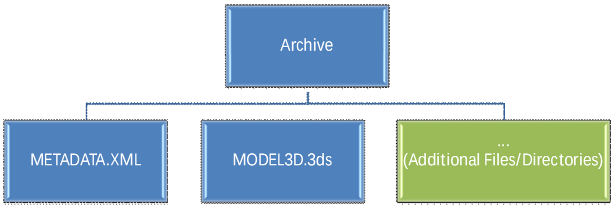


Figure 3. Archive file structure for the SOAP-based web service

The archive is then extracted to the file system for further processing. Beside the upload of new data the web service also supports modifying the 3D model and/or metadata of already uploaded data.

Metadata Processing

The metadata provided by the contributor (METADATA.XML) are stored in the metadata database of PROBADO3D. Since the PROBADO project concentrates on content-based indexing, the only required metadata is information about the contributor. An excerpt of the optional metadata is listed in Table 1.

Conversion of the Source 3D Input

For easier indexing, searching and viewing, a copy of each model is normalized and automatically converted into a uniform format for indexing and different formats for preview and delivery.

Due to the number of input formats it is not feasible for the content-based indexer to implement support for all these formats. The OBJ format from Alias Wavefront was chosen as the uniform format for indexing. For previewing, the PDF format from Adobe was selected, because of the wide distribution of the Adobe Reader.

Table 1. Optional information included in the contributors' metadata

Name	Description
Event	Contains data about events (e.g. competition, seminar, presentation)
Title	The title of the 3D model (not the filename)

Description	A textual description of the 3D model
Subject	Keywords or classification of the 3D model
Location	A geographic reference to a real building
Object	A reference to a real building
ExternalInfo	Contains additional information about external providers, events
Relation	To define relations between models

Most of the tools used for this task during the design phase supported only semi-automatic conversion or had license restriction in terms of usage as a service. In order to provide a fully automatic conversion for the workflow, DeepServer from Right Hemisphere was chosen. The following requirements for the conversion module were verified:

- Automatic conversion By using a watched folder DeepServer can execute a custom defined workflow on adding files to this folder. For the evaluation the input formats were converted to OBJ, PDF, and PNG (thumbnail).
- Support of all input formats DeepServer supports a large number of input formats. Even proprietary formats like the MAX format can be used if the appropriate software is installed.
- Licence The licence of DeepServer allows usage and deployment as a service.

Extraction of Technical Metadata

Technical metadata of the 3D model are automatically extracted during the conversion process. These include number of vertices, polygons, textures, etc. which can provide an approximate estimate about the model complexity. This task is also performed by DeepServer (see Figure 4); the resulting technical metadata are stored in the metadata database of PROBADO3D.

<keyframeanimated>False</keyframeanimated>	<TextureLinks>True</TextureLinks>
<NumObjects>39</NumObjects>	<UVMapped>True</UVMapped>
<NumPolygon>17819</NumPolygon>	<XMin>-434.3582</XMin>
<NumTextures>2</NumTextures>	<XMax>3345.902</XMax>
<SoftbodyAnimated>False</SoftbodyAnimated>	<YMin>-724.8393</YMin>
<AnimationLength>0</AnimationLength>	<YMax>2269.562</YMax>
<NumVertices>10171</NumVertices>	<ZMin>-0.04620994</ZMin>
	<ZMax>722.0472</ZMax>

Figure 4. Technical metadata of a 3D model extracted by DeepServer

Content-based Indexing

Indexing of architectural 3D models is a prerequisite for content-based query by example and document browsing. By creating a concise object description, the similarity between two 3D models can be computed. In the last years, automatic content-based indexing research resulted in the detection of many indexing characteristics especially in the lower layers of semantic, i.e. characteristics relying on rather pure geometrical shape content. For a detailed overview of these algorithms we refer to [8]. We compute global and local low-level semantic features for characterizing architectural components. In a query-by-example scenario where the user either uploads an existing 3D component or uses a sketch-based PROBADO3D interface to generate one, the search for similar objects in the database is conducted using global shape features based on spin-images [9] which are easy to compute and guarantee fast response times of the query engine. For browsing the PROBADO3D repository based on shape similarity, we comprehensively characterize the components using high quality local shape descriptors and special distance measures tailored to the requirements of architectural 3D models [6].

While low-level semantic features are an effective means to characterize the geometric shape of an object, they are not well-suited to describe the structure of building models which is mainly defined by the topology of rooms and floors. To overcome this drawback, we introduced the concept of Room Connectivity Graphs [4]. These graphs are especially designed to capture the topology of buildings. Rooms are represented by vertices, and connections between rooms like doors, windows, stairs, etc. are represented by edges. The graph is additionally enriched by semantic attributes like the dimension of rooms or the type of the connection. By that, users can search for building models that contain a certain spatial configuration of rooms. For the definition of such configurations we provide a graphical user interface. Additionally, the search can be further constrained regarding e.g. the area of certain rooms.

Content-based indexing of components using global and local low-level features works completely automatic and does not require manual data processing. The extraction of Room Connectivity Graphs however requires the building models to be oriented in a consistent way, i.e. the object's positive Z-axis must point towards the virtual sky. Additionally, the scale of the object must be known. In our experience so far it shows that this requirement does not lead to an increased amount of manual preprocessing interaction. In many cases, the scale is contained as metadata in the underlying 3D model data. Additionally, when presented a new charge of models from a content-provider, orientation and scale is usually consistent within one batch and therefore only requires minimal manual interaction.

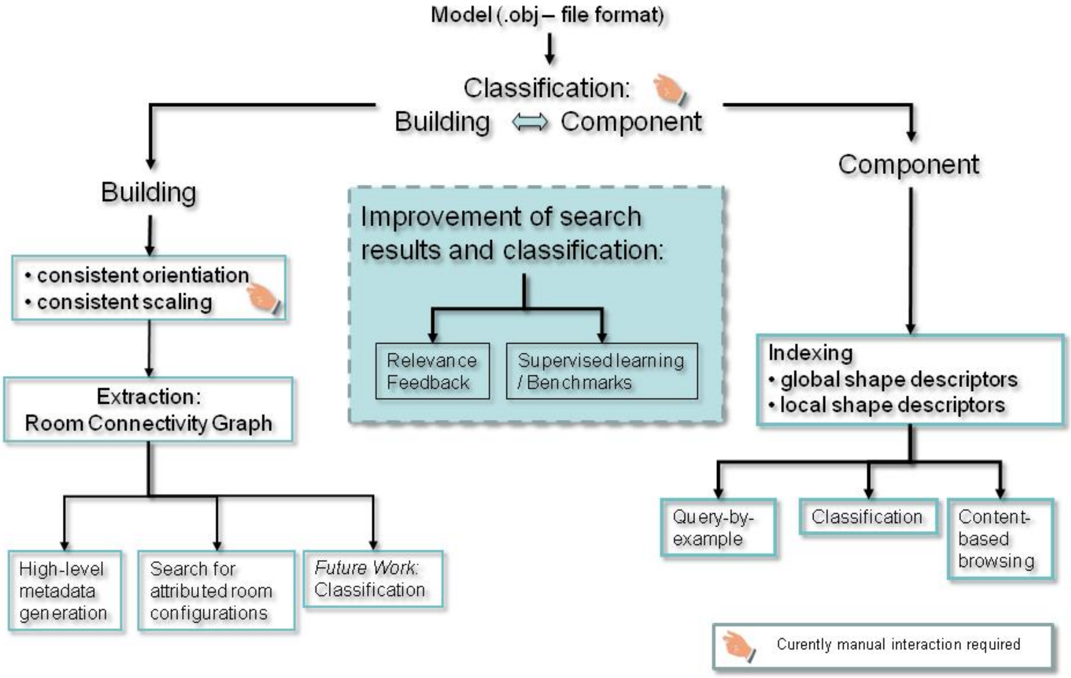

Model (.obj – file format)
Classification:
Building ⟺ Component
Building
• consistent orientiation
• consistent scaling
Extraction:
Room Connectivity Graph
High-level metadata generation
Search for attributed room configurations
Future Work: Classification
Improvement of search results and classification:
Relevance Feedback
Supervised learning / Benchmarks
Component
Indexing
• global shape descriptors
• local shape descriptors
Query-by-example
Classification
Content-based browsing
Curently manual interaction required

Figure 5. Content-based indexing workflow

Generation of High-Level Metadata

The afore described indexing techniques involving local shape descriptors and Room Connectivity Graphs serve as a starting point for fully automatic generation of high-level 3D object metadata. For predicting the object category of architectural components, we developed a supervised learning framework [6]][5][7] that classifies components according to their associated local shape descriptors. To this end it incorporates shape knowledge about a large number of manually classified architectural components contained in the Architecture Shape Benchmark [7]. The shape classification in this benchmark was created according to common architectural shape taxonomies.

The extracted Room Connectivity Graphs provide a large amount of information about building models that is important to architects. For example, we automatically extract the number of building floors, room areas, gross floor area, window areas per room and per floor, number of rooms per floor etc.

The resulting high-level metadata is finally stored in the PROBADO3D metadata database, allowing the user to textually search for components belonging to certain object categories as well as to search for building models fulfilling certain spatial specifications, e.g. concerning the number of floors or the gross floor area.

6. Conclusion and Future Work

Most parts of the processing pipeline have already been implemented. The integration of DeepServer is planned for the next project phase. The proposed workflow will then be established as a service at the German National Library of Science and Technology.

Future work will concentrate on improving the automatic classification and processing of the 3D models. Similar to the above described categorization of components, we will additionally examine how building models can be automatically classified according to their Room Connectivity Graph.

One additional content-based indexer using semantic enrichment methods based on procedural shape representations [10] is currently implemented and integrated into the PROBADO3D system. By fitting a procedural description to the target model the semantic information carried with the generative description can then also be applied to the target model (e.g. number of columns, stairs, etc).

Acknowledgments

PROBADO is a joint research project supported by the German Research Foundation DFG under the LIS program. PROBADO started in February 2006 with a tentative duration of five years. Partners are the University of Bonn, Technische Universitaet Darmstadt, Graz University of Technology, the German National Library of Science and Technology in Hannover, and the Bavarian State Library in Munich.

The work presented in this paper was partially supported under grants INST 9055/1-1, 1647/14-1, and 3299/1-1. For further information, please visit the project website at http://www.probado.de/.

References

[1] BERNDT, R.; KROTTMAIER, H.; HAVEMANN, S. and SCHRECK, T. *The PROBADO-Framework: Content-based Queries for Non-Textual Documents*. In Proceedings 12th International Conference on Electronic Publishing (ELPUB 2009), 2009,pp. 485-500.

[2] SHILANE, P.; MIN, P.; KAZHDAN, M.; FUNKHOUSER, T. *The Princeton Shape Benchmark*. In Shape Modeling International (June 2004).

[3] FALCIDIENO, B. *Aim@Shape Project Presentation.* In Proceedings of Int. Conference on Shape Modeling and Applications, 2004, pp. 329.

[4] WESSEL, R.; BLÜMEL, I.; KLEIN, R. *The Room Connectivity Graph: Shape Retrieval in the Architectural Domain*. In Proceedings of The 16-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2008, UNION Agency-Science Press, February 2008.

[5] WESSEL, R.; KLEIN, R. *Learning the Compositional Structure of Man-Made Objects for 3D Shape Retrieval*. To appear in Proceedings of 3rd EUROGRAPHICS Workshop on 3D Object Retrieval, May 2010.

[6] WESSEL, R.; BARANOWSKI, R.; KLEIN, R*. Learning Distinctive Local Object Characteristics for 3D Shape Retrieval* In Proceedings of Vision, Modeling, and Visualization 2008 (VMV 2008), pp. 167-178, Akademische Verlagsgesellschaft Aka GmbH, Heidelberg, October 2008.

[7] WESSEL, R.; BLÜMEL, I; KLEIN, R*. A 3D Shape Benchmark for Retrieval and Automatic Classification of Architectural Data* In Proceedings of Eurographics 2009 Workshop on 3D Object Retrieval, pp. 53-56, March 2009.

[8] TANGELDER, J. W. H.; VELTKAMP, R.C. *A survey of content based 3D shape retrieval methods.* Multimedia Tools and Applications, volume 39, 2008, pp. 441-471.

[9] JOHNSON, A. *Spin-Images: A Representation for 3-D Surface Matching.* PhD thesis, Robotics Institute, Carnegie Mellon University, August, 1997.

[10] ULLRICH, T.; SETTGAST, V. and FELLNER, D.W. *Semantic Fitting and Reconstruction*. In: Journal on Computing and Cultural Heritage 2008(1), pp. 1201-1220.

WHAT WE BLOG?
A QUALITATIVE ANALYSIS OF RESEARCHERS' WEBLOGS

Helena Bukvova; Hendrik Kalb; Eric Schoop

Chair of Business Informations, esp. Information Management,
Faculty of Business and Economics,
Dresden University of Technology
Dresden, Saxony, Germany
e-mail: {helena.bukvova, hendrik.kalb, eric.schoop}@tu-dresden.de

Abstract

Traditionally, academic researchers have been using platforms such as conferences, academic journals and books to present their findings and engage in academic discourse with their peers. The development of Information and Communication Technologies provides researchers not only with new tools, but also with new means of interaction. Among the new platforms are also weblogs (blogs). Formerly defined as `online logbooks', blogs can be used for a variety of purposes. A small but growing number of researchers write research related blogs. In this paper we present a qualitative, explorative study, carried out with the aim of describing and structuring information provided by academic researchers in their blogs. We describe a framework for categorising blogs and blog posts as well as patterns of blogging behaviour we have observed in research blogs.

Keywords: weblogs, blogs, research blogs, scientific blogging, scientific communication;

1. Introduction

Academic research as part of science has a long tradition. It has developed over centuries to gain today's form and rules and it is being developed still. The contact to other researchers, the academic discourse, plays an important role in academic research [1]. This exchange is often triggered by the publication of research results. By making their findings and their expertise public, academic researchers invite the opinions and criticisms of their peers. Through the issuing discussion, ideas and concepts can be refined and developed. Furthermore, the academic discourse and the social exchange among the researchers influences personal development and career paths. The platform for the communication of scientific finding depends on the progress of the inquiry. Traditionally, researchers were using conferences, academic journals and books for their publishing. The recent development of Information and Communication Technologies provides researchers not only with new tools, but also with new means of interaction [2, 3]. How far does this change the established practices? How do researchers use the new tools and platforms? In this article, we focus on the use of weblogs by academic researchers.

The term 'weblog' emerged in 1997/1998 and shortly afterwards it was abbreviated to blog [4, 5, 6]. Blogs are web pages with a list of dated entries that are typical displayed in a reverse chronological order [7]. Most blogs combine text, images, and links to other blogs and web pages and allow the readers to comment blog postings, generally in a mediated manner, where the blog host retains control [6]. Beside the reverse chronological order, other typical features of blogs are an individual ownership, a hyperlinked post structure, and an archival of postings [8]. The aims and target groups of blogs can differ and the entries vary from short opinions or references to large reports with citations. Herring [6] names as the purpose of blogs filters (with postings about other web pages), personal journals (with the blogger's thoughts and internal workings), knowledge logs (with relevant references about a particular knowledge domain), and mixed purposes. In the scientific world, blogs are well established for a fast dissemination of information. Portals like ScienceBlogs [9] or Scientificblogging [10] have evolved, which aggregate and organize scientific blogs about different subjects.

The use of blogs in science has already been a subject to some interest. The existing literature focuses on the potentials and risks of blogs as well as the motives of bloggers [11, 12, 13]. Yet this field is comparatively new and blogging scientists still represent only a fragment of the scientific community. A systematic inquiry about the use of blogs by researchers is therefore still missing. In this paper we present a qualitative, explorative study, carried out with the aim of describing and structuring information provided by academic researchers in their blogs. Our findings show that other factors have to be considered when viewing research blogs besides just the type of produced content [6]. We also describe patterns observed in the studied sample that illustrate the blogging behaviour of the blog authors. Although still to be seen as a research in progress, our findings can offer a new approach to analysing the role of blogs in scientific communication.

2. Method

This article describes an empirical study carried out to provide better understanding of the use of weblogs by researchers. At the moment, blogging behaviour particularly in science and research is still unexplored. We have therefore carried out an explorative study analysing the information that researchers publish in blogs. The study is a part of a larger research design, exploring what information researchers publish about themselves on the Internet.

To focus our work, it was necessary to operationalise and further to define the facets of this topic. Firstly, when defining who will be seen as researcher, we chose to concentrate on academic researchers. Academic research (also scholarly research, scientific research) is a crucial part of science. Science uses research, a process of systematic inquiry, as means of gaining new knowledge [14, 15]. However, not all research takes place in science. Academic research has to fulfil very specific criteria [16, 17], e.g. be public, replicable, unprejudiced, independent and it must advance the state of the art. Research outside these restrictions is non-scientific. An academic researcher, thus, is an individual professionally engaged in academic research. In the first place, we considered individuals working at academic institutions (mainly universities) to be academic researchers. As described below, we have sampled German researchers. In Germany, the understanding of academia is very much influenced by the Humboldian ideal of unity of research and teaching [18]. Viewing German academic staff as academic researchers thus appears appropriate. We have further viewed as academic researchers individuals working at scientific institutes. Also, individuals engaged in academic qualification (dissertation and habilitation [19]) and affiliated to academic institutions were considered academic researchers. Secondly, we viewed only the activity in blogs directly connected to the researchers. These were those blogs, where the researchers figured as authors or co-authors. Finally, we focused on research-related information and selected the blogs accordingly. Clearly private blogs (e.g. travel blogs, hobby blogs) were excluded from the study. Private information in blogs (e.g. blog post related to private interests or activities) were noted, but not analysed. The study was based on the following research question:

What professional information do academic researchers publish in their blogs?

The research design method was derived mainly from the Grounded Theory Method, also using aspects of analytic induction [20] and matrix analysis [21]. The research was based on the constructivist understanding of reality. We assume, that individuals create subjective reality in a process of construction [22]. The understanding of the subjective reality of other individuals is limited and can take place only through communication [23]. As researchers, we are thus not objective entities, but actively influence the findings by our interpretations. With regard to this position, we have based the exploration on the principles of constructivist grounded theory [24]. The Grounded Theory Method is "a systematic qualitative approach to data collection and analysis, that is concerned with generating theory" [25]. Key features of grounded theory are a structured, but highly iterative procedure of simultaneous data collection and analysis, based on constant comparison between already coded and new data samples [26].

Our study can be divided into two distinct though interconnected stages. In both stages, a sample of researchers was selected and analysed. In the first stage, a sample of n = 5 researchers has been purposively selected [27, 28]. Our aim was to find researchers, who not only had a research blog, but whom we also expected to be actively engaged online. We therefore chose researchers, who also had a Twitter account. We understood the engagement in both blogging and microblogging as an indicator of high level of online engagement. The Twitter streams were, however, not a part of our analysis. The researchers were all engaged at an academic institution in different positions (research assistants as well as professors). The sample included both male and female researchers, coming from three distinct areas: linguistics, literature, and cultural studies, social sciences, and natural sciences and mathematics. These areas were defined according to the German Federal Statistical Office. For each researcher, we viewed the blog and collected the last 15 blog posts and analysed. We chose not to collect more, because we assumed that blogging behaviour develops and changes through time. For our study, we preferred to take a snapshot of blogging activities. This data was analysed and coded using the qualitative-data-analysis software AtlasTI [29]. We started with in vivo coding (i.e. using terms or phrases used by the researchers) [30, 26], later developing further codes and categories. The analysis process was highly iterative. The results of the analysis were two categories describing information in blogs based on the researchers' engagement in the virtual environment (see Results). At this point a second sample was drawn using dimensional sampling [31]. Using the dimensions sex, academic position and area, we drew a sample of n = 12 blogging researchers, each representing one combination of the considered factors. Again, we viewed their blogs and collected and analysed 15 last blog posts for each researcher, leaning on the identified categories. As a result of this second stage, we have described five patterns of blogging behaviour.

When selecting the researchers for our analysis, we have also considered ethical issues. Although the data in the researchers' blogs is publicly accessible and thus technically public, we consider it a private property of the authors. As such, it has been created with a certain purpose. Although research blogs are generally addressed to the broad public, the authors still may not be comfortable with the use of their blogs for research purposes [32]. Therefore, all selected researchers were contacted per e-mail, informed about the research and asked for permission to use the content of their blogs. Three researchers did not give their permission and two did not reply. These were then removed from the samples and replaced, to gain the samples described above. Furthermore, to protect the researchers' privacy, we have used a very broad classification of the considered research areas and we refrained from using any direct quotes or material from their blogs.

3. Findings

The findings discussed here are a result of an integrative analysis of the first sample and the subsequent analysis of the second sample with the developed categories. They are to be seen as preliminary results of a research still in progress (see Conclusions).

Before exploring the blog posts, we examined the blog itself, including static pages and widgets. Each blog included in the study was clearly attributed to an individual researcher. The researchers stated their names and nearly all of them also stated their affiliation to an academic institution or an institute. A photo was also mostly included. More than a half of the researchers also provided links to other online profiles (web pages, social networking profiles, Twitter and others). All researchers write their posts in the first person and address an audience, more or less directly. This includes directly stating the audience in the blog description, ordering the posts according to the audience as well as addressing or questioning the audience in the posts. Where researchers stated the purpose of their blogs, they often noted the desire to share. Even for researchers who did not make the purpose of their blogs explicit this desire was visible in their blog posts: to share knowledge, experience or simply interesting information.

In our analysis, we have recognised a key category, describing the researchers' engagement in the virtual world, in this case in their blogs. The engagement is defined by the type of the content the researchers produce and its verbosity. Content and verbosity can be each described by three subcategories.

Content. The content of the blog posts varied greatly not only among the blogs but also within each blog. We have isolated three types of content authored by the researchers: expertise, activity and identification. Expertise-related content provides information on particular topic. We have termed it 'expertise', assuming it is typically related to the author's research area or area of interest. This assumption is in most cases confirmed by the expression of the author within the blog post. Further, researchers often report about the activities they engage in, both related to research and teaching. Very typical are reports from conferences and workshops. Finally, some content is apparently dedicated to describing the researcher as a person. These include descriptions of interests, personal information or purely reflexive posts. Content of this type identifies the author as a 'real' person, existing outside the virtual platform. Although the three content categories are sufficiently clear-cut, they are often combined in single blog posts. For example, an activity related content triggers a fluent shift towards an expert explanation of a particular topic. Similarly, a description of experience can lead the authors to reflect on themselves as individuals.

It was interesting to observe the role of external resources in blog posts. These are resources (content, media, events, people) outside the blog. Blog posts often contain links or references to external resources. In some manner, this reminds us of the citation practice in scientific publication. However, in some posts, the external resource plays the chief role in shaping the content of the post. We have distinguished two cases: either the resource was the apparent trigger of the content (these are typically placed in the beginning of the post) or it is used as a major illustration (appearing further on). In both cases, such posts appear to be written with the purpose of presenting these resources. We have termed such crucial resources 'Fundstücke' (German for finds). Connected to their use is typically an explicitly declared wish to share them. Often the authors note that they have found them and wished to share them. A common note is also, that the authors had this Fundstück for some time and wished to show it to others. Fundstücke are very typical for expertise-type content and they appeared to play a major role in the blogs we have examined.

Verbosity. Besides viewing the content type, we have also recognised, that the researchers show different forms and level of involvement. This verbosity of content can be viewed and measured with regard to three areas: level of detail, personalisation, and interaction. Level of detail describes how much information do the researchers provide about the particular topic. Personalisation measures how far the authors relate to themselves (e.g. give their opinions, judgements). Interaction gives the intensity of the authors' exchange with the audience as well as the potential for such an exchange. None of these measures is simply quantitative. Each form of verbosity leads to the production of more words and thus longer posts. A text analysis is necessary to describe them in each post. The 'measuring' is thus subjective.

Patterns. Viewing the content and the verbosity, we have attempted to describe the distributions. Although this was not very meaningful, given the limited size of our sample, by combining the dimensions content and verbosity, we were able to identify patterns. These patterns described the individual use of blogs by researchers. Figure 1 illustrates these patterns within the content-verbosity portfolio. The content types are given on the vertical axis and the verbosity on the horizontal axis. The width of the bars represents the frequency of the content type and the colour intensity depicts the level of verbosity. Not all blog posts have to follow the descriptions. The names of the patterns have a metaphoric meaning and should support quick association with the pattern description.

Presence. The first pattern is characterised through low levels of verbosity. The researchers (we found two in our sample) produced mainly expertise-type posts as well as some activity posts. These posts, although in one case numerous, however contained a low amount of information. Most of them were based on external resources and did hardly more than disseminate what was written elsewhere. Although the authors we present in the blogosphere, they showed little engagement.

Knowledge base. This pattern was most common particularly among the researchers on the level of research assistant as well as those coming from natural sciences. The posts are mainly expertise-oriented, with a high level of detail, but lower level of personalisation and interaction. The focus appears to be on the dissemination of information and the authors often mention the motive of sharing.

Expose. Particularly researchers on professorial level provided expertise- and activity-oriented posts, providing not only high level of detail, but also higher levels of personalisation. Researchers within this pattern focused not only on dissemination, but apparently also sharing their opinions and thoughts.

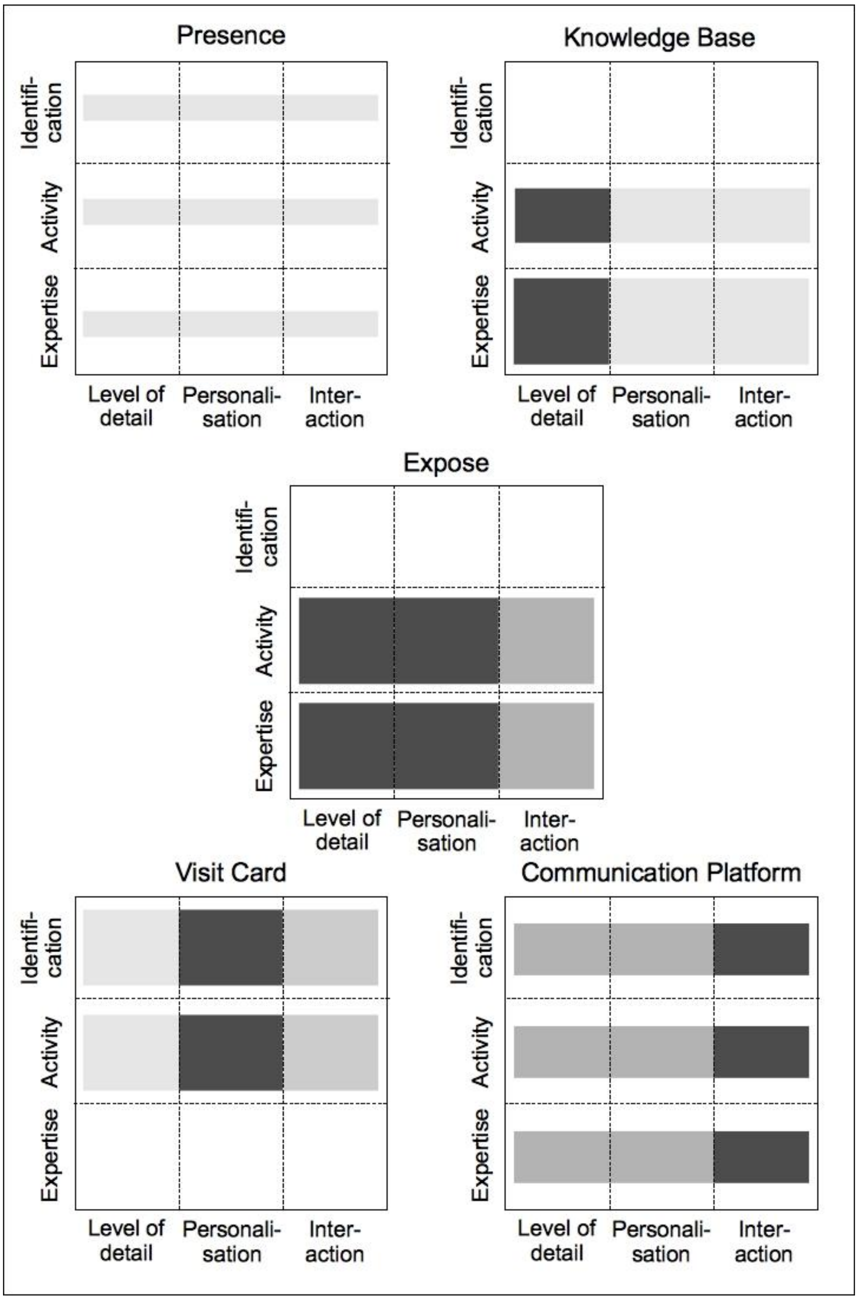


Figure 1: Blogging patterns

Visit card. Although we did not find this pattern in full, we have seen tendencies of development towards this pattern. A researcher writing a visit card blog would focus on activity and identification-oriented posts, with a high level of personalisation. Posts matching this pattern showed surprisingly low levels of detail - for more detail, the readers were supplied with links to external resources.

Communication platform. Again, we did not find a full demonstration of this pattern. However, some researchers would very actively call for interaction and attempt to actively interact with their readers. A blog that would fully follow this pattern would be expected to contain posts of different content, but highly personalised and interactive.

4. Conclusions

First, we have to point out, that given the nature of our sampling procedures, the data does not allow any conclusions about the population of blogging academic researchers. Although some conclusions might be drawn through analytical generalisation, we refrained from it in this article. The research presented here is to be still seen as a research in progress. The small sample was acceptable in the described first stages of the research. It allowed a very detailed, iterative analysis, resulting in an analytical framework and a first typology of blog and blog contents. However, it does not allow further verification of the results. To verify and further develop the findings, further research blogs and bloggers have to be included in the analysis. It is also insufficient to include only German researchers. Only a fraction of German researchers uses either blogs or other Social Media and Web 2.0 applications [33]. We will therefore include international researchers in further research.

The purpose of our study was not to describe the blogging behaviour of the population of blogging academic researchers, but to explore the types of information that they provided in their blogs. In our study, we have viewed blog as a publication, possibly presentation platform of existing researchers. By seeing the author of the blog as a 'real' person engaged online, have been led to distinguish between the content and its verbosity. This is a different approach from existing typologies, which focused mainly on content [6, 34]. The verbosity can be interpreted as the authors' engagement with the posts as well as the blog itself. The patterns we have derived rely strongly both on the content type and the verbosity, supporting the importance of both factors. The patterns we have observed, though they cannot be viewed as verified or generalisable, underline the focus on the individual bloggers an their blogs. This approach could be inappropriate in analysing private blogs, because the offline identity of the authors is uncertain [35]. In case of research blogs, however, the authors' identity can be traced. Approaching blogs as platforms for *presentation* and not just for *publishing* offers new views for discussing blogs in research.

Acknowledgements

The authors would like to thank all researchers, who have agreed to have their weblogs examined in this study.

Notes and References

[1] P. A. DAVID, Patronage, reputation, and common agency contracting in the scientific revolution: From keeping 'nature's secrets' to the institutionalization of 'open science', SIEPR Policy paper No. 03-039, 2004.

[2] H. KALB, H. BUKVOVA, E. SCHOOP, The digital researcher: Exploring the use of social software in the research process, *Sprouts: Working Papers on Information Systems*, vol. 9, no. 34, 2009.

[3] E. LEE, D. MCDONALD, N. ANDERSON, P. TARCZYHORNOCH, Incorporating collaboratory concepts into informatics in support of translational interdisciplinary biomedical research, *International Journal of Medical Informatics*, vol. 78, pp. 10–21, January 2009.

[4] M. ALCOCK, Blogs - what are they and how do we use them?, *Quill*, vol. 103, no. 8, 2003.

[5] R. BLOOD, Weblogs: a history and perspective, 2000.

[6] S. C. HERRING, L. A. SCHEIDT, S. BONUS, E. WRIGHT, Bridging the gap: A genre analysis of weblogs, in *Proceedings of the 37th Hawaii International Conference on System Sciences* - 2004, Jan 2004.

[7] A. WILLIAMS, Internet-based tools for communication and collaboration in chemistry, *Drug Discovery Today*, vol. 13, Jan 2008.

[8] J. W. S. SIM, K. F. HEW, The use of weblogs in higher education settings: A review of empirical research, *Educational Research Review*, Jan 2010.

[9] http://scienceblogs.com/

[10] http://www.scientificblogging.com/

[11] S. A. BATTS, N. J. ANTHIS, T. C. SMITH, Advancing science through conversations: bridging the gap between blogs and the academy., *PLoS biology*, vol. 6, September 2008.

[12] L. BONETTA, Scientists enter the blogosphere, *Cell*, vol. 129, pp. 443–445, May 2007.

[13] J. WILKINS, The roles, reasons and restrictions of science blogs, *Trends in Ecology & Evolution*, vol. 23, pp. 411–413, August 2008.

[14] K. BORDENS, B. B. ABBOTT, Research Design and Methods: A Process Approach. New York, NY, USA: McGraw-Hill Humanities/Social Sciences/Languages, 7 ed., July 2007.

[15] A. M. GRAZIANO, M. L. RAULIN, Research Methods: A Process of Inquiry. Boston, MS: Allyn & Bacon, 7 ed., February 2009.

[16] L. J. HEINRICH, Wirtschaftsinformatik: Einführung und Grundlegung. München, Germany: Oldenbourg, 1993.

[17] S. M. SHUGAN, Consulting, research and consulting research, *Marketing Science*, vol. 23, no. 2, pp. 173–179, 2004.

[18] B. R. CLARK, The modern integration of research activities with teaching and learning, *The Journal of Higher Education*, vol. 68, no. 3, pp. 241–255, 1997.

[19] Habilitation is the highest academic qualification available in Germany as well as in a number other European countries, where this qualification enables the individual to get a professorship. A habilitation can be earned after obtaining a doctoral degree, i.e. dissertation. The habilitation process is similar to the dissertation, however the level of scholarship is expected to be much higher.

[20] J. LATHLEAN, Qualitative analysis, in *The research process in nursing* (K. Garrish and A. Lacey, eds.), pp. 417–433, London, UK: Blackwell Publishing, 5 ed., 2006.

[21] M. B. MILES, A. M. HUBERMAN, Qualitative Data Analysis - An Expanded Sourcebook. London, UK: Sage, 2nd ed., 1994.

[22] E. VON GLASERSFELD, Thirty years constructivism, *Constructivist Foundations*, vol. 1, no. 1, pp. 9–12, 2005.

[23] G. RUSCH, Understanding – the mutual regulation of cognition and culture, *Constructivist Foundations*, vol. 2, pp. 118–128, March 2007.

[24] J. MILLS, A. BONNER, K. FRANCIS, The development of constructivist grounded theory, *International Journal of Qualitative Methods*, vol. 5, no. 1, 2006.

[25] I. HOLLOWAY, L. TODRES, Grounded theory, in *The research process in nursing* (K. Garrish and A. Lacey, eds.), pp. 192–207, London, UK: Blackwell Publishing, 5 ed., 2006.

[26] K. CHARMAZ, Constructing Grounded Theory - A practical Guide Through Qualitative Analysis. London, UK: Sage, 2006.

[27] J. M. MORSE, Sampling in grounded theory, in *The SAGE Handbook of Grounded Theory* (A. Bryant and K. Charmaz, eds.), pp. 229–244, London, UK: Sage, 2007.

[28] D. A. DE VAUS, Surveys in Social Research. London, UK: Routledge, 5 ed., 2002.

[29] http://www.atlasti.com/

[30] K. CHARMAZ, Qualitative interviewing and grounded theory analysis, in *Inside Interviewing: New Lenses, New Concerns* (J. A. Holstein and J. F. Gubrium, eds.), pp. 311–330, Thousand oaks, CA, USA: Sage, 2003.

[31] L. COHEN, L. MANION, K. MORRISON, Research Methods in Education. Abigdon, UK: Routledge, 6th ed., 2007.

[32] M. BAKARDJIEVA, A. FEENBERG, Involving the virtual subject, *Ethics and Information Technology*, vol. 2, pp. 233–240, December 2000.

[33] D. KOCH, J. MOSKALIUK, Onlinestudie: Wissenschaftliches arbeiten im web 2.0, 2009.

[34] W. MACIAS, K. HILYARD, V. FREIMUTH, Blog functions as risk and crisis communication during hurricane Katrina, *Journal of Computer-Mediated Communication*, vol. 15, no. 1, pp. 1–31, 2009.

[35] N. HOOKWAY, 'Entering the blogosphere': some strategies for using blogs in social research, *Qualitative Research*, vol. 8, pp. 91–113, February 2008.

Writeslike.us: Linking people through OAI Metadata

Emma Tonkin

UKOLN, University of Bath, Bath, United Kingdom
e.tonkin@ukoln.ac.uk

Abstract

Informal scholarly communication is an important aspect of discourse both within research communities and in dissemination and reuse of data and findings. Various tools exist that are designed to facilitate informal communication between researchers, such as social networking software, including those dedicated specifically for academics. Others make use of existing information sources, in particular structured information such as social network data (e.g. FOAF) or bibliographic data, in order to identify links between individuals; co-authorship, membership of the same organisation, attendance at the same conferences, and so forth. Writeslike.us is a prototype designed to support the aim of establishing informal links between researchers. It makes use of data harvested from OAI repositories as an initial resource. This raises problems less evident in the use of more consistently structured data. The information extracted is filtered using a variety of processes to identify and benefit from systematic features in the data. Following this, the record is analysed for subject, author name, and full text link or source; this is spidered to extract full text, where available, to which is applied a formal metadata extraction package, extracting several relevant features ranging from document format to author email address/citations. The process is supported using data from Wikipedia. Once available, this information may be explored using both graph and matrix-based approaches; we present a method based on spreading activation energy, and a similar mechanism based on cosine similarity metrics. A number of prototype interfaces/data access methods are described, along with relevant use cases, in this paper.

Keywords: formal metadata extraction; social network analysis; spreading activation energy; OAI-PMH metadata; informal scholarly communication

1. Introduction

Social network analysis is frequently applied to study 'community' structures. Web 2.0, with its social nature, is expected to contribute changes to scholarly communication. Community data mining, i.e., mining real world-data such as scholarly articles in order to characterize community structures, is considered a top data mining research issue [1]. Automated inference through mining now represents a plausible approach to extracting candidate information from data that is already publicly and openly available in institutional repositories.

The Writeslike.us system was initially intended to support future work with the University of Minho in establishing informal links between researchers [2]. It also constitutes an exploration of a general area of interest, which is that of processing OAI metadata to extract author identity and to link it with external data (as in Wikipedia) and linked data.

Understanding the relationship between individuals and their communities is a very old problem. Several excellent tools exist that attempt to support the exploration of community structures within research publication information; a prominent example is the application of RKBExplorer [3] to support the exploration of OAI-PMH and DBLP metadata [4]. The problem has received widespread interest via bibliometrics and the potential for their application in the area of impact assessment for academic publications, as well as the widespread use of FOAF (Friend-of-a-Friend) to encode machine-readable information about individuals and their relationships.

One problem often encountered when making use of this sort of data is, in the case of FOAF, that the data is too sparse to provide an overview of the community as a whole; in the case of citation analysis, the information is much more exhaustive, but instead suffers from the difficulty that the data is not sufficiently extensive or accurate to enable a clean graph to be drawn. For example, FOAF information is necessarily limited to the data provided by individual users and their contacts on a given web site. Sometimes it is also possible to merge FOAF data from multiple web sites, but this is complicated by the need to identify equivalence between users.

FOAF information is generated by direct input from users; for example, creating an account on LiveJournal or Facebook creates a new individual identity. As the account holder seeks out individuals to add to their ‘friends’ list, they create new links between themselves and others. Characterising those links is sometimes difficult; some systems therefore ask users to specify the nature of the link. It is a simplification, based on the assumption that individuals play clearly definable roles in each others’ lives – but a social graph can nonetheless be a useful resource. The problem, however, is that not all users make use of any given site, and a large percentage will make use of none at all. A recent report by Connell [5] offered a literature review of several studies demonstrating user opinion to social networking sites; one, from 2005, showed that over 50% of study participants responded that Facebook had no potential as an academic outreach tool, 12% said that it had potential, and the rest were unsure.

Reported demographics of use of social networks are largely consistent with the popular image of a bias towards younger users, although the picture is changing rapidly over time. A recent study [6] of the use of online social networks demonstrated that younger users were associated with higher levels of usage of Facebook, as well as a greater number of ‘friends’. The types of use of the site also vary greatly between user groups, and therefore all social graphs are not created equally (nor equally detailed). Finally, the characteristics of different social networks vary in that some social networks predominantly reflect offline connections, whilst others are predominantly places where people meet for the first time online. Facebook, according to Ross et al. [7], is ‘offline-to-online’ – that is, Facebook friends are mostly met offline and then added at a later date to the online social network.

The outer edges of the Social Web

Citation analysis-based bibliometrics tends to privilege those papers that have a large number of citations or are published in journals or proceedings that are indexed in large citation databases such as the domain-specific PubMed, DBLP or ACM, or journal-specific but widely accessible citation indexes. Indeed, Mimmo & McCallum [8] point out that it is ‘natural to ask which authors are most influential in a given topic’. Because citation is essentially a ‘voting’ process, giving those who are better-known or more influential or key to a given area of research a higher profile, less well-known authors and those who disseminate on a more local level will be almost invisible by those metrics, unless indexes also explore user-driven opportunities for deposit such as institutional repositories.

The majority of online bibliographic research tools, understandably, focus on establishing the primary figures within a field, and de-emphasise encouragement of collaboration between the 'foot-soldiers' of the research world. Yet there are reasons to explore this territory. The aim of this project is not to identify the most popular or well-cited individuals in a field, as there are many existing methods that enable this to be done. Rather, we aim to explore a mixture of factors; matching expertise and interest and enabling a multifaceted browse model for information about individuals.

We chose to make use of a mechanism that depends only on data that is already publicly available, and as a consequence the startup cost was small. Additional data was extracted from publicly available sources, and is to be republished for others to reuse in the same spirit of encouraging innovation.

Service design

A number of well-understood components are required in order to create a system of this type. A data source and parser are required in order to extract the essential information – for example, author names, institutions, and other formal metadata. In the majority of cases citation analysis over a large corpus of electronic versions of papers is applied for this purpose, perhaps along with a formal metadata extraction system. We replace this step with extraction from OAI-PMH records.

Extracting strings from OAI-PMH records is extremely simple, but the difficulty lies with their interpretation. A perfect example of this is the author name disambiguation problem, which is to say, the question of identifying, from a pool of uses of a given string, which instances refer to a given individual.

The obvious conclusion when seeing five hundred papers by John Smith is that John Smith is a prolific author. However, in reality there are in all probability several John Smiths at work and writing peer-reviewed papers. The question becomes how to tell the works of a given individual from those of another individual with the same name. If there appears to be a Smith working in ethnography with a second working in quantum physics, then we are perhaps fairly safe in assuming that they are different people – but it is nonetheless possible that a single individual named Smith has moved from physics to HCI in the last few years. If one Smith works for Harvard and the other works for MIT then it is reasonable to assume that they are different people, but it is not by any means certain. Authors may be affiliated to several institutions; they may have moved from one institution to another, or taken a sabbatical to work in another and then returned to their parent institution. They may be working as a visiting fellow.

Author name disambiguation is a complex problem. Many unsupervised methods exist, such as calculating the distance between strings. We explored the use of several, including a vector-based cosine distance approach applying similarity between authors' coauthor lists, institutions and subjects (calculated using simple text analysis to extract approximate ‘noun phrases’, and then reducing these further using Wikipedia as a text corpus).

The most promising methods, according to Laender et al. [9], are generally based around supervised machine learning techniques, which require initial training. Examples of these include Naive Bayes [10], and support vector machines (SVM) [11].

On et al. also describe various heuristics to be applied across the names themselves; spelling-based heuristics, based on name spellings, token-based blocking, n-gram based blocking, and tokens similarity [11]. Laender et al. [9] describe a heuristic-based hierarchical clustering method that offers comparable results. It has been applied for a wide number of purposes; one that relates closely to our own is described by Minkov et al. [12], who applied contextual search and name disambiguation in order to relate name mentions in emails to a given identity.

Collecting relevant background information

During the development of the project identifying and utilising suitable data sources led to the need to make use of a whole variety of publicly available resources, such as Wikipedia; this required enhancement for effective re-utilisation of the data.

Many of the possible functions of this prototype depend on having access to appropriate datasets. For example, in order to improve the accuracy of a formal metadata extraction algorithm designed to identify the institution with which an author is affiliated, it is useful to have both a gazetteer of institution names and variant forms, and a list of the domains and sub-domains associated directly with that institution.

Again, this is greatly simplified when the data exists in a well-formatted, well-structured form, and to an extent this is true; for example, there exists partial data on DBPedia – a ‘community effort to extract structured information from Wikipedia and to make this information available on the web’ [13]. DBPedia is built up of the subset of information on Wikipedia that is well-structured and well-formatted, which is to say, predominantly information that is placed inside structured templates. However, although already a useful resource, there is not enough information to cover all of our requirements. In terms of institutions, for example, only a small subset of institutional Wikipedia pages contained a ‘legible’ dataset, and even in these cases, the information did not – by design – cover any of the variant forms of institutional name, identifier, domain, etc. that are useful for our purpose.

Therefore we chose to make a less subtle use of the Wikipedia resources, by spidering the pages directly, extracting relevant terms and URLs from the page source, and attempting to characterise them by means of application of a small set of heuristics. This was particularly difficult in shorter Wikipedia articles, articles that contained information about institutions that were not, themselves, present on the Web, and articles that were written in languages other than English; however, we found that the simplest possible heuristic alone – that the top 1/3 of external links present on the Wikipedia page were likely to represent the institution – was correct over 65% of the time. We chose to take a permissive approach to information collecting and to add a confidence rating to each data point, reasoning that it is preferable to store too much data than too little at this early stage in the prototype's development.

Scenarios of use

Usage scenarios are real-world examples of how people can interact with the system being designed, and are often collected as part of development processes, especially user-centred processes. In general, usage scenarios are written with specific user personas in mind; in this case, they have been generalised for publication. Initial usage scenarios were developed through consultation with the University of Minho and institutional repository managers elsewhere. A shortened summary is included in this paper. Scenarios varied from supporting a student in seeking a supervisor working in an area of interest, to exploring influences of relevance to the modern-day concept of the impact factor. Several example scenarios that informed the design of the writeslike.us prototype are given here.

Scenario 1: Classifying events and forums by listed participants

A researcher in the field of evolutionary linguistics has become increasingly very interested in possible mathematical mechanisms for describing the nature, growth and adaption of language, as he has heard that others have done some very interesting and apparently relevant work in this area. Unfortunately, the researcher finds that some of the detail is hard to follow. He decides to seek out an appropriate event and/or online forum, and finds some people who might be interested in exploring links between his specialist area and their own. He is concerned about the potential cost of attending several events, so he chooses to look up possible events and web forums, intending to look through the participant lists for names that he recognises. This is greatly simplified by an automated system enabling him to identify papers and authors that he considers most relevant; with this information it is possible to parse through lists of participants in events or online communities in order to provide him with a rough classification of how relevant the group is likely to be to his ideas.

Scenario 2: Building a ‘dance card’ for a Research Council event

One of the purposes of a Research Council event is to encourage serendipitous meetings. Rather than simply assuming that synchronicity at the coffee-table will carry the day, the Research Council decides to produce a ‘dance card’ that suggests several other individuals that you might like to meet. Whilst elements of the composition of this ‘dance card’ are resultant from program managers’ knowledge of the individual’s interests and character, a service that is able to identify, characterise and compare researchers from their existing work and affiliations can be used to quickly build some interesting (and at times amusing) meeting suggestions, based on the individuals’ papers and output, and/or on the names and Research Council-held descriptions of the projects on which the individuals work.

Scenario 3: Facilitating collaboration in a multidisciplinary research environment

An anthropologist with a particular interest in the area of paleolithic archaeology, who works in the Department of Humanities, is very interested in exploring likely patterns of migration, and particularly in the idea that this activity may have been driven by climate change. However, the Department of Humanities has limited funding for the purpose of data collection and interpretation regarding modeling of climate change, so it is not possible for him to develop a paleoclimate simulation system. Therefore he decides that it is more appropriate for him to look for other people who have other reasons to be interested in modeling of this kind, particularly during the time period in which he is interested. This is not a trivial problem for several reasons; firstly, he does not usually publish in the same area as paleoclimatologists and therefore is unlikely to make chance acquaintances. Secondly, he and they have very different ways of describing their areas of interest, and therefore there is quite a lot of interpretation required in order to ascertain that the datasets they require are (or are not) closely related. Successfully establishing that these groups could usefully share data with each other is a non-trivial problem. However, it is an important goal for all concerned, not only because it is likely to help the data *consumer* – the anthropologist – but also because the data *creator* – the palaeoclimatologist – will benefit from wider impact and reach of their research.

Engineering from a series of scenarios

It is noticeable that the majority of the scenarios depend on information taken from several sources; much of the benefit of this system requires the data extracted to be enriched with externally sourced information.

For example, geographical information is necessary in order to successfully complete some of these tasks, such as finding local academics with similar interests. In principle, this information can be found from several sources, such as the address of the institution that the author places on his/her conference submissions. However, affiliation with an institution does not necessarily require that the author spends a great deal of time physically present at that institution. In practice, reliably eliciting a researcher's workplace or present location is difficult, requires considerable information to be made available – such as calendar data, GPS, etc – and is not really practical without finding a solution for several major infrastructural and social issues.

This additionally marks an intersection with well-known research themes in the field of ubiquitous and pervasive computing, in particular the many research projects that have sought to enhance social networking by means of additional information gleaned from context-sensitive mobile computing. For example, Kortuem and Segall [14] describe a system that supports augmentation of social networking through mobile computing. Information such as the individuals with whom people have met and spoken can be collected and stored [14]; individuals can 'pledge' to work together towards a given aim, and discovery of other individuals in the group can be managed accordingly. This form of mobile computing requires devices that are always on and running (*constant*), aware of presence of nearby devices and people, able to communicate with other collocated devices (presumably a heterogeneous environment of devices), and proactive – operating without explicit user interaction. All this goes to underline the point that, although there is great potential in this sort of work, it is non-trivial as an engineering problem.

Because so much data is involved – or at any rate, sought – there is great potential for this sort of work to be implemented via the reuse of data from several other sources, in what is sometimes known as a mash-up. That said, much of the data is not sufficiently cleanly formatted or complete to be accessible as simply as well-formatted and structured FOAF, meaning that the major challenge here is one of extracting as much as possible from the information available and storing it in a normalised form. The second challenge is to take the data and attempt to establish equivalence between entities, which is to say, explore whether multiple mentions of the same name refer to the same individual.

As a result, this problem can be seen as a useful step in the general problem of author name/identity disambiguation. Much of the data mined could potentially be used to enrich formal data sources such as that offered by the NAMES project [15].

2. Methodology

The writeslike.us project contained several specific stages; harvesting, analysis, user-level interface development and evaluation. . The project had as its final goal the development of a functional prototype intended to extract as much information as possible, making it accessible for further research in the area.

Source data was retrieved from OAI-PMH metadata repositories, but was expected to require supplemental information from several sources; identification of appropriate sources was itself an aim of this work. As such, one focus of the project was to identify and use existing services wherever possible. The quantity of data was also expected to give rise to a number of data management issues.

2.1 Data collection

The dataset is harvested from OAI-PMH – the Open Archives Initiative Protocol for Metadata Harvesting [16]. According to the Repositories Support Project [17] about 75% of institutional repositories worldwide, and ~85% in the UK, provide an OAI-PMH interface. The dataset is harvested via UKOLN's OAI data harvesting service, RepUK, which performs regular UK-wide data harvests and stores the resulting information in an XML dump, available for reuse by other services and applications. The system currently takes in data from across the UK - a future evaluation methodology will be to explore the applicability of the system at an international level. The data is initially input into a database, from which it is processed in the following manner:

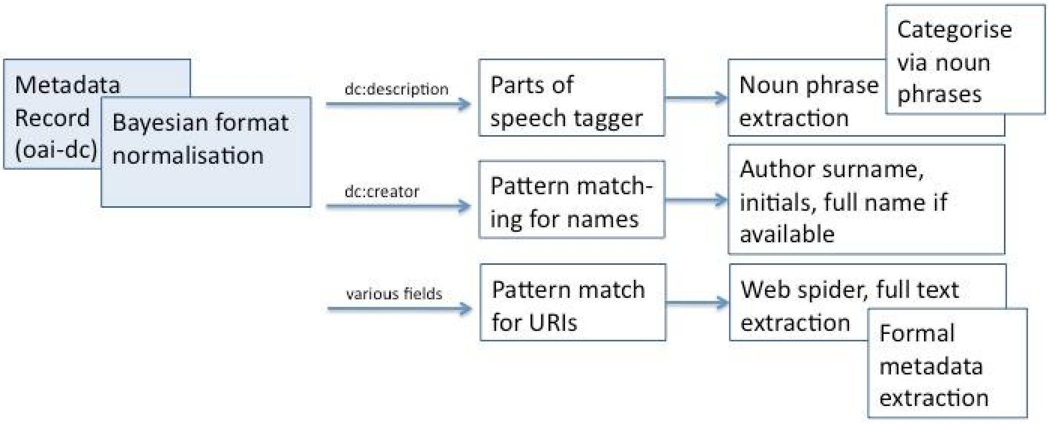

Metadata Record (oai-dc)
Bayesian format normalisation
dc:description
Parts of speech tagger
Noun phrase extraction
Categorise via noun phrases
dc:creator
Pattern match-ing for names
Author surname, initials, full name if available
various fields
Pattern match for URIs
Web spider, full text extraction
Formal metadata extraction

Figure 1: Published articles per year

2.2 Collecting information from the metadata record

The metadata record itself is collected in oai-dc, which is to say that it contains up to twelve elements, most of which are unlikely to be filled out. The actual placement of information within the record may vary a great deal depending on the interface used to create the record and the design decisions of its originators. Often the information will have initially been input as qualified Dublin Core rather than the simpler format, and as such there are likely to be several instances of certain fields, containing different refinements of a given field, but not marked as such. Hence, a useful initial step in analysing this information is to attempt to characterise its qualified representation.

This step can be achieved in a number of ways; for example, a simple heuristic can often identify certain fields such as well-formatted dates (ie. validation against a schema, regular expressions capable of identifying common citation formats, etc). Content-level feature analysis can also help to identify certain common types. The result, however, will not be as clean as collecting the qualified information directly in the occasions on which it is available.

2.3 Format normalisation

Knowing what type of information is contained in a given field is only the beginning. Following that, it becomes necessary to come to an understanding of the format of that information; for example, author names may be given in any one or more of a number of different formats:

Smith, John and Richard, Peter
J. Smith and P. Richard
John Smith, Peter Richard.
Smith, John and Peter Richard
SMITH, J. ; RICHARD, P.
…and so on.

It is unlikely that all of these can be convincingly, uniformly and accurately parsed, especially since there are certain cases in which there is an essential ambiguity that would stop even a human user from having any certainty of his or her conclusions – for example, Peter Richard and Richard Peter are both acceptable, valid names. A Bayes filter designed to make use of a knowledge base of information regarding the prevalence of different first names and surnames would have a similar problem in this case; making use of the last published set of US census data, a classifier might make use of Table 1 to find the following statistics:

Table 1: US Census data for surnames Richard and Peter

Name	Rank in US	Approx. number in US	Freq. of occurrence per 100,000
PETER	3758	8662	3.21
RICHARD	581	52138	19.33

True, ‘Peter Richard’ is a more statistically probable match than ‘Richard Peter’, but neither solution has an overwhelmingly convincing lead over the other.

Very similar problems exist with many other data formatting areas, for example, dates; the well-known discrepancy between preferred forms of date formatting in the United States versus the preferred mechanisms in use in the United Kingdom mean that:

10-11-2009
11-10-2009

may be very difficult to parse. Ideally, these formatting and encoding problems would be avoided by appropriate choice of standard and strict adherence to that decision. As these issues do occur in practice, we are left with the requirement of developing a mechanism for solving such problems with reasonable degrees of success. Fortunately, there are often simple heuristics that can be applied. In both of the cases described here, there is a solution to do with the observation that these variations are often regional and often have a link to the choice of interface in use. If this is the case, then the convention is likely to be at least somewhat systematic. It then becomes desirable to identify the convention in use within that context, the frequency with which it applies and the likelihood that a similar pattern holds in our particular case.

2.4 Categorisation via text analysis

The means of categorisation was described briefly above as a method of extracting noun phrases from text, followed by dimensional reduction using Wikipedia as a corpus. Here, we will explore how this works in more detail.

Noun phrases are defined by Wordnet as ‘a phrase that can function as the subject or object of a verb’. We do not precisely look for ‘noun phrases’, but content ourselves by looking for a more loosely defined phrase that is either a noun phrase or a descriptive phrase of a somewhat similar nature. This is achieved by making use of a part of speech tagger to analyse the textual content that is available to us. This is naturally language-dependent, and marks the point at which it is no longer possible to work in a language-agnostic manner. The particular tagger that we used is available in a number of common European languages (English, French, German, Spanish, Portuguese).

This is a relatively processor-heavy and slow part of the analysis process. To complete a single run of data analysis on the full textual content of the metadata records alone takes over 24 hours – this is on UK content only, which is to say, a few hundred thousand records. To perform a similar task on a global scale would take weeks. To run a similar analysis on the full-text of each document spidered (see the following subsection) would take much longer. As such, it is important to consider both efficiency and the possibility of caching results – if a piece of data has been processed once, that information should be stored for the future.

In practice, the data set proved to be large enough to feed quite an extensive collection of terms, meaning that there was enough information to use this information for browsing purposes. Smaller datasets offer difficulties both to interface designers and for those intending to reuse the information for categorisation purposes, not least because there is not sufficient information regarding similarities between documents to enable an effective categorisation process to occur.

However, the data did suffer from excessive specificity; that is, there are similarities between the concepts underlying ‘superstring theory’ and ‘Higgs boson’, but it is not necessarily obvious from the content of the term or of their application. This problem can be tackled in a number of ways. One common solution is to make use of the LSI approach (latent semantic indexing) to reduce the dimensionality of the large matrix of terms, in the hope that the similarities between term domains would become evident. However, we chose to make use of Wikipedia as a 'crowdsourced' text corpus on which to look up and attempt to identify classifications for each of the terms that we had extracted. This meant that terms related to, for example, physics, would be clearly identified as such.

Once the popular phrases are extracted, they can also be used as navigational elements supporting a subject-level browse mechanism. The same is true of dimensionally reduced categories, although in this case it may be preferable to treat them as a simpler set of categories, graphically presented in a similar manner to a standard 'breadcrumb trail' navigational element.

2.5 Formal metadata extraction

Before formal metadata can be considered, it is necessary to identify appropriate candidate data objects from which that metadata can be taken. For example, an OAI record may refer to a paper, a presentation (.ppt, etc) file or a dataset contained within the institutional repository. These digital objects may not be accessible to the outside world – indeed it is not uncommon for records to be placed online without depositing a data object. Alternatively, the object(s) may be hidden for a limited time (embargoed).

Because direct links to digital objects are rarely given within an OAI record (only ~600 records contained actionable, externally accessible links directly to digital objects), it is often necessary to use a web crawler (spider) to identify candidate links to the fulltext record. About a quarter of the records surveyed contained a link to a page from which the originating record could be retrieved. To harvest digital objects effectively can be time-consuming and the details are outside the scope of this paper, so we will merely comment here that whilst there is considerable variation to be handled, handling the most common is easily achieved.

We found that on average just under half of the pages retrieved through crawling of links provided within DC records contained one or more accessible documents (Fig. 1). Around 15% of linked pages resolved to a variety of journal endpoints – 'paywalls' - (Ingenta, Taylor & Francis, Wiley, Sage, IOP, etc). These sometimes contain additional useful metadata about the document, but do not contain the document themselves. However, the copyright ownership is in itself a useful data point. Around 40% of institutional repository links were found to contain no accessible data.

A further finding from this work was the number of DOIs and handle.net persistent links that were present in metadata records, including the percentage that were broken. 240,000 records were harvested. Out of the 62,000 records containing an actionable http dc:identifier, 35,000 contained a handle.net (15,500) or dx.doi.org (20,000) actionable persistent identifier. DOIs and handles appear to have a similar prevalence in UK institutional repositories.

In principle, persistent actionable identifiers are useful in part because they permit URLs to be assigned, managed, and reassigned when something causes that persistent identifier to break. In practice, we found that there was a noticeable proportion of broken persistent identifiers. We found that out of a test run of 20,000 URLs retrieved from URL records, almost 400 were unresolvable DOIs or handle.net persistent identifiers – 2% of URLs were invalid persistent identifiers, of which two-thirds were DOIs.

Overall, harvested data is retrievable for about 1/8 of the indexed objects. Additional metadata (for example, details about a journal publication, author, affiliation, citations, and so on, that are not reflected in the metadata) may potentially also be retrievable from journal item pages, so there is a case to be made for this approach.

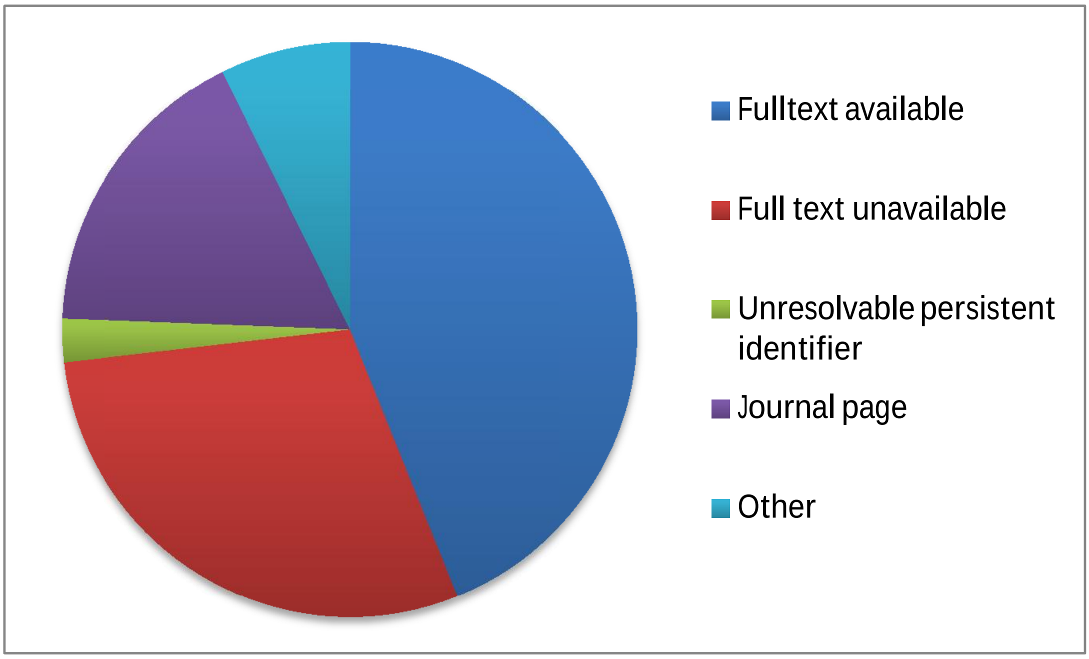


Figure 2: Findings from web crawl

The sorts of information that can usually be retrieved from this approach include, for eprints: title, author name, sometimes a date or range of dates, citations, format information, format metadata, software from which the document was created, and perhaps additional information such as keywords, abstracts, etc. This provides a useful check against the OAI metadata to ensure that the correct document has been retrieved. For most digital objects it is possible to retrieve some form of format information and/or object-specific metadata. Object provenance information such as software used in its creation can say a great deal about author identity, as specific authors may choose different routes to object creation.

2.6 Name disambiguation in metadata graphs

The result of the data collection process is to develop a set of relations between entities, author-name strings and articles. Before the information can be effectively used, the distinction between author name strings and identity must be applied. Characteristics of authors for a given paper may often be extracted directly from the article text. These pieces of information may be added to what we already know about a given set of instances of use of a string. Several name disambiguation approaches were explored using test datasets taken from the real-world OAI-PMH dataset.

The information that is available to us can be displayed as a matrix of terms; a given author-string *a* in authorlist *A* has created i_a digital objects out of an overall set of objects *I*. Each object has a set of subject characteristics s_i, a title t_i, author affiliation information f_i, provenance from repository r_i, and so forth. So the problem becomes one of ascertaining the most likely number of individual authors identified by string *a*. These characteristics can be treated either as links within a graph, or a sparse matrix of characteristics; both graph-based approaches such as spreading activation energy and matrix-based methods such as cosine similarity metrics may be applied.

Table 2: Document object data in matrix form

I	*t*	*f*	*r*	*s (a, b, c…)*
i_2	t_2	f_2	r_2	$s_2(a,b,c…)$
i_n	t_n	f_n	r_n	$s_n(a,b,c.…)$

2.6.1 Relatedness metrics: Cosine similarity

The first metric explored was a very simple cosine similarity calculation. This is a measure returning the similarity between vectors of *n* dimensions, and is a staple of search engine design [18]. Cosine similarity states that two identical vectors are exactly identical (eg. Eq. 1 returns 1) or that they are dissimilar (Eq. 1 returns 0).

$$similarity = \cos(\vartheta) = \frac{\mathbf{A}.\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Equation 1: Cosine similarity

Cosine similarity can be applied to the document object characteristics almost directly, although there is a need to 'unpack' listed characteristics into a simpler numeric form if they are to be treated as term frequencies. Applying this directly to information such as that provided in Table 2 returns a modified form of document object similarity, weighted with additional factors such as provenance metadata. To relate this to the author requires an assumption that there is a relationship between document object similarity and author identity. In principle this would seem to be a strong assumption, - authors primarily write about topics that they know well and seldom write outside their field. However, in practice this assumption may not hold up over time, as authors may change research groups, areas of interest, and even subject areas. It may also presuppose a meaning to authorship that is not there; for example, co-authorship of a paper may imply that the author has produced data that was used within the paper, but the paper itself may not be in the author's core area of interest – a palaeoclimatologist may co-author a paper with an archaeologist, published in a journal with a focus on archaeology and written for an audience consistent with the journal's focus. Authorship represents contribution, which is traditionally expected to be textual, but often this may be an experimental collaboration, inspiration, supervision, perhaps even an administrative link.

This method provides a reasonable metric for comparison between items. There is, however, considerable computational overhead in calculating this for each document and metadata set, so this was done periodically in exploring the link between author name and unique identity - that is, this method provided one data point for the following question, essentially a clustering problem without the advantage of any definitive knowledge of the number of authors involved: given a set of objects created by a person or people with the author name *a*, how likely is it that objects *i* and *j* were created by the same author? Note, however, that grounding author identity as a subset of author name ignores a large number of complications, such as authors who change their names or anglicise their names inconsistently.

The exploration of similarity provides an approximate notion of identity, which can also be represented as nodes in the graph – which, indeed, supersede author name string similarity in calculating the relatedness between papers.

2.6.2 Relatedness metrics: Spreading activation energy

A search algorithm based on spreading activation energy over a contextual network graph modelled over a series of timesteps [19] [20] was applied to support graph-based searching. This is not a pre-calculated method, but is rerun on each search. This method is appropriate where it is possible to search from the starting point of a prechosen node – given a unique node id as a search key, this algorithm could be seen as spilling a little ink on one node of the graph, which then spreads a predefined distance through the graph of relations between authors, objects, roughly calculated identities, classifications, and other metadata, in a manner defined by the way in which the implementation is tuned. The result is a ranked list of matching nodes and their types, which can then be presented to the user.

Modification of this approach can be used to reflect the relative relevance of different types of connection or to tune a search to prioritise different types of relation (eg. topicality, similar location of publication, similar physical location). The search may be weighted according to specific search types. In terms of efficiency, frequent calculation via a contextual network graph algorithm is observed to be relatively inefficient on a dense graph, by comparison to alternative methods (eg. latent semantic indexing); intuitively, this is reasonable, the number of links to process per timestep is related to the density of the graph. The decision of whether to use a contextual network graph/spreading activation energy method or whether to precalculate is also linked to the number of changes expected to be made to the graph – frequent change and hence frequent recalculating negates any benefit to be gained from what is essentially a caching mechanism. Furthermore, the contextual network graph approach is memory-hungry [19]; in a production environment it may be preferable to pre-process the data in a persistent (cached) form.

3. Evaluation

Initial qualitative and quantitative evaluation studies have been completed, and preliminary results are encouraging. In particular, there is significant diversity between information and authors indexed into well-structured datasets such as DBLP, ACM, and so forth, and the world as viewed through

OAI-PMH. From a random sample of authors, it is very visible that authors with few publications have little visibility in the formal indexes. Figure 2 contrasts authors who have published between six and twenty papers with the indexing visibility of authors who have deposited five or fewer documents (see Figure 2).

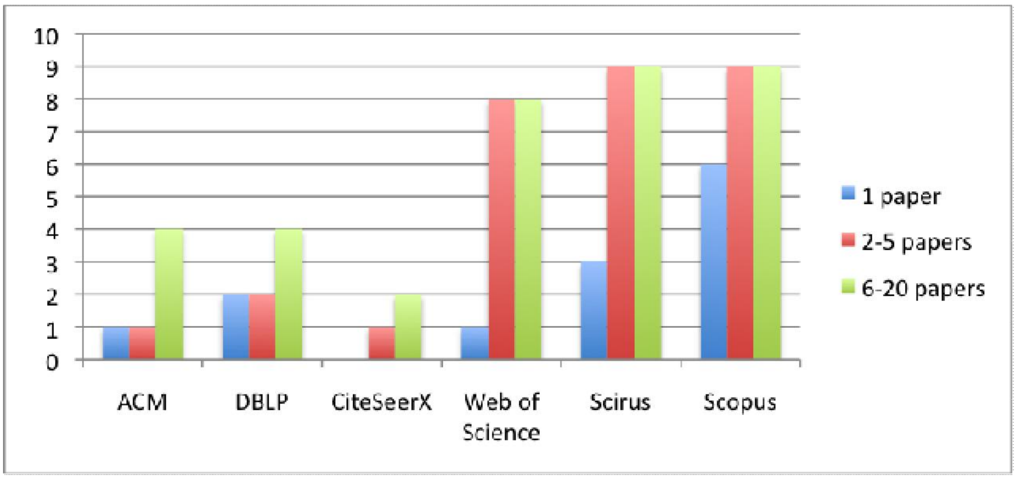


Figure 2: Author appearance in popular indexes (average values)

One potential interpretation of this limited visibility is the following: it could be taken to indicate that the data is relatively low-quality, as it contains so much information that was, for whatever reason, not published in a popularly indexed publication – and hence perhaps it is not published in a peer-reviewed form. However, for the purpose of encouraging informal collaboration between researchers, this may not prove to be a significant impediment.

Certain areas of functionality are key, in particular those areas concerning author identity, and it is clear that alternative methodologies, the use of additional data sources, and user-level amendment functionality could improve things greatly.

4. Discussion

During this work, we established a database extracted primarily from metadata, a knowledge base derived from Wikipedia, a variety of classification information derived variously from part-of-speech tagging and the use of Wikipedia as a classification dictionary. We used this in order to retrieve further information where available, and extract what metadata was available from these sources. We then used simple mechanisms from text analysis to classify and provide a mechanism for exploring this data.

The major difficulty in the project was simply one of interface design and access; it is one thing to develop a database, and quite another to create an interface that supports the aim of encouraging informal collaboration – an aim that depends a great deal on factors that are very difficult to identify or represent digitally, such as trust and organisational culture [21]. The problem is as much social as technical. We expect to release parts of the information extracted as linked data for future reuse by ourselves and by others. One likely future avenue for reuse may be the application of this data in supplementing more formally generated information sets – 'filling in the gaps'.

The course of this experimental development has been a journey of discovery, and not least an opportunity to challenge our own assumptions about appropriate interpretation of apparently straightforward data, even those as simple as document creation and authorship.

5. Conclusion

OAI-PMH metadata alone provides sufficient information to collect basic information about authorship data. However, the quality and completeness of that data is greatly improved if the full-text document is also available and may be analysed. Another means of supplementing basic data about authors is to compare and contrast with information derived from networks of citations; for popularly-cited texts, this permits the retrieval of additional information such as authors' full names, relevant dates and so forth. The effectiveness of this approach is dependent on the number, style and quality of existing citations - so again, a synthesis of available approaches is likely to provide the best overall result. Methods of disambiguating unique author identity vary greatly in effectiveness depending on the available data, as well as factors such as resource type, format and language. Standardised benchmarking requires the establishment of and testing against a ground truth. Bibliographic networks are often used to identify 'star' researchers working in each field. The problem explored here is to enable the ability to browse for others (who may be at any of various stages in a research career) working on a given topic area.

In our future work, we intend to widen the availability of writeslike.us as a pilot service, to develop a clearer set of requirements for practical usage of the interface, API and data, and to explore questions such as automated classification of authors' likely primary occupations (eg. primary investigator, researcher, technical writer, student). We also intend to explore the possibility of bringing in other sources of data – and publish relevant segments of existing information for wider reuse, as a clearly and consistently formatted linked-data resource.

Acknowledgements

The author would like to acknowledge the input of Alexey Strelnikov, the primary developer of the web interface and database, Andrew Hewson, database guru, and Wei Jiang, for his work in the area of identifying and testing possible author name disambiguation methods.

Notes and References

[1] VARDE, A. Challenging Research Issues in Data Mining, Databases and Information Retrieval. In ACM SIGKDD Explorations Journal, 11(1) 2009, p. 49 - 52.

[2] BAPTISTA, A; FERREIRA, M. Tea for Two - Bringing Informal Communication to Repositories. D- Lib 13 (5/6) (2007). Retrieved 10th Jan, 2010, from www.dlib.org/dlib/may07/baptista/05baptista.html

[3] GLASER H; MILLARD I; JAFFRI A. RKBExplorer.com:A Knowledge Driven Infrastructure for Linked Data Providers. In: European Semantic Web Conference, 1-5 June 2008, Tenerife, Spain. pp. 797-801.

[4] GLASER H; MILLARD I; CARR L. RKBExplorer: Repositories, Linked Data and Research Support. In: Eprints User Group, Open Repositories 2009, 20/05/2009, Atlanta, GA, USA.

[5] CONNELL RS. Academic Libraries, Facebook and MySpace, and Student Outreach: A Survey of Student Opinion. Portal: Libraries and the Academy. Volume 9(1), January 2009. E-ISSN: 1530-7131

[6] JOINSON AN. 'Looking at', 'Looking up' or 'Keeping up with' People? Motives and Uses of Facebook CHI 2008, April 5–10, 2008, Florence, Italy.

[7] ROSS, C; ORR, ES; SISIC, M; ARSENEAULT, JM; SIMMERING, MG; ORR, RR. Personality and Motivations associated with Facebook use. Computers in Human Behaviour. Volume 25 (2), March 2009. Pp. 578-586.

[8] MIMNO, D; MCCALLUM A.s Mining a Digital Library for Influential Authors. JCDL 2007.

[9] LAENDER AHF; GONCALVES MA; COTA RG; FERREIRA AA; SANTOS RLT; SILVA AJC. Keeping a Digital Library Clean: New Solutions to Old Problems. DocEng '08.

[10] HAN H; GILES L; ZHA H; LI C; TSIOUTSIOULIKLIS K. Two Supervised Learning Approaches for Name Disambiguation in Author Citations. JCDL 2004.

[11] ON, B-W; LEE, D; KANG, J; MITRA, P. Comparative Study of Name Disambiguation Problem using a Scalable Blocking-Based Framework. JCDL 2005.

[12] MINKOV E; COHEN WW; NG AY. Contextual Search and Name Disambiguation in Email Using Graphs. SigIR '06, Seattle, Washington, USA.

[13] JENTZSCH, A; et al. About DBPedia. Retrieved March 20, from http://dbpedia.org/About

[14] KORTUEM G; SEGALL Z. Wearable Communities: Augmenting Social Networks with Wearable Computers. IEEE Pervasive Computing, volume 2(1) 2003, p. 71-81, ISSN: 1536-1268

[15] NAMES project introduction. Retrieved March 18 from http://names.mimas.ac.uk

[16] LAGOZE C; VANDESOMPEL H. "The Open Archives Initiative: Building a Low-Barrier Interoperability Framework". Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01). pp. 54–62.

[17] Introduction to the RSP project. Retrieved March 18 from http://www.rsp.ac.uk/usage/harvesters

[18] TAN, P.-N; STEINBACH, M; KUMAR, V. *Introduction to Data Mining*. Addison-Wesley (2005), ISBN 0-321-32136-7, chapter 8; page 500

[19] CEGLOWSKI M; COBURN A; CUADRADO J. Semantic search of unstructured data using contextual network graphs. 2003.

[20] BOLLEN, J; VANDESOMPEL, H; ROCHA LM. Mining associative relations from website logs and their application to context-dependent retrieval using spreading activation. 1999.

[21] DODGSON, M. Organizational Learning: A Review of some Literatures. Organization Studies, 14(3), 375-394 (1993). DOI: 10.1177/017084069301400303

Authors publication strategies in scholarly publishing

Paola Dubini[1]*; Paola Galimberti*[2]*; Maria Rita Micheli*[1]

1 Ask Research Center; Università Bocconi
Via Roentgen 1 20136 Milano Italy
{paola.dubini, maria.micheli@unibocconi.it}
2 Biblioteca di Scienze dell'antichità e Filologia moderna,
Università degli studi di Milano
Via Festa del Perdono 7 20121 Milano Italy
paola.galimberti@unimi.it

Abstract

In this exploratory study, we analyze publishing patterns of authors from different disciplines, as part of a broader analysis of the transformation of the scholarly publishing industry. Although a growing body of literature analyses the author's role within the process of research production, validation, certification and dissemination, there is little systematic empirical research on publishing patterns; little therefore can be said on relevant issues within the current debate on the future of scholarly publishing such as authors' responses to (or even awareness of) the growing array of publication possibilities or the speed of adaptation to the increasing series of incentives by funding agencies or academic institutions. On the basis of the analysis of three years of publications gathered in the institutional repository of Università degli Studi di Milano, we highlight trends of publication strategies and different responses to incentive systems. Preliminary results indicate that publication outcomes and intensity differ across disciplines, while similarities occur mainly in terms of choice of preferred outcomes by seniority. Open access is still uncommon among the authors in our sample and it is more utilized by relatively senior authors and active authors.

Keywords: scholarly publishing; publishing strategies; industry changes;

1. Introduction

The process of knowledge creation in the academic field follows a quite rigidly codified pattern. Academic and scholarly knowledge is in fact systematic, premeditated, reflective and continuously submitted to the scrutiny of a community of experts.[1] Creation is therefore a long, time consuming process for academic authors, and several steps have to be overcome, to reach the final moment of knowledge delivery to the audience [1].[2] Publications as well as academic affiliation contribute to strengthen authors' reputation, which is a critical element for economic and social professional growth in the academia. Over time, as scholars build their reputation and become visible within their community, publication occurs on increasingly more prestigious journals. As personal prestige increases, authors are likely to orient research trajectory development and to influence publication patterns of younger colleagues.

The tangible starting point of academic knowledge creation is identified in the existing body of scholarly literature, which constitutes the background of all academic scientific works [2],[3]. The central stage is the moment of design, in which the social process of knowledge becomes tangibly represented. After the designing stage, concepts are integrated into a particular body of knowledge, whose choice is influenced by several factors, with a special weight of the discipline of interest.

The choice of the body of knowledge to refer to in a particular research design coincides with the choice of the viable publication outcome, which is widely considered a determinant step for the evaluation of the scholarly work and the resulting academic assessment within academic institutions [4], [5].

Research patterns are influenced by the necessity to conform to reward mechanisms of institutions to obtain career advancements; authors may therefore choose where to publish on the basis of specific incentive structures, deriving from national, institutional or community indications and explicit or implicit incentives. In recent times academic institutions are progressively increasing control mechanisms on faculty members in order to enhance a greater visibility and major prestige at the international level. As competition for research funding becomes more intense, and institutions and

[1] It is also necessary to underline that knowledge production activities in different areas entail different epistemic cultures (Knorr Cetina, 1999), and consequently different patterns of results delivery.

[2] This conceptualization of science as a knowledge production system (Latour and Woolgar 1986) is functional for understanding the possibilities for the inclusion of new publication channels in the editorial chain of scholarly publishing.

funding agencies are increasingly interested in the visibility of the outcomes of the research process by authors associated with these institutions, attention on what and where publication occurs becomes higher and possibly influences authors' behaviour.

Publication outcomes by academic authors is therefore a good dependent variable of strategies put in place by authors to ensure maximum visibility, reputation and personal achievement. In spite of a growing body of literature analysing the author's role within the process of research production, validation, certification and dissemination, there is little systematic empirical research on publishing patterns; little therefore can be said on relevant issues within the current debate on the future of scholarly publishing such as authors' responses to (or even awareness of) the growing array of publication possibilities or the speed of adaptation to the increasing series of incentives by funding agencies or academic institutions.

In this exploratory paper, we are interested in analysing publishing patterns by academic authors as part of a broader research project on the evolution of scholarly publishing. In recent years, digitization and technological advancements have indeed contributed to a structural redefinition of the scholarly publishing industry and contributed to an increase in publishing and diffusion of scholarly output. While traditional publishers have developed a digital strategy and upgraded their offering, a variety of digital only publishers and repositories have emerged with a multiplicity of innovative business models, covering all phases of the scholarly publishing process (from idea discussion to publishing to research dissemination and communication), different revenues streams and intellectual property protection régimes. While publication tools are constantly evolving, authors' strategies remain sometimes unaffected [6], [7], reflecting the established norms of the traditional academic environment.

Based on the systematic analysis of three years of publications by authors from different disciplines but from the same institution, we wish to highlight to what extent recent publication patterns mirror the changes occurring within the scholarly publishing industry and whether similarities and differences occur in publishing strategies across different disciplines. Although descriptive in nature, we think that our study contributes with fact based hints to the current debate on the assessment of the quality of research activity and on the future of scholarly publishing, while giving evidence to all parties involved on how academic authors from different disciplines build their reputation and visibility, while strengthening that of their institution.

2. Literature review

There is a growing body of literature describing why academic authors publish and how they choose where to publish. Broadly speaking, literature addresses individual drivers to publication, the incentive system put in place at the institutional level, the patterns of research diffusion and certification across different disciplines.

The willingness to contribute to science's advancements is undoubtedly a leading motivation both to undertake the academic career and to publish [8], [9]; moreover, authors publish to be promoted and advance in their institutions. [10], [11], [12], [13], [14] show a correlation between journal rankings and tenure and promotion decisions. Last but not least, authors publish as part of their legitimization process: consensual evaluations of publication channels have the potential to impact on research quality assessment and individuals' promotion prospects and publishing strategies [15]. Recognition of personal contribution to journals' articles can be used as a *currency* to obtain reputation and being accredited by the scientific community. In recent years international collaborative research projects have increased [16], as a consequence of the rising competition to publish in top quality journals. Collaborative research represents a way to improve data availability, and collaboration with foreign researchers is very attractive for those who face difficulties in data collection when trying to conduct studies across countries [16]. Researchers are often invited by institutions to collaborate, in order to include more features to the final paper and to increase the probability of publication in high ranked journals. Although authors tend to publish in the same channels their senior peers and their scientific community deem appropriate, young researchers may benefit from a less conformist behaviour and the choice of more radical journals [17].

As competition for research funding increases, authors are pressured on the one hand to accelerate publication of results and increase visibility and to publish on key refereed journals for purposes of promotion and tenures on the other, respecting constraints and strict rules [6].

The choice of where to publish at the individual level parallels the effort academic institutions are making to build their reputation at the international level and their degree of acceptance of new forms of publication. Attitudes of disciplines toward scholarly communication in general is affected by the institutional setting of departments [18], [19], [20]; the use of new electronic media is influenced by the academic field [21 [22], [23], [24], [25]. Moreover, the reward in terms of reputation that authors from different disciplines achieve from the publication on specific channels varies and influences authors' decision for results delivery [26], [27], [28].

Web 2.0 tools are already considered as essential means for creating users community networks for commercial businesses[3]; they are also increasingly used to accelerate knowledge production and diffusion in the scientific fields[29], [30], [31], [32]. Scientists consider wikis and collaborative tools in general as a convenient place to post ideas and comments but not to publish freely, because of the possibility of being scooped and lose credit [29]. The advocates of Science 2.0 affirm that Web technologies have to be sustained in order to move researchers toward the kind of openness and community that were supposed to be the hallmark of science in the first place and these new interactive technological forms are conceived to support traditional research, with the aim of facilitating scientific communication.

In the meantime, economic constraints have made research funding very competitive, stimulating research and funding institutions to put pressure on the research community to be effective in the dissemination of research results; therefore, several research institutions have put in place incentive systems on research publication outcomes, whereas funding agencies have been increasingly committed to maximise visibility and public access of research outcomes financed with public resources [33], [34], [35], [36], [37], [38], [39].

Publication strategies are influenced by the specific research field [40]. Communication strategies and mechanisms for the creation of trust among authors vary across disciplines [41], [42], [20], [43], [44], [45]. In 1999, Kling and McKim called for the need of systematic studies across disciplines, because past literature tended to homogenize publication strategies across different disciplines, thus inevitably promoting quantitative studies and methodologies associated with the most prolific disciplines in terms of publication, typically medicine and life sciences. In spite of a growing number of studies comparing publications from different research fields, the topic is still underexplored, particularly on the coexistence of traditional and alternative publishing tools. Many studies have looked at researchers in different fields, but without disaggregating results in a systematic way [46], [47], [48].

Of particular interest for the current debate of the future of scholarly publishing is the attitude towards digital tools. Kling [23], analysed how authors were facing the transition from paper to digital tools. Allen [49] focused on the differences among authors from humanistic disciplines in terms of engagement in depositing in institutional repositories. Antelmann

[3] E.g. *see* Vickery G., Wunsch-Vincent S, *Participative web and user-created content: Web 2.0, wikis and social networking*, 2007, OECD Publisher, available at http://www.oecd.org/document/40/0,3343,en_2649_34223_39428648_1_1_1_1,00.html (April 2010)

[50], [51] focused instead on how authors from different disciplines approached technological tools; other studies addressed the degree of acceptance of digital publications and new forms of publication and research diffusion, from open access journals to repositories [52], [53], [54]. Yet, publication strategies across disciplines as a response to institutional pressures and differences in behaviour among more productive and less productive researchers are issues still largely unexplored.

3. Methodology

In this paper, we wish to describe publication strategies of authors with different seniority and from different research fields; more specifically, we are interested in their choices of publication outcomes, their attitude towards new forms of publication and diffusion of scientific results (such as open access journals and repositories), their response to institutional incentives to publication. We claim that most studies on scholarly publishing take a "one size fits all" approach, in that they do not adequately consider differences among publications and differences among authors in terms of reputation and attitude to research. In any given academic institution, only a limited portion of faculty is devoted to research and only a limited portion of such faculty is highly productive, visible and targeting to top tier journals. Moreover, it is likely that publication patterns change with seniority, as authors reach a higher level of reputation and status on the one hand and are on the other less pressured to publish. Moreover, it is still unclear how different disciplines approach the coexistence of traditional and digital channels for publishing their works and if differences are due to the presence of specific norms of the field of belonging or to the different scientific framework in which authors work.

In the next paragraphs we wish to answer to the following questions:

- are there differences among authors in different disciplines as of where to publish and how much to publish?
- are these differences driven by discipline or by academic seniority?
- to what extent open access journals are being exploited as a viable publication channel? What drives their utilisation?

We feel that answers to these questions, although still preliminary, contribute to the current debate on the future of scholarly publishing as they start systematically comparing outcomes from different academic disciplines. More specifically, they help understand the current acceptance of open access journals as viable alternatives to traditional journals and under which conditions they are most appreciated.

Our empirical base consists of the institutional repository of Università degli Studi di Milano, for the years 2006-2009. The repository has been active since 2005 and currently holds a stock of 43,264 publications by 7,646 authors from 14 research areas.

The choice of this repository as the empirical base of our research is driven not only by convenience and accessibility, but also by the fact that it is the most complete institutional database in Italy, as the University made it mandatory to its faculty to archive scientific outputs since 2008 and the mandate has been effectively enforced[4], thus making the institutional repository a good reference to understand publication patterns across disciplines for scholars of different seniority. Table 1 shows the percentage of faculty members complying with the repository; as part of the faculty is not involved in publishing activities (particularly at a very young age, as it is the case with research assistants and first year PhD students), we feel that the repository is a good representation of the situation in this particular university.

Table 1: Percentage of faculty in institutional repository.

	Tenured professors	Full researchers	PhD and temporary researchers
In repository	1,037	703	337
Not in repository	339 (24.6%)	272 (27.9%)	332 (49.6%)
Total faculty	1,376	975	669

Moreover, Italy is characterised by huge differences in reputation and scientific productivity of universities, and the debate on the evaluation of scientific outcomes is very strong, as universities have to comply with national standards in the evaluation of scholars for career advancements.[5] Yet, the assessment of research outcomes[6] does not take into consideration authors' performances.[7] Milan University is a good starting point to address opportunities and difficulties in evaluating authors' performance, as it is characterised by big variety in terms of disciplines[8], level of authors'

[4] The IR has been defined primary source for every internal and external research assessment

[5] For further details see Reale, E. (2007), La valutazione della ricerca pubblica. Un'analisi della valutazione triennale della ricerca, Milano, Franco Angeli.

6 E.g. see http://www.crui.it/valutazione/HomePage.aspx?ref=1176.

[7] Research assessments involve only institutions and departments.

[8] All disciplines except engineering are represented.

productivity, international reputation of authors. More specifically, Medicine departments enjoy a long standing reputation at the national and international level for the quality of the research and education.

The repository archives publications authored by at least one faculty member; faculty was classified according to the following categories:

- tenured faculty (associate and full professors);
- permanent researchers
- temporary researchers (PhD students, research assistants…)

Publications were classified in the following categories

- books
- chapters of books
- journal articles
- conference proceedings.

The analysis was conducted in two steps. First an analysis of the overall database was performed, in order to assess:

- the percentage of faculty involved in research activity;
- the number of published outcomes;
- the number and type of published outcomes by seniority;
- the relative importance of different publication outcomes;
- the impact of career opportunities on research productivity;
- the language used to publish.

A subsequent analysis for a limited number of disciplines allows an analysis by author, carried on to highlight specific publication strategies for more active authors in terms of attention to more recent forms of publications (namely open access), language used, types of publication outcomes. As we were particularly interested in the penetration of open access, we chose to focus on the following disciplines: computer science, medicine, humanities, chemistry and physics. Medicine was chosen as being traditionally important and prestigious department within the university; the others were chosen as literature on open access stresses the importance of new forms of publications in these disciplines. Data were analysed using SPSS.

4. Analysis of results

Table 2 shows the distribution of publications by faculty and by year; for each faculty, announcements of permanent positions available within the university are highlighted.

Table 2: Publications and announcements of tenured positions per year.

		2006	2007	2008	2009
Agriculture	Publications	937	1225	1175	653
	Announcements	1	3	14	0
Pharmacy	Publications	930	981	924	400
	Announcements	0	3	6	0
Law	Publications	321	333	306	192
	Announcements	4	5	5	0
Letters	Publications	597	694	645	450
	Announcements	0	8	19	0
Medicine	Publications	3980	4433	4549	3719
	Announcements	11	15	15	0
Veterinary	Publications	1333	1443	1405	635
	Announcements	4	3	6	0
Mathematics, Natural Sciences, Physics	Publications	2311	2612	2425	1368
	Announcements	6	11	35	0
Sport Sciences	Publications	162	140	124	81
	Announcements	0	1	7	0
Political Sciences	Publications	431	534	488	328
	Announcements	1	6	7	0

Table 2a shows the breakdown of the announcements by role.

Table 2a: Announcements for positions available at Università degli Studi di Milano.

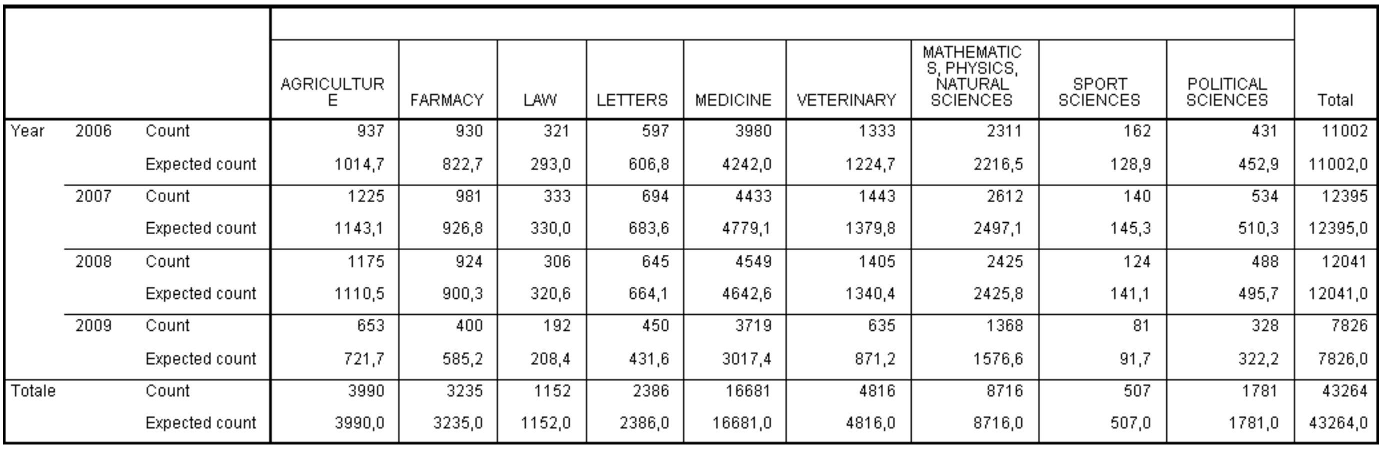
	2006	2007	2008
Full professor	3	0	15
Associate	0	0	55
Researchers	24	55	49

In 2006, 27 new academic positions were announced; this is a small number if compared with the 55 positions announced in 2007. We see a generalized increase of publication records deposited for all the faculties, except for the Faculty of Sport Science, for which no positions were announced. In 2008, there has been an increase of positions announced, in particular for the faculties of Physics, Mathematics and Natural Sciences and the Faculty of Medicine for the role of researchers and associate professors. If we look at publication patterns of the different roles, we see that there is an increase of publication for researchers, even though, in general, there is a decrease in the total publication stock for 2008. In 2009, no positions have been announced and we see a consistent decrease of publications for all the roles.

As it can be expected, there is an increase in the number of publications deposited over time, as the awareness about the repository increases and enforcement policies for its use are more effective. Moreover, both in 2007 and 2008, the university opened several tenure track positions, which are not surprisingly correlated with an increase in publication outcomes for those years. Although it is not necessarily true that positions will be occupied by the university faculty members (as these positions are opened at the national level), it is reasonable to expect that authors will try to put themselves in the position of becoming eligible candidates. For the Italian system, this is the highest incentive to publication; as positions are announced by law, potential candidates normally get ready one year in advance by increasing the number of their publications, so as to have higher chances to be admitted to the evaluation procedures; this explains why there is a strong drop in the publication rate between 2008 and 2009.

Table 3 shows the outcome of a cross tabulation analysis performed on the number of publications per faculty per year; the "expected count" row for each year shows the number of publications that one could expect in the hypothesis that there were no relationship between year and discipline. Broadly speaking, the table shows that Pharmacy, Law, Veterinary, Mathematics and Sport Sciences show a high publication activity in 2006; except Medicine, Law and Sport sciences all disciplines respond actively to incentives in 2007 in view of 2008 positions; Agriculture and Veterinary show active publication rates in 2008. Medicine is the only discipline with higher than expected publications in 2009. Apart from the opening of positions, all disciplines show a cyclical publication pattern.

Table 3: Publication pattern per discipline per year



			AGRICULTURE	FARMACY	LAW	LETTERS	MEDICINE	VETERINARY	MATHEMATICS, PHYSICS, NATURAL SCIENCES	SPORT SCIENCES	POLITICAL SCIENCES	Total
Year	2006	Count	937	930	321	597	3980	1333	2311	162	431	11002
		Expected count	1014,7	822,7	293,0	606,8	4242,0	1224,7	2216,5	128,9	452,9	11002,0
	2007	Count	1225	981	333	694	4433	1443	2612	140	534	12395
		Expected count	1143,1	926,8	330,0	683,6	4779,1	1379,8	2497,1	145,3	510,3	12395,0
	2008	Count	1175	924	306	645	4549	1405	2425	124	488	12041
		Expected count	1110,5	900,3	320,6	664,1	4642,6	1340,4	2425,8	141,1	495,7	12041,0
	2009	Count	653	400	192	450	3719	635	1368	81	328	7826
		Expected count	721,7	585,2	208,4	431,6	3017,4	871,2	1576,6	91,7	322,2	7826,0
Totale		Count	3990	3235	1152	2386	16681	4816	8716	507	1781	43264
		Expected count	3990,0	3235,0	1152,0	2386,0	16681,0	4816,0	8716,0	507,0	1781,0	43264,0

Table 4 compares outcomes by faculty and by type of output; a crosstab analysis shows the expected distribution per row and per column, should the row and column variables independent of each other.

Table 4: publications patterns by faculty; different work types.

		Article	Book chapter	Conference Proceedings	Books	
AGRICULTURE	Count	1988	330	1551	121	3990
	Expected count	2233,9	314,5	1266,9	174,8	3990,0
	% Faculty	49,8%	8,3%	38,9%	3,0%	100,0%
	% work type	8,2%	9,7%	11,3%	6,4%	9,2%
	% total	4,6%	,8%	3,6%	,3%	9,2%
PHARMACY	Count	2405	97	702	31	3235
	Expected count	1811,2	255,0	1027,2	141,7	3235,0
	% Faculty	74,3%	3,0%	21,7%	1,0%	100,0%
	% work type	9,9%	2,8%	5,1%	1,6%	7,5%
	% total	5,6%	,2%	1,6%	,1%	7,5%
LAW	Count	453	337	99	263	1152
	Expected count	645,0	90,8	365,8	50,5	1152,0
	% Faculty	39,3%	29,3%	8,6%	22,8%	100,0%
	% work type	1,9%	9,9%	,7%	13,9%	2,7%
	% total	1,0%	,8%	,2%	,6%	2,7%
LETTERS	Count	570	778	367	671	2386
	Expected count	1335,8	188,1	757,6	104,5	2386,0
	% Faculty	23,9%	32,6%	15,4%	28,1%	100,0%
	% work type	2,4%	22,8%	2,7%	35,4%	5,5%
	% total	1,3%	1,8%	,8%	1,6%	5,5%
MEDICINE	Count	9962	576	5945	198	16681
	Expected count	9339,1	1314,8	5296,5	730,6	16681,0
	% Faculty	59,7%	3,5%	35,6%	1,2%	100,0%
	% work type	41,1%	16,9%	43,3%	10,4%	38,6%
	% total	23,0%	1,3%	13,7%	,5%	38,6%
VETERINARY	Count	2654	98	2022	42	4816
	Expected count	2696,3	379,6	1529,2	210,9	4816,0
	% Faculty	55,1%	2,0%	42,0%	,9%	100,0%
	% work type	11,0%	2,9%	14,7%	2,2%	11,1%
	% total	6,1%	,2%	4,7%	,1%	11,1%
MATHEMATICS, PHYSICS, NATURAL SCIENCES	Count	5325	567	2663	161	8716
	Expected count	4879,8	687,0	2767,5	381,8	8716,0
	% Faculty	61,1%	6,5%	30,6%	1,8%	100,0%
	% work type	22,0%	16,6%	19,4%	8,5%	20,1%
	% total	12,3%	1,3%	6,2%	,4%	20,1%
SPORT SCIENCES	Count	270	8	220	9	507
	Expected count	283,9	40,0	161,0	22,2	507,0
	% Faculty	53,3%	1,6%	43,4%	1,8%	100,0%
	% work type	1,1%	,2%	1,6%	,5%	1,2%
	% total	,6%	,0%	,5%	,0%	1,2%
POLITICAL SCIENCES	Count	595	619	168	399	1781
	Expected count	997,1	140,4	565,5	78,0	1781,0
	% Faculty	33,4%	34,8%	9,4%	22,4%	100,0%
	% work type	2,5%	18,2%	1,2%	21,1%	4,1%
	% total	1,4%	1,4%	,4%	,9%	4,1%
Total	Count	24222	3410	13737	1895	43264
	Expected count	24222,0	3410,0	13737,0	1895,0	43264,0
	% Faculty	56,0%	7,9%	31,8%	4,4%	100,0%
	% work type	100,0%	100,0%	100,0%	100,0%	100,0%
	% total	56,0%	7,9%	31,8%	4,4%	100,0%

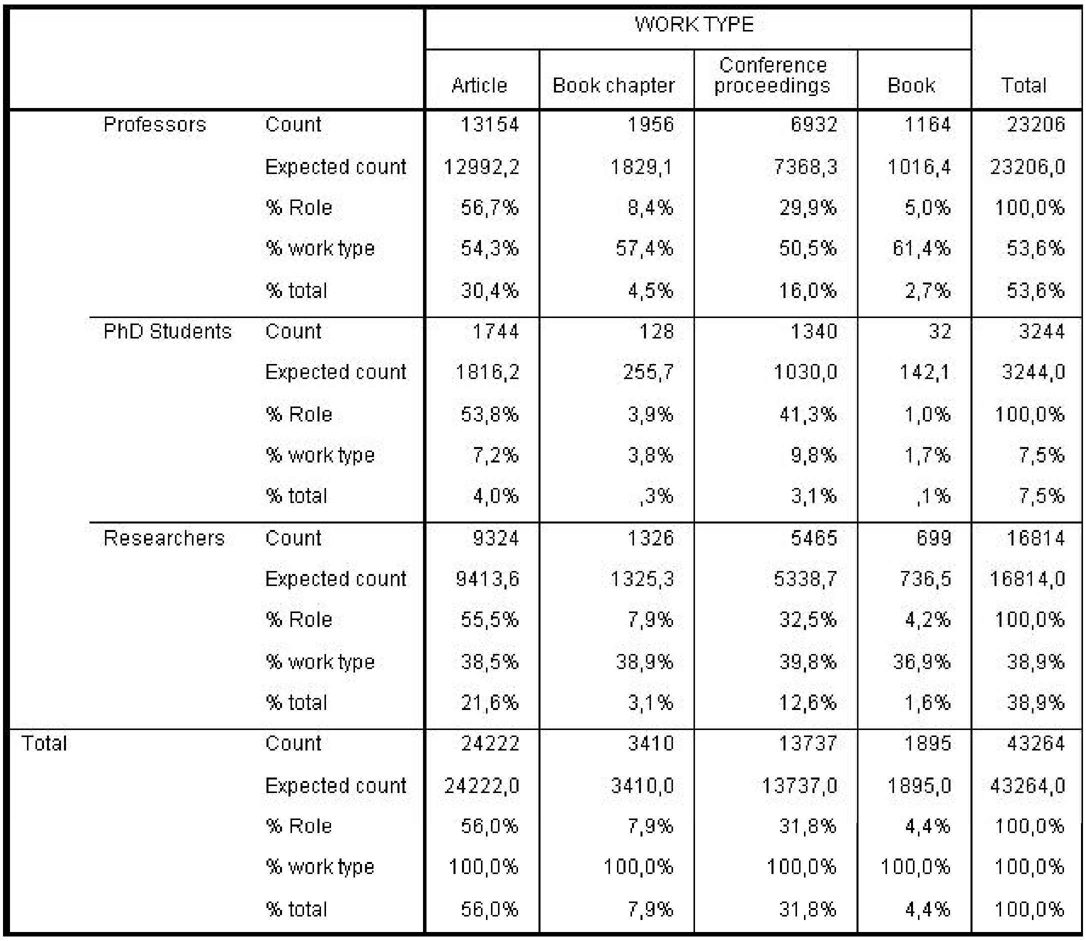
If we look at the differences in terms of publication stocks for different faculties, we see not surprisingly that the most productive faculty in terms of publication stock is the Faculty of Medicine, with a stock of 16.681 publications, accounting for 38% of the total references archived in AIR. If we add Veterinary Medicine, the percentage rises to 48% (4.816 publications). The second Faculty in term of publication stock is Physics, Mathematics and Natural Sciences, with 8.716 works archived. After Veterinary Medicine, we find the Faculties of Agrarian Studies and Pharmacy. Faculty members of Humanistic Faculties contribute for a minor part to the publication stock of AIR. Letters, Law and Political Sciences represent together 12.3% of the repository. Sport Sciences Faculty, although cannot be classified with humanistic faculties, follows the same path and accounts for 1.2%. On the whole the least represented Faculty is Law (2.7% of the references in AIR).

Concerning the kinds of works published, an important distinction emerged from data analysis is between faculties more used to write articles and faculties more focused on book chapters publication.

The hard writers of articles are the members of scientific faculties, Medicine at the first place, representing more than 50% of the total articles (together with Veterinary Medicine) archived in AIR repository in the period 2006 – 2009. Almost 60% of the works published by faculty members of Medicine are articles. Also the faculty members of Physics, Mathematics and Natural Sciences write a consistent number of articles with respect to their colleagues of other faculties. Their articles represent 22% of the total articles archived in AIR and 61.4% of their works are articles. The least productive faculty in terms of articles is the Sport Science Faculty, which represents 1.1% of the total articles. A similar pattern is present for the faculties of Law, Letters and Political Sciences. Medicine and Physics, Mathematics and Natural Sciences have a consistent number of publications also for other kinds of works; in particular their works represent respectively 20% and 16.9% of book chapters in AIR. Concerning this kind of publication, the Faculty of Letters is the most productive, representing almost 23% of the total of book chapters. For this Faculty, even tough the publication of books chapters has higher percentage respect to that of articles (32,6% vs. 23,9%), the difference among publication channels is less evident with respect to scientific faculties.

Contrary to common wisdom, the publication flow of faculty does not stop once tenure is attained. Quite the contrary, tenured faculty are responsible for a high number of publications. For all the three categories (Researchers, PhD Students, Tenured Professors), journal articles represent the most used publication channel, representing almost half of the publication stock of the three categories. Not surprisingly, tenured faculty tend to publish an increased number of books, while PhD students, tend to be overrepresented as far as the incidence of conference proceedings (41.3%) is concerned.

Table 5 : distribution of outcomes per academic seniority

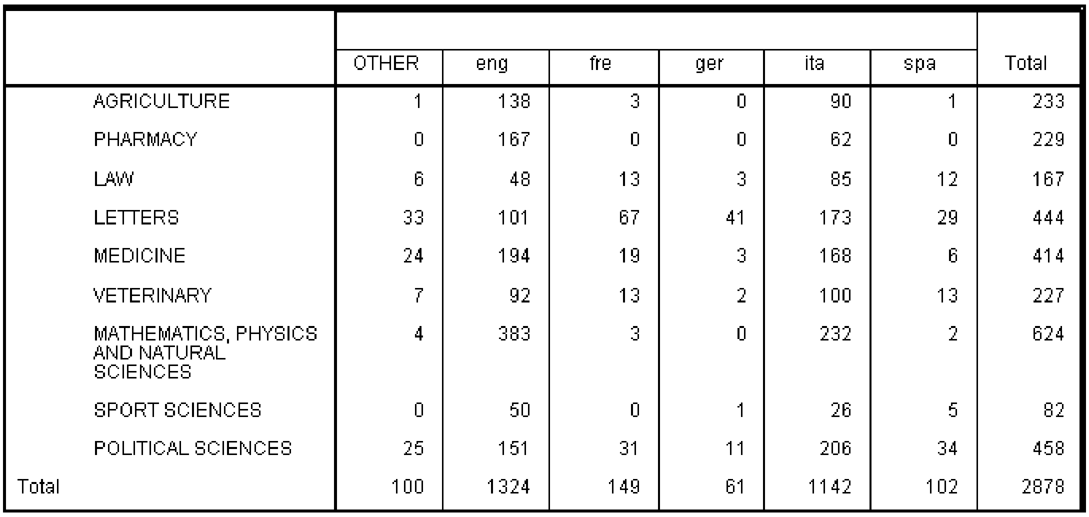


			WORK TYPE				Total
			Article	Book chapter	Conference proceedings	Book	
	Professors	Count	13154	1956	6932	1164	23206
		Expected count	12992,2	1829,1	7368,3	1016,4	23206,0
		% Role	56,7%	8,4%	29,9%	5,0%	100,0%
		% work type	54,3%	57,4%	50,5%	61,4%	53,6%
		% total	30,4%	4,5%	16,0%	2,7%	53,6%
	PhD Students	Count	1744	128	1340	32	3244
		Expected count	1816,2	255,7	1030,0	142,1	3244,0
		% Role	53,8%	3,9%	41,3%	1,0%	100,0%
		% work type	7,2%	3,8%	9,8%	1,7%	7,5%
		% total	4,0%	,3%	3,1%	,1%	7,5%
	Researchers	Count	9324	1326	5465	699	16814
		Expected count	9413,6	1325,3	5338,7	736,5	16814,0
		% Role	55,5%	7,9%	32,5%	4,2%	100,0%
		% work type	38,5%	38,9%	39,8%	36,9%	38,9%
		% total	21,6%	3,1%	12,6%	1,6%	38,9%
Total		Count	24222	3410	13737	1895	43264
		Expected count	24222,0	3410,0	13737,0	1895,0	43264,0
		% Role	56,0%	7,9%	31,8%	4,4%	100,0%
		% work type	100,0%	100,0%	100,0%	100,0%	100,0%
		% total	56,0%	7,9%	31,8%	4,4%	100,0%

Finally, Table 6 analyses the internationalisation pattern of faculty publications.

Even in this case, there is a clear distinction between scientific and humanistic disciplines. Faculty members from scientific disciplines are used to publish in English. For the Faculty of Physics, mathematics and natural sciences, as well as Pharmacy, the majority of contributions are written in English. If we report the numbers of table 6 to the total publication stock, we see that 61.4% of faculty members' publications are in English and only 37.2% are in Italian. The rest is published in other languages. A similar situation characterises faculty members of Sport Sciences. For Medicine we find a major balance between English and Italian publications: 46.9% of publications are in English and 40.6% are in Italian.

Table 6: publications for different languages.



	OTHER	eng	fre	ger	ita	spa	Total
AGRICULTURE	1	138	3	0	90	1	233
PHARMACY	0	167	0	0	62	0	229
LAW	6	48	13	3	85	12	167
LETTERS	33	101	67	41	173	29	444
MEDICINE	24	194	19	3	168	6	414
VETERINARY	7	92	13	2	100	13	227
MATHEMATICS, PHYSICS AND NATURAL SCIENCES	4	383	3	0	232	2	624
SPORT SCIENCES	0	50	0	1	26	5	82
POLITICAL SCIENCES	25	151	31	11	206	34	458
Total	100	1324	149	61	1142	102	2878

For humanistic disciplines, the majority of contributions is written in Italian (39%) and only a minor part in English (22.7%). What is interesting is that faculty members of Letters have a good number of publications in other languages too, showing a remarkable international attitude. The least international Faculty is that of Law, with almost 60% of contributions written in Italian.

Results so far confirm the existence of different publication patterns across disciplines and different publication strategies related to seniority. Journal articles are increasingly becoming the most common publication outcome across disciplines, although books are more relevant in humanistic disciplines and are generally published when faculty reach academic maturity. Younger scholars start building their reputation through participation to conferences across all disciplines and gradually publish on academic journals and edited books. Medicine is the most prolific discipline in terms of publication outputs, while Law (but also Veterinary and Political Sciences) is the least international. For hard sciences English is more common than Italian, while humanistic disciplines show a broader spectrum of languages covered.

The second step of our analysis looks at the acceptance of open access publications as viable alternatives to traditional journals for authors across different disciplines. Due to the characteristics of the repository we could only track gold open access journals and not other forms of open repositories.

We therefore identified five disciplines and tracked the evolution of open access publication between 2006 and 2009. The five disciplines are chemistry, physics, letters, medicine and computer science.

Chemistry, like physics, is considered an advanced scientific discipline for the use of alternative publishing routes [23], whereas the opposite can be said for letters, whose authors are traditionally less appreciative of the digital features of journals [52] and favour books to articles in journals as the preferred mode of publication [40]. Medicine, and scientists in general, are used to conduct systematic directed searches in aggregated databases on line to validate their findings and to look for early visibility for their works. Concerning Computer Sciences, faculty members are supposed to have the necessary skills to use IT tools and, for what concerns their publishing strategies, they are influentially driven by monetary return and this fact could have deep influences on their approach to open access [40].

Four years is not a very long time span, but it allows looking at the growth in acceptance of Open Access across different disciplines. In order to analyse authors' publication strategies with respect to the introduction of Open Access, we looked at individual authors' behaviour in five disciplines over the time span analysed.

On the whole, open access publications represent 2.3% of the total publication records in AIR by the considered disciplines, which is quite modest in absolute terms. If we look at the faculties who have the greatest percentages of OA articles, Computer Science is the discipline with the highest number of publications (3%), followed by Medicine (2,7%). This result can be explained by the international orientation of some of the faculty in these disciplines and by the presence of high reputation open access journals both these disciplines have been early exposed to changes of scientific publication tools and have developed an early aptitude to openness.

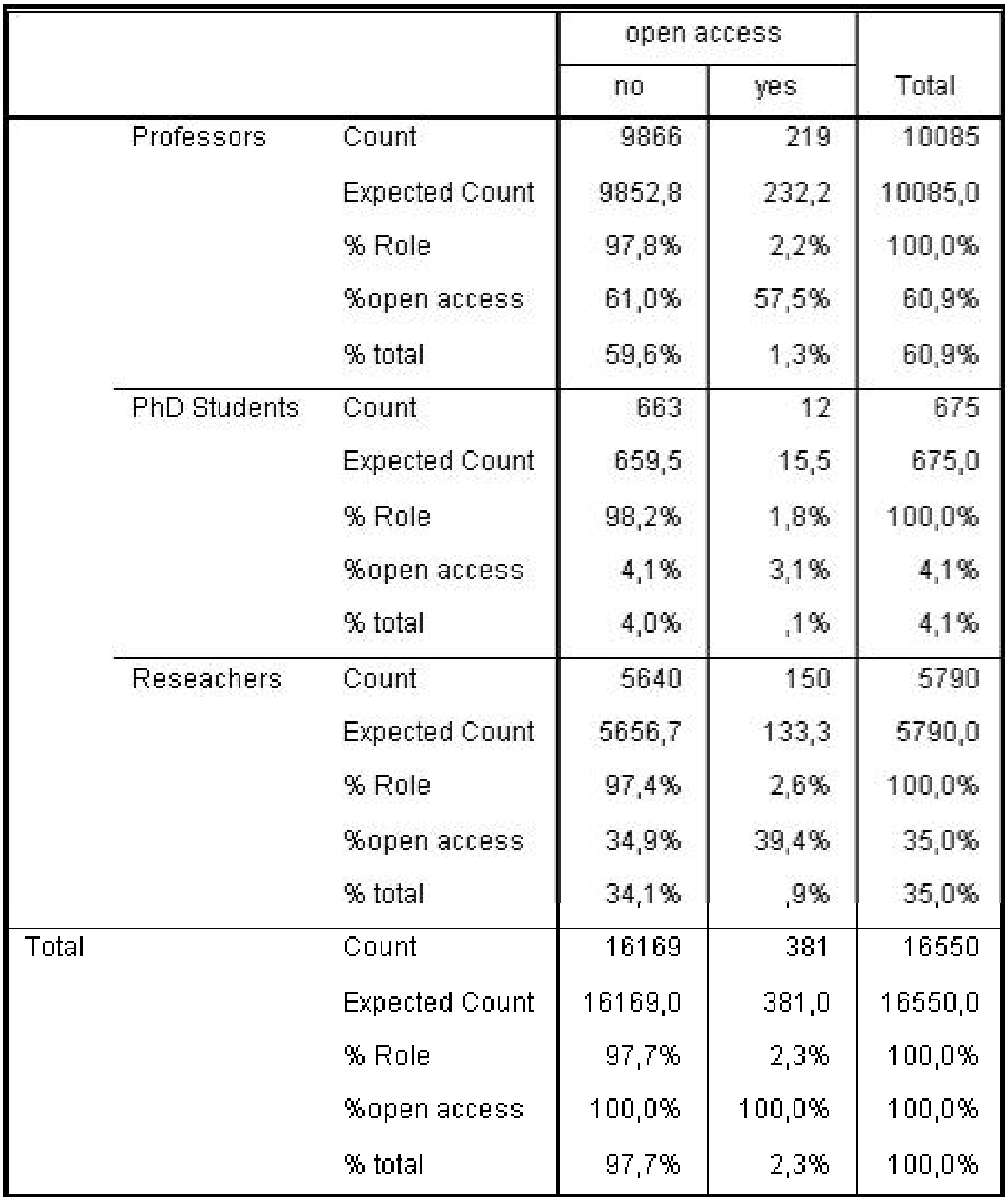
Chemistry follows the same pattern while, contrary to letters, physics is underrepresented among open access journals in our sample. University of Milano has opened several positions in physics for researchers and tenured professors between 2006 and 2008. Given the fact that open access journals are a relatively recent phenomenon, it may be that authors have preferred more traditional publication outcomes so as to maximise their chances of complying with the criteria set by evaluation committees.

Table 7: evolution of open access articles for the five disciplines considered over the time span considered.

	2006	2007	2008	2009
Number of OA articles	85	74	142	80
Number of articles	4,227	4,294	4,567	3,462

If we focus on the number of Open Access publications deposited by scholars of different academic positions, we see that tenured faculty is more keen to publish OA; their works represent 57,5% of the total OA publications in AIR by the considered disciplines. Researchers represent 39.4% and PhD students only 3.1%.

Table 8: number of open access articles.



			open access		Total
			no	yes	
	Professors	Count	9866	219	10085
		Expected Count	9852,8	232,2	10085,0
		% Role	97,8%	2,2%	100,0%
		%open access	61,0%	57,5%	60,9%
		% total	59,6%	1,3%	60,9%
	PhD Students	Count	663	12	675
		Expected Count	659,5	15,5	675,0
		% Role	98,2%	1,8%	100,0%
		%open access	4,1%	3,1%	4,1%
		% total	4,0%	,1%	4,1%
	Reseachers	Count	5640	150	5790
		Expected Count	5656,7	133,3	5790,0
		% Role	97,4%	2,6%	100,0%
		%open access	34,9%	39,4%	35,0%
		% total	34,1%	,9%	35,0%
Total		Count	16169	381	16550
		Expected Count	16169,0	381,0	16550,0
		% Role	97,7%	2,3%	100,0%
		%open access	100,0%	100,0%	100,0%
		% total	97,7%	2,3%	100,0%

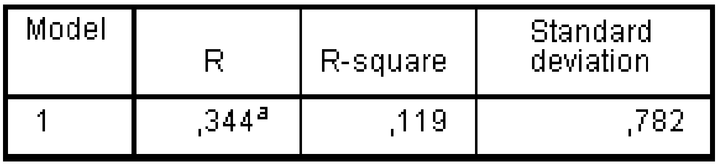
This result is consistent with findings of past researches on this topic: younger faculty tend to prefer more established channels in order to get legitimization; senior faculty, who has also access to higher funding, has more degrees of freedom and therefore can experiment with a wider range of publication activities.

Table 9: averages per discipline.

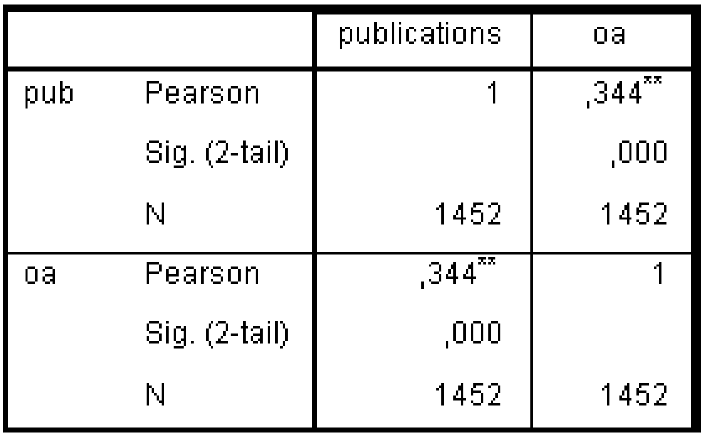
	chemistry	physics	letters	computer sciences	medicine
Average number of articles per author	11,32	11,32	11,31	11,37	11,39
Average number of publications for the top 20 of authors in terms of articles published	31,05	33,3	18,45	15	100,15
% of OA articles	2%	0,5%	1,1%	3%	2,7%

Not surprisingly, OA publications tend to concentrate among authors with the highest publication rate. We find a mild correlation between the number of publication and the number of OA publications. Yet, the numbers are too small for a generalisation.

Table 10: regression output.



Model	R	R-square	Standard deviation
1	,344[a]	,119	,782



		publications	oa
pub	Pearson	1	,344[**]
	Sig. (2-tail)		,000
	N	1452	1452
oa	Pearson	,344[**]	1
	Sig. (2-tail)	,000	
	N	1452	1452

5. Conclusions

In this paper, we were interested in exploring different publication strategies put in place by authors from different disciplines and seniority, in order to identify common trends and peculiarities within the same institutions; not surprisingly, the first result is that each discipline shows idiosyncratic patterns, particularly as far as the preferred publication outcomes are concerned: books are the preferred form of publication among scholars in the humanities, whereas journal articles are preferred in science. Moreover, internationalisation patterns are quite different across disciplines, with some (namely physics) being a global academic field with English as the language of reference, while others show the need to address both the local and the international audience; in the humanities, English is not a lingua franca, rather, other languages are also used to communicate scientific outcomes.

Yet, the analysis of results shows remarkably common patterns across disciplines. Within the academia, only a limited number of faculty members publish and an even smaller number of them publishes regularly and a significant number of contributions per year; yet, those who publish are quite active even when they reach tenure, and this is true across disciplines. Incentives do play a role: opening of positions within the university is correlated to a sharp increase in the number of publications.

As publications alternatives multiply, it is becoming increasingly important for the author to be aware of what rights, opportunities and limitations are associated with different channels. At the same time, the increased variety of juridical options associated with each of them makes publication decisions more complex than in the past. In this respect, authors in our sample tend to follow a quite conservative approach in choosing where to publish; younger faculty members tend to be more active in conference participation, while more senior faculty progressively publish journal articles and books. Open access is still a very small percentage of outcomes and there is a mild correlation between tendency towards open access and intensity of publications.

Incentives seem to be the most effective way to modify publication strategies: as scientific communities tend to be quite resilient, changes in the patterns are likely to be introduced either by relatively senior and active authors or by specific policies put in place at the faculty level.

From this emerging perspective, new publication ways can be integrated in the knowledge creation process of science and they can be considered important vectors for the final steps of diffusion, in parallel with traditional channels. For some of these channels there is a lack of transparency for what concerns the consideration by academic communities and the evaluation for promotion and tenure decision within departments and Universities. Current incentive mechanisms of universities can therefore represent an obstacle to a wider diffusion of new publication models [55], when they are not aligned with the trends of the scholarly publishing sector, reflecting the established norms of the traditional academic environment.

Authors seem to be rational in their publication strategies: apart from the selected few who are systematic authors "no matter what", academic authors in our sample respond to career advancement opportunities and publish in established channels defined up by their departments / communities of peers. Although we could not measure it, impact factor most likely drives journal selection. New alternative models for early visibility or publication (such as repositories like SSRN or PLoS) are clearly showing that it is possible to offer an alternative to traditional journals, provided that they are able to attract significant numbers of readers, offer high IF in addition to open access. Should they be able to comply with academic requirements, they will undoubtedly succeed in attracting to authors.

We do not have enough data to statistically verify this datum. Yet, it is likely that gold open access is more available to senior faculty, which is likely to have more access to financial resources. This needs to be taken into consideration in resource allocation, should the gold open access model be encouraged.

Notes and Bibliography

[1] Cope, B., and Kalantzis, M. (2000). Designs for social futures. In B. Cope, and K. Mary (editors). *Multiliteracies: Literacy learning and the design of social futures.* London:Routledge, pp. 203–234.

[2] Cope B. and Kalantzis, M. (2009) Signs of epistemic disruption: transformation in the knowledge system of academic journals. First Monday, 14(4) http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2309/2163 (April 2010)

[3] Kress, G. (2000). Design and transformation: New theories of meaning. In: B. Cope, and M. Kalantzis (editors). *Multiliteracies: Literacy learning and the design of social futures.London: Routledge*, pp. 153–161.

[4] Weiss, Y. , and Lillard, L. (1982). Output Variability, Academic Labor Contracts, and Waiting Time for Promotion. *Research in Labor Eco- nomics, 5* : 157-188.

[5] O' Neill G.P., and Sachis P.N. (1994). The importance of refereed publications in tenure and promotion decisions: A Canadian study. *Higher Education*, *28*(4): 472-435.

[6] Houghton, J. (2000). Economics of Scholarly Communication. Discussion paper. Center for Strategic Economic Studies, VictoriaUniversity, available at http://www.caul.edu.au/cisc/EconomicsScholarlyCommunication.pdf (April 2010)

[7] *Guedon J. C. (2001)* In Oldenburg's long shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing*, ARL. Available at:* *http://www.arl.org/resources/pubs/mmproceedings/138guedon.shtml (April 2010)*

[8] Merton, R. (1973). *The Sociology of Science: Theoretical and Empirical Investigation*. Chicago, University of Chicago Press.

[9] Dasgupta, P. and David, P.A. (1994). Towards a new economics of science. *Policy Research, 23*(5) 487–521. available at http://ideas.repec.org/a/eee/respol/v23y1994i5p487-521.html (April 2010).

[10] Coe*,* R.*,* *&* Weinstock, I. (1969*).* Evaluating journal publications*:* Perceptions versus reality. AACSB Bulletin, 1, 23-37

[11] MacMillan, I. C., and Stern I. (1987), Delineating a forum for business policy scholars, Strategic Management Journal, *8*: 183- 187.

[12] MacMillan, I. C. (1991).The emerging forum for business policy scholars, Journal of Business Venturing, 9(2): 85-89.

[13] Gordon, M. E., and Purvis, I.E. (1991). Journal publication records as a measure of research performance in industrial relations. Industrial and Labor Relations Review, 45(1): 194-201

[14] Park, S. H., and Gordon, M.E. (1996).Publication records and tenure decisions in the field of strategic management, Strategic Management Journal, 17(2): 109-128

[15] Lowe, A., and Locke, J. (2005), Perceptions of journal quality and research paradigm: results of a web-based survey of British accounting academics, Accounting, Organizations and Society 30(1): 81–98.

[16] Baden-Fuller C., and Hwee Ang S. (2000). Building Reputations: The Role of Alliances in the European Business School Scene, Long Range Planning 34(6): 741–755

[17] Doran, J.S., and Wright, C. (2007). So You Discovered an Anomaly… Gonna Publish It? An investigation into the rationality of publishing market anomalies. Working Paper. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=956105 (April 2010)

[18] Kling, R., and Iacono, S. (1989) The institutional character of Computerized Information Systems. *Office: Technologies and People 5* (1): 7-28.

[19] Kling, R., and McKim, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. Journal of the American Society for Information Science, 51: 1306- 1320.

[20] Fry, J. (2006). Scholarly Research and Information practices: a domain analytic approach. *Information Processing and Management, 42*: 299-316.

[21] Curtis, K.L., Weller, A.C., & Hurd, J. (1997). Informationseeking behaviour of health sciences faculty: The impact of new information technologies. Bulletin of Medical Library Association, 85, 402-408.

[22] Rogers, E. M. (1995) Diffusion of Innovation. New York. Free Press.

[23] Kling, R., and McKim, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. Journal of the American Society for Information Science, 51: 1306- 1320.

[24] Foster, N.S., and Gibbons, S. (2005). Understanding Faculty to understand content Recruitment for Institutional Repositories. D-lib Magazine, 11(1). Available at http://www.dlib.org/dlib/january05/foster/01foster.html. (April 2010)

[25] Talja, S., Vakkari, P., Fry, J., and Wouters, P. (2007). Impact of Research Cultures on the use of Digital Library Resources. *Journal of the American Society for Information Science and Technology, 58*(11): 1674-1685.

[26] Bauer, K., and Bakkalbasi, N. (2005). An Examination of Citation Counts in a New Scholarly Communication Environment. *D-lib Magazine, 11*(9). Available at http://www.dlib.org/dlib/september05/bauer/09bauer.html (April 2010)

[27] Clarke, R. (2005). A proposal for an open content licence for research paper (Pr)ePrints, *First Monday, 10*(8). Available at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/issue/view/187 (April 2010).

[28] Bar – Ilan, J. (2008). Which H-index? A comparison of Wos, Scopus and Google Scholar. Scientometrics, 74(2): 257-271.

[29] Waldrop, M. (2009). Science 2.0: Great Tool or great risk?, Scientific American online. available at http://www.sciam.com/article.cfm?id=science-2-point-0-great-newtool-or-great-risk&page=5 (April 2010)

[30] Harnad, S. (2003),Maximizing University research impact trough self-archiving.available at http://jcom.sissa.it/archive/02/04/A020401/ (April 2010)

[31] Schneiderman, B. (2007), Science 2.0, Science, 319: 1349-1350

[32] Hooker B. (2006). The future of Science is Open. available at http://3quarksdaily.blogs.com/3quarksdaily/2006/10/the_future_of_s_1.html (April 2010)

[33] Katz, D. A. (1973). Faculty Salaries, Promotions, and Productivity at a Large University. American Economic Review, 63(3) : 469-477.

[34] Weiss, Y. , and Lillard, L. (1982). Output Variability, Academic Labor Contracts, and Waiting Time for Promotion. Research in Labor Economics, 5 : 157-188.

[35] MacMillan, I. C., and Stern I. (1987), Delineating a forum for business policy scholars, Strategic Management Journal, 8: 183-187.

[36] Gordon, M. E., and Purvis, I.E. (1991). Journal publication records as a measure of research performance in industrial relations. Industrial and Labor Relations Review, 45(1): 194-201.

[37] O' Neill G.P., and Sachis P.N. (1994). The importance of refereed publications in tenure and promotion decisions: A Canadian study. Higher Education, 28(4): 472-435.

[38] MacMillan, I. C. (1994).The emerging forum for business policy scholars, Journal of Business Venturing, 9(2): 85-89.

[39] Park, S. H., and Gordon, M.E. (1996).Publication records and tenure decisions in the field of strategic management, Strategic Management Journal, 17(2): 109-128.

[40] Kingsley, D. A. (2008). PhD thesis. Australian National University.

[41] Becher T. (1981). Towards a definition of disciplinary culture. Studies 6(2). 109-122.

[42] Becher T. (1984). The significance of disciplinary differences. Studies in higher education 19(2) 151-161.

[43] Sparks S. (2005). *JISC disciplinary differences.* JISC Scholarly Communication Group

[44] Walsh J. P. and Bayma T (1996). Computer network and scientific work. Social studies of Science 26(3), 661-703.

[45] Whitley R. (1984) *The intellectual and social organization of the Sciences*. Oxford University press

[46] Bergstrom, T. (2007) and Lavaty R. . How often do economists self-archive?, EScholarship Repository. Available at http://repositories.cdlib.org/ucsbecon/bergstrom/2007a/ (April 2010).

[47] Cozzarelli, Nicholas R., Kenneth R. Fulton, and Diane M. Sullenberger. "Results of a PNAS Author Survey on an Open Access Option for Publication." Proceedings of the National Academy of Sciences of the United States of America 101, no. 5 (2004): 1111. http://www.pnas.org/cgi/reprint/101/5/1111.pdf (April 2010)

[48] Gadd E. at al. (2003) Romeo Studies 2: Haw academics want to protect their open access research papers. Journal of information science 29(5) 333-356.

[49] Allen J. Interdisciplinary attitudes towards deposit in institutional repositories http://eprints.rclis.org/archive/00005180/ (april 2010)

[50] Antelmann K. (2006) Self archiving practice and the influence of publisher policies in the Social Sciences. Learned publishing 19(2) 85-95.

[51] Talja S. Et al. (2004) Field differences in the use and perceived usefulness of scholarly mailing lists. Information research 10(1).

[52] Swan, A., and Brown, S. (1999) What Authors Want: the Motivations and Concerns of contributors to Learned Journals, Learned Publishing 12(3): 170-172. Available at http://www.alpsp.org/ForceDownload.asp?id=142 (April 2010)

[53] Swan, A., and Sheridan, B. (2005), Open Access and self – archiving: an author study. available at http://cogprints.org/4385/ (April 2010)

[54] Kennan, M.A. (2007), Academic authors, scholarly publishing and open access in Australia. *Learned Publishing, 20*(2): 138-146. Available at http://docserver.ingentaconnect.com/deliver/connect/alpsp/09531513/v20n2/s7.pdf?expires=1245422858&id=50857584&titleid=885&accname=Guest+User&checksum=1638DD04FB019AFB8514FA48903C9B02 (April 2010)

[55] Dubini, P., and Giglia E. (2009), Economic Sustainability during transition: the case of scholarly publishing. available at http://portale.unibocconi.it/wps/wcm/connect/resources/file/eb137a0a686d8e3/Dubini%20Giglia%20%20Economic%20sustainability%20during%20transition.pdf (April 2010)

The changing scholarly information landscape: Reinventing information services to increase research impact

Linda O'Brien

Information Services, Griffith University, Queensland, Australia
Linda.OBrien@griffith.edu.au

Abstract

In an increasingly competitive global knowledge economy the role of the university in a nation's innovation agenda has taken on greater prominence. Innovation through knowledge creation and application is seen to be the driver of long term national economic and social prosperity. With this recognition comes a growing interest by government in quality assuring and measuring the value of their universities. University league tables have become an accepted part of this landscape, as nations seek to position themselves in a globally competitive environment. A university's research impact - the extent to which its research informs further research and practice - is a significant component of the innovation system, and of the league table measures. Citation impact is often used as a proxy for research impact, though it only tells part of the story. Against this backdrop the research lifecycle is being transformed by information and communication technologies, fundamentally changing the scholarly information and communication landscape. What once appeared to be a linear process, from research through to publication, has become more complex, more collaborative, challenging the boundaries between disciplines, organisations, nations [1]. Emerging opportunities to leverage research data to increase research impact have yet to be realised. Universities, as long lived institutions, must balance short term utilitarian demands driven by national innovation agendas and league table positioning, with their fundamental mission of knowledge creation, synthesis, transmission and preservation. This is a mission that aligns strongly with the traditional place of the library in providing access to scholarship for current and future generation for all who wish to learn, a role that has been challenged by apparently ubiquitous access to digital content. The complexity of the current environment offers new opportunities for a university's information service providers to further the university, and the nation's aspirations – both short and long term. Information service providers are ideally positioned to navigate the complexity of the scholarly information landscape to achieve university imperatives within a national context, through collaboration within and across organisational boundaries; to achieve short term imperatives whilst staying true to the long term mission of universities in knowledge creation, dissemination and preservation for future generations of scholars and practitioners. Griffith University, a research intensive, innovative university situated within the south east corner of Queensland, Australia enjoys the benefit of an integrated information services division. Information Services brings together library, information and IT professionals to provide the information leadership, services, systems and infrastructure which underpin the University's research, learning, teaching and administrative activity. Over the last twelve months the division has built on its strengths to re-shape its services to tightly align them with University's aspirations. A significant part of this re-shaping has been the implementation of new service models, new services and systems, and strengthened partnerships, to increase the University's research impact. This initiative has been welcomed by the academy. More complex measures will be required to indicate the success of this initiative over time.

Keywords: research impact; university information services;

1. Introduction

In an increasingly competitive global knowledge economy nations are looking to their universities to drive the economy and prosperity. University leagues tables seem to now be a permanent part of the higher education landscape. A university's research impact - its impact on future research and practice - is a key driver of national innovation and a core component of the leagues tables. We are seeing profound change in the scholarly information and communication lifecycle as technology facilitates new ways of researching, communicating, collaborating, and sharing scholarly outcomes. It is still unclear how this will evolve as conventional policies and practices are challenged by the opportunities offered through technological innovation. Taken collectively these trends have a profound impact on the way in which a university can, or should, seek to increase its research impact.

This paper explores these issues, concluding by showing how one Australian university is seeking to increase research impact through the role played by its information services. It begins by outlining Australia's national innovation agenda as it relates to universities, the research quality agenda and then explores how the scholarly information and communication landscape is changing. These observations are framed within the context of broader international trends. The concept of research impact is explored, a concept which is used differently by different stakeholders. It concludes with an exploration of how a university's information services can serve to increase research impact, using Griffith University as a case study.

2. Universities and the innovation agenda

"Over the last decade or so there has been firmly established among governments around the world the view that high quality, internationally competitive research and higher education, mostly contained within universities, are prerequisites for long-term success in globalised knowledge economies."[2]

The generic social and economic benefits of universities – through educating the population and generating knowledge - have long been recognised as an important source of industrial innovation [3]. More recently, in a world where knowledge and its application is seen as the key to global competitiveness, the world's developed and developing nations have renewed their focus on knowledge innovation as a driver of national prosperity, advocating a central role for universities [4]. Australia is no exception, following in the path that the UK and others have already travelled, though from a perspective relevant to the national context [5].

A logical consequence of governments' viewing universities as sources of highly specific benefits, as drivers of innovation and national prosperity, is a tendency to then regulate and stimulate to drive specific behaviours, and a consequent desire to measure the success of these policy drivers. The significant government investment in research infrastructure over the past decade, in e-science and cyber-infrastructure across North America, Europe and Australia; has been to stimulate national performance and competitiveness [6] - whilst the obsession with university research quality assessment and rankings is a consequence of the desire to measure performance (and often to provide a regulatory measure through performance driven funding).

The Australian Minister for Innovation, Industry, Science and Research commissioned a review of the national innovation system in 2008 [7] leading to publication of a Government innovation strategy in 2009 [8]. The review examined the way in which Australia's national innovation system was positioned in a globally competitive Internet-enabled world. Whilst Australia is small and geographically remote, with less than 1% of the global economy, we manage to produce 2% of the world's scientific literature [9]. Australia has more Nobel prize winners per capita than any other nation [10]. Yet a range of indicators showed that we had slowed in terms of productivity growth, despite an enviable record. The review raised concern that investment levels in research in Australian universities was low as compared to the OECD. The scope of the review was broad ranging – looking at the three highly interdependent aspects of a national innovation system: the development of new knowledge and ideas, the deployment of those ideas in a real world context and the diffusion and adoption of applied knowledge. The important contributions made by the social sciences and humanities to the health and prosperity of the nation were acknowledged- the review wasn't purely science and technology focused.

The breadth of the review meant that universities featured prominently – being seen as the repositories of existing knowledge and the hub for generation and exchange of new knowledge [11]. There was recognition that our understanding of innovation had changed – what is being referred to as the concept of 'open innovation' [12]. Innovation increasingly relies on distributed inter-organisational networks rather that innovation within an organisation. Universities form part of multi-faceted social or information channels or mechanisms through which information, knowledge and other resources are shared or co-produced- a much richer picture of university engagement than that of the traditional university concept of knowledge transfer [13]. The critical value of the nation's information infrastructure to the national innovation system was therefore central: from high speed networks and collaboration tools, through to the value of unlocking public information and content, the importance of the national collections held by libraries, museums and other agencies. Specifically the review acknowledged the need for a high level of interaction between knowledge providers and knowledge users – particularly given that productivity growth in Australia will require the capability to adopt and adapt the 98 percent of new knowledge which is generated by the rest of the world [14].

The subsequent national strategy: Powering Ideas adopts many of the review's recommendations. More than A$3.1 billion in funding is to be made available through the strategy. Of particular relevance to this paper are that over the next four years, there will be more than doubling of funding for the indirect costs of research, a A$1.1 billion investment in science infrastructure with A$312 targeted at e-research infrastructure funding. This includes A$97 million for data storage and collaboration tools through the Australian Research Collaboration Service (ARCS), A$48 million to establish a national research data commons through the Australian National Data Service (ANDS) A$48 million, A$37 million to enhance the Australian Research Network, A$130 for national high performance computing initiatives and A$37 million for enhancement to the Australian research and education network [15].

Geoffrey Boulton and George Lucas critically examine the role of universities, questioning the current obsession of governments with universities as drivers of innovation [16]. They clearly articulate why, in a world of globalisation, universities are crucial national assets: “they research into the most theoretical and intractable uncertainties of knowledge yet also seek the practical application of discovery; they test, reinvigorate and carry forward the inherited knowledge of earlier generations; they seek to establish sound principles of reasoning and action which they teach to generations of students” [17]. They regard a national innovation system as ecology, a set of systems, premising that the way in which universities contribute to innovation varies according to the regional economy, the business sector involved and the nature of the university. The definition of the utility of universities is often too narrowly drawn from their perspective - the useful knowledge and skills generated by universities are a derivative of a much deeper capability than that of driving innovation. “It is a capability deeply embedded in the fundamental role that universities have in creating new knowledge and transmitting it to successive generations together with the knowledge which has been accumulated by predecessors and which in each generation is subjected to renewed tests of verification.” [18]. Their paper is a plea for the autonomy and freedom of universities to “do what they do best”, without oppressive mechanisms which seek to drive short term utility. It is the flexibility and adaptability of universities which enables them to stay true to their core mission in pursuing and explaining knowledge whilst being sensitive to the needs of the contemporary world. Courant, when considering the impact of disruptive technologies on universities, would concur, proposing that universities be both conservative and revolutionary: conservative in terms of mission and revolutionary in the way in which they attain their mission [19].

3. Defining research impact

Together with the stimulation strategy there is the consequent regulation strategy. Governments seek to measure the quality of their universities and the contribution they make to the nation’s prosperity. The impact of a university’s research is a significant element of a university’s contribution.

They wish to maximise the economic and social returns from any public investment in research. Within Australia the federal government is in the midst of rolling out a national research evaluation framework, the Excellence in Research for Australia, to measure research quality against international benchmarks [20]. This is in keeping with overseas initiatives such as those in the UK [21]. The measures will be used to inform funding decisions based upon performance, though the detail of how this will be done is not yet known [22].

The obsessive interest in university league tables is similarly a symbol of the international interest in measuring university quality and impact. International league tables, such as the Shanghai Jiao Tong and Times, are now a permanent and significant feature of the higher education landscape. In a much more competitive global knowledge economy, with a more mobile, and valuable, international student market, universities are competing to attract the best students, the best teachers and researchers and the best grants. Global rankings of universities are a familiar and increasingly visible part of the higher education landscape, as universities compete to promote their value, status and attractiveness.

Within Australia the Research Quality Framework (RQF) was introduced in 2005 to follow in the footsteps of other nation's research quality frameworks [23]. The RQF differed from existing international research assessment exercises in that it sought to measure 'research impact'. 'Research impact' was defined as "the beneficial application of research to achieve social, economic, environmental and/or cultural outcomes." [24]. Measures of impact included analyses of patents, cost-benefit assessments, social returns and citations [25]. In its infancy the RQF was replaced by the Excellence in Research for Australia (ERA) initiative, which no longer seeks to measure research impact in the same way, instead examining more quantifiable measures of citation impact and esteem.

Whilst the broader definition of research impact measurement has been dropped in the ERA process, the concept is still part of the national landscape in Australia. One aspect of the Australian national innovation system and research landscape are the Cooperative Research Centres (CRCs) [26]. They are funded by the Australian government to build critical mass in specific research ventures which link universities and industry. The government commissioned Deloitte to develop a framework to evaluate the performance of CRCs [27]. The framework was to help CRCs assess their outcomes through examining the impact chain from inputs, through to activity, outputs, usage and impact. They note that quantifying the final impact of research is necessarily the most uncertain of the stages [28]. Impact types may include productivity gains, industry development, environmental, health and social benefits which are not easy to quantify and which are highly contingent in nature. This broad definition of research impact plays to the role of universities in achieving the national innovation agenda in its most complete sense.

In a knowledge economy it is the generation and exploitation of knowledge that plays the predominant part in the creation of wealth. Scholarly publishing plays a key role in the effective dissemination and diffusion of knowledge and research findings [29] and conventional publishing is still the main form of research dissemination [30]. It is therefore not surprising that the more complex, difficult to measure, yet valuable, definition of research impact as outlined above is often abbreviated to a measure of publishing quality as measured through journal ranking and citation impact [31]. Butler reasons that research quality is best judged by peers. Peer reviewed prestige publication is sill the route to academic success [32]. A stellar publication record and citation impact is integral to promotion and tenure. Hence the importance of a published research paper as judged by academic peers through journal quality and citation has become an agreed quantitative measure of research quality [33] and of research impact. Missing from these measures is the evidence that the research has had a positive economic or social impact.

4. The changing scholarly communication landscape

The mission of a university's library is intertwined with that of the university – making the world's knowledge accessible to current and future scholars [34]. Libraries have traditionally seen their role as providing free access to the world's scholarship. "This freedom gave us something real. It gave us freedom to research, regardless of our wealth; the freedom to read, widely and technically, beyond our means. It was a way to ensure that all of our culture was available and readable" [35]. This role is now challenged by a scholarly information and communication landscape which has changed profoundly and irrevocably.

The scholarly information lifecycle is transforming as advances in information and communication technologies enable new ways to create, contribute to, access and use scholarly outputs of all types. The creation of a university's scholarly output, whether published works, research data, working papers, teaching materials or multimedia of a variety of kinds, is increasingly digital. Scholarly books are published in digital form with some predicting that virtually all new scholarly titles will be digital within 10 years [36]. More than 30% of Amazon's titles are now sold in digital form [37]. Scholarly information is published by individuals, institutions and large corporations and delivered via a multitude of business models which continue to evolve and change in unpredictable ways.

"The environment in which research is being conducted and disseminated is undergoing profound change, with new technologies offering new opportunities, changing research practices demanding new capabilities, and increased focus on research performance." [38]. As Borgman notes [39], every stage in the research lifecycle can be facilitated, or complicated by technology. Research practice is radically changing as large-scale, distributed, collaboration in research projects is facilitated through the capacity of digital technologies, enabling the study of complex problems across organisational and national boundaries [40]. Collaboration in the social sciences and humanities is increasing as rich data sets and previously difficult to access texts and objects are made accessible through digitisation. Just as new content is being created digitally, large collections of printed text and other objects are being made accessible globally and freely. Scholarly output now includes not only the published works but the research data, tools and techniques associated with the research. Existing research data can be re-mined and re-used, research algorithms, tools and techniques can be easily shared, large data sets can be visualised to render complex findings in useable ways.

An unknown amount of this research data will have value for the future as an important part of scholarly output. A recent Intersect study of four New South Wales universities found that more than 87% of researchers' collect or create research data, more than 50% said their data was almost all digital and a further 23% said it was more than 60% digital [41]. Almost half the respondents allowed access to the data from outside their research team. Fewer than half the respondents believed they faced data management or preservation issues and 20% weren't sure. Yet with appropriate stewardship research data has the potential to significantly increase research impact.

Australia has been well served by the Australian Partnership for Sustainable Repositories (APSR) [42], the agency that has led national thinking on the research data issue. In 2006 APSR released a report on Australian e-research sustainability. The report explores the issues surrounding research data stewardship, the incentives and disincentives for appropriate stewardship of research data. There were clusters of issues, some of which are now being actively addressed at a national level. The report suggests that from a policy perspective the research funding bodies lack guidelines for clear administrative responsibility for data stewardship, yet there is interest in maximising research outcomes from the public dollar. It was not difficult, therefore, to convince funding agencies to encourage more open access to research data. The policy framework has now been changed with the responsibility for research data access clearly resting with the university as a long lived institution. This mirrors international trends. Though whilst agencies require data management plans and deposit, enforcement if often inconsistent [43]. Policy is necessary, but not sufficient, requiring the addition of "carrots and sticks" if behaviour is to change.

The APSR report also found that there are strong disincentives for researchers to engage with long-term data management. They are funded to do the research, research groups come and go, there is no funding for stewardship, no rewards or recognition. Good research data stewardship will not, at least in the immediate future, impact on their ranking in ERA, nor in league table positioning. The universities themselves are one of the enduring features of the research landscape and hence arguably a logical home for long term commitment to data stewardship. But the report notes that whilst universities want an environment that maximises research outcomes, this is currently established by publishing and citation metrics – it is not in the university's interest to follow policy prescriptions if there are no rewards and/or penalties [44]. One of the policy problems with data curation and preservation is that the costs persist long after the project ends [45]. Researchers may generate very long-lived and substantial financial responsibilities for the institution.

Universities have invested significant sums of money in building and sustaining library collections for future generations of scholars. They have done so based on a belief that the library plays a key role in supporting their research and learning through preserving and making accessible scholarly output- though arguably this is currently under challenge. Borgman notes that whilst libraries are a logical steward for research data management, libraries are no better placed to take on an unfunded mandate [46]. Lynch also notes that 'With data storage services, campus cyberinfrastructure design and deployment begins to interconnect with fundamental campus policies and culture about the stewardship responsibilities of scholars, about contracts and grants compliance issues, and about risk management" [47].

APSR also found that there was no systemic sustainable infrastructure available to broadly support research data management. It is in this area that we have seen significant national investment since the report, as noted in the earlier section of this paper. Australia has made a significant financial commitment to development of a national research data fabric, data storage, high performance computing and networks.

5. New challenges and opportunities

Swan [48] questions whether we would invent our present system of scholarly communication in our current context and decides not. If scholarly communication is to aid the progress of science, then, arguably, some of our current mechanisms act as barriers. Swan persuasively argues the case for open access, showing that it increases citation impact, shortens the research lifecycle, advances science by enabling use of new technologies to mine and manage science and opens the way for greater collaboration across discipline and geographic boundaries. Similarly Houghton and Sheehan [49] have sought to examine the economic benefit offered through increased access to research findings, afforded by new models of scholarly communication. They explore different publication models, examining their potential for greater research impact (as measured by citation). They analyse the literature and quantify the potentially measurable impacts of enhanced access to research findings, for researchers, government and the wider community, including:

- more timely access to both accelerate and widen opportunities for more timely, collaborative research, and for adoption and commercialisation
- greater access leading to improved learning outcomes, a greater opportunity to inform professional practice, improve the capabilities of practitioners, future researchers and research users
- the potential to create more informed citizens and consumers with implications for better use of health care, social benefits and education, and potentially improved productivity.

Their modelling shows significant economic benefit from open access to publicly funded research, with, for example, a 5% increase in access and efficiency in Germany worth USD 3 billion. This work was extended through a further study commissioned by JISC [50] to examine the economic implications of alternative scholarly publishing models. This paper posits that if the aim is to have the most cost-effective scholarly publishing system, then both costs and benefits must be quantified. All costs and benefits associated with the scholarly communication lifecycle are modelled in an attempt to understand the increasingly complex scholarly publishing landscape. They demonstrate that research and research communication are major activities with substantial costs and conclude that a preliminary analysis of the potential benefits of more open access to research findings suggest that the returns to research can be substantial. Different models for scholarly publishing can make material differences to the returns realised and the costs faced. Whilst the paper refers to the UK context (which produces 10% of the world's scientific papers [51]), the importance of promoting greater use of open access on an international scale is even more relevant to Australia if it means that the 98% of the world's scholarly output produced elsewhere will be more accessible.

A range of recommendations are made to overcome the barriers and realise the benefits of more open access publishing. Among these are to ensure that research evaluation is not a barrier to moving toward more cost-effective scholarly publishing models and that incentive and reward systems are aligned. Arguably these barriers still exist.

In a highly competitive and complex environment a scholar's competitiveness is still judged by the quality of their publication and citation record. Whilst the scholarly communication and dissemination landscape if changing dramatically, it is within the context of relatively conservative value and reward system for scholars, a system which has the practice of peer review at their core [52]. A recent study by the Center for Studies in Higher Education [53] found that from a researcher's perspective one of the greatest challenges for disseminating research is choosing where to publish. Scholars are concerned with the stature and selectivity of the publication outlet but also its appropriateness for the target audience. The study suggest that the primary motivation of a scholar is to choose an outlet that will have the highest visibility with the specific audience they want to reach, even if that audience is small, preferring a prestigious commercial publisher over an open access publication without a prestigious imprimatur. Interestingly a recent article on marketing publishing is Australia questions whether, in fact, scholars are publishing for other scholars at the expense of improving professional practise [54]. The CSHE study found, perhaps unsurprisingly, that young scholars were particularly conservative in their research dissemination behaviour whereas established scholars could afford to be more innovative [55]. Scholars remain under pressure to publish in high impact journals, many of which are still subscription access only, finding older business models profitable in an environment where national research quality schemes can serve to reinforce their market position.

Within ERA the concept of research impact is judged through citation impact and esteem measures. In calculating these measures the Australian Research Council has worked with the academic community to rank journals, including Australian titles, based on their assessment of quality. These rankings will inform the way in which publication quality is judged. This has been a highly contentious, and arguably flawed, process [56]. Butler is concerned that impact, as measured through publication quality and citations, is becoming the proxy for research quality in total, rather than only one aspect. She also expresses concern that a system which impacts on prestige and/or funding (which ERA will do on both counts) will affect the behaviour of researchers and administrators. There is a risk of goal displacement, where increasing the measure becomes the imperative. Within the context of ERA, where the scholarly community has ranked journals for the purposes of measuring research impact, it is already clear that there is now an unspoken imperative to seek to publish almost exclusively in journals ranked A and A* in order to drive ERA quality outcomes.

Arguably we are at risk of reducing our ability to achieve the more aspirational notion of research impact, of contributing to national innovation, as universities, faculties and/or individual researchers seek to maximise ERA outcomes at the expense of getting their research into the best place to maximise its real social and economic impact.

6. Reinventing the role of information services

The changing scholarly communication landscape increases the potential to increase research impact, and also increases the complexity. The once apparently linear process of research, communication and application of the results has become more much complex. Advances in information and communication technologies are disrupting the traditional models of publishing [57].

At all stages of the research lifecycle there are opportunities for information services providers to enhance their university's research impact.

6.1 Information access

Studies have shown that increasingly researchers use Google for everything, that they are confident they can manage their information seeking, though many are less certain that they are managing their research data well [58]. Haglund [59], in a study of young university researchers at three universities in Sweden, found that Google was the first choice of information seeking, search methodologies were haphazard at best, yet researchers feel they are competent information searchers. Convenience was important – if an item wasn't "one click away" they didn't bother seeking it, and they were receptive to new technologies such as PDAs.

6.2 Becoming part of the research endeavour

Personal networks were important to researchers and collaboration was widespread yet they appeared to have no working relationship with the library. They rarely went to the library and did not see how the Library could assist them with instruction or IT support. Haglund proposes that the paradigm shift, wrought through the Internet, digital publishing and reinvention of libraries as the "living room" for undergraduates, has served to make libraries and librarians more removed from the world of the researchers.

6.3 Research data services, generic and tailored

In assessing the future of the scholarly communication landscape the recent Center for Studies in Higher Education study [60] found that support structures and organisations available for the preservation and storage of a researcher's own data are uneven at best, with most institutions approaching the issue in a piecemeal manner. They found five key areas that need immediate attention:

- More nuanced tenure and promotion practices that did not rely exclusively on publication and 'easily gamed' citation metrics
- A re-examination of peer review – meaning, timing, mechanisms, locus
- Competitive high quality affordable publishing platforms
- New models of publication with institutional assistance to manage copyright
- Support for managing and preserving new research methods and products- GIS, visualisation, complex distributed databases etc.

The study found that the scope of support needs by the different disciplines was starkly different, with scientists wanting bigger 'pipes', new ways to store, manage, process and visualise large data sets and mechanisms to support 'grand challenge' research. Social scientists and humanists needs were more modest though they included interest in integrated complex data mining, computational analysis and visualisation. Arguably the differences are of scope rather than of substance. All disciplines identified the problem of data storage and preservation (the authors noted that it appears that the EU had prioritised this ahead of the US) [61].

The need for specialist support, particularly IT support, was prevalent though the preference was for technology-savvy scholars who work in collaboration rather than a model of "academic computing services" who are unaware of the scholarly questions and methodologies that drive a discipline. In many cases the library was seen as the locus of support for archiving, curation and dissemination of scholarly output. They conclude by noting that "although robust infrastructure are needed locally and beyond, the sheer diversity of scholars' needs across the disciplines and the rapid evolution of the technologies themselves means that one-size-fits-all solutions will almost always fall short. As faculty continue to innovate and pursue new avenues in their research, both the technical and human infrastructure will have to evolve with the ever-shifting needs of scholars. This infrastructure will, by necessity, be built within the context of disciplinary conventions, reward systems, and the practice of peer review, all of which undergird the growth and evolution of superlative academic endeavours." [62]

6.4 Clear leadership in research information services, internally and externally, with strong collaborative links

Within the Australian context the Intersect study [63] found that the vast majority of researchers had not heard of any of the major national bodies involved in developing and providing research information infrastructure services. When asked what support they most needed, scholars identified data management, expertise in data analysis, collaboration platforms, data management and storage, access to research software and the need for more IT personal.

APSR found that "the immediate critical issue for the stewardship of research data in Australia is the lack of administrative responsibility for the task" [64]. The report noted that "There are boundaries between research groups, data providers, repositories and data centres. These boundaries lead to duplication or capability gaps. It is important to identify responsibilities and opportunities across these groups where possible. Data management requires greater cooperation between the players" [65]. No administrative group has responsibility for research data sustainability – to create and manage policies, understand cost benefit, accept funding and harvest the benefits.

6.5 Publishing and curation

The Center for Studies in Higher Education study found that from a researcher's perspective one of the greatest challenges for disseminating research is choosing where to publish. One response to this challenge has been that of the University of New South Wales. It introduced RIMS, the research impact measurement service, in 2005 to realign its services to support the university's goals [66]. Recognising the increasingly competitive nature of the research environment and a renewed emphasis by the University on research outcomes the Library provided a new bibliometric service providing comparative publication and citation data to schools and faculties. Knowledge gained through this process informed collection development, training opportunities for the academy on higher-impact publishing.

A 2010 study [67] showed that scholars across a broad range of disciplines had a growing interest in electronic publication and that scholars embraced the potential of linking final publications directly to data sets and/or primary sources material. Though most of those interviewed believed they didn't have access to easy-to-use tools or to the expertise required. Publishing is seen as an emerging role for libraries as it becomes easier to implement e-press services. Hahn [68] found that in most cases libraries were assisting scholars to move existing journals into the digital world or into open access publishing; in some cases they were publishing new titles. The overlap of expertise and demands of publishing with the knowledge and skills required by libraries made it a natural progression.

It is against this backdrop that scholarly information services providers within the university context: libraries, information and communication technology units, must position themselves as valued partners in the scholarly and research endeavours of their universities. Lynch [69] questions how the cyber-infrastructure challenge differs for universities as compared with the national challenge. He believes there is a strong obligation and mandate for base level of universal service across all campuses: all researchers need to be able to apply IT in their research, to access and build on cyber-infrastructure services including data management, data curation, to get help in learning how to use the services, particularly those without specialist IT support. He notes that the campus perspective is concerned with the 'average' rather than the 'extreme' scholar. "One of the key challenges - politically, financially, and technically - is defining the demarcation between free universal service and the more specialized package of support services offered to extreme users, a package that may be predicated on such users' ability to obtain funds or other resource allocations" [70.] His recommendation – that campuses create a support organisation that can reach out to scholars early in the data lifecycle to assist with data management and curation/preservation strategies, involving IT professionals, librarians and archivists and maintaining a close relationship with the research and grants office and that perhaps the Library take responsibility for the long term curation of the data at an appropriate point in the lifecycle.

Borgman [71] suggest that data may become the new 'special collections' for libraries. Noting that strategies for data curation will require involvement from academics, the campus research office, the library and instructional and information technology services.

7. A case study

Griffith University is a university of some 38,000 students from 124 countries studying at undergraduate through to doctoral level in one of four broad academic groups: arts, education and law; business; science, engineering, environment and technology; and health. Griffith is a large multi-campus institution spanning Australia's fastest growing corridor from Brisbane to the Gold Coast in Queensland. Griffith's strategic research investment strategy positions it to be a world leader in the fields of Asian politics, trade and development; climate change adaptation; criminology; drug discovery and infectious disease; health; sustainable tourism; water science; music and the creative arts.

Griffith is regarded as one of Australia's most innovative tertiary institutions and one of the most influential universities in the Asia-Pacific region. This innovation is carried through into the provision of information services, with e-learning, e-research, library, information and communication technology services, systems and infrastructure offered through a single integrated division, Information Services. This provides a distinct advantage to the University in an increasingly complex scholarly information and communication environment.

In response to the University's strategic intent to build its research impact, informed by the rapid changes to the scholarly information landscape and the increased competitive nature of research measurement, Information Services created a unique service portfolio, Scholarly Information and Research (SIR), to provide an integrated end-to-end service, offering support to researchers at all stages of the research cycle. Information services already had established relationships with the academic community, with academic librarians working closely with disciplines where the library is their "research laboratory" and research computing services well connected to specific researchers and research groups. Research computing services initially focused on the provision of high performance computing services and specialist software development. More recently much of their work had involved not only development of research portals and research analysis tools but assistance with research data management. This particular service has grown through fee for service work, often with work undertaken under service level agreement, enabling us to recruit discipline specific specialists. We also had a thriving digital repositories team which had built a strong working relationship with the office of research, working under service level agreement to collect the data for all university publications for input into the research quality assessment and funding process. Our academic librarians, whilst providing traditional library research support services, felt they could be doing more to support the University's research mission. Across the division there was also a sense that no single Director provided leadership in support the University's research endeavours within Information Services. The creation SIR in 2009 brought together our academic librarians, digital repositories, acquisitions, cataloguing and metadata services and research computing under a single leader, providing the catalyst for a renewed focus on research. Our aim was to focus on driving the University's core research mission through service innovation and collaboration.

7.1 Information access

We are currently seeking more creative ways to expand access to scholarly content by adopting different purchasing models, fine tuning our selection processes to acquire relevant content and by moving to an e-preferred format. A new library system will go live mid-year, increasing discoverability of our collections. Through our involvement in relevant state and national bodies we will continue to be strong advocates for improved access to content of relevance to our scholars.

7.2 Becoming part of the research endeavour

In 2005 I noted that "we must bring our know-how forward and actively engage in strengthening our partnerships with each other [library, information and IT professionals] and with the researchers within our own institutions if we are to continue to be a relevant and important part of the research endeavours of our institutions." [72]. At Griffith we have created contact librarian roles as part of the new portfolio. Their role is to build and maintain relationships with the academic community, referring them to specialist librarians and IT professionals as required. They are required to develop a clear understanding of the academics' requirements, ensuring we deliver services to meet academic needs and expectations and that we continue to evolve services over time to meet changing requirements.

All universities were awarded funding (scaled according to publication record) to contribute to the Australia Research Data Commons, an initiative sponsored by the Australia National Data Service. The funding is to be used to describe research data collections produced by, or relevant to, Australian researchers, with the view of making research data more widely accessible. We used this opportunity to strengthen and build new relationships with the academy through the contact librarians. They have been progressively visiting every active researcher with a current national research grant, seeking their assistance in identifying and describing any research data associated with Griffith research projects. Whilst ostensibly their visit is to elicit the data required to meet the criteria set by the Australia Research Data Commons project, they are using the opportunity to explore a broader range of questions to better position our services to meet the researcher's needs. Their questioning is free flowing, as the librarian seeks to understand the researcher's environment, their research practices, how they currently use our services and to suggest some services we could provide to gain their level of interest in these. This process will be complete in the coming months, at which time the remainder of the academy will be interviewed. The results will be invaluable in shaping our services to meet University requirements.

The contact librarians will remain an important part of our new strategy as we seek to build stronger relationships with the academy. Whilst academic librarians have traditionally been invited to academic boards, this role is now strengthened as they are able to represent the broad base of services we provide to support the University's research endeavour.

7.3 Research data services, generic and tailored

We already have a good working model for tailoring services to specific researchers or research group. The challenge now is to extend this service, building a baseline of service for all researchers whilst still meet specific research groups or researcher needs. We are seeking to learn from our understanding of particular needs to build baseline university-wide services and infrastructure. Planned increases in federal research infrastructure funding to universities over the coming years provide an opportunity to raise policy and funding questions at a University level. A paper on the development of a University research data management service will be an early candidate for discussion.

From the work of the contact librarians we will know the types of research data our academics produce, the kinds of storage practices used for maintaining research data, how the research data is managed, what access permissions are in place or are required any legal requirements in respect to the data. This information will be used to develop a repository of metadata about the University's research data as part of the national data commons.

We are also working in collaboration with a partner university on another federally funded project to build tools to harvest metadata from commonly used institutional repositories to populate the national data commons.

Planning is underway for the development of a university research data management service. We plan to provide a baseline of service for all academics - a service which leverages national data storage services whilst providing complimentary local services - from policy, management and technical advice through to provision of infrastructure.

7.4 Clear leadership in research information services, internally and externally, with strong collaboration

Under the leadership of their Director, Scholarly Information and Research, information services staff: librarians, business analysts, information architects, programmers, advanced computing specialists; are developing into contemporary information workers, strengthening their capabilities in the areas of content, technology and the disciplines to build support services to allow the researchers to thrive in this demanding, competitive and rapidly changing environment. The gaps and overlaps that might occur with distributed units can be managed internally- a full information service offering can be provided akin to that proposed by Lynch [73]. The University has welcomed the clarity of leadership around research from an Information Services perspective. Building on the existing strong relationship with the Office for Research, and building strong relationships with other University research leaders, is much simpler. Library and IT domains can be represented by a single role – it leverages relationships which each professional group already had, drawing on different strengths and different expertise. The division now has a seat on the University’s main research committee – something that can be more difficult for a library or IT unit alone. Many of the potential gaps and overlaps in supporting research are internalised within a single organisational unit, allowing them to be managed, whilst also making it easier to collaborate at a university level as fewer units must work together.

Another significant benefit is the ability to better manage the complexity, and leverage potential benefits, of the national and regional research information environment to get the best outcome for the University. Having a single division as the relationship manager on the University’s behalf makes it easier to build to develop stronger and mutually beneficial relationships with the state and federal bodies. It removes some of the complexity for the external agency when dealing with the University and the complexity for our academics who no longer need to navigate through a complex environment.

7.5 Publishing and curation

The academic community is increasingly time poor, with heavy teaching loads, reduced administrative support and increasing pressure to generate high quality research. Research success is increasingly ranked by complex measures created from within this rapidly changing scholarly information landscape, evolving into a new discipline of research management and measurement. The new environment rewards researchers who profile themselves and their work most effectively. We are building an integrated service offering to facilitate effective research information management across the institution with the specific goal of building the University's research impact, balancing short term utility requirements with the long term requirement to preserve the work of our scholars for future generations.

We are well positioned to assist researchers with their publishing decisions, providing journal trend data and potentially high citing alternatives to traditional publishing. It is increasingly necessary for researchers to consider a large range of factors when disseminating their research outputs to ensure that their work gains the highest possible impact. To assist them with this we are providing seminars, workshops and/or presentations to support researchers to manage their research for maximum impact. This can include information on changing journal trends, publishing choices, impact factors, research management, discoverability, research data management, profile management and any legislative requirements for reporting research outputs. Bibliometric analysis will be used to identify researcher performance and to inform researchers of their personal, school, group or institutional publishing impact.

To complement our strong institutional repository which enables all academics to deposit an open access copy of their work to increase accessibility and discoverability, an ePress service has been established. This will further extend the reach and impact of the University's research. The ePress provides a range of tools to manage author submissions through to managing peer-review and publication. It supports audio, video and image capabilities as well as text, enabling opportunities for deeper engagement with journal content and the potential to link research data to published output. Journals published by the Griffith ePress are harvested by major search engines, indices and citation services which will increase discovery and dissemination of Griffith research.

Assistance can be given to ensure Griffith researchers grow their profile to attract partners of international standing both domestic and international. We are working in close collaboration with the University's research office to replace our existing research management system with one which integrates with our digital repositories and other systems. This will provide an opportunity to more effectively profile Griffith's researchers and their scholarly output.

8. Conclusion

Universities are integral to a nation's innovation agenda. The impact of their research has the potential to significantly improve a nation's economic and social outcomes. With this comes increased national interest in stimulating and regulating universities to drive potentially utilitarian aims, and an interest in measuring their quality. Universities must stay true to their core mission of knowledge creation, dissemination and preservation not just for current, but for future generations. They cannot afford to adopt tactical responses to government imperatives or international league tables. As the scholarly information lifecycle transforms, the ability for a university to enhance its research impact is greater than ever, but it is also a much more complex environment. This complexity offers new opportunities for a university's information service providers to further the university's, and the nation's, aspirations – both short and long term. Information service providers are ideally positioned to navigate the complexity of the scholarly information landscape to achieve university imperatives within a national context, through leadership and expertise and collaboration within and across organisational boundaries; to achieve short term imperatives whilst staying true to the long term mission of universities in knowledge creation, dissemination and preservation for future generations of scholars and practitioners.

Acknowledgements

I would like to acknowledge the leadership, innovation and professionalism shown by the incoming Director, Information Services (Scholarly Information and Research) Ms Lyn Bosanquet.

Notes and References

[1] See for example OFFICE OF SPECIAL PROJECTS, NATIONAL RESEARCH COUNCIL. Issues for Science and Engineering Researchers in the Digital Age. Washington: National Academies Press, 2001 and MARKAUSKAITE, J; et al. Co-developing eResearch infrastructure: technology enhanced research practices, attitudes and requirements, Full technical report, Sydney: The University of Sydney & Intersect, 2009 which found that less than 23% of

researchers across 4 universities said nearly all their research is individual and 53% with collaborated outside Australia.

[2] BOULTON, G; LUCAS, L. What are universities for? League of European Research Universities, September 2008. Available at *www.leru.org/file.php?type=download&id=1323* (March 2010)

[3] PERKMANN, M; WALSH, K. University-industry relationships and open innovation: towards a research agenda. Loughborough: Wolfson School of Mechanical and Manufacturing Engineering Loughborough University, 2007. Available at http://ssrn.com/abstract=154532 (March 2010)

[4] See BOULTON and LUCAS and PERKMANN and WALSH

[5] See for example PERKMANN and WALSH

[6] LYNCH, C. The institutional challenges of cyberinfrastructure and e-research. EDUCAUSE Review Nov/Dec 2008, pp.74-88

[7] CUTLER, T. Venturous Australia: building strength through innovation. 2008. http://www.innovation.gov.au/innovationreview/Documents/NIS_review_Web3.pdf (March 2010)

[8] Powering ideas: an innovation agenda for the 21st century http://www.innovation.gov.au/innovationreview/Documents/PoweringIdeas_fullreport.pdf (March 2010)

[9] DELOITTE – INSIGHT ECONOMICS. Impact monitoring and evaluation framework background and assessment approaches. Cooperative Research Centres Association Inc. June 2007. p.4

[10] FINKEL, A. Innovation rests on simulating challenges. Focus no. 148 February 2008, pp.11-12 http://www.atse.org.au (March 2010)

[11] See CUTLER p.67

[12] See PERKMANN

[13] See PERKMANN p.4

[14] See CUTLER p.41

[15] See Powering Ideas and https://www.pfc.org.au/bin/view/Main/SuperScience

[16] See BOULTON and LUCAS

[17] See BOULTON and LUCAS p.4

[18] See BOULTON and LUCAS p.8

[19] COURANT, P. Scholarship: the wave of the future in a digital age. Chapter in Katz, R (Ed.), The Tower and Cloud: Higher Education and Information Technology (in press). Boulder, Colorado: EDUCAUSE, 2007.

[20] http://www.arc.gov.au/era/default.htm

[21] BUTLER, L. Assessing university research: a plea for a balanced approach. Science and Public Policy 34(8) Oct 2007 pp. 565-574

[22] See for example the speech made by the Minister at the Australian Technology Network of Universities Conference in Feb 2010 http://minister.innovation.gov.au/Carr/Pages/AustralianTechnologyNetworkofUniversities.aspx (March 2010)

[23] See BUTLER

[24] DURYEA, M; HOCHMAN, M; PARFITT, A. Measuring the impact of research. Research Global Feb 2007 p. 8

[25] JOHNSON, R. 1995 Research impact quantification. Scientometrics 34(3): p.415

[26] https://www.crc.gov.au/Information/default.aspx

[27] See DELOITTE-INSIGHT ECONOMICS

[28] See DELOITTE-INSIGHT ECONOMICS p.10

[29] HOUGHTON, J. et al. Economic implications of alternative scholarly publishing models: exploring costs and benefits, a report to the Joint Information Systems Committee. Jan 2009. http://www.jisc.ac.uk/media/documents/publications/rpteconomicoapublishing.pdf (March 2010)p.IX

[30] See MARKAUSKAITE

[31] See for example BUTLER, HOUGHTON et al

[32] HARLEY, D. et al. Assessing the future landscape of scholarly communication: an exploration of faculty values and needs in seven disciplines. UC Berkeley: Center for Studies in Higher Education, 2010 available at http://escholarship.org/uc/sche_fsc (March 2010)

[33] See BOULTON p.569

[34] See O'BRIEN, L. et al. Scholarly information in a digital age; choices for the University of Melbourne, a consultation paper that invites involvement and response. Melbourne: University of Melbourne, Feb 2008. Available at http://www.jisc.ac.uk/media/documents/publications/rpteconomicoapublishing.pdf

[35] LESSIG, L. For the love of culture. The New Republic Jan 26 2010. http://www.tnr.com/print/article/the-love-of-culture (March 2010), p.9

[36] BRINDLEY, L. Quoted in CHRISTENSEN, L. British Library predicts 'switch to digital by 2020.' Media release. London: British Library, 2005. Available from www.bl.uk/news/2005/pressrelease20050629.html (Feb 2008)

[37] http://www.digitaltrends.com/computing/amazon-sees-71-percent-profit-boost-from-holidays/

[38] HOUGHTON, J; SHEEHAN, P. The economic impact of enhanced access to research findings, CSES Working Paper no. 23, Melbourne: Centre for Strategic Studies, 2006. Available at http://www.cfses.com/documents/wp23.pdf p.1

[39] BORGMAN, C. Scholarship in a digital age: information, infrastructure, and the Internet, London: MIT Press, 2007

[40] O'BRIEN, L. E-research: an imperative for strengthening institutional partnerships, EDUCAUSE Review Nov/Dec 2005, pp.65-76

[41] See MARKAUSKAITE

[42] See http://www.apsr.edu.au

[43] BORGMAN, C. Supporting the 'scholarship' in E-scholarship, EDUCAUSE Review Nov/Dec 2008, pp.32-33

[44] BUCHHORN, M; MCNAMARA, P. Sustainability issues for Australian research data, http://www.apsr.edu.au (March 2010), p.45

[45] See LYNCH

[46] See BORGMAN (2007), p.247

[47] See LYNCH, p.82

[48] SWAN, A. Open access and the progress of science. American Scientist vol 95 May-June 2007, pp.198-200

[49] See HOUGHTON and SHEEHAN

[50] See HOUGHTON et al

[51] See HOUGHTON et al, p.233

[52] See HARLEY

[53] See HARLEY

[54] See for example http://www.campusreview.com.au/pages/section/article.php?s=Comment&idArticle=15060

[55] See HARLEY, p.12

[56] See for example BUCKLE, S. Philosophy betrays its first principles, The Australian, Wed March 31 2010, Higher Education, p.29

[57] See HOUGHTON et al

[58] See for example HAGLUND, L; OLSSON, P. The impact on university libraries of changes in information behaviour among academic researchers: a multiple case study, The Journal of Academic Librarianship vol. 34 no.1 pp. 52-59 and O'BRIEN et al

[59] See HAGLUND

[60] See HARLEY

[61] See HARLEY, pp..24-25

[62] See HARLEY, p.26

[63] See MARKAUSKAITE

[64] See BUCHHORN, p.1

[65] See BUCHHORN, p.46

[66] DRUMMOND, R; WARTHO, R. RIMS: the research impact measurement service at RIMS the University of New South Wales, Australian Academic & Research Libraries vol.40 no.2 June2009, pp.76-87

[67] See HARLEY

[68] HAHN, K. Publishing services: an emerging role for libraries, EDUCAUSE Review Nov/Dec 2008, pp.16-17

[69] See LYNCH

[70] See LYNCH, p.78

[71] See BORGMAN 2008

[72] See O'BRIEN. p.68

[73] See LYNCH

[74] BOSANQUET, L. Building relevance in the content revolution, Library Management vol 31, issue 3 2010, pp.133-144

[75] BOSANQUET, L. Beyond digital repositories- a role for University libraries, in press 2010.

The PEG-BOARD Project: a case study for BRIDGE

Gregory Tourte[1]; Emma Tonkin[2]; Paul Valdes[1]

[1] School of Geographical Sciences.
University of Bristol,
Bristol, United Kingdom
{g.j.l.tourte,p.j.valdes}@bristol.ac.uk;
[2] UKOLN,
University of Bath,
Bath, United Kingdom
e.tonkin@ukoln.ac.uk

Abstract

With increasing public interest in the area of historical climate change and in models of climate change in general, comes a corresponding increase in the importance of maintaining open, accessible and usable research data repositories. In this paper, we introduce an e-Science data repository containing extensive research data from palæoclimatology. Initially designed to support internal collaboration and organise data, the sharing of research outputs became an increasingly significant role for the service over several years of practical use. We report on a data preservation and interoperability assessment currently under way. Finally, we discuss the ongoing significance of open research data and capacity for analysis in the area of climate research, with palæoclimatology as a case study.

Keywords: palæoclimate modelling; data management; data curation.

1. Introduction

The BRIDGE research group, or Bristol Research Initiative for the Dynamic Global Environment, focuses on the emerging area of ‘Earth System Science’ exploring the complex interactions between the Earth’s components: the oceans; atmosphere; ice sheets; biosphere; and the influence of human activity on global change. This approach requires the input of multidisciplinary teams drawn from across Bristol University Glaciology, Hydrology, Biogeochemical Cycles, Chemistry, Earth Sciences, Mathematics, Engineering, Biological Sciences, Archæology, Personal Finance Research) and beyond (Hadley Centre, British Antarctic Survey, UK Met Office, DEFRA, Environment Agency, Centre for Global Atmospheric Modelling, Oil Industry).

Climate, 'the synthesis of atmospheric conditions characteristic of a particular place in the long term', is 'expressed by means of averages of the various elements of weather'; climatology, then, is the scientific study of climate [1]. The main research effort of the group is to improve the understanding of the causes of climate change, by testing the computer climate models used to predict future climate change. Major themes include:

- quantifying environmental and climate change in the distant past through the combined use of data and models;
- evaluating climate models with accurate proxy climate records, especially during periods of rapid climate change;
- improving climate models by incorporating additional components of the Earth System and detailed analysis of these processes for past, present and future change;
- assessing the impact of future climate change on spatial and temporal scales relevant to society and including timescales from decadal to millennial.

Many of these activities require—and produce—many terabytes of data. Making this data widely available is therefore a complicated and non-trivial process.

Researchers worldwide in both the sciences and humanities reuse BRIDGE data in their work. The project developed and applies de-facto preservation and data compression policies. Since the types of information required by users from areas as diverse as evolutionary biology, archæology and earth science very greatly, the project also developed an in-house interface designed to support tailored information extraction from climate model information.

Despite the complications associated with open access to large scientific datasets, openness in procedure and output is a priority for BRIDGE, and has been for many years. The importance of open data in climatology research in general has recently been highlighted, due to the high profile of the research area in the media and politics.

1.1 Background

Sweet [2] divides climate modelling into theory, empirical work, and modelling, and notes that modelling attracts the most attention since this area most directly assesses impact and produces predictions. It is expensive; simulations can take up to three months to run on high-performance computers ('supercomputer' clusters) and can equate to up to a hundred thousand pounds worth of computer time, excluding the cost of storage. The existing archive of resulting data sets consists of over 2,000 simulations and represents several million pounds worth of CPU time. The cost of CPU time has reduced; however, the scale of models has increased as a result. In terms of data requirements, a single model simulation can produce up to 2 TiB of raw model output data. A smaller subset of 2 to 50 GiB per simulation is retained.

Adopting Sweet's approach, we view the area as containing three areas of endeavour: empirical work, including data collection and preservation, theory, and modelling. In practice, these areas are difficult to divide; Edwards [3] qualifies the model/data relationship in climate science as 'exceptionally complex'. The boundaries between a global climate model (GCM) and data are 'fuzzy', and the interaction between model and theory is supple and ongoing. A model inspired by theory may apply initial conditions taken from measured data points. Data generated via a GCM may be compared with observed data points to evaluate the *validity* of the model. This demonstrates that model results agree with observations and that no detectable flaws exist, rather than that the GCM is essentially correct, but is nonetheless a significant step in establishing realism.

e-Science has a strong tradition in climate science. In data collection, for example, Benford et al. [4] describe the use of a Grid-based networked device to enable remote monitoring of Antarctic freshwater lakes and explore the potential for distributed collaborative research based on the resulting dataset. Benford et al. [4] highlight Anderson and Lee's [5] four phases of software fault tolerance as key to ensuring confidence in the resulting data: error detection, damage confinement and assessment, error recovery and fault treatment. Data, then, is only part of the story; provenance and context are required to ensure confidence.

Climate modelling software, too, is increasingly designed in order to make use of e-Science concepts and facilities. The SciDAC-supported Earth System Grid Center for Enabling Technologies (ESG-CET), for example, enabled all of the simulation data from the IPCC's AR4 to be made available to climate scientists worldwide [6]. The GENIE—Grid ENabled Integrated Earth modelling system—also applies a Grid-enabled architecture, in this case designed with the intent to 'build simplified and faster-running models of the Earth's climate system, and make them easier to use and more widely available to other people who want & need to use them' [7] . GENIE is designed to facilitate cyclic improvement of models through comparison with available datasets; to improve traceable integration between various model types, and to integrate multiple representations of the natural Earth system. GENIE enables large ensemble studies on the Grid, supports complex workflows and provides Grid-based data handling and post-processing facilities [8]. In each of these applications, as is generally true with Grid-based approaches [9], rich and descriptive metadata, including extensive information about data provenance, is required to enable effective use of available data.

The political significance of climate modelling as a research area is currently such that openness is absolutely key. With publicly funded research, the 'citizen scientist' should be considered as a stakeholder, and ultimately this is dependent on working with the user community [10].

1.2 The case for open access to data

The importance of open data in climatology research in general has been highlighted in recent years, due to the high profile of the research area in the media and politics. Climate modelling, particularly in the area of climate prediction, is subject to a high level of scrutiny.

Consider for example a recent news article [11], discussing the open review of a recent report, the 4th Assessment Report or AR4, published by the Intergovernmental Panel on Climate Change (IPCC). The process described is a review conducted by 'climate "sceptics", […] busy searching the rest of the panel's report for more mistakes'. One statement queried is described as 'basically correct but poorly written, and bizarrely referenced'; the process of establishing accuracy has highlighted issues regarding appropriate referencing and clarification of the distinction between 'grey', or non-peer-reviewed, literature, and peer-reviewed sources. Harrabin suggests 'a need for much greater transparency'. A further famous example are the international repercussions (both political and scientific) surrounding the recent 'leak' of emails from the Climatic Research Unit at the University of East Anglia, dubbed 'Climategate' by many.

Access to data and modelling resources is variable. For example, the UM (Unified Model), the popular suite of atmospheric and oceanic numerical modelling software developed and used at the UK's Met. Office has limited availability, being primarily available to UK academic researchers. Availability of the GENIE software is currently limited, as the software remains work-in-progress. A great deal of data is available, from sensor data released by the

British Antarctic Survey, the Australian Antarctic Division and others to the OpenGeoscience service offered for non-commercial use by the British Geological Survey; a great deal of open-access data may be discovered via the NERC Data Services initiative (http://ndg.nerc.ac.uk/) that gathers together the NERC data centres. Data centres typically hold collections of empirical data (e.g. observations and measurements).

Open procedure and open access are priorities for BRIDGE, and a software platform has been developed over many years to support this aim, allowing modellers to publish datasets along with relevant experimental metadata. Although the present iteration of the software predates recent best practice in the area, the service has been widely used for those requiring secondary data, to the mutual benefit of BRIDGE and external users of the data.

1.3 The PEG-BOARD project: Palæoclimate & Environment data Generation – Building Open Access to Research Data

In response to the community's need for openly accessible research data, we need to make sure that the data generated as part of our research remains accessible and preserved for a certain amount of time after its creation and original use.

However, preservation of digital information is a very complex subject. Su-Shing Chen in the Paradox of Digital Information [12] explains why it is difficult to come up with a simple definition of what 'to preserve digital information' means. He says that 'on the one hand we want to maintain digital information intact as it was created' (one facet of preservation) 'on the other, we want to access this information dynamically and with the most advanced tools' (preserving access to the data).

This is extremely relevant to our data as the models used to generate it as well as the hardware architecture on which the models are run evolve and change over time. A particular experiment run five years ago may not run on current hardware or if it runs, may not produce the same results. We have seen recently that the implications of publications and data may be seen and questioned decades later. However, from a more pragmatic point of view, the benefit of keeping old data can easily be questioned. The cost of storing large dataset is very high, despite the raw cost of storage going down dramatically with time, archival enterprise grade storage is still very expensive and the long-term maintenance cost of keeping a storage system working and up-to-date may well be higher than the cost of re-running the experiment, especially when computers speed is also increasing with time. Another point to consider is the fact that the science included in the models evolves as well.

With computers becoming more and more powerful, the complexity of the models have increased, adding CO_2, NO_2 and H_2O exchanges to atmospheric models as well as vegetation over the last 15 years [13]. This means that old experiments will be inaccurate compared to our current understanding of the earth system and therefore may as well be re-simulated to get a result more in line with the current science.

With that in mind, the PEG-BOARD project has several aims, targeting every aspect of our data and our user base :

- assist the work of modellers by facilitating data processing, manipulation and analysis by the modellers and scientists who generate data as part of their research;
- facilitate data reuse by modellers and by any consumers of the data by providing methods to search and browse through the data;
- discover and characterise modes and means of data reuse, and identify relevant user groups;
- identify current patterns of metadata use, the standards used and the extent to which they comply with relevant data types;
- describe current data retention policies and relevant standards;
- provide clear guidelines to research groups and researchers to help manage their data;
- ensuring proper data retention and curation policies based on both the research and the data life cycle;
- disseminate documents and software to wider community to provide better understanding and better accountability for the research communities to the wider public and stakeholders.

We are now in the requirements-analysis stage of a new project, PEG-BOARD, designed to support the curation of historical climate data within BRIDGE’s large global consortium of palæoclimate researchers, and to ensure ongoing availability of this data for reuse within research, teaching and the media. This work is carried out in the context of the UK e-Science infrastructure [14]. The project focuses on providing the community with a better understanding of the data and the limits of its validity, and defining a clear policy structure for palæoclimate data. An improved data management infrastructure is expected to improve availability and accessibility of data, as well as providing a stabler structure for collaborative reuse. Open availability of well-structured and documented research data is key, enabling open and easy creation of malleable prototypes, adaptable to relevant research or interest communities.

2. Methodology

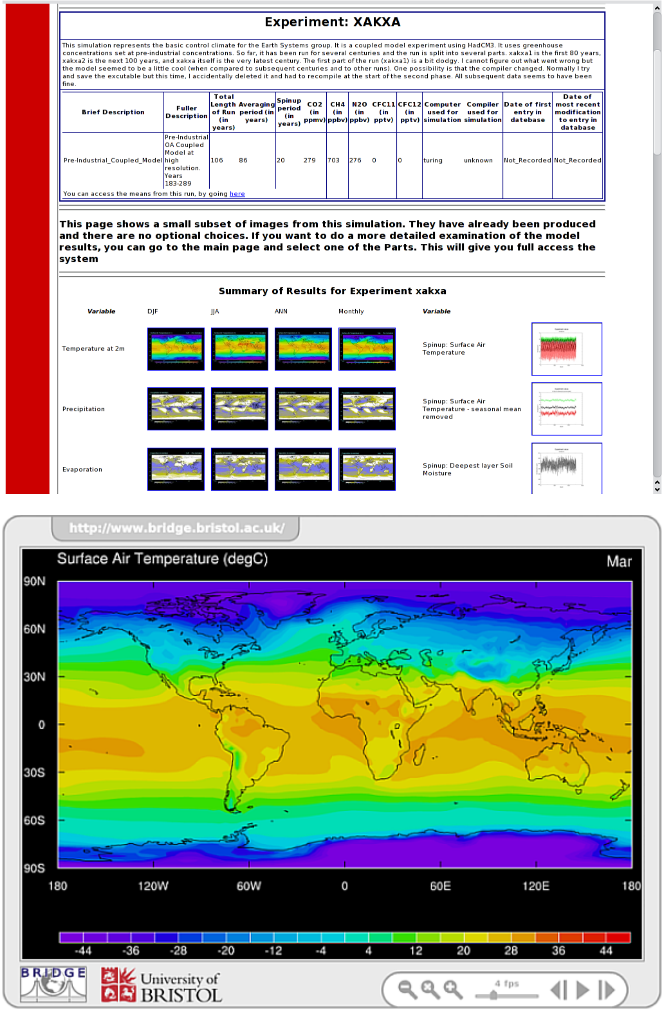
Due to the strong user-analysis component of these aims, we chose to begin with a phase of user analysis of the present system. Various mechanisms exist for exploring user requirements; indeed, the field of requirements engineering has over time attracted a large and very active research community. Requirements engineering is described by Laplante [15] as 'a subdiscipline of systems engineering and software engineering that is concerned with determining the goals, functions, and constraints of hardware and software systems'. Nuseibeh & Easterbrook [16] describe requirements engineering as follows:

'The primary measure of success of a software system is the degree to which it meets the purpose for which it was intended. Broadly speaking, software systems requirements engineering (RE) is the process of discovering that purpose, by identifying stakeholders and their needs, and documenting these in a form that is amenable to analysis, communication, and subsequent implementation.'

RE is not a single operation but a sequence of operations. Stakeholder analysis is a necessary precursor, a part of the process that in our case has been explored for previous developments in the BRIDGE area, but which due to the nature of the problem area is necessarily an ever-shifting target. Nuseibeh & Easterbrook describe the core areas of RE as: *eliciting* requirements, *modelling and analysing* requirements, *communicating* requirements, *agreeing* requirements and *evolving* requirements. The mechanisms used in the PEG-BOARD project thus far can be fitted into this overall model of the process of requirements engineering, although some aspects were explored prior to the beginning of the project (stakeholder identification in particular).

The processes of eliciting, modelling and communicating requirements are all touched on in this paper. Requirements are elicited initially by the exploration of existing systems in use as part of the task decomposition process – via interface surveys (see Section 2.2), and then via the use of structured interviews with selected users. This is completed in two areas; with users internal to the BRIDGE project, and with a case-study of an external consumption of BRIDGE data. 'Data Sharing Verbs' are used as part of the modelling and communication of requirements.

We chose to begin with a series of interviews, exploring a number of 'characteristic' individual users' perceptions of their interactions with the BRIDGE services. The results of this process form part of the background material for the Results section of this paper (Section 3).



Experiment: XAKXA

This simulation represents the basic control climate for the Earth Systems group. It is a coupled model experiment using HadCM3. It uses greenhouse concentrations set at pre-industrial concentrations. So far, it has been run for several centuries and the run is split into several parts. xakxa1 is the first 80 years, xakxa2 is the next 100 years, and xakxa itself is the very latest century. The first part of the run (xakxa1) is a bit dodgy. I cannot figure out what went wrong but the model seemed to be a little cool (when compared to subsequent centuries and to other runs). One possibility is that the compiler changed. Normally I try and save the excutable but this time, I accidentally deleted it and had to recompile at the start of the second phase. All subsequent data seems to have been fine.

Brief Description	Fuller Description	Total Length of Run (in years)	Averaging period (in years)	Spinup period (in years)	CO2 (in ppmv)	CH4 (in ppbv)	N2O (in ppbv)	CFC11 (in pptv)	CFC12 (in pptv)	Computer used for simulation	Compiler used for simulation	Date of first entry in datebase	Date of most recent modification to entry in database
Pre-Industrial_Coupled_Model	Pre-Industrial OA Coupled Model at high resolution. Years 183-289	106	86	20	279	703	276	0	0	turing	unknown	Not_Recorded	Not_Recorded

You can access the means from this run, by going here

This page shows a small subset of images from this simulation. They have already been produced and there are no optional choices. If you want to do a more detailed examination of the model results, you can go to the main page and select one of the Parts. This will give you full access the system

Summary of Results for Experiment xakxa

Variable
DJF
JJA
ANN
Monthly
Variable
Temperature at 2m
Spinup: Surface Air Temperature
Precipitation
Spinup: Surface Air Temperature - seasonal mean removed
Evaporation
Spinup: Deepest layer Soil Moisture
http://www.bridge.bristol.ac.uk/
Surface Air Temperature (degC)
Mar
90N
60N
30N
0
30S
60S
90S
180
120W
60W
0
60E
120E
180
-44
-36
-28
-20
-12
-4
4
12
20
28
36
44
BRIDGE
University of BRISTOL
4 fps

Figure 1: The BRIDGE Data Access Portal

2.1 Exploring existing software development

We continued by exploring the current software system put in place to manage palæoclimate research data as this system has been and continues to be extremely dynamic, in order to follow the science involved and the needs of the scientists who use it. This is therefore an extremely valuable source of information on user requirements, technological requirements and preliminary insight into the de-facto research and data lifecycle evolution.

However, the system is currently very much designed to simplify the work of the climate modeller in that the interface really helps a scientist to work on his/her own experiments: the metadata describing the experiments usually references other experiments on the system which were used to create it, as well as parameters used in the first place by the central UM interface. Within each experiment, the variables shown on the web interface are taken verbatim from annotations stored within the file itself, each of which follow the CF metadata standard.

There is currently no requirement for the modellers to describe their experiments in a way an external, non-modeller, user could understand, or for that matter a way a computer could interpret. The use of CF metadata is a very good start but it is embedded within the file and only describes that specific file in which it is embedded with no references to the experiment to which it belongs. There is therefore a need to work on an experiment-level metadata schema that would describe the experiment as a whole and enable proper indexing on values that all users of the system could understand and not only the original modeller who created the data.

We have started looking at several metadata formats, such as the DIF (Directory Interchange Format) schema created by NASA [17] and the currently on-going work on the Scientific Data Application Profile [18].

2.2 Describing the Research Lifecycle

The process of creating, disseminating, storing and reusing research data is part of the overall research lifecycle. In order to come to an understanding of how this works, therefore, it is useful to characterise the research lifecycle that underlies it. There are considerable potential benefits to this process; if the process as it is today is well understood then it becomes possible to support the process as it stands, and potentially to find social, process-oriented and technical means to improve the speed, ease, and cost-effectiveness of that process further.

There are a number of models, mechanisms and proposed methods designed to support this process, a few of which we will briefly discuss here. Swann, for example, designed a model that was used for some time by the UKRDS (UK Research Data Service). This focused on separation of individuals involved in the research lifecycle into a set of possible types, notably data creator, user and viewer [19]. This was useful as a method of decomposition, but focused on categorizing people into one of a number of types. It was later suggested that individuals might more usefully be seen as involved in a number of different activities, and hence a later model focussed on individuals' roles at given times within a give research workflow.

'Data Sharing Verbs' represent one such model, a mechanism described by the ANDS as a 'structuring device', to support discussion about the technology and process of the data sharing aspect of the research lifecycle. The key insight underlying this is the assertion that thinking about the 'what' rather than the technical details of the system is useful — that user experience can be described through a description of what is being done from the user perspective. This mechanism is described by Burton & Treloar as 'Data Sharing Verbs' [20]; the candidate terms offered include Create, Store, Identify, Describe, Register, Discover, Access and Exploit, although additional verbs are likely to be required for specific use cases and as time passes.

This approach can be effectively compared to relatively traditional methods drawn from human-computer interaction and design methodologies, such as task analysis and decomposition. According to Kieras [21] task analysis is the process of understanding the user's task thoroughly enough to help design a system that will effectively support that user in doing the task. Task analysis aims to systematically analyse a task based on the knowledge and goals of the user, system, information and functionality (that is, social, organisational, technical factors). The 'Data Sharing Verb' idea could be described as a user-focused subset of this overall set of aims, specifically characterising an accessible researcher-level viewpoint on that overall area of endeavour. The fundamental aim of Data Sharing Verbs is as a structuring device, high-level architectural approach and descriptive mechanism [20]; they are described as 'one way of thinking about the things that need to take place', and it is noted that they 'encourage a focus on the functionality [and] result'. They can therefore be seen as an approach to collaborative representation and design. However, little information is provided regarding the mechanisms by which they are assigned to a novel usage context, so that is an area of interest for our ongoing work.

The work reported here was achieved using methods derived primarily from classical task analysis, with modifications designed to take in the useful idea of accessible data sharing verbs. There are many formalized methods in existence for the purposes of requirements gathering and task analysis in particular, but these do not in general provide a novel mechanism of analysing or understanding a task. In fact, much like the Data Sharing Verbs representation described above, most formal methods are ways to represent the results of a task analysis [21].

According to Kieras [21] the process of task analysis itself is usually based around some or all of the following methods:

- observation of user behaviour – a thorough, systematic and documented overview of observations with the aim of understanding the user's task. This may use a think-aloud protocol (ie. the user is invited to vocalise his/her observations about a task while working through it).

- Review of *critical incidents* and *major episodes* – rather than discussing the full span of user experience, a subset of particularly informative case studies are discussed.
- Questionnaires: these often suffer from difficulties with accuracy limitations, but are economical to use and can collect some types of user and task data.
- Structured interviews: talking to users or domain experts about a task is a good way of gaining some idea of the basics, and a more structured interview series at a later time can be an effective means of systematically exploring the area.
- Interface surveys: exploring existing interfaces, scripts, and so on, can provide useful information about interface characteristics, explanations, interface issues as perceived and annotated by users, and so forth.

Due to the inevitable time constraints of a relatively short-term project we chose to limit the use of observational/ethnographic methods to the latter phases of exploration of our system. Instead we looked towards the use of, initially, unstructured interviews, supplemented by an intensive interface survey series of the various visual and script-oriented interfaces that have been developed to serve the day-to-day needs of BRIDGE users of various types over the fifteen years of its operation. We then used this information to build a questionnaire, the results of which will be used to develop our initial findings as presented here into a second iteration.

We do, however, feel that ethnographic methods and think/talk-aloud workthroughs are likely to be of importance, particularly when exploring the cost-value propositions underlying our interface and those of other data providers/data centres in which the data is deposited. For example, it is often the case that users perceive deposit processes in particular as excessively lengthy and something of a waste of time, and in some cases there are very different ways to present that task to alter the value proposition as presented to the user.

3. Results

We begin by describing what has been elicited so far regarding data generation, storage, administrative and descriptive metadata, and reuse. We then present a candidate research data lifecycle model. Because the findings demonstrated emphatically that data consumption and reuse was a very significant part of the lifecycle, and indeed proved to represent the proximate cause of a great deal of the effort historically applied to this data collection, we found the need to place a far greater emphasis on it than was originally predicted.

BRIDGE data is generated via global climate models simulations (GCM), run on several national and international high performance computing (HCP) facilities. Our main tool is the Met Office Unified Model (UM) which runs a number of standard models such the Hadley Centre HADCM3 and HADGEM, or more recently FAMOUS, but we also use the European oceanography model NEMO or GENIE (Grid ENabled Integrated Earth). The majority of our output comes from the UM which uses a proprietary output file format. However the industry standard for such large data sets is NetCDF.

NetCDF, currently maintained by University Corporation for Atmospheric Research (UCAR), is a widely used open standard. It is an extremely flexible format optimised to store large multidimensional arrays of numerical data such as those describing high resolution planet-wide data.

When the data is created, it is moved and converted to NetCDF to a storage and processing farm of server where the data is processed. Climatology involves running weather simulation and then averaging the output to obtain the climate information. There is a number of default processes that are always running on the data to produce defaults sets of plots. It is then up to the modeller to add the specific output required for a specific project.

Due to the large amount of data created (around 2TB per day of raw output), it is not possible to store and keep everything, so raw output (from the UM) is discarded after conversion to NetCDF and calculation of intermediary averages generated from the converted files. Only the directly converted NetCDF files and the final averages and plots are kept. No expiration date is currently mandated for the data.

3.1 BRIDGE Service Design

The BRIDGE project at present has over 100 research groups spread over approximately 10 countries—see stakeholder analysis, figure 2 and figure 3. The multidisciplinary reach of palæoclimatology data presents some unique challenges in data dissemination. Historically, this diversity in user communities has meant that direct interaction with expert users of the BRIDGE environment is a necessary component in enabling access to, and reuse of, research data. However, as the number and diversity of background of stakeholders has continued to widen, these manual processes have become increasingly unfeasible. Enabling computer-supported scientific collaboration is at the intersection of Computer Supported Cooperative Work (CSCW) and e-Science [22], and the specific problem of data curation is a recent addition to the area.

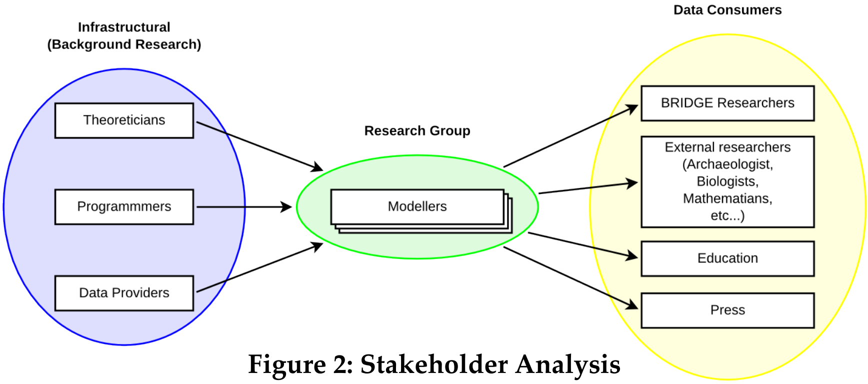


Figure 2: Stakeholder Analysis

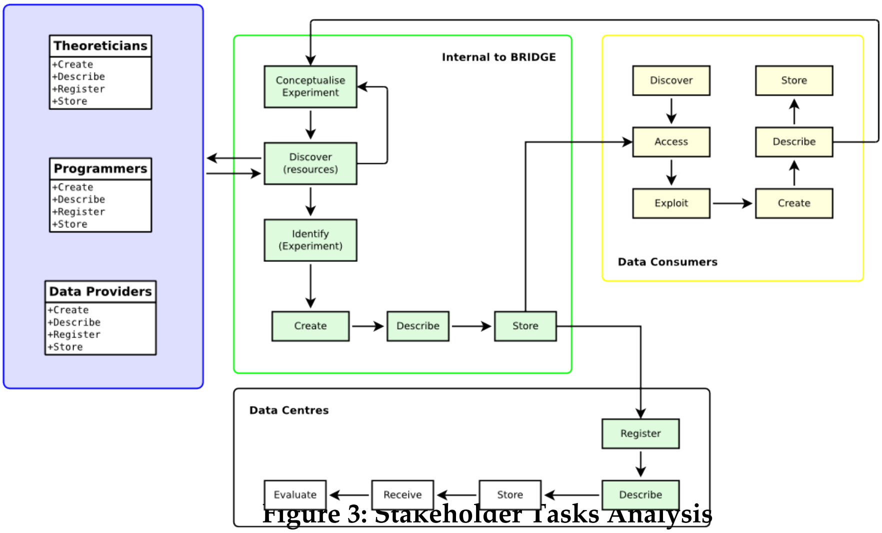


Figure 3: Stakeholder Tasks Analysis

The first challenge for those working in interdisciplinary research is to locate relevant data repositories and databases [23]. The second is to get 'up to speed' with the nature of the data and with its practical uses, metadata and it's provenance.

3.2 BRIDGE Systems Architecture

The current BRIDGE infrastructure only supports UM data which constitutes 99% of the data utilised. Compatibility with non-UM data is under consideration.

The current BRIDGE facilities provide services for the groups of stakeholders described here as the research group and the data consumers. Data providers are accessed by the modellers independently as the sources providing boundary conditions are rarely computer readable and usually come in the form of results published in scientific papers. These have to be 'translated' by the modeller before being added to the models.

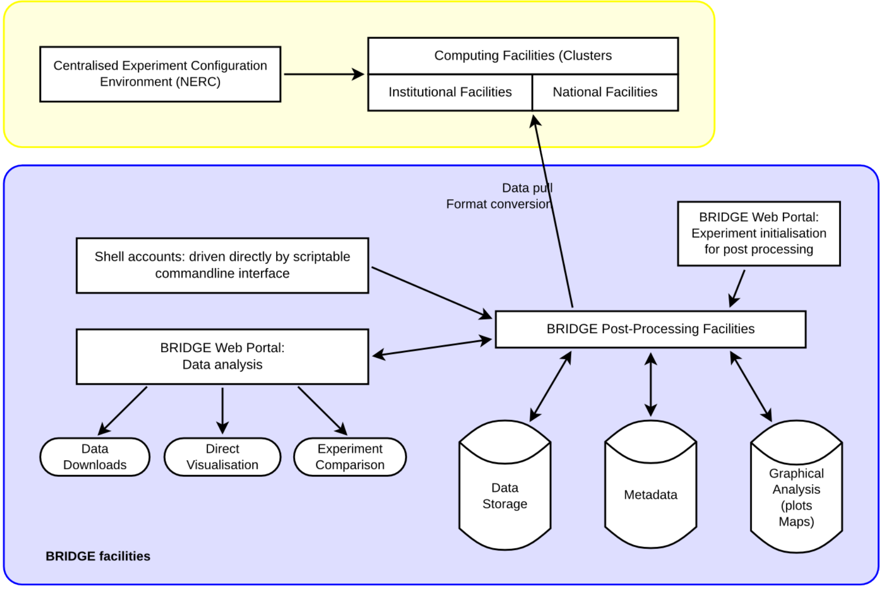

Centralised Experiment Configuration Environment (NERC)
Computing Facilities (Clusters
Institutional Facilities
National Facilities
Data pull
Format conversion
BRIDGE Web Portal: Experiment initialisation for post processing
Shell accounts: driven directly by scriptable commandline interface
BRIDGE Post-Processing Facilities
BRIDGE Web Portal: Data analysis
Data Downloads
Direct Visualisation
Experiment Comparison
Data Storage
Metadata
Graphical Analysis (plots Maps)
BRIDGE facilities

Figure 4: Architecture Diagram of the BRIDGE facilities

In figure 4, we show the overall architecture of the services provided by the BRIDGE portal. Experiments are configured on a centralised national facility provided by NCAS and run on national and institutional HPC facilities. In parallel, BRIDGE modellers need to initialise their experiments on the

BRIDGE facility by inputting details and metadata of the experiments. Once this is done, the modellers just have to let things run; the system is fully automated (unless the experiment fails in some way).

In order to avoid straining the limited storage capacity of HPC facilities, generated data is pulled regularly by the post-processing servers which then check and convert it to NetCDF from the proprietary UM format. The original UM files are then discarded. This process runs during the entire time taken by the experiment to complete, which can be up to several months. On completion, the modeller is given the choice to apply predefined averaging algorithms to the data, or define his own, in order to create an initial set of plots, maps and animations for a preliminary analysis of the data. The predefined algorithms are updated regularly by the research group to suit its evolving needs.

Once the experiment is processed, most of the data is archived and only the post processed data, enough to generate most graphics is kept available and shared on the portal. From this point, the experiment is available to 'data consumers'. The portal allows users to to view pre-generated graphics as well as creating new ones from the data, either to change the output format, to use different variables, or to combine multiple variables. The option is also given to compare the results of two experiments. This is made possible by the fact that extensive work has been done to make all graphics of the same type use the same colour scaling (visual conventions). All generated graphics outside the default predefined ones are cached for a limited time period so will not have to be re-generated at every access (a very time consuming and resource intensive process). Users also have access to post-processed data, either in NetCDF or converted into other formats such as CSV.

3.3 Case study: BRIDGE in Archæology

Archæology researchers at the University of Southampton make use of the BRIDGE software as part of their research. In this case study, their interest is in data regarding the climate in which a group of early Neanderthals lived. The specific information that can be provided as a result of BRIDGE palæoclimate simulations includes wind speed, temperature, and rainfall. Palæoenvironmental information can help archæologists understand likely patterns of migration as well as providing contextual information surrounding artefacts, etc. In particular, palæoclimatology may be key to our understanding of the extinction of the Neanderthals [24].

Originally, very few NetCDF viewing applications existed for non-UNIX environments. Therefore, the use of BRIDGE resources was required. Even now, the level of computer literacy required to analyse NetCDF data is very high. Our data uses meteorological units (temperature in Kelvin, precipitation in kg/m^2/s and wind in m/s) whereas what is usually required is the more every day units (temperature in Celsius, precipitation in mm/day and wind in mph or kph.) Doing a single numerical unit conversion may not be a complex process, however, the overall process of extracting thousands of values from a number of files and then performing type-appropriate batch conversion is relatively challenging and time consuming. It was therefore decided to add data conversion and merging services to the BRIDGE service.

Another issue regarding the interdisciplinary use of climatology raw datasets is the terminology used to describe the data variables contained in each files. This is even an issue for a glaciologist trying to use palæoclimate datasets. The netcdf files are all CF compliant (Climate and Forecast Metadata convention, as required for data generated as part of NERC funded projects) which includes over 30 variables describing some sort of air temperature—eg. *air_temperature, air_temperature_at_cloud_top, surface_temperature, surface_temperature_where_land, surface_temperature_where_open_sea, …*— as well as over 15 names describing types of air pressure (*air_pressure, air_pressure_at_cloud_base, air_pressure_at_cloud_top, air_pressure_at_convective_cloud_base, air_pressure_at_sea_level, …*). This multiplicity of terms which for some disciplines would be described as air temperature and air pressure makes exchange and reuse of data particularly difficult without very close collaboration with a scientist. An individual acting as a 'gateway' between disciplines would ordinarily be from the same field as the original data creator but who also understands the requirements of the scientist who is trying to use the data.

Issues brought up during this work included the difficulty of discovering appropriate datasets—finding experiments that contained relevant data. This was solved by requesting that appropriate experiments were recommended by BRIDGE team members. This, coupled with the need to automate common tasks, meant that the collaboration had a significant cost in terms of time. Hence, changes made to the service at the time included a concept of 'typed' data—for example, precipitation—to which a number of standard conversions may be applied. The need for appropriate metadata is also very clear, but with a legacy of over a decade of datasets (over 2000 simulations), the problem of introducing an improved standard includes the need to deal with a large amount of legacy content. Metadata applied to the data should also enable the cross-disciplinary browsing, discovery and use of the data, by the use of some sort of description table or translation table to either provide this translation automatically or provide the user with a plain english description of the term to allow him or her to choose the right one.

4. Discussion

The task analysis/preservation hybrid approach, making use of the 'data sharing verbs' to support discussion, has fitted well into our environment. Furthermore, it offers a strong theoretical basis in both preservation and HCI.

So far, we have successfully completed an investigation into the research lifecycle of research data from the BRIDGE project. We have built up an understanding of the existing software and hardware infrastructure that has been built up to support this lifecycle, and explored the rules associated with data creation and reuse, both external and internal in nature. We have also explored a case study of the reuse of palæoclimate data, in which archæology researchers at the University of Southampton make use of the BRIDGE software to access relevant datasets for the purpose of exploring patterns of migration. From this case study, we note a need for clear and consistent metadata, as well as for metadata to be applied to existing and older datasets – and we note that such collaborations often have a significant cost in terms of time, which can be reduced by enabling the development of software that supports ongoing collaboration by accessing consistent and well-defined data-access services or APIs.

The next stage for us is to ground our existing work with further detailed analysis of:

- the path(s) to completion of common tasks; for example, the time taken to complete a task, technical and knowledge-organisational issues and dependencies.
- technical infrastructure/system
- related infrastructural dependencies, such as the requirement to deposit information in data centres
- patterns of reuse of the data; impact, review and overall benefit to the community
- the costs and benefits of each aspect of the system.

4.1. Updating BRIDGE

Initially, we chose to focus on data management requirements analysis, exploring requirements for named stakeholders. Following the work described here, we have greatly improved our understanding of the broad technical and social processes that take part around the BRIDGE data. Now, however, we will need to identify appropriate methodologies for developing an improved understanding of the practical implications of the system as it is described here. For example, the time taken to complete any given process is very relevant to the question of the total cost of that process. For example, the time taken to develop an archival copy of a dataset (depending on the definition of the term 'archival'; this necessarily depends on the choice of archiving method, so that the costs of putting data into a data repository and that of storing it locally are very different) may be measured.

We will also continue to explore the practical issues and opportunities surrounding the reuse of BRIDGE data both in local formats and in the data-centres' preferred representations and formats.

4.2 Requirements analysis: Preservation, accessibility and metadata extraction

We intend to continue by consolidating our work with further questionnaires, observational studies and interviews. Tor this purpose, we have identified relevant components of the JISC Data Audit Framework (DAF) [25], DRAMBORA [26], the Planets project-preservation planning workflow [27], and similar tools to help identify and develop a formal data management strategy for palæoclimate model data, taking into account the requirement that consistency with the NERC Data Grid is a critical factor.

This should enable us to analyse the workflow described above in more details. In particular, we are looking to gain further information about users' (data creators and consumers alike) viewpoints and experiences with the data, its administration, access issues and potential enhancements. To this end, we have developed a questionnaire adapted from the Data Audit Framework, which is generally expected to be used primarily within an organisational context. The use of the DAF to explore data reuse externally constitutes a change from the usual way in which the framework is used, so it will also be an opportunity to explore and evaluate this approach.

In the following phase we expect to apply these components to the problem area described in this paper.

4.3 Requirements analysis: Automated metadata extraction.

We expect to explore the use of relevant metadata standards—PREMIS [28], STFC, etc.—to enhance the structures currently in use, as well as exploring the use of metadata extraction in order to supplement the file-format specific metadata currently used as the primary data management tool. The scale of data generated in palæoclimate research means that, wherever possible, metadata will need to be automatically generated.

Automated metadata extraction is the process of mechanically extracting metadata from a source document [29]. A completely automated process is unlikely to give perfect results; however, augmenting a manual metadata extraction process with an automated mechanism, even one that has an error rate of perhaps 10% to 20% of cases can nonetheless increase the speed and, potentially, the consistency of a metadata generation process. It can also increase user satisfaction with the interface; that the system has tried to support the user, even if it has not totally succeeded, can lead to a less frustrating user experience than a totally manual system.

In this instance, automated metadata extraction may explore the datasets and their associated files and format metadata as sources. Additionally, one may use the paper-based outputs of the research process as a source of information about the simulations that took place. One particularly relevant point to this process is the problem of data citation; what should a data citation look like, and what does it resemble at present? Informal exploration of the problem area has suggested that a co-author relationship is often used as an alternative to dataset citation, acknowledging the contributor of the research data in an implicit manner.

4.4 Supporting user reuse of data: Accessibility and visualisation

Exploration of the extent and diversity of the user community surrounding the BRIDGE dataset has demonstrated that data reuse is widespread and diverse. Much of this is data reuse is formally uncharted, which is to say that although it appears in individual researchers' records, often as citations, it is not always acknowledged as such. The nature of the data makes many different representations possible; as geographical data it can be directly explored using software such as NASA's WorldWind [30]. However, radically different representations may be appropriate for different user groups – so the collection of end user requirements is key to scoping out relevant activities such as developing appropriate recommendations for APIs, services or policies relating to preferred data storage formats.

A few specific cases that we expect to explore in the near future include the requirements for development of clear, high-quality visualisations, suited for high definition broadcasting in the media, and the requirements for the development of simulations that support haptic rendering, which is to say, that augment visual representations with tactile feedback. Provision of an application programming interface that can support this work is expected to facilitate this sort of development in future, as it should reduce the cost, complexity and learning curve associated with making use of the dataset. Because ongoing reuse of the data is an important part of this research data lifecycle, making it as easy as possible for developers to work with the information is likely to be an effective way of increasing the impact of research data publication in the area.

4.5 Expectations

The key assertion underlying this project states that adoption of appropriate data management strategies, appropriate to partner institutions across the various research disciplines involved, will have several benefits. The most visible initially is expected to be improved accessibility for potential users of the dataset. We additionally assert that the sustainability of a research data curation programme is dependent on the existence of data management strategies with a robust approach to appraisal. Finally, a strong data management strategy should improve traceability, reducing the difficulty of answering questions such as data origin and confidence levels.

5. Conclusion

BRIDGE software is already being used to support a wide range of reuse patterns, including those described above. From exploring practical usage patterns, we have developed a number of updated requirements. The first is the need to provide high-quality metadata, enabling us to develop means for searching or browsing—exploring—the data, in an appropriate manner for specific end-user groups, be they archæologists, statisticians or biologists. It requires a viewpoint on metadata that is not excessively prescriptive or restrictive in terms of form or interface, but that enables the base dataset to be presented to many different user groups, in their own terms. The current BRIDGE software review will take into account these disparate user requirements in designing a flexible architecture that can support the generation of a wide variety of data representations.

Secondly, preservation is a key issue, along with provenance and the ability to precisely cite a given data set. Climate science is not a subject in which the ‘fire and forget’ philosophy can be adopted. However, it is also an area of e-Science that generates very large quantities of data. Data curation and preservation in this area is reliant upon the development of appropriate data retention policies; as part of the PEG-BOARD project we will explore data management requirements and develop appropriate policies along with any infrastructural dependencies.

Finally, better-quality visualisations and tools able to support accessible exploration of data are very important enablers for data reuse and widening the impact of completed research. This is a rich and open field for further research and development, particularly but not exclusively for educational purposes; high-quality visualisations are also sought after in many other fields, including audiovisual broadcast.

Notes and References

[1] LINACRE, E. *Climate data and resources : a reference and guide*. Routledge, 1992.

[2] SWEET, B. Three Cultures of Climate Science. *IEEE SPECTRUM*, 2010.

[3] EDWARDS, PN. *Global Climate Science, Uncertainty And Politics: Data-Laden Models, Model-Filtered Data. Science as Culture,* 8:437–472, 1999.

[4] BENFORD, S. et al. *e-Science from the Antarctic to the GRID*. In 2nd UK e-Science All Hands Meeting, 2003.

[5] LEE, PA; ANDERSON, T. *Fault Tolerance: Principles and Practice (Second Revised Edition).* Springer-Verlag, 1990.

[6] ANANTHAKRISHNAN, R. et al. *Building a global federation system for climate change research: the earth system grid center for enabling technologies (ESG-CET)*. Journal of Physics: Conference Series, 78, 2007.

[7] *Geniefy: Creating a grid enabled integrated earth system modelling framework for the community.* http://www.genie.ac.uk/GENIEfy/, 2009.

[8] LENTON, TM; et al. *Using GENIE to study a tipping point in the climate system.* Phil. Trans. R. Soc. A, 367(1890):871–884, 2009.

[9] SIMMHAN, YL; B. PLALE, B; GANNON, D. *A survey of data provenance in e-Science.* SIGMOD Rec., 34(3):31–36, 2005.

[10] DE ROURE, D. *e-Science and the Web. IEEE Computer*, August 2009.

[11] HARRABIN, R. *Harrabin's notes: IPCC under scrutiny.* http://news.bbc.co.uk/1/hi/sci/tech/8488395.stm, 2010. Retrieved Jan 30, 2010.

[12] SU-SHING, C. *The Paradox of Digital Preservation,* IEEE Computer, 34, p. 24–28, 2001.

[13] MCGUFFIE, K; HENDERSON-SELLERS, A. *A Climate Modelling Primer, Third Edition*, p. 47–79, 2005

[14] LYON, L. *Dealing with data: Roles, rights, responsibilities and relationships.* Technical report, UKOLN, Bath, UK, June 2007.

[15] LAPLANTE, P. *Requirements Engineering for Software and Systems (1st ed.),* CRC Press, 2009.

[16] NUSEIBEH, B; EATERBROOK, S. *Requirements engineering: a roadmap,* Proceedings of the Conference on The Future of Software Engineering, p.35-46, June 04-11, 2000, Limerick, Ireland

[17] OLSEN, L. *A Short History of the Directory Interchange Format (DIF),* http://gcmd.nasa.gov/User/difguide/whatisadif.html.

[18] BALL, A. *Scientific Data Application Profile Scoping Study Report,* http://www.ukoln.ac.uk/projects/sdapss/, June 2009

[19] SWAN, A; BROWN, S. *The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs,* 2008.

[20] BURTON, A; TRELOAR, A. *Designing for Discovery and Re-Use: the ? ANDS Data Sharing Verbs? Approach to Service Decomposition.* International Journal of Digital Curation, 4(3), 2009.

[21] KIERAS, D. *Task Analysis and the Design of Functionality*, In CRC Handbook of Computer Science and Engineering, p. 1401–1423, CRC Press, 1996.

[22] KARASTI, H; BAKER, KS; HALKOLA, E. *Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER).* Network. Comput. Supported Coop. Work, *15(4):321–358, 2006.*

[23] LUCE, R. *Learning from e-databases in an e-data world.* EDUCAUSE Review, 43(1):12–13, 2008.

[24] VAN ANDEL, TH; DAVIES, W; edts. *Neanderthals and modern humans in the European landscape during the last glaciation.* Cambridge: McDonald Institute for Archæological Research, 2004.

[25] JONES, S; BALL, A; EKMEKCIOGLU, Ç. *The Data Audit Framework: A First Step in the Data Management Challenge.* International Journal of Digital Curation, 3(2), 2008.

[26] MCHUGH, A; ROSS, S; RUUSALEEP, R; HOFMAN, H. *The Digital Repository Audit Method Based on Risk Assessment (DRAMBORA).* HATII, 2007.

[27] FARQUHAR, A; HOCKX-YU, H. *Planets: Integrated Services for Digital Preservation. International Journal of Digital Curation, 2(2), 2008.*

[28] *PREservation Metadata: Implementation Strategies Working Group.* PREMIS Data Dictionary, 2005.

[29] TONKIN, E; MULLER, H. *Semi automated metadata extraction for preprints archives,* JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, p. 157–166, 2008.

[30] NASA World Wind Java SDK, http://worldwind.arc.nasa.gov/java/

MOTIVATIONS FOR IMAGE PUBLISHING AND TAGGING ON FLICKR

Emma Angus[1]; Mike Thelwall[2]

School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1LY, UK
[1] e-mail: emma.angus@wlv.ac.uk
[2] e-mail: M.Thelwall@wlv.ac.uk

Abstract

Changes in photographic and internet technology have revolutionised the way people create, process and share digital images. This paper investigates people's motivations for image publishing and tagging on the web 2.0 site Flickr. Using an online pilot survey, 33 participants answered questions about their uploading and tagging practices, and whether or not they hope to make a commercial gain from their images. The results show that most people have two main motivational reasons both for using Flickr, and for the tagging of their images. However, whilst a person may be motivated to use Flickr for both personal and social reasons, tagging motivation tends to focus more exclusively on either one or the other of these two factors. Overall it was found that social organisation and social communication are the most popular motivational factors for both using Flickr and for tagging images, suggesting that Flickr is enjoyed for the community environment it provides rather than as a place to store images. However despite people's desire to share their images, most users are not hoping to make a commercial gain from the items they upload.

Keywords: motivations; digital images; tagging; Flickr.

1. Introduction

Advancements in photographic technology have resulted in a renewed interest in the role of the image in fields such as information and computer science, anthropology, economics, sociology and visual studies. Increasing numbers of individuals carry a camera with them every day: either a digital camera or a mobile phone with an inbuilt camera. Free from the limitations of a 24/36 exposure film, people are now able to point and click almost endlessly. Coupled with this advancement in photographic technology, there are significant changes in the way the population is using internet technology; the main change being the web 2.0 pronounced emphasis on collaboration and user contribution. Kirk *et al.* [1] point out that 'the traditional role of the user from one as picture "taker", into picture editor, developer and printer' is also contributing to the pervasive photography landscape and this has been supplemented with the development of web 2.0 image sites such as Flickr, PhotoBucket and Picasa. Such sites allow users to upload, store and share images either with selected friends and family or with the public at large. Images uploaded to such sites are generally annotated with 'tags', which are freely chosen keywords [2], assigned by the user, ostensibly to aid with subsequent search and retrieval.

These changes have had a dramatic effect on attitudes towards self presentation and publishing. The Web 2.0 revolution is 'enabling Internet users to author their own content….[a] technology platform [that] will radically democratize culture, build authentic community, [and] create citizen media.' [3] As a result of this influx of user-generated content on the Web, there are some big businesses and organisations which are now looking to take advantage of this new model of creation and authoring. One such example is Getty Images, the world's largest distributor of pictures and videos. Getty joined forces with Flickr in July 2008 and their editors will now be regularly browsing through Flickr for images they like and inviting selected users to become paid contributors to their team of professionals.

However, does Getty's new business model actually complement the desires of Flickr users and their user-generated content? Why do people actually publish images on Flickr and what do they hope to achieve from doing so? This pilot study investigates what motivates individuals to publish their images, what motivates them to tag their images, and whether or not people are seeking to make a commercial gain from the images they publish on Flickr.

2. Related work

Digital images are at the core of Flickr's existence (despite the fact that Flickr's creators originally intended it as an online game [4]). However, Flickr's attraction now lies in its ability to act as both an image storage site and as a place for people to share images and cluster in communities of like-minded people in order to converse, share tips and advice on photographic techniques and to gain comments and feedback on photos which have been uploaded. Most of the work to date which has looked at Flickr describe it as a social site and a place for sharing images rather than as a place for merely storing and backing up collections of digital images [5, 6, 7, 8].

Reasons for taking and publishing images

Kindberg *et al.* [9] carried out an in-depth investigation into camera phone use and differentiated between *social* and *individual* intentions behind image capture. In their investigation they found that two thirds of all images taken on camera phones are captured with the intent to share (i.e., taken with *social* intentions). Whilst the nature of taking pictures on a phone may be different to that of using a standard digital camera due to the ease with which images can be sent simultaneously to contacts in the phone's address book, Kindberg also found that the subjects in their study only knew on average eight people who had compatible camera phones who they could actually send images to. Subjects also expressed the intent to permanently save a selection of their images on either a PC, with the subsequent intention to perhaps share certain images with friends via email or by posting onto a webpage.

The traditional reason for taking photographs on a standard camera (whether it be an instamatic, an SLR, or a digital camera) is to document memories and events and to store them so that family, friends and future generations can look back on them. In an investigation into how people manage their collections of photographs, Rodden and Wood [10] found that the organisation of traditional photos requires significant effort, and is not usually done to facilitate searching but to create an attractive 'presentation' of photos for keeping as part of a 'family' or 'personal' archive. Digital organisation on the other hand requires much less effort and is much more likely to be carried out with the intent of sharing the photos and allowing others to view them in the near future.

Cox *et al*. [8] carried out open-ended telephone interviews with 11 Flickr users in an attempt to 'explore the use of the system within the context of the interviewees' photographic practices.' One of the questions which was asked of participants was: 'Why do you use Flickr?' Overall, the interviewees expressed that they used Flickr as, 'part of a wider nexus of self presentation or communication through the web' and their collection of photos on Flickr was, 'usually a selection of the best or most appropriate to be shared.' Flickr itself was also found to be an important motivation for taking photos in the first place.

Similarly, Van House *et al*. [11] in their interviews and observations with 60 participants found that sharing is an important use of photos on cameraphones, and the authors argue that 'cameraphones will soon be the dominant platform for low end consumer digital imaging.'

Reasons for tagging images

Once images have been placed on the Web, if they are in a Web 2.0 archive then they may be tagged by the owner or others. Tagging is the process of adding keywords to something as a form of metadata. There is much debate concerning whether people tag their resources primarily for personal organisation or to aid in sharing and discovery [5, 6, 12, 13, 14]. Although categorised differently by authors and researchers, these tagging motivations can be largely grouped together to form two distinct bodies of motivational practices: organisational, selfish, personal, intrinsic; and social, altruistic, public, extrinsic, evangelical - or put another way, information management vs information sharing [15, 16].

Marlow *et al.* [6] claim that motivations to tag can be split into two high-level practices: *organisational* and *social*. Organisational motivations are associated with users attempting to develop a personal digital filing system, whereas social motivations are associated with users attempting to express themselves with other users of the system. Hammond *et al.* [12] similarly define these two practices as *selfish* and *altruistic*.

Ames and Naaman [14] extend the notions of *organisational* and *social* in an investigation which explored 'the various factors that people consider when tagging their photos' and the authors offer a taxonomy of tagging motivations based along the two dimensions of: sociality and function. The authors conducted in-depth semi-structured interviews with 13 Flickr users and they found that users generally had one or two primary motivations for tagging their images rather than solely one motivation, and that the motivations could be placed along the dimensions of sociality and function, rather than fitting into a mutually exclusive category (See Table 1.)

The sociality dimension relates to the tag's intended audience (i.e., for oneself, or for others: friends/family/public). The function dimension relates to the actual purpose of the tag (i.e., is it to aid in organisation: placing the image into a category or classifying it somehow according to when/where it was taken or perhaps grouping images into common themes. Or, is it to aid in communication: providing context about the image content, or perhaps tagging it as a way of drawing attention to it from other Flickr users).

Table 1. A taxonomy of tagging motivations (Ames and Naaman [14])

		Function	
		Organisation	Communication
Sociality	Self	• Retrieval, directory • Search	• Context for self • Memory
	Social	• Contribution, attention • Ad hoc photo pooling	• Content descriptors • Social signalling

From their findings, the authors suggest that most of the participants were motivated to tag by organisation for others (social organisation), with self organisation (adding tags for later retrieval) and social communication (adding context for friends, family and the public tied for second). This offers a more complex insight into motivation than previous research which has tended to crudely split tagging intention into either a manifestation of organisation for the self, or having the intention to share with others. The work of Ames and Naaman [14] proves that organisational tagging can often be carried out more so for the benefit of others than for the self.

As part of the telephone interviews carried out by Cox *et al.* [8], the authors also asked their participants, 'how do you choose descriptions, tags etc?' They found that a key motivation for tagging was in order to increase the amount of people who could find and view the interviewee's photos.

In a study which looked at whether users of social tagging systems use such platforms for the purposes of personal information management or for information sharing [16], 48 Flickr participants were recruited from the Mechanical Turk service. From qualitative judgements taken from free text comments, these Flickr users showed a strong tendency towards information sharing with friends and family, although personal information management still played a big factor in their motivations. Flickr users also perceived tags as helpful for information retrieval and users often search through image collections other than their own.

In a study into the use of Flickr, Van House [5] interviewed 12 Flickr users and found that most participants saw Flickr as, 'a social site, a place for sharing images...and since they rarely searched back over their own images, tagging was almost exclusively for others.'

The findings of Nov *et al.* [7] and Ames and Naaman [14] indicates that social presence plays a role in tagging behaviour. It could be hypothesised that if people are motivated to use and publish their images on

Flickr in order to share them with others, then this ‘social presence’ should motivate them to tag in a way which is socially orientated.

The previous research presented in this paper provides an excellent framework of motivational factors from which to base future studies on. However such research has either analysed Flickr image tags [17, 18, 19]; or motivations to tag [5, 7, 14, 18] and these studies have tended to adopt either a wholly quantitative (tag analysis) or qualitative (open ended in-depth interviews) approach. An investigation combining both qualitative and quantitative methodologies may help us to better understand people’s motivations behind image publishing and image tagging, so that conclusions can be drawn about the potential uses of web 2.0 image sites. To date there has been no empirical research which has investigated if users of Flickr wish to make a commercial gain from the images they publish there.

3. Research questions

This pilot study which is part of a programme of research into tagging with Flickr aims to combine a qualitative and quantitative approach via the use of a structured online questionnaire and it aims to directly compare motivation to use Flickr with motivation to tag within Flickr.

Using an information science and webometric approach this research paper addresses the following questions:

- What motivates people to publish their images on Flick?
- What are the key motivational factors for tagging images?
- Are people seeking to make a commercial gain from the images they publish?

4. Methods

In order to investigate what motivates people to publish and tag their images on Flickr, a pilot questionnaire was developed and administered on the Web to a sample of Flickr users utilising both a direct and indirect approach. This will be followed-up with a larger sample in a future study.

Questionnaire Design

As the target sample for the questionnaire was Flickr users, it was decided that an online questionnaire would be more appropriate than a paper based version. The questionnaire was designed using the online survey software and questionnaire tool, surveymonkey.com. Utilising the SurveyMonkey software, a custom designed questionnaire could be created fairly quickly and assigned its own unique URL. In order to try to increase the response rate of the questionnaire and also to make the questionnaire as user friendly as possible, a number of measures were taken:

- The questionnaire was kept short and consisted of only 1 page of questions with minimal scrolling needed
- A clean, simple and uncluttered layout was used

The questionnaire was comprised of four main sections:

- A series of question statements relating to a respondent's motivations for tagging their images (using a 5 point Likert scale)
- A free text box asking respondents to explain why they upload their images to Flickr
- A question asking if respondents hope that their images will be picked up by a commercial stock photography organisation or the media
- Demographic questions such as age, gender, and nationality

Question construction, wording and order

Based on the findings from the literature review, motivations for image tagging seem to naturally align with the two dimensions as put forward by Ames and Naaman [14]; the first dimension being *sociality* (relating to whether the tag's intended usage is by the individual or others i.e., self or social) and the second dimension being *function* (referring to a tag's intended uses of either facilitating later organisation and retrieval or to communicate some additional context to viewers of the image). In light of these two prominent dimensions, it was decided that the survey questions relating to motivations for tagging would be based on these two constructs. Therefore four questions were developed, one for each of the two main tagging motivations within each of the two dimensions, thus creating four main possible reasons for tagging. In order to increase reliability, a further set of four questions were then developed which could be paired with the first set. Respondents were asked to express their level of agreement/disagreement with these statements using a 5 point Likert scale.

It was decided that the demographic questions would be placed at the end of the questionnaire as the respondents may be more likely to disclose information such as their age once they had already answered some questions and felt a greater sense of involvement with the questionnaire as a whole. However whereas the motivational statement questions were a compulsory aspect of the questionnaire, the demographic questions were not, and a respondent could skip these questions if they felt uncomfortable disclosing such information. The researchers tried to ensure that all questions were worded in a short and concise manner in order to reduce ambiguity

Data collection

In order to try to increase the response rate of the questionnaire, both a direct and an indirect method of data collection were utilised. For the direct approach, the URL of the questionnaire was posted to the discussion forums of two public Flickr groups (*Flickr Social* and *Surveys&Quizzes)*. The indirect approach utilised advertising the questionnaire URL on the researchers' Facebook and Twitter profiles, and also on their personal web pages. In all instances, the questionnaire URL was accompanied by a small paragraph of explanatory text, briefly stating the purpose of the questionnaire and advising that all responses would remain confidential and any published results would be anonymised. A URL was also provided which linked to the first author's webpage where further details on the questionnaire and the study as a whole could be found. The questionnaire was available for a period of 3 weeks during March 2010.

5. Results

A total of 33 valid responses to the questionnaire were received. 51.5% of the respondents were female, and the mean average age of the respondent was 30 years. The majority of the respondents originated from the UK and Denmark. See Figure 1 for a full breakdown of nationalities

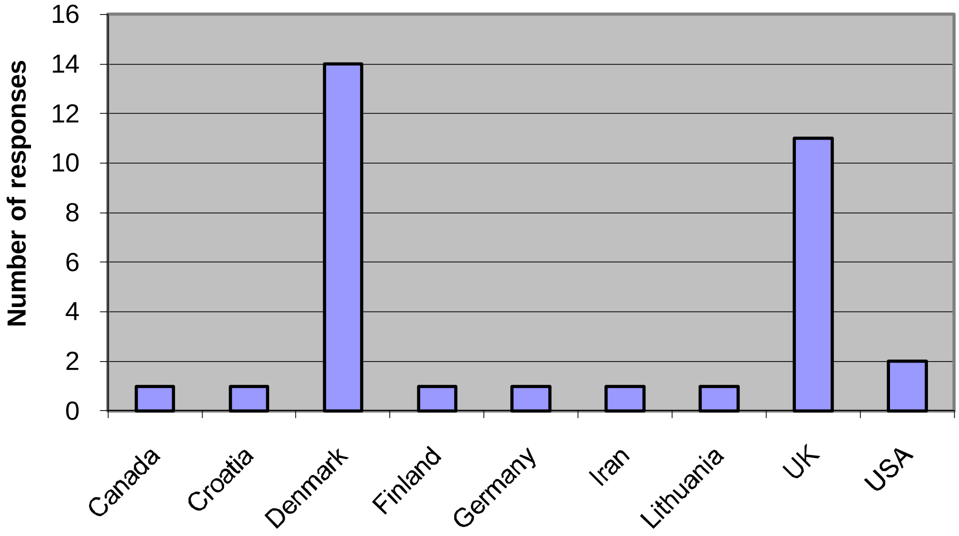


Figure 1. Respondent's country of origin

Why people upload their images to Flickr

Participants were asked to briefly explain why they upload their images to Flickr. The responses were then broken down into the reasons stated and these reasons were grouped together according to the motivational factors as put forward by Ames and Naaman [14] (i.e., social organisation, social communication, self organisation, self communication).

Most respondents (48%) reported two main reasons behind their use of Flickr, with the two most predominant reasons being to share images with friends and family (social organisation), and to promote their work and connect with other people in the photography community (social communication).

P22: "I use Flickr to promote my creative work, get feedback, and share with friends/family."
P31: "To keep a nicely presented, easily shared record of my photography and to get feedback, encouragement and advice from other users about technique."

45% of respondents reported that they had only one main motivation for using Flickr, and 6.5% reported that they had three main reasons. Figure 2 shows respondent's overall preferences between each of the four main motivational factors.

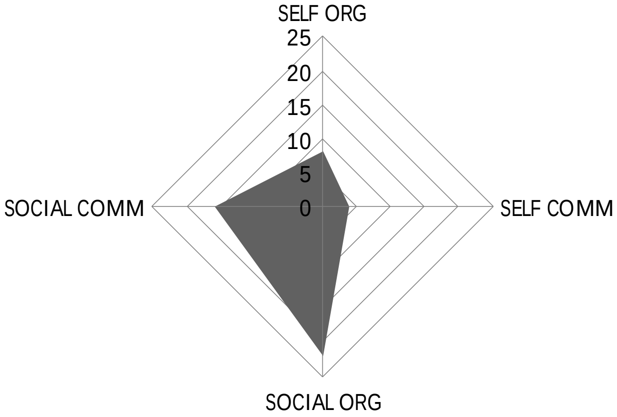


Figure 2. Number of respondents who exhibited each of the four motivational practices

The results support the general consensus that people are drawn to Flickr because of the social aspects and the 'community environment' it provides, rather than using it solely as a place to store and archive images.

Do people hope to make a commercial gain from their images?

Despite the fact that 51% of the respondents in this investigation specifically mentioned using Flickr as a way of promoting their work and receiving feedback on their images, 75.8% of respondents said that they did not use Flickr with the hope that their images would be picked up by either a commercial stock photography organisation or by the media. So whilst the 'sociality' element is a big factor for many Flickr users, people are predominantly interested in having their images found so that they can gain feedback and encouragement from other Flickr users, rather than hoping their images will be picked up by a commercial agency or the media.

What motivates people to tag their images?

Motivation to tag images slightly differs from people's motivations in using Flickr to publish their images. Whereas people strongly state that social organisation is the main factor in using Flickr, social communication comes out slightly on top in terms of people's motivations for tagging their images (see Figure 3). Social organisation and communication are the top two motivational factors in both instances. This finding differs from the work of Ames and Naaman [14] and Cox *et al.* [8] who found that social organisation was the top motivating factor in tagging practices.

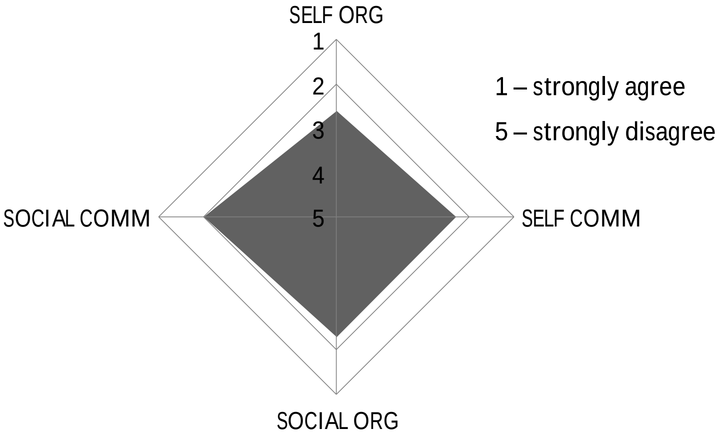


Figure 3. Motivations for tagging images

Similar to the finding which suggests that most people have two main motivations for using and publishing their images on Flickr, most people were also found to have two main motivations for tagging their images (42.4% of respondents).

Whilst people seem to be primarily drawn to Flickr because of the social function and community environment that it provides, tagging practices don't necessarily follow this primary motivation, with self organisation and self communication reasons appearing as fairly high motivational factors overall. It would seem that people are much more dominantly motivated by the desire to either please themselves or others when it comes to describing and adding context to their images.

Using a Spearman correlation and a Mann-Whitney test it was found that age and gender had no influence on tagging motivation.

Factor analysis

Despite a fairly small sample size, a factor analysis was performed on the survey items relating to the motivational constructs of self, social, organisation and communication. The correlation matrix shows that people gave similar answers to the two survey statements relating to social motivations, suggesting that this was a particularly coherent construct.

Table 2. Factor analysis correlation matrix

		Self1	Self2	Social1	Social2	Comm1	Comm2	Org1	Org2
Correlation	Self1	1.000	.387	-.351	-.514	.253	.265	.168	-.026
	Self2	.387	1.000	-.821	-.654	.162	.044	.095	-.125
	Social1	-.351	-.821	1.000	.622	.037	.085	.043	.165
	Social2	-.514	-.654	.622	1.000	-.077	-.100	-.172	-.162
	Comm1	.253	.162	.037	-.077	1.000	.354	.100	.127
	Comm2	.265	.044	.085	-.100	.354	1.000	.169	.164
	Org1	.168	.095	.043	-.172	.100	.169	1.000	.418
	Org2	-.026	-.125	.165	-.162	.127	.164	.418	1.000

However, the 'social' and 'self' statements tended to pair up with each other, so that someone scoring high on one would tend to score low on the other. This means that there are three main types of motivation rather than the predicted four.

This finding is further corroborated by the results shown in Table 3. Factor 1 is a social factor – the two social factors load on it and the two self factors negatively load on it (so are strongly not associated with it). Factor 2 is an organizational factor. Factor 3 is a communication factor, with negative loading on the self questions, suggesting social and self motivations are polar opposites in Flickr.

In summary, the factor analysis suggests that sociality (self vs. social), organization and communication factors are the three main independent types of motivation.

	Factor		
	1	2	3
Social1	.910		
Self2	-.869		
Social2	.765		
Self1	-.469		.438
Org2		.848	
Org1		.482	
Comm2			.610
Comm1			.541

Table 3. Rotated factor matrix

6. Discussion

The results from this investigation suggest that whilst it is possible to have a number of different motivations for using Flickr, as well as a number of different motivations for tagging images, tagging motivation will tend to be driven by only one direction of sociality (i.e., for oneself or for others) even if a person states that their motivation for using Flickr in the first place is for a mixture of self and social reasons. Tagging tends to be driven exclusively by either self or social reasons, with the factors of organisation and communication being less exclusive and in many cases playing a dual role.

The results of this investigation give a valuable insight why people publish and tag their images on Flickr, however the results cannot be generalised too widely due to the small sample size. Whilst some literature suggests that it usually takes no more than 12-25 cases to reveal the major difficulties and weaknesses in pre-test questionnaires [20] this is referring more to the design of the questionnaire and the discovery of things such as suppositions, awkward wordings or missing categories. In order to test the underlying assumptions of the information contained within the variables being questioned, it is suggested that, ‘a minimum of five subjects per variable is required for factor analysis.' [21]. This investigation was therefore seven subjects short of the 40 required in order to fully test the 8 statements included in the factor analysis. However the results from the factor analysis were clear and conclusive in suggesting that there were three main factors which made up the motivational statements rather than the predicted four. Therefore as a pilot investigation this proved to be a worthwhile finding, which could be further tested using a larger sample.

Despite the heavy bias towards UK and Danish participants, no noticeable differences were found in the motivational intentions of these two nationalities, so the main factor is the European bias, which could be further investigated by having a larger sample from a more internationally representative set of countries.

As with all surveys there is the possibility that participants may have lied when answering questions. People often answer questions in the way that they think they are expected to answer, and people also often answer questions quickly, without giving much thought to their answers. In order to try to overcome this problem, the main motivational statement questions were paired up, to test the assumption that people should answer similarly on the pairs of questions.

As stated in the Results section of this paper, it is possible for someone to have more than one main motivation to use Flickr, as well as more than one main motivation for tagging their images. However, whilst motivations to use Flickr can be for a mixture of both self and social reasons (i.e., using Flickr as a personal archive as well as using it to share images with friends and family), tagging motivation was found to be exclusively for either self or social reasons. This is particularly interesting given that a number of participants in this investigation specifically stated both self and social reasons for using Flickr:

P17: "I use Flickr to archive for myself and also to promote my work."
P25: "to store my images and to share with friends."
P30: "as storage and for displaying my images to friends and family."

These statements would suggest that perhaps people are not fully aware of how much their tagging practices differ from their main motivations for using Flickr in the first place.

7. Conclusion

Whilst motivations for using Flickr and uploading images can be for a number of different reasons at the same time, motivations for tagging images tends to have a more predominant role. People may use Flickr as both a personal archive and as a place to share images with friends and family, but their reasons behind choice of tags will tend to be very distinctly either a ‘self’ or a ‘social’ action, with less hesitation in the mind of the tagger as to who will ultimately benefit from their choice of tag. People don’t appear to want to use a mixture of highly personal and social tags; they will adopt one strategy or the other, regardless of if they are tagging for archive and storage or communicative purposes.

However in support of much of the previous work carried out on Flickr, the respondents who took part in this investigation seem to use Flickr for the social aspects and the community environment which it provides with social organisation and social communication being the two most popular motivational factors overall. Despite people’s desire to have their images found and commented upon, as a general rule, people aren’t interested in making a commercial gain from the images they upload – the community spirit of Flickr and its ability to connect people both known and unknown to the image uploader is its most appealing feature.

The responses from the pilot questionnaire have given a valuable first insight into why people publish and tag their images on Flickr, and also into the changing nature of self-publishing in the world of user-generated content.

Acknowledgements

The authors would like to thank Dr Ann O’Brien (Loughborough University) and David Goda (University of Wolverhampton) for their input and advice on the development of the questionnaire. A big thank you also goes to all of the participants who took the time and effort to complete the questionnaire and for providing such useful comments and feedback on their use of Flickr.

Notes and References

[1] KIRK, DS; SELLEN, A; ROTHER, C; and WOOD, K. Understanding Photowork’, In *Proceedings of CHI Conference on Human Factors in*

Computing Systems, 22-27 *April 2006*, ACM, Montreal: Canada p. 761-770.

[2] GUY, M; and TONKIN, E. Folksonomies: Tidying up Tags? *D-Lib Magazine,* 12(1), 2006. Available at http://www.dlib.org/dlib/january06/guy/01guy.html (March 2006).

[3] KEEN, A. The second generation of the Internet has arrived. It's worse than you think, *The Weekly Standard,* 15th Feb, 2006.

[4] FITZGERALD, M. *How we did it*: Stewart Butterfield and Caterina Fake, Co-founders, Flickr, December 2006. Available at http://www.inc.com/magazine/20061201/hidi-butterfield-fake.html

[5] VAN HOUSE, N. Flickr and Public Image-Sharing: Distant Closeness and Photo Exhibition, *CHI 2007 Proceedings, April 28 – May 3, 2007*, San Jose, CA, USA.

[6] MARLOW, C; NAAMAN, M; BOYD, D; and DAVIS, M. Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead, *Proceedings of the 15th International World Wide Web Conference*, *(Collaborative Web Tagging Workshop), May 22, 2006*, Edinburgh, UK.

[7] NOV, O; NAAMAN, M; and YE, C. What Drives Content Tagging: The Case of Photos on Flickr. *Proceedings of the 26th annual SIGCHI conference on human factors in computing systems*, 2008, p. 1097-1100.

[8] COX, A; CLOUGH, PD; and MARLOW, J. Flickr: a first look at user behaviour in the context of photography as serious leisure, *Information Research*, 13(1), March 2008. Available at http://informationr.net/ir/13-1/paper336.html (May 2008).

[9] KINDBERG, T; SPASOJEVIC, M; FLECK, R; and SELLEN, A. The Ubiquitous Camera: An In-Depth Study of Camera Phone Use, *IEEE Pervasive Computing,* 4(2), 2005, p. 42-50.

[10] RODDEN, K; and WOOD, KR. How do people manage their digital photographs?, *ACM Conference on Human Factors in Computing Systems (ACM SIGCHI 2003) Fort Lauderdale, Florida, USA, April 2003*, p. 409-416.

[11] VAN HOUSE, N; DAVIS, M; AMES, M; FINN, M; and VISWANATHAN, V. The Uses of Personal Networked Digital Imaging: An Empirical Study of Cameraphone Photos and Sharing, *CHI 2005 Proceedings, April 2-7 2005*, Portland, Oregon, USA.

[12] HAMMOND, T; HANNAY, T; LUND, B; and SCOTT, J. Social Bookmarking Tools (I): A General Review", *D-Lib Magazine*, 11(4), 2005. Available at http://www.dlib.org/dlib/april05/hammond/04hammond.html (October 2006).

[13] GOLDER, SA; and HUBERMAN, A. Usage Patterns of Collaborative Tagging Systems, *Journal of Information Science*, 32(2), 2006, p. 198-208.

[14] AMES, M; and NAAMAN, M. Why We Tag: Motivations for Annotation in Mobile and Online Media, *CHI 2007 Proceedings*, *April 28 – May 3, 2007*, San Jose, CA, USA.

[15] THOM-SANTELLI, J; MULLER, M. J; and MILLEN, DR. Social tagging roles: publishers, evangelists, leaders. In *Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, April 05 - 10, 2008,* ACM: New York. Available at http://doi.acm.org/10.1145/1357054.1357215

[16] HECKNER, M; HEILEMANN, M; and WOLFF, C. *Personal Information Management vs. Resource Sharing: Towards a Model of Information Behaviour in Social Tagging Systems*. *May 2009*, San Jose, California, USA.

[17] RAFFERTY, P; and HIDDERLEY, R. Flickr and Democratic Indexing: dialogic approaches to indexing, *Aslib Proceedings: New Information Perspectives,* 59(4/5), 2007, p. 397-410.

[18] ZOLLERS, A. Emerging Motivations for Tagging: Expression, Performance and Activism, *WWW 2007*, *May 8-12, 2007,* Banff, Canada.

[19] ANGUS, E; THELWALL, M; and STUART, D. General patterns of tag usage among university groups in Flickr, *Online Information Review*, 32(1), 2008, p. 89-101.

[20] SHEATSLEY, P. Questionnaire Construction and Item Writing. In "*Handbook of Survey Research*", ed. Peter Rossi, James Wright, and Andy Anderson, 1983, p. 195–230. New York: Academic Press.

[21] COAKES, SJ; and STEED, LG. *SPSS: analysis without anguish: version 10.0 for Windows.* Milton, Qld: John Wiley & Sons Australia, Ltd. 2001.

The HKU Scholars Hub: Unlocking collective intelligence

David T Palmer

University Libraries, The University of Hong Kong, Hong Kong SAR

dtpalmer@hku.hk

Abstract

In 2009 The University of Hong Kong (HKU) wrote mission and vision statements, and strategic plans highlighting Knowledge Exchange (KE). The HKU Scholars Hub, the institutional repository of HKU, was chosen to be the chief vehicle to forge the necessary culture for KE within HKU and engage staff in delivering the desired outcomes of KE. Development work to create this vehicle serendipitously created other desired outcomes. Chief amongst these is a collective knowledge system, created from the interaction between machine and data, author and institution, and, local authority and remote indexing. The result is a bootstrapped "intelligence", greater than the sum of its parts.

Keywords: institutional repository; The University of Hong Kong; knowledge transfer; knowledge exchange; Scopus; ResearcherID; ORCID; Collective Knowledge Systems; Collective Intelligence;

1. Introduction

The research funding body for Hong Kong academic research, The University Grants Committee (UGC), in 2009 charged and funded all of its eight tertiary education institutions to begin programmes for Knowledge Transfer (KT). Their definition of KT:

> the systems and processes by which knowledge including technology, know-how, expertise and skills are transferred between higher education institutions and society, leading to innovative, profitable or economic or social improvements [1].

Upon receiving this UGC directive, The University of Hong Kong (HKU) set several initiatives in motion to uniquely do and show KT, with HKU characteristics. They re-articulated their mission and vision statements, showing three themes; 1) research, 2) teaching and learning, and 3) knowledge exchange (KE) – HKU's local interpretation of KT, emphasizing bilateral exchange between HKU and its community. They created a five year strategic plan [2, 3] based upon these three themes, showing strategic initiatives and operational priorities, two of which read:

- … setting up a database to record knowledge exchange activities in the University and improving communication within and outside the University, we will facilitate dissemination of information and service as an exchange hub.
- … implementing a sustainable web-based expertise directory which draws upon research output, research grant records, contract research, media expertise and community service databases, we will facilitate inbound enquires that seek to identify expertise [4].

Finally, key indicators were assigned to determine the success of HKU's KE initiative [5]. These include:

- Item count of HKU theses in open access (OA)
- Item count of HKU research in OA

- Applications for patents
- Download counts of the above
- Number of staff available for media contact
- Number of collaborative researches
- Number of contract researches
- Number of consultancies, and income thereby generated
- Number of invited public lectures, symposia, exhibitions, performances & honorary degree speeches
- Number of University staff invited to be mentors
- Number of positive media impact related to knowledge transfer coverage, including print, online and electronic media
- Number of appointments of external members to HKU advisory boards

The Hub. The HKU Scholars Hub [6] (The Hub) is the institutional repository (IR) of HKU. It began in 2005 with a mission to collect, preserve, and provide OA to the intellectual output of HKU. The HKU Knowledge Exchange Office (KEO) realized that the goal of The Hub -- OA on HKU research -- largely already aligned with that of KE, and with development could be used as the key vehicle to enable, and show KE at HKU. The Hub could directly measure, and answer several of the key indicators above, and actively promote the increase of several others. The Hub was therefore designated to be an "exchange hub" to make HKU research and expertise highly visible.

With encouragement from the KEO and others on campus, Hub administrators began to plan, how best to make The Hub,

- An expert finder, showing many relevant particulars on each HKU scholar, enabling their discovery by searchers in government, industry and academia
- A supplier of metrics to evaluate these experts and their research
- An "exchange hub" to show and measure all relevant HKU KE activities

2. Methodology

Item-centric. Traditional IRs are item-centric, with metadata records for each separate item. Similar to IRs at other universities, we were collecting the published and grey literature of our academics, and the theses of our postgraduates. Because of a policy requiring thesis deposit, and a program of retrospective digitization, HKU was the first university in Asia to show online theses for all students – 17,000 theses in 2009. However the published and grey literature of our academics proved more difficult to collect – we were ingesting perhaps only 10% of the total.

Metrics in IRs will normally cumulate for each item. In this regard we used APIs from Scopus and Web of Science, to show citation metrics for each Hub item, if also available in Scopus or Web of Science. More article-level linking is planned, such as trackbacks, links to corresponding entries in the blogosphere, etc [7].

Author-centric. Besides IR items in OA, the KE initiative also called for an online locus where particular details of one HKU author could be found. We therefore planned the creation of ResearcherPages (RPs) for each current HKU scholar with these particulars, and metrics that will cumulate to this individual scholar. To build these RPs, fortunately there were several data silos internal to HKU and external, from which we could extract and quickly build a mash-up.

After tendering for this work to be done, we chose Cilea Consorzio Interuniversitario [7] as our partner in development. The first round of enhancements completed in 2009, with others still in progress, and many others still in planning.

On Campus. HKU has long required authors to input data regarding their research output and professional activities; albeit in several disparate and often overlapping databases. Chief amongst these is the Research Output System (ROS), which allows data input directly, but also receives data from the Academic Performance Appraisal (APA), used for performance appraisal, and the Research Grants Committee Administration System (RGCAS), used for applying for, and reporting on, grants given by the Research Grants Committee (RGC) – the research arm of the aforementioned UGC.

We set up a process in 2007 to receive publication data for journal articles, conference papers, and books from ROS. ROS input screens ask for author permission to post to the Hub, and for attachment of the author's manuscript. These items were posted to The Hub, only if publisher and author permission allowed. However with the KE initiative, we began to plan to import all relevant data from ROS, to repost to The Hub. If relevant permissions cannot be obtained, citation data only will be posted.

There are several other sources which we plan to harvest. The following chart shows data to be obtained and from where.

Table 1: Data Elements & Sources (HKU)

Data Element	Source (HKU)
Publications, awards, prizes, etc	ROS
Contact details	Communications Directory
Professional qualifications	Central Personnel Database
Supervision of Research Postgraduate Students	Postgraduate Student Systems
Expertise / Research Interest	APA
Public & Community Service	APA
Research grants received & project undertaken as principal or co-principal investigator	RCGAS
Patents applied for & granted	Technology Transfer Office
Picture	Departmental / personal web pages
Subjects for media comment	Media Content Directory

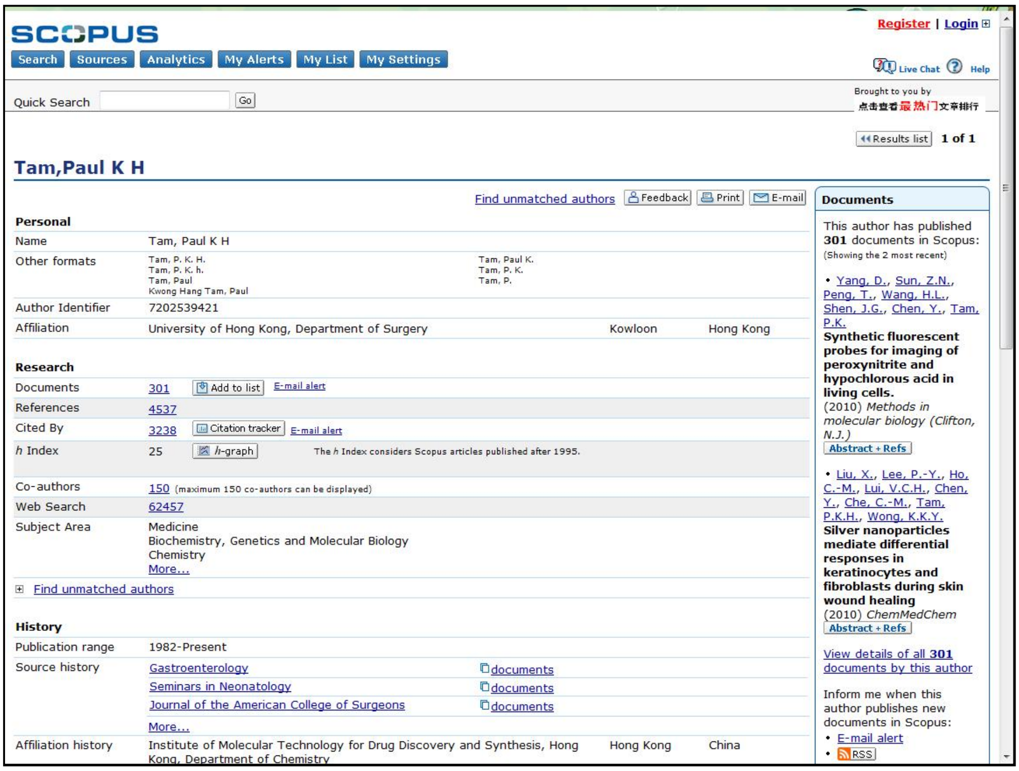
KEO arranged permission for The Hub administrators to extract initially and then for weekly updates. In 2009 administrators set up unattended batch processes to extract data from the Media Content Directory and Communications Directory, and to marshal into Excel files. Cilea created a procedure that would load the data from these Excel files into The Hub. Future work will use similar processes to load data from the other sources.

External to HKU. Working with the various publication lists, it became apparent that sources within HKU and externally all had different subsets of the total output for any one staff; the reasons for which are many and varied. Therefore, it is valuable to show all sources with their different publication lists and citation metrics. The two largest providers for paper counts and citation metrics are Elsevier's Scopus, and Thomson Reuters' Web of Knowledge (TR WoK). Although for certain subjects, other sources are perhaps better, these two provide the widest coverage across the most disciplines.

We have long used each of their APIs to show article level metrics on Hub items, and therefore quickly thought of them to also provide author-cumulated metrics. Before we could begin however, there were several problems to overcome, chief of which is disambiguating names variants and like-named individuals.

Scopus. This database [9] provides a unique AuthorID and page for each author, on which paper counts and metrics will cumulate for this author. However these pages, which are created by machine algorithm, are frequently in error, cumulating two or more like named individuals into one AuthorID, or, creating two or more separate AuthorIDs for one individual when he or she used two or more variants to publish. The affiliation information was frequently in error also because, 1) most of the UGC universities in Hong Kong have very similar names, and, 2) second and subsequent authors frequently showed erroneously, the affiliation of the first author.

Although authors themselves can request these changes, they rarely do. We therefore hired research assistants to search out these problems, and report them to Elsevier. Elsevier has committed to fixing these reported errors and that changes once made will not need to be made again.



Tam,Paul K H

Personal

Name	Tam, Paul K H		
Other formats	Tam, P. K. H. Tam, P. K. h. Tam, Paul Kwong Hang Tam, Paul	Tam, Paul K. Tam, P. K. Tam, P.	
Author Identifier	7202539421		
Affiliation	University of Hong Kong, Department of Surgery	Kowloon	Hong Kong

Research

Documents	301	Add to list E-mail alert	
References	4537		
Cited By	3238	Citation tracker E-mail alert	
h Index	25	*h*-graph	The *h* Index considers Scopus articles published after 1995.
Co-authors	150 (maximum 150 co-authors can be displayed)		
Web Search	62457		
Subject Area	Medicine Biochemistry, Genetics and Molecular Biology Chemistry More...		

Find unmatched authors

History

Publication range	1982-Present		
Source history	Gastroenterology	documents	
	Seminars in Neonatology	documents	
	Journal of the American College of Surgeons	documents	
	More...		
Affiliation history	Institute of Molecular Technology for Drug Discovery and Synthesis, Hong Kong, Department of Chemistry	Hong Kong	China

Documents

This author has published **301** documents in Scopus:
(Showing the 2 most recent)

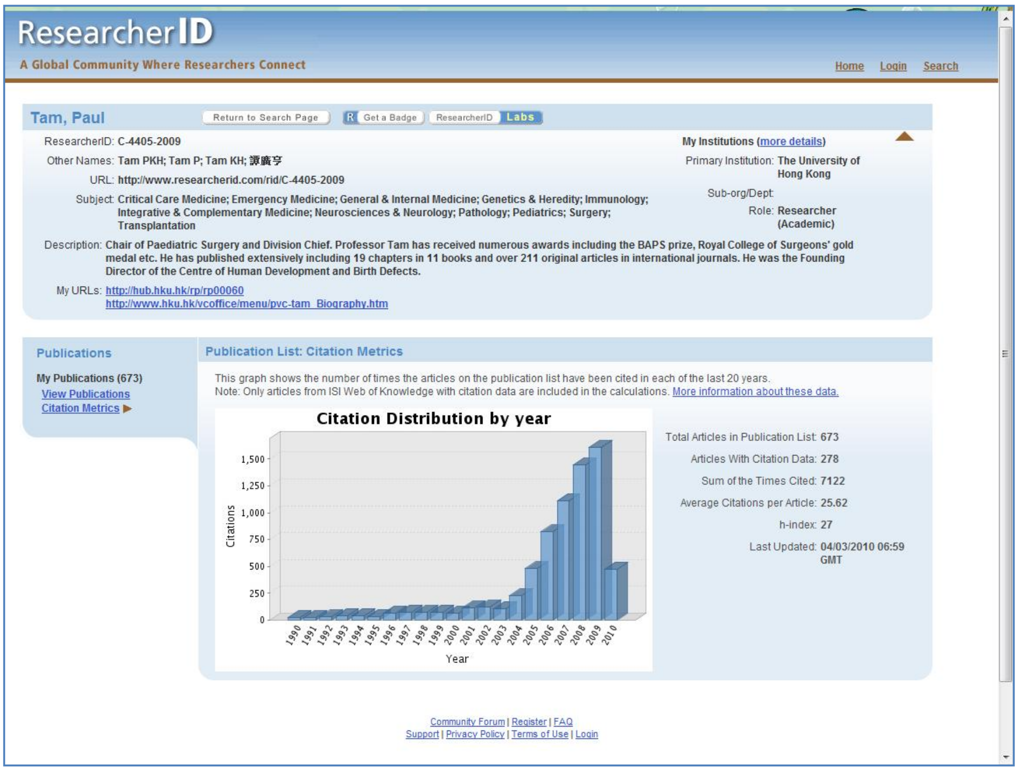
- Yang, D., Sun, Z.N., Peng, T., Wang, H.L., Shen, J.G., Chen, Y., Tam, P.K.
Synthetic fluorescent probes for imaging of peroxynitrite and hypochlorous acid in living cells.
(2010) *Methods in molecular biology (Clifton, N.J.)*
- Liu, X., Lee, P.-Y., Ho, C.-M., Lui, V.C.H., Chen, Y., Che, C.-M., Tam, P.K.H., Wong, K.K.Y.
Silver nanoparticles mediate differential responses in keratinocytes and fibroblasts during skin wound healing
(2010) *ChemMedChem*

View details of all **301** documents by this author

Inform me when this author publishes new documents in Scopus:
- E-mail alert
- RSS

Figure 1. Example of Scopus AuthorID (April 2010)

ResearcherID. We could not find similar procedures to correct data in Web of Knowledge. We then happily learned of Thomson Reuters ResearcherID [10]. Although researchers themselves could create these accounts, they rarely do. We therefore used XML files to create in batch mode, ResearcherID accounts for each of the approximately 1,500 HKU professional staff. We used publication data from the HKU ROS, placed into XML, and uploaded them to these accounts. We then gave the unique ResearcherID and password to each individual researcher, who can now personally edit this information. If the data matched upon entries in WoK, citation metrics from WoK will accrue in real-time to the entry in ResearcherID and cumulate to its author. Using ResearcherID, we generated an “R” badge and HTML code to place on each scholar’s ResearcherPage. MouseOver on this badge will show the author’s top three cited articles. ResearcherID is public, needing no paid subscription to view.



Tam, Paul

ResearcherID: C-4405-2009
Other Names: Tam PKH; Tam P; Tam KH; 譚廣亨
URL: http://www.researcherid.com/rid/C-4405-2009
Subject: Critical Care Medicine; Emergency Medicine; General & Internal Medicine; Genetics & Heredity; Immunology; Integrative & Complementary Medicine; Neurosciences & Neurology; Pathology; Pediatrics; Surgery; Transplantation
Description: Chair of Paediatric Surgery and Division Chief. Professor Tam has received numerous awards including the BAPS prize, Royal College of Surgeons' gold medal etc. He has published extensively including 19 chapters in 11 books and over 211 original articles in international journals. He was the Founding Director of the Centre of Human Development and Birth Defects.
My URLs: http://hub.hku.hk/rp/rp00060
http://www.hku.hk/vcoffice/menu/pvc-tam_Biography.htm

My Institutions (more details)
Primary Institution: The University of Hong Kong
Sub-org/Dept:
Role: Researcher (Academic)

Publications
My Publications (673)
View Publications
Citation Metrics ▶

Publication List: Citation Metrics

This graph shows the number of times the articles on the publication list have been cited in each of the last 20 years.
Note: Only articles from ISI Web of Knowledge with citation data are included in the calculations. More information about these data.

Total Articles in Publication List: 673
Articles With Citation Data: 278
Sum of the Times Cited: 7122
Average Citations per Article: 25.62
h-index: 27
Last Updated: 04/03/2010 06:59 GMT

Community Forum | Register | FAQ
Support | Privacy Policy | Terms of Use | Login

Figure 2. Example of ResearcherID (April 2010)

With both AuthorID and ResearcherID ready for extraction, The Hub administrators used Visual Basic Application (VBA) in Excel to build scripts to extract data from both repositories. Scopus AuthorID and ResearcherID were input into the Excel file, whose VBA script then returned 15 fields of relevant data.

There are several other repositories from which we hope to do similar.

- BiomedExperts
- MathSciNet
- Mathematics Genealogy Project
- ACM Digital Library
- Social Sciences Research Network (SSRN)
- Research Papers in Economics (RePEc)
- Google Scholar

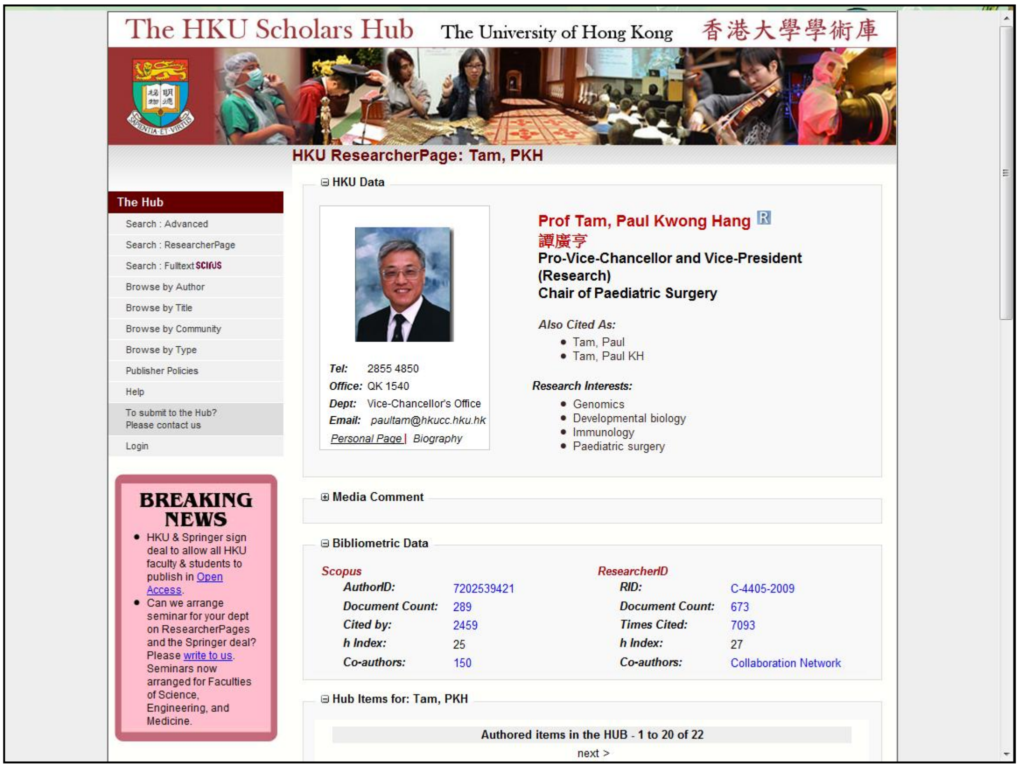
As Google Scholar does not provide an author page, or a way of cumulating citations to one author, we must rely upon the authors themselves to run software such as Publish or Perish [11] on their Google Scholar data. We will build a function for them to input this data themselves, with an accompanying RIS file made from Publish or Perish, into The Hub.

Matching Publications with ResearcherPages. Cilea built procedures for cumulating author name variants to one established name, or ResearcherPage. They then built procedures to make preliminary matches between ResearcherPages and item records whose author names matched the established name heading or any of its cumulated variants. Hub administrators confirmed or rejected these preliminary matches. Authority control and preliminary matches now work on the DSpace Dublin Core element, "dc.contributor", and any of its qualified variants; "dc.contributor.author", dc.contributor.editor", etc.

3. Results

In the initial 2009 round of developments, The Hub administrators, Cilea, and the data providers produced these results,

- TR ResearcherID accounts for each HKU scholar populated with their HKU research publication lists. Thomson Reuters will begin to use ResearcherID numbers in Web of Knowledge entries later this year, to reduce search noise, and to create a more exact method of author name searching. Some researchers use their publicly accessible accounts as online CVs. HKU uses them to harvest Web of Knowledge metrics and display in the Hub.



HKU ResearcherPage: Tam, PKH

⊟ HKU Data

Prof Tam, Paul Kwong Hang
譚廣亨
Pro-Vice-Chancellor and Vice-President (Research)
Chair of Paediatric Surgery

Also Cited As:
- Tam, Paul
- Tam, Paul KH

Tel: 2855 4850
Office: QK 1540
Dept: Vice-Chancellor's Office
Email: *paultam@hkucc.hku.hk*
Personal Page | *Biography*

Research Interests:
- Genomics
- Developmental biology
- Immunology
- Paediatric surgery

⊞ Media Comment

⊟ Bibliometric Data

Scopus		*ResearcherID*	
AuthorID:	7202539421	*RID:*	C-4405-2009
Document Count:	289	*Document Count:*	673
Cited by:	2459	*Times Cited:*	7093
h Index:	25	*h Index:*	27
Co-authors:	150	*Co-authors:*	Collaboration Network

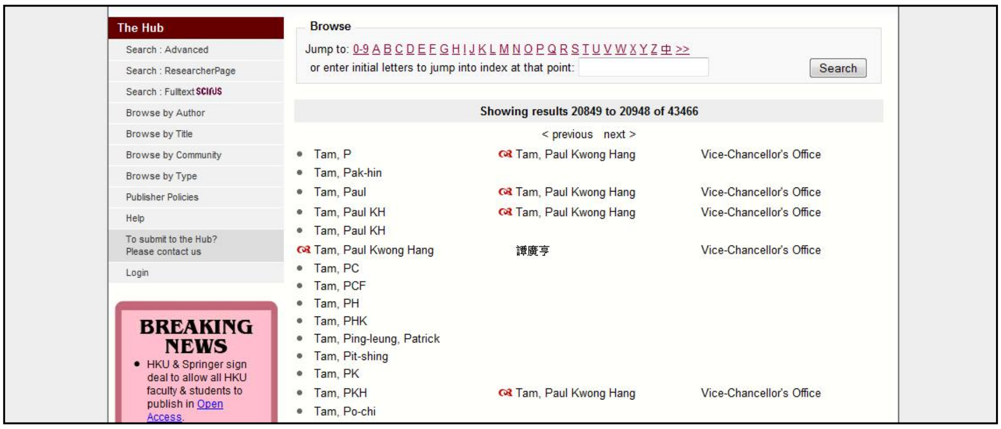
⊟ Hub Items for: Tam, PKH

Authored items in the HUB - 1 to 20 of 22
next >

BREAKING NEWS
- HKU & Springer sign deal to allow all HKU faculty & students to publish in Open Access.
- Can we arrange seminar for your dept on ResearcherPages and the Springer deal? Please write to us. Seminars now arranged for Faculties of Science, Engineering, and Medicine.

Figure 3. Example of a ResearcherPage (April 2010)

- Clean AuthorID records in Scopus. Users of Scopus, search and retrieve on HKU author names with greater accuracy. Government reports and university rankings using Scopus data will show greater accuracy of HKU research output (described below). HKU can harvest cleaner and more accurate data, and display in the Hub.



Showing results 20849 to 20948 of 43466

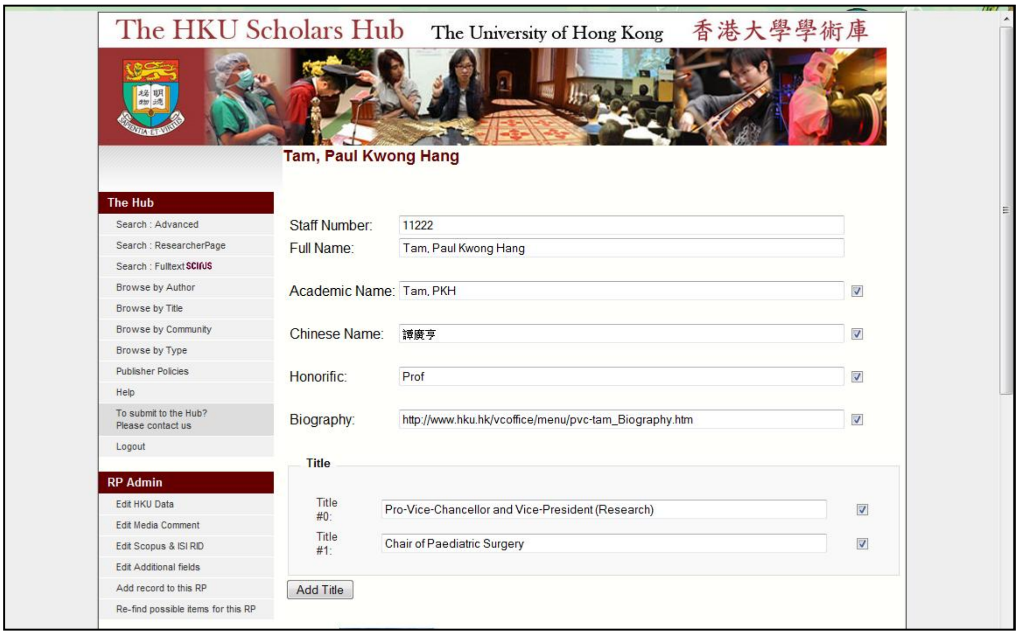
< previous next >

Tam, P	Tam, Paul Kwong Hang	Vice-Chancellor's Office
Tam, Pak-hin		
Tam, Paul	Tam, Paul Kwong Hang	Vice-Chancellor's Office
Tam, Paul KH	Tam, Paul Kwong Hang	Vice-Chancellor's Office
Tam, Paul KH		
Tam, Paul Kwong Hang	譚廣亨	Vice-Chancellor's Office
Tam, PC		
Tam, PCF		
Tam, PH		
Tam, PHK		
Tam, Ping-leung, Patrick		
Tam, Pit-shing		
Tam, PK		
Tam, PKH	Tam, Paul Kwong Hang	Vice-Chancellor's Office
Tam, Po-chi		

BREAKING NEWS
- HKU & Springer sign deal to allow all HKU faculty & students to publish in Open Access.

Figure 4. Example of Authority Control. The " GR" denotes established heading.

- HKU ResearcherPages with author-centric details and metrics, and linked publication lists with fulltext articles. Searchers in government, industry, the media, and academia use Google and other search engines to find HKU experts for media comment, contracted research, research collaboration, supervision of graduate students, speaking engagements, etc. RP owners use them for online CVs, publication lists validated by the institution, reputation management, publication list export (explained below), etc. The University uses them to highlight its research talent, and for KE with its community.



Tam, Paul Kwong Hang

Staff Number: 11222
Full Name: Tam, Paul Kwong Hang
Academic Name: Tam, PKH ☑
Chinese Name: 譚廣亨 ☑
Honorific: Prof ☑
Biography: http://www.hku.hk/vcoffice/menu/pvc-tam_Biography.htm ☑

Title
Title #0: Pro-Vice-Chancellor and Vice-President (Research) ☑
Title #1: Chair of Paediatric Surgery ☑

Figure 5. Example of Edit Page for RP Owner

- Authority control indexing to gather variant names together and disambiguate like-named individuals, in Roman scripts and in Chinese, Japanese, and Korean (CJK) scripts. Initiatives such as the Bibliographic Knowledge Network use this HKU established authority control, along with the ResearcherID and Scopus AuthorID (manually matched by HKU staff) to ascertain identity and link to corresponding records in other sources. Readers within HKU and externally, use them as a finding aid for further publications by the same author, past history of the author, co-authors, etc.
- Individual login authentication using the HKU LDAP (CAS). RP owners login to edit or add details, personalizing their individual RP. Data extracted from other sources – Media Content Directory, Scopus, and ResearcherID – must be changed in the source silo, and not The Hub.
- Unique author identifier, to further disambiguate each current HKU author. This number appears in the ResearcherPage URL; for example, http://hub.hku.hk/rp/rp00060. Elsevier and Thomson Reuters in the future will allow this number to be written in the Scopus AuthorID and ResearcherID records, respectively, and thus increase the trust of all three sources of disambiguated identity.
- Procedure for matching Hub items with RPs. Hub administrators examine each potential match before confirming.
- High visibility of HKU research and authors in Google and other search engines.

HKU administrators in KEO have lauded these developments and asked The Hub staff to take a "road show" to each of the departments. Elsevier and Thomson Reuters have been enthusiastic, and each claims that HKU is the first in the world to achieve these results. Scopus and ResearcherID add value to the Hub, and The Hub adds value to Scopus, and ResearcherID, driving traffic to both.

Individual scholars at HKU have, for the most part, also been enthusiastic. Upon seeing their results in The Hub, Scopus and ResearcherID, several have begun to take an active interest in showing their metrics in the best possible light; adding missing citations, variant names used, and asking for corrections. Several have suggested other sources of data, especially good for their discipline. This reputation management done by the individual also enhances the reputation of the institution.

Not all of the results are in. Development continues, and an authoritative survey has yet to be done. However along the way, a few more purposes for this work have appeared.

RAE. The UK and Australia report that bibliometric data from Scopus, WoK, or both will be used in their upcoming research assessment exercises (RAE): Research Excellent Framework (REF), and Excellence in Research for Australia (ERA), respectively. Hong Kong is heavily influenced by both. A subgroup in the UGC is now considering whether bibliometrics will be used in Hong Kong's own RAE, to be done in either 2011 or 2012.

Re-Positioning the Library. A recent report on research assessment in five countries by OCLC highlighted the role of libraries in this process. It noted that in those countries where bibliometrics are central to RAEs, academic libraries and librarians are often pivotal. The author writes:

> ,In terms of information infrastructure, the libraries that are playing a central role in the research assessment process – particularly Australia – are those which have been able to leverage the value of the institutional repository, which is typically managed and populated by librarians [12].
>
> He describes the role of librarians in each of the five countries surveyed. For Denmark, he observed:
>
> There is a general sense that the traditional library business of books on shelves is being consigned to the past and that librarians see their libraries as having an institutional information infrastructure role within the universities [13].

A companion report to the above, gave seven recommendations on how libraries can, "provide a researcher-centered view" [14]. Curating the institution's research output and providing expertise in bibliometrics for RAEs and other purposes are clearly directions in which libraries can move, to their benefit. This will increase their usefulness to researchers and the institution, and correspondingly align libraries with the mission and vision of the hosting institution.

University Rankings. Though many criticize these studies, they have taken on ever more importance in recent years. Parents use them to decide which school to choose for their child. Governments use them to distribute research and education funding, etc. Research metrics from Scopus, WoK, or both play a large part in these rankings. HKU is taking charge of its own reputation, and applying resources to ensure proper accounting of its research output.

ORCID. A new worldwide initiative was announced in December 2009, the Open Researcher & Contributor ID (ORCID) [15]. Members include Elsevier, Thomson Reuters, major publishers, and large universities. The ORCID will be based upon, or perhaps use, the ResearcherID, and will be operational in June 2010. Authors will use the ORCID when submitting articles to publishers. Publishers will record the ORCID in the metadata for each article, and pass to third parties such as Scopus and WoK. Therefore institutions, publishers, and database managers will finally have a way to disambiguate authors and assign unambiguous identity. At this time, HKU is the only institution in the world to have ResearcherIDs for all of its authors. In June 2010, we expect to announce to our HKU authors, that they must begin using the ResearcherID / ORCID to submit articles to publishers, record their publications in the HKU ROS, place in their CVs, etc. With each member already having a ResearcherID, we expect almost full, and immediate compliance. This will finally consign problems of HKU author name ambiguity to history,.

Publicaton List Export. Although only minimal publication lists are now available in The Hub, authors still could envision and suggest to us, a use very relevant for them. We will build a procedure for authors and their readers to select on publications, and export them in their desired format; RIS, EndNote, CSV, etc. HKU scholars are presented with dozens of requests for publication lists during the year for various purposes; grant application, conference papers, postgraduate student supervision, etc. The present HKU Research & Scholarship pages do not allow this. In future work, publication lists in The Hub will be made complete with each scholar's full HKU record.

4. Discussion

Traditional IRs are item-centric, and done for the purpose of OA. Most suffer low population rates. The Hub has luckily enjoyed the attention of HKU's policy on Knowledge Exchange, which has meant a change in focus and alignment. The Hub has therefore grown beyond its initial scope as an institutional repository. Besides cumulating data around the item, it now also does the same around authors. It serves other purposes besides that of OA. Our primary goal is to create a system that will forge the necessary culture for KE within HKU and engage staff in delivering the desired outcomes of KE.

However, whether done for OA or for KE, the results are much the same in many cases [16]. Work done for either OA or KE mutually contribute to each other. Recent months have seen these developments in OA at HKU:

- The Vice-Chancellor of HKU signed the Berlin Declaration on Open Access in November 2009 [17].
- With funding again from HKU KEO, HKU Libraries came to agreement in March 2010 with Springer, to allow HKU faculty and students to publish using Springer's Open Choice [18] option for one year. All such articles will be open access, and posted to SpringerLink, The Hub, and relevant ones to PubMed Central.
- In February 2010, HKU Libraries created a mandate for its staff to deposit authored items in The Hub [19].

E-Science. "E-science" has been defined as, "shorthand for the set of tools and technologies required to support collaborative, networked science" [20]. Although this was not a consideration in our initial planning, The Hub has indeed become, serendipitously, such a tool.

The Hub is now a unique locus for HKU researchers to interact with the web, and for remote services to interact with HKU researcher data. Initially the individual data was supplied from various HKU and remote data silos. HKU researchers then edit, delete, or extend this data. This data is then exposed to remote web services, which may also enhance this data, for greater value to the individual and his or her institution. An example of a remote web service doing this, is the Bibliographic Knowledge Network People (BKNpeople) hosted by UC Berkeley, and funded by the (US) National Science Foundation (NSF) Cyber-enabled Discovery and Innovation (CDI) Program. Still in an experimental stage, BKNpeople displays HKU data, with links to corresponding records held by other data suppliers, such as MatchSciNet and the Mathematics Genealogy Project [21].

Collective Intelligence. In The Hub paradigm, the institution loads relevant data to create ResearcherPages, which RP owners can then edit and otherwise control for their own purposes.

This symbiosis between machine and data, author and his or her institution, local authority and remote indexing, creates a "collective knowledge system". Tom Gruber's well cited article on Collective Knowledge Systems, argues that these systems, "… unlock the 'collective intelligence' of the Social Web with knowledge representation and reasoning techniques of the Semantic Web" [20]. Other descriptors for this process are "synergy", and "bootstrapping". Well known examples are of course, Wikipedia, Facebook, etc. An interesting observation by Chris Dixon of Hunch, writes, "I think you could make a strong argument that the most important technologies developed over the last decade are a set of systems that are sometimes called, 'collective knowledge systems'" [23].

The last player in this symbiosis is Google, in which RPs are highly visible and discoverable. Searching on the interlinked documents of the planet, Google's page ranking, "provide[s] a very efficient system for surfacing the smartest thoughts on almost any topic from almost any person" [24]. Because of tagging done by machine loads, and manually by the RP owner, on any relevant Google search, RP pages are at, or near the top of the hit list.

Authority Control. An example of this symbiosis is the authority control in The Hub. The traditional paradigm has been a central or national library maintaining an authority file, to which remote libraries can add; for example, the US Library of Congress and its member NACO libraries. However, The Hub has added a third, and perhaps more important player; the author who is the subject of this authority work. The Hub begins with a full name, an academic shortened name, and a Chinese name extracted from the HKU Registry's files. Variant names are loaded from Scopus. Hub administrators can edit or add more. Finally the owning author can also edit or add more. Each of these parties has incentive to create and maintain an accurate record, with perhaps the author holding the most incentive. Once this record of name, name variants, and publication list is created, it is valuable to many researchers and web services within HKU and beyond.

In this regard, a draft report by the Working Group on the Future of Bibliographic Control to the Library of Congress, which included librarians and representatives from Google and Microsoft stated:

> The future of bibliographic control will be collaborative, decentralized, international in scope, and Web-based. Its realization will occur in cooperation with the private sector, and with the active collaboration of library users. Data will be gathered from multiple sources; change will happen quickly; and bibliographic control will be dynamic, not static [25].

Indeed the authority control exhibited by The Hub on its limited data set of HKU authors, appears to be in the vanguard of what the Bibliographic Knowledge Network project calls a,

> *bibliographic revolution*, whereby responsibility for bibliographic control (the organizing and cataloguing of metadata associated with publications) will shift from centralized agents, such as the Library of Congress, OCLC and large abstracting and indexing services, to an aggregation of many smaller [virtual organizations] which will contribute discipline-specific expertise on a collaborative basis [26].

Linked Data. Although we have seen many purposes in which The Hub can serve, there will be many more, as yet unintended. In the concept of "linked data", data once identified can be re-purposed by many other players. Data in The Hub now carries the imprimatur of HKU authority. This data can be used in many future mash-ups, at HKU and beyond, whose purpose is as yet unknown. Future Hub development will produce APIs or widgets for the purpose of extracting Hub data.

5. Conclusion

The 2009 development to re-purpose The Hub into a vehicle for KE has produced favourable results, and more besides; 1) infrastructure that can be used for RAE, ORCID, etc., 2) a method for the Library to re-position itself with its institution, and 3) an e-science tool, or a collective knowledge system. This latter is slowly beginning to be understood, and used to unlock collective intelligence within HKU, and beyond. It presents a great challenge, and perhaps the area of most reward; how best to extract from the several sources (author included) and structure it in a way that invites interaction with all partners, for present purposes, and for those as yet unknown?

Notes and References

[1] University Grants Committee. "Knowledge Transfer". Available at http://www.ugc.edu.hk/eng/ugc/activity/kt/kt.htm (April 2010).

[2] The University of Hong Kong. "2009 – 2014 Strategic Development". Available at http://www3.hku.hk/strategic-development/eng/index.php (March 2010)

[3] The University of Hong Kong. "Promoting Knowledge Exchange and Demonstrating Leadership in Communities across the Region". Available at http://www3.hku.hk/strategic-development/eng/strategic-themes-for-09-14/promotion-knowledge-exchange-and-demonstrating-leadership.php (March 2010)

[4] Ibid.

[5] Knowledge Exchange Office, The University of Hong Kong. "Recurrent Funding for Knowledge Transfer Activities in the 2009/10 to 2011/12 Triennium, Initial Statement", The University of Hong Kong, Hong Kong, June 2009.

[6] The University of Hong Kong Libraries. "The HKU Scholars Hub". Available at http://hub.hku.hk (April 2010).

[7] Binfield, P. "Plos One: Background, Future Development, and Article-Level Metrics", Electronic Publishing 2009, Milan, 10-12 June 2009. Available at http://conferences.aepic.it/index.php/elpub/elpub2009/paper/view/114 (April 2010).

[8] Cilea Consorzio Interuniversitario. Available at http://www.cilea.it/ (April 2010).

[9] Elsevier. "Scopus". Available at http://www.scopus.com (April 2010)

[10] Thomson Reuters. "ResearcherID". Available at http://www.researcherid.com (April 2010)

[11] Publish or Perish. Available at http://www.harzing.com/index.htm (April 2010).

[12] Key Perspectives Ltd. "A comparative Review of Research Assessment Regimes in Five Countries and the Role of Libraries in the Research Assessment Process" (Report commissioned by OCLC Research), 2009. Available at http://www.oclc.org/research/publications/library/2009/2009-09.pdf (April 2010).

[13] Ibid.

[14] MacColl, J. "Research Assessment and the Role of the Library" (Report produced by OCLC Research), 2010. Available at http://www.oclc.org/research/publications/library/2010/2010-01.pdf (April 2010).

[15] "Open Researcher & Contributor ID". Available at http://www.orcid.org/ (April 2010).

[16] The Canadian Institutes of Health Research (CIHR) used their policy on "Knowledge Translation", similar to HKU's KE, to create a mandate for OA deposit of their grantees published literature into PubMed Central. "Knowledge Translation at CIHR", 2007. Available at http://www.cihr-irsc.gc.ca/e/35412.html (April 2010). "Policy on Access to Research Outputs", 2007. Available at http://www.cihr-irsc.gc.ca/e/34846.html (April 2010).

[17] The Max Planck Society for the Advancement of Science". "Berlin Declaration on Open Access, Signatories", Available at http://oa.mpg.de/openaccess-berlin/signatories.html (April 2010).

[18] Springer. “Springer Open Choice @ University of Hong Kong”. Available at http://www.springer.com/HKUAuthors (April 2010).

[19] The University of Hong Kong Libraries. “The HKU Libraries Open Access Policy”. Available at http://hub.hku.hk/local/oaPolicy.jsp (April 2010)

[20] Hey T, Hey J. “E-science and its implications for the library community”, *Library Hi Tech*, 2006; 24(4): 515–28. Available at http://conference.ub.uni-bielefeld.de/2006/proceedings/heyhey_final_web.pdf (April 2010).

[21] Bibliographic Knowledge Network. “BKNpeople”. http://people.bibkn.org/ (March 2010).

[22] Gruber, T. “Collective Knowledge Systems: Where the Social Web meets the Semantic Web”, *Journal of Web Semantics*, 2007. Available at http://tomgruber.org/writing/collective-knowledge-systems.htm (April 2010)

[23] “Chris Dixon’s Blog”. 17 January 2010. Available at http://cdixon.org/category/hunch/ (April 2010).

[24] Ibid.

[25] “Working Group on the Future of Bibliographic Control”. Draft Report to the Library of Congress. November 30, 2007. Available at http://www.loc.gov/bibliographic-future/news/draft-report.html (April 2010).

[26] Pitman, J; et al. “Bibliographic Knowledge Network” (Proposal submitted to the NSF Cyber-enabled Discovery and Innovation Program), 2008. Available at http://www.stat.berkeley.edu/users/pitman/bkn-proposal.pdf (April 2010).

Social networks and the national art gallery (Dublin |…|Sofia)

Mícheál Mac an Airchinnigh; Glenn Strong

School of Computer Science and Statistics
University of Dublin, Trinity College. Dublin, Ireland
{mmaa, Glenn.Strong}@cs.tcd.ie

Abstract

To publish is to make public. And one sense of being public is surely to be accessible? Today it is not only the writing and the images that are published formally, that is to say through official channels, but also the casual human artefacts, the chat, the blog, the quick pic, the self-made music and dance and film, and all of the latter through the medium of the social network. In the World-Wide Web (WWW), to be published is to have a unique resource identifier (URI) and usually a unique resource locator (URL). But to be visibly published on the WWW one needs to be found (much in the same way that one might be found say, 200 years ago, through the library catalogue). Hence at the very core of electronic publishing is to be found the metadata nucleus. In olden times the scholar/reader would have to travel to that place, the Library, if it were accessible, to read/study the work. Today, (s)he travels electronically to those places which are accessible. E-publication does not necessarily entail accessibility. For example, many scholarly works are behind pay walls, costs are borne by institutions of would-be accessors; someone has to pay for maintenance, security, and accessibility. Works of art are in a peculiar and particular category. A work of art is considered to be unique, by which one understands that there is no other copy, properly understood. There may be thousands of prints of the unique piece authorised. But the digitization of an artwork forces a categorical change. The digital artwork is, by nature different. It can be seen, not by reflected light but by transmitted see-through light! In this specific regard it is completely other vis-à-vis the book qua text. In this paper we consider the typical state of the "digital art" as e-publication and explore the extent to which such art is freely accessible to the public, whether on social network or otherwise, with respect to four chosen "National Art Galleries" on the circumference of the European Union.

Keywords: access; art gallery; augmented reality; el-pub; folksonomy.

1. Introduction

Let us imagine a time of scarcity of trees, of materials that might be used to make paper? Let us imagine a time in the not too distant future when all publication is necessarily electronic? Let us imagine a time when great paintings will be electronically freed from their museums? Let us imagine… ?

> «The world has changed recently, yet again, in January 2010. The Guardian Newspaper [1], famous for its establishment of an online presence [2], that was distinctly different from its physical newspaper print presence, abandoned its Technology Insert that always appeared every Thursday. There is, of course, the now much richer web presence that provides the Technology News, all the time, around the clock, independently of the newsprint presses.
>
> Today, "Google puts off launch of mobile phone in China" [3], and yesterday, "Apple confirms date for its 'event': we know it's a tablet, but what else?" [4].
>
> At a slower pace one can read the Technology Editor's blog [5]. Want to keep up to date? Then get the tweets [6].
>
> mihalorela ElPub2010 has not yet happened; signed up to learn to tweet for it before 16-18 June, Helsinki. Maybe will have iTablet with me... then? :)» [Mihal Orela 2010-01-19].

The foregoing extended block quotation is a conceived, imagined, *mashup* [7] of text from Charles Arthur, editor of the online Guardian Technology section, and a related *tweet* on the same date by a *follower* of Charles Arthur commenting on the possibility that Apple's January 2010 event might just be related to the (un)expected iTablet aka iSlate. On the 19th of November 2009, Charles Arthur gave advance warning of this revolution in the making:

> "What you are holding in your hands — assuming you're reading this in print form, which a substantial number of you are — is a collector's item. Guardian Technology, in its print incarnation, is to cease publication. The last edition will be on 17 December" [8].

Since the date of the *mashup* [7] one now knows that the mooted "Apple device" for e-publications (such as newspapers) has the simple name of iPad.

In this paper (destined in context to become an e-paper) we shall present state of the art "electronic publication" with respect to the Fine Arts, using illustrations/examples from a "National Art Gallery" (NAG) at the "extremities" of Europe: to the west, the National Gallery of Ireland (NGI), Dublin [9], to the east, "Национална художествена галерия" (National Art Gallery), Sofia [10], to the north, Valtion Taidemuseo, (we focus on the Ateneum), Helsinki [11], and to the south, Museo Nacional del Prado, Madrid [12], thus avoiding the "usual suspects" in Art discourse. Before proceeding let us mention our use of the buzz-word. For "electronic publication" we will use the abbreviation el-pub (or elpub). Naturally it corresponds with the name of the conference series. But more importantly, it has significant ambiguity. In other words, el-pub is multi-referent on the Web. Specifically, due to the treatment of non-letters, el-pub may be interpreted as "el público" in Spanish; Pub in many languages is taken to be a drinking establishment, and extracted from the more formal English name "Public House". Similarly, we introduce here the abbreviation soc-net (socnet) for social network. A quick search will show just how "popular" and ambiguous this buzz-word is.

The deliberate focus in this paper, is Art in the classical and traditional sense. In particular, we include photos of paintings, sketches, drawings and photos of sculptures; we exclude photos per se. "Art is notoriously hard to talk about" [13] and if it is hard to talk about it, or even to write about art, naïvely (i.e., not formal critical discourse), then we may pose a basic research question. Is it harder or easier to ontologize the art rather than to talk about it? And having ontologized it, how easy or difficult is it for the machines to make sense of the ontologization? Let us make a first pass to test this hypothesis by 1) restricting ourselves to the Dublin Core 15 tag elements [14] and 2) exploiting folksomonic tagging such as used in Flickr. Furthermore let us use a simple tool, DCdot [15], to extract the Dublin Core metadata and present it in a readable fashion. Our research will show just how little has been accomplished in just over 15 years [16]. However, our main focus in this paper will be elsewhere: on the electronic access to the Art, whether *in situ* or on the Web.

With respect to Art in situ, we note the potential for wireless devices to be used to inform the "visitor" [17] to an art gallery/museum, whether the technology be classical RFID [18] or NFC [19]. An Art object appropriately tagged in its immediate environment would become an entity within the "Internet of Things" [20]. By environment we mean primarily, for example, that for a painting in a given physical setting, there would be at least 4 wireless tags on: the canvas, the frame, the wall labeling (etiquette), and the wall itself. Use of a wireless device such as a mobile phone would facilitate reading of the wall label in one's own language, for example.

One year later after the formal submission on the "Internet of Things" to an EU "request for response", we found that there was a company called Plink [21] which released an Android app [22] called "PlinkArt" and which works just as we have theorized. They have a server-side database with around 50,000 works and have plans to pilot the app as a replacement for the "clunky tape-recorder audio tours". They also have a developer API coming, they say, that allows other apps to link in to their server-based recognition engine. Plink appears to be the output of the PhD thesis in machine vision of the two lead developers, Mark Cummins and James Philbin. A version of PlinkArt for the iPhone is expected soon (perhaps in time for the Conference itself in June). And (perhaps no surprise) Google itself subsequently released a similar sort of app with the appealing title of "Google Goggles" [23] for the Android phone, which brings us right back to the issue of "Google and China and the Google phone world" with which we started. Finally, a surprise, just as we went to press Nokia announced its own lookalike app by the name of "Augmented Reality" with features, in some respect, similar to those of Google Goggles [24]. Let us now turn our attention to real electronic publishing within the established formal Art world.

2. Methodology

How can one know something of the effectiveness of the use being made of an electronic publication (el-pub) within a social network (soc-net)? And how might one distinguish between effectiveness and simple popularity? By effectiveness we mean the taking root of the el-pub within the distributed community. For example, the circulation within a soc-net to certain trusted Wikipedia (WP) pages would be a strong indicator of effectiveness. The primary soc-net of Wikipedia itself consists of the registered editors. The first author is a member of this WP soc-net and belongs to "Wikipedia language communities" in English, Bulgarian, French, German, and Irish, meaning he edits pages in these languages.

Our research methodology is characterized quite succinctly by the well-known phrase:

"By indirections find directions out" (Act II, Scene 1, Hamlet)

This is a theatrical or artistic way of explaining that we belong to the great methodological school before the "time of the separation of the arts and sciences"[25]. Or… to put it differently we are here dedicated to reunite the "Sciences" with the human reality of the experience of most humans, the people, the non-digitally connected, the people who feel at home with the… arts, with the feelings of life.

<u>Scenario</u>: — In keeping with the general research strategy in the domain of the digital re-discovery of culture [26] ·(DrDC), one works outwards from a grounded scenario (a playlet, in other words) which consists of a short one

page backstory… For example, one might first come across the art work of Hieronymus Bosch: “The Garden of Earthly Delights” as an illustration in an art book, such as “How to Read a Painting” [27], or on the Web through Wikipedia [28], or directly through the web presence of the holding gallery, “Museo Nacional del Prado” [12], or even courtesy of Google [29]. The visual impact of the art work usually provokes a desire in the viewer to read up on the background and to ask oneself fundamental questions: who is the artist? Why did (s)he paint it? When was it done? For whom was it done? Where is it now? And so on. In the context of the art book cited above, many of these questions are answered. The backstory is given on two facing pages. In the English version of the text (the original was Dutch) there are 4 key (words/phrases) marked out in bold font and which we list here in order top to bottom, left to right: “image of paradise or a world of debauchery”, “union of Adam and Eve”, “paradise”, “musical intruments”. These key phrases or tags clearly belong to a folksonomy. They are formally recorded in the Index of the book in order for the reader to see/lookup the “persons, themes, or motifs referred (or alluded) to in the titles of the illustrated works” p.369.

One of the most significant features of the “old-fashioned” art book is simply that high quality images with accompanying erudite text opened a door to another world. In the case of the text and example cited, there was the added suprise of accounts of two more triptychs by the same painter: “The Temptation of St Anthony” (key words/phrases: “Anthony”, “trio”, “kneeling hermit”, “naked woman”) p.96-7 and “The Haywain” (key words/phrases: “Adam and Eve”, “central panel”, “on top of the cart”, “risen Christ”, “pilgrim”) p.99. Today, the art enthusiast of the works of Hieronymus Bosch will find a (complete?) list of his paintings on Wikipedia [30]. In the context of the electronic publication of artworks and associated commentary and folksonomy within the social network community in 2010 it will not be surprising that we rely on Wikipedia as a substantive part of our research methodology.

<u>Wikipedia</u> ·(EN and at least one other language): — Searching for and finding relevant information is a difficult task in any medium. For books, the solution was/is the provision of an Index (in addition to any front matter such as chapter and/or section headings). For the Web (aka Internet) it is the chosen search engine. Where once one relied on the encyclopedia (English, French, German,…) for terse erudite scholarly information, today one is more likely to go first to Wikipedia. The soundness or otherwise of Wikipedia is not the issue. In a Social Network context, it has proven itself to be an el-pub resource consulted by (very big number of?) people every day, and not only in English.

For example, in our context of National Gallery [31], there is a page for the National Gallery of Ireland (NGI), in English, as expected; nothing in Irish; the other languages, for which there is a description, are: Català (stub) Español, Esperanto (stub), Français (stub), Italiano (stub), Nederlands (stub), Русский. Those pages which are "inadequate" with respect to content are marked here as (stub). Excluding stubs, we conclude that the NGI has significant presence in 2 languages other than English. There are certain measures available by which one might wish to judge the page. Details are given on Wikipedia. For comparison El Prado has a page in (roughly) 39 languages.

Dublin core metadata: — There are 15 key tags that one might want to use for e-publications, whatever the nature of these latter might be. Our research has already shown that in the category of Newspapers online, very few of this basic set of 15 metadata tags are ever used. Notice we speak of "newspapers online" and not "online newspapers". It seems to be the current paradigm that newspapers "go online" while retaining their existing print production form. It is our considered opinion that the burden to produce the Dublin Core metadata for each section of each issue is too great either from a commercial or a practical point of view. An alternative hypothesis might be, simply, that nothing out there makes use of such Dublin Core metadata when it comes to search, to access! The trend has been for news aggregators to come to the fore, the greatest of which is (still) Google.

Folksonomy: — Instead of the exploitation of the use of formal metadata tags, such as is used in the Dublin Core, it appears to be the case that informal folksy tagging has predominated. This is tagging by social networkers. From the point of view of Art, the social networker's folksy description of a particular piece will be dramatically different from the description given by that one with the "trained eye to see." One of the key texts that developed this "Ways of Seeing" of Art was produced as a TV programme on the BBC [32], episodes of which are publicly available on YouTube [33].

Augmented reality personal devices: — first was the iPhone which was not a phone at all! The basic character/nature/role of a "mobile phone" was changed. There were earlier intimations of what was to come; one of the key additions was the camera. Who would have imagined a telephone that would take pictures? And yet looking back at the initial beginnings of the mobile phone, we now realize that few if any predicted that the main purpose of the phone would be to use a certain redundancy in the mobile signalling, to wit the coming into being of "text messaging". In 2010, another revolution is taking place — the camera on the phone is connecting with pictures of the world and augmenting the touch/display screen with another layer of hypertext reality.

Given the nature of the iPhone (app) it is no longer surprising to note the kinds of marriage between the el-pub and the soc-net. In the context of ElPub 2010 we note the ART app [34] which provides images of paintings in the public domain and supplies corresponding biographies of artists, incorporating information from Wikipedia, using the Creative Commons Attribution-shareAlike License. One has access, in the palm of one's hand, to the art of Vincent van Gogh, Hieronymous Bosch,… There are currently "204 artists" represented. (To a certain extent it appears that one may remove or add artists from the list at will). A closer examination shows that although Marc Chagall (1887-1985) is listed, there are no paintings in the gallery. One is invited to "add them oneself" using "this button in browser to save images to your gallery", with suggested links to 1) Wikipedia.org, 2) Artst.org, 3) ABCGallery.com, and 4) ShowMeArt.info. A good social network test might be to obtain some "freely available" Chagall images of painting, add them to one's own gallery, and then see if they become available to others with the same app, either automatically or upon request.

3. The Experiments and the results

We now present some of the details of the 4 experiments that we carried out. Each experiment is introduced with a different type of backstory. For the National Gallery of Ireland the backstory is built around a specific Exhibition of Finnish art, deliberately chosen in order to ground the paper with respect to the location and culture of the hosting city of the ElPub 2010 conference—Helsinki. For the National Art Gallery in Sofia our backstory ties the Art of the Slav to the Language of the Slav, and so also to the tradition of the Byzantine Art. It seemed natural that Google Earth would play the major role in the backstory for El Prado in Madrid. Lastly, to complete the "grand tour", we return to the Ateneum Museum in Helsinki where we needed to construct a backstory that would provide closure for our work and at the same time provide breakout for further experimentation and research into the state of the other National Art Galleries (currently listed on Wikipedia in 2010).

Scenario 1: The National Gallery of Ireland (Dublin, Ireland).

Backstory: From 8th November 2008 until 1st February 2009, there was an exhibition of *Finnish* Art, organized in collaboration with the Ateneum Art

Museum, Helsinki [11], entitled "Northern Stars and Southern Lights: The Golden Age of Finnish Art 1870-1920" [35]. On p.53 of the published catalogue one can see a photograph of the painting *Virginie* (1883) by Albert Edelfelt, Cat. 28. The Catalogue number is an index to p.115 where further information on the painting may be read:

> "Signed and dated: A. EDELFELT/PARIS 83;
> Oil on canvas, 73.5 x 92.5 cm;
> Joensuu Art Museum, Joensuu, JTM 71;
> Bequest 1962, Arla Cederberg collection.

A google search for "Albert Edelfelt" will lead to a brief biography [36, 37] and the art enthusiast will eventually obtain some idea of what the actual painting looks like [38], in this case found on Flickr.

The story told is that of the classical tale in the context of Art. Specifically, physically presence, both of the painting and the observer, are required. This is the way it always has been. If the (modern) Art Gallery has "easy" access to "professional" publication facilities and sufficient resources then a record, a book with full colour plates, can be produced and sold to the interested art viewer. The National Gallery of Ireland excels in this way.

However, in the context of Social Networks, one now needs to examine to what extent said Gallery and Collections have virtual presence. In other words, what is the current state of its electronic publications? On a scale 1 to 10, the Gallery gets 1. To "see" what images are available is practically non-existent at present. We are aware that extensive "computerization of the Art Work" is underway , since at least two years and it would not surprise us were the job to be completed by the time ElPub 2010 unfolds in June. On the other hand the truly persistent Networker with a passion for art will note that "In September 2010, the National Gallery will present an exhibition celebrating the Dutch seventeenth-century artist Gabriel Metsu (1629-1667) and his exquisite scenes of daily life, which rank among the finest of the Dutch Golden Age. It will bring together some 40 of his paintings and drawings from public and private collections around the world. An accompanying catalogue will be published, edited by Dr Adriaan Waiboer, NGI curator of the exhibition and author of the catalogue raisonée on Metsu. Following its showing at the National Gallery of Ireland, the exhibition will travel to the Rijksmuseum, Amsterdam and on to the National Gallery of Art, Washington."

There are two thumbnail images given: 1) Man writing a Letter, c.1664-1666, 2) Woman reading a Letter, c.1664-1666. For the record (2010-04-08), Dublin Core subject metadata is "Press Release; Exhibitions; Johann Zoffany (1733-1810; 13 MARCH - 25 JULY 2010; Taispeántais; Pierre Bonnard

(1867-1947; FORTHCOMING EXHIBITIONS; Talks & Tours; Gabriel Metsu 1629-1667 4 SEPTEMBER - 5 DECEMBER 2010; Acquisitions 2000-2010; Exhibition Catalogue; Roderic O'Conor (1860-1940; Taking Stock; Gabriel Metsu (1629-1667"

The Social Networker will quickly find that there is a website dedicated to Gabriel Metsu [39] on which there are reasonably high-quality images of 33 of his paintings. It is from this website we learn that the two paintings with thumbnails shown on the NGI site belong to the NGI Collection. Now the issue for our e-times becomes the nature of the quality and source of the "digital images" and their accessibility whether on a large high quality display computer or a small high quality mobile device.

Scenario 2: The National Art Gallery (Sofia, Bulgaria).

Backstory: "Orthodox painting has its own peculiar language… (today) impenetrable to the understanding of the worshipper as well as to the common spectator… this subordinate function of the landscape only characterizes the starting of its understanding." [40]

We would like to illustrate one aspect of this concept of the "landscape" in the Orthodox painting. We choose an icon from the late 16th century (originally from Nessebar) and now in "The Crypt" of the National Art Gallery Sofia [41]. A full color plate is available [40, 42 Icon 50]. The landscape aspect in question is the "architectural detail of Jerusalem." Unfortunately, although there does not appear to be any picture of the Icon publicly available, there is a photo of the said architectural detail [43].

A detailled analysis of the current Web site of the National Art Gallery, Sofia (NAGS) [44] will reveal a considerable amount of inconsistency between the pages in Bulgarian and the "corresponding" pages in English. Indeed it is only in the last six months or so that 65 images of paintings from its collection became officially publicly available. These can now be seen also on the Social Network Flickr [45]. There is considerable difference between the ways in which the paintings are presented on the official web site and on Flickr. In the latter, the image occurs once (uniquely) with information given in both Bulgarian and English. In addition there is a link back to the two sources on the official site. In contrast, the official Web site is divided linguistically (Bulgarian and English), clearly a significant technological failing. (At the time of writing (2010 April 8), the picture [46] is upside down at http://www.nationalartgallerybg.org/index.php?l=55&id=43).

On February 10, 2010 the Bulgarian Cabinet announced a major re-alignment of the "Art Museums" of Sofia: "Cabinet approved a proposal presented by the Culture Ministry for four metropolitan museums, a ministry media statement said on February 10 2010." [47]. In the light of this information, it is difficult to foresee and assess the nature of forthcoming National Gallery el-pubs. On the other hand, the 65 images which are now on Flickr may give rise to interesting mashups of all kinds.

Scenario 3: Museo Nacional del Prado (Madrid, Spain)

Backstory: One of the most interesting surprises of 2009 was the announcement that anyone could use Google Earth to travel to El Prado (There are two places in Google Earth with such a name.) and see 14 of its paintings in exquisite detail [29]. This set of 14 was a subset of the 15 images already available online [48]. The missing image is readily explained. It is a photograph of a sculpture "Offering by Orestes and Pylades (San Ildefonso Group)", not a painting [49].

El Prado avails of the Social Networks: Facebook and Twitter [50]. In other words, "Social Networks" is an official and explicit part of the Museum's presence online. The first author is signed up on both. The language is unsurprisingly, Spanish.

The DC data for the˜ sculpture page [49] is extensive.

The DC Subject or keywords: Museo del Prado; Prado; Museo; Madrid; España; Spain; Velázquez; Goya; Tiziano; Rubens; Juan de Flandes; El Greco; Ribera; Fra Angelico; Rafael; Tiepolo; van der Weyden; el Bosco; Meninas; la Crucifixión; el caballero de la mano en el pecho; el sueño de Jacob; el tres de mayo de 1808; el 3 de mayo de 1808; los fusilamientos en la montaña del Príncipe Pío; la Anunciación; el Cardenal; el emperador Carlos V a caballo en Mühlberg; Inmaculada Concepción; el Descendimiento; el Jardín de las Delicias; las Tres Gracias; Artemisa; ofrenda de Orestes y Pílades; el arte de educar; tienda prado; holandeses en el prado; la obra invitada; Richard Hamilton; Las hijas de Edward Darley Boit; John Singer Sargent. ˙Whether or not this choice of DC Subject/keywords is appropriate for this page is a matter of judgment; our judgment is NO! In other words, it is clear to us that the Dublin Core data used is generic! It is applicable to the entire web site; it is not specific to the web page!

The DC Description data: ágina web oficial del Museo Nacional del Prado (Madrid, España). Información sobre visita al museo, obras maestras, colección, exposiciones, actividades, educación, investigación, enciclopedia, la institución, sala de prensa, acción corporativa, empleo, licitaciones

The DC Date: 2009-09-15

Our judgment? This is a superb combination of Museum Art and

Social Networking. It goes without saying that access to this great Art Gallery via Google Earth is a first in World History? But, in reality, all that Google Earth gives to the experience is a virtual geographical surrounding… for the Art.

Scenario 4: The Ateneum Museum (Helsinki, Finland)

Backstory: “The exhibition focuses on the cultural life of young women in 1910s Helsinki through the eyes of writer and critic L. Onerva (1882–1972) [51]. She studied art history at university, lived on her own, enjoyed the cultural scene of the city, had an active social life, got married, ran away, got divorced, and had a secret affair. She made her living and supported her writer's career by teaching and translating, and above all by journalism and art reviews. In this exhibition, Onerva introduces us to her Helsinki: art galleries, theatre premieres, films, cafés, restaurants, concerts, and other social events. She also reveals the flipside of an independent life: debts, limits to her freedom, and moral judgment. The exhibition features plenty of art from the era, from Ateneum's own collections as well as other museums. Pioneers of early Finnish modernism, such as Helene Schjerfbeck, Sulho Sipilä and Yrjö Ollila, depicted modern man and the urban culture of the time. The curator of the exhibition is PhD Anna Kortelainen. In connection to the exhibition there will also be a book coming out, published by Tammi.”

We will be at ElPub 2010 in Helsinki. We will be able to see the Onerva exhibition. We will be able to demonstrate “live” the interplay of Social Networks and the (Finnish) National Gallery, 16-18 June 2010. There is a Wikipedia article on L. Onerva and although it is currently available only in Finnish, accessing it through the Chrome browser permits instant translation into English (and there are the usual sorts of blunders one expects from such automatic machine translation; but one can grasp the sense of the original Finnish text).

In comparison with the National Gallery of Ireland, the Finnish National Gallery is outstanding with respect to its e-presence [52].

4. Discussion and Conclusions

In the paper we have attempted to blend the static with the dynamic. We have sought to bring together the classical “this will appear in print” type of material (ordinary pub type)—dated instantly at the time of release, whether in paper copy form or as an el-pub (just like ordinary pub type in modern medium). Such static forms then become a matter of public record—history. At the same time we wished to express the dynamic, to note that the technology unfolds continuously in our times. We wished to indicate this sense of the dynamic by the use of present and future tenses. The core of that dynamic was grounded in the 4th Scenario above on the "Onerva Exhibition," already opened in the Ateneum, Helsinki (2010 March 25) and with a promise to illuminate this text in the ElPub conference in the same city two months later.

In a similar way we wish to make the current text "dynamic-like" by referring back to the technological developments announced and unleashed circa January 2010 and reported on in the introduction. We do not engage in futurology. Rather we wish to discuss the future of the Social Network and the (National) Art Gallery within the context of the Art-sensitive mobile devices.

Technology: The computer science and engineering

In many Art Galleries one is allowed to take photographs provided that the "Flash facility" is turned off. In such galleries individual art works might be tagged by the "universal no photography allowed for this work" icon of a camera with a red X. In many Art Galleries photography is strictly prohibited. In most Art Galleries photographs may be permitted by application in advance and the signing off of a memorandum of agreement.

But we are currently researching into the use of the camera phones which "capture" the image of the picture, not as photograph as such, but rather as image to be recognized in order that it may be identified. This falls into the category of content-based image retrieval, a computer vision problem in which a program is given an input image of some subject and attempts to locate further images of the same subject in a collection. The difficulty of this problem is clear; lighting conditions, camera angles and perspectives will likely all be different in the images. Let us imagine 3 people with camera phones standing side-by-side (with usual comfort zone separation between them) in front of a picture, such as Carravagio's "The Taking of Christ" (1602) [53] in the National Gallery of Ireland? The perspective view of each will be different. The computer vision technology must facilitate such "minor" differences in the view. Algorithms in this field usually rely on identifying invariant elements ("interest points'") in the image using a variety of techniques [54-56]. A compounding difficulty in some domains is the reality that many images of different real-world scenes contain incidental similarities due to repeated manufactured elements. Similar such elements are often a constituent part of Modern Art. For example, the works of Bridget Riley rely extensively on repeated elements [57].

This last point of repeated elements is a particular concern in the field of robot navigation; work by Cummins [58, 59] presents an improved technique for allowing a robot navigation system to take observations (images) of the environment and assign probabilities that any two images had been taken at the same location, and thus recognise its own location. The author of this work noted that "Our model is also applicable to some types of image retrieval task." [59] Indeed, "this author" is now the author of the PlinkArt application for Android-based mobile phones, which makes use of the camera on the mobile device to capture an image of an artwork. The image is then processed and uploaded to a remote server where image retrieval is performed and attempts to identify the original artwork are carried out. The mobile device can then display relevant information.

A similar approach is taken by the Google Goggles application (also available for the Android platform). This particular application is more general in its reach. It also attempts to identify books and DVD's (by cover), landmarks, corporate logos, and a number of other elements. In this application domain the presence of a large and well categorised corpus seems to be critical to the success of the application [60, 61].

We conclude with a brief short story. The first author made an appointment with the Director of the National Gallery of Ireland in order to discuss some of the technical details concerning the digitization of the Gallery's Collection, for this paper. The meeting was subsequently cancelled. Unfortunately, the Director had to go to Rome "with the Carravagio" — a colloquial name for famous painting "The Taking of Christ (1602)" for an exhibition. One deduces by the phrase "with the Carravagio" that the National Gallery of Ireland has just the one work by him. It was the time, if memory serves well, when the Catholic Bishops of Ireland had been assembled by the Pope to discuss the major problem of the handling of clerical pedophiles in the country. The painting itself is very big. Fortunately a digital copy of the painting is available under Creative Commons Licence at Wikimedia Commons. Consequently, the first author has a copy (as well as 54 other digital images of Carravagio's works) on the iPhone. These el-pubs of Art are everywhere on the soc-nets. And now as we go to press (2010-04-13) it has just been announced that Plink has been acquired by Google and consequently PlinkArt will be absorbed by/within Google Goggles.

Notes and References

1. *guardian.co.uk*. 2010 [cited 2010 January 19]; Available from: http://www.guardian.co.uk/.
2. *The Guardian/Technology*. 2010 [cited 2010 January 19]; Available from: http://www.guardian.co.uk/technology.
3. Arthur, C. *Google puts off launch of mobile phone in China*. 2010 [cited 2010 January 19]; Available from: http://www.guardian.co.uk/technology/2010/jan/19/google-nexus-one-phone-china.
4. Arthur, C. *Apple confirms date for its 'event': we know it's a tablet, but what else?* 2010 [cited 2010 January 19]; Available from: http://www.guardian.co.uk/technology/blog/2010/jan/18/apple-tablet-date-confirmed.
5. Arthur, C. *Charles on... anything that comes along*. 2010 [cited 2010 January 19]; Available from: http://www.charlesarthur.com/blog/.
6. Arthur, C. *Charles Arthur (charlesarthur) on Twitter*. 2010 [cited 2010 January 19]; Available from: http://twitter.com/charlesarthur.
7. Wikipedia. *Mashup (web application hybrid)*. [cited 2010 March 25]; Available from: http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid).
8. Arthur, C., *Our technology coverage: what happens next*, in *The Guardian/TechnologyGuardian*. 2009. p. 2.
9. *National Gallery of Ireland*. 2010 [cited 2010 January 20]; Available from: http://www.nationalgallery.ie/.
10. *National Art Gallery Sofia*. 2008 [cited 2010 January 18]; Available from: http://www.nationalartgallerybg.org/.
11. *Ateneum Art Museum*. 2010 [cited 2010 January 20]; Available from: http://www.ateneum.fi/.
12. *Museo Nacional del Prado*. 2010 [cited 2010 January 21]; Available from: http://www.museodelprado.es/.
13. Geertz, C., *Local Knowledge: Further Essays in Interpretive Anthropology*. Third ed. 2000, 1983: Basic Books.
14. DCMI. *Dublin Core Metadata Initiative*. [cited 2009 June 26]; Available from: http://dublincore.org/.
15. Powell, A. *DC-dot*. 2000 [cited 2009 June 26]; Available from: http://www.ukoln.ac.uk/metadata/dcdot/.
16. *Dublin Core History*. 2010 [cited 2010 January 21]; Available from: http://dublincore.org/about/history/.
17. Mac an Airchinnigh, M. *Towards a vision of an Internet of Cultural Things*. 2009; Available from: http://ec.europa.eu/information_society/policy/rfid/library/index_en.htm#iotcons.
18. Wikipedia. *Radio Frequency Identification*. 2010 [cited 2010 January 21]; Available from: http://en.wikipedia.org/wiki/RFID.

19. Wikipedia. *Near Field Communication*. 2010 [cited 2010 January 21]; Available from: http://en.wikipedia.org/wiki/Near_Field_Communication.
20. Wikipedia. *Internet of Things*. 2010 [cited 2008 November 14]; Available from: http://en.wikipedia.org/wiki/Internet_of_Things.
21. Plink. *PlinkArt*. 2010 [cited 2010 January 20]; Available from: http://www.plinkart.com/.
22. Wikipedia. *Android operating system*. 2010 [cited 2010 January 21]; Available from: http://en.wikipedia.org/wiki/Android_(operating_system).
23. Google. *Google Goggles*. 2010 [cited 2010 January 21]; Available from: http://www.google.com/mobile/goggles/#landmark.
24. Nokia. *Augmentation Reality App: Point and Find*. 2010 [cited 2010 January 22]; Available from: http://pointandfind.nokia.com/.
25. Snow, C.P. *The Two Cultures*. 2010 [cited 2010 April 6]; Available from: http://en.wikipedia.org/wiki/The_Two_Cultures.
26. Mac an Airchinnigh, M., K. Sotirova, and Y. Tonta, *Digital re-discovery of culture game of inquiry & the physicality of soul.* Review of the National Center for Digitization, 2006. 2006(9): p. 19-37.
27. De Rynck, P., *How to Read a Painting*. 2004: Thames & Hudson.
28. Wikipedia. *The Garden of Earthly Delights*. [cited 2010 March 21]; Available from: http://en.wikipedia.org/wiki/The_Garden_of_Earthly_Delights.
29. *The Prado in Google Earth*. [cited 2010 March 23]; Available from: http://www.google.com/intl/en/landing/prado/.
30. Wikipedia. *List of paintings by Hieronymus Bosch*. [cited 2010 April 1]; Available from: http://en.wikipedia.org/wiki/List_of_paintings_by_Hieronymus_Bosch.
31. Wikipedia. *National Gallery*. 2010 [cited 2010 April 1]; Available from: http://en.wikipedia.org/wiki/National_gallery.
32. Berger, J., *Ways of seeing : based on the BBC television series with John Berger*. 1972, London: British Broadcasting Corporation and Penguin. 166.
33. Berger, J. *Ways of Seeing*. 1972 [cited 2009 August 11]; Available from: http://www.youtube.com/watch?v=LnfB-pUm3eI.
34. ADS Software Group Inc., *ART Version 4.4*.
35. Waiboer, A.E. and L. Ahtola-Moorhouse, *Northern Stars and Southern Lights: The Golden Age of Finnish Art, 1870-1920*. 2008, Dublin: National Gallery of Ireland.
36. Wikipedia. *Albert Edelfelt*. 2010 [cited 2010 January 20]; Available from: http://en.wikipedia.org/wiki/Albert_Edelfelt.
37. Wikipedia (Suomi). *Albert Edelfelt*. 2010 [cited 2010 January]; Available from: http://fi.wikipedia.org/wiki/Albert_Edelfelt.
38. Edelfelt, A. *Virginie*. 1883; Available from: http://www.flickr.com/photos/floridagirl7/4180173503/.

39. Metsu, G. [cited 2010 April 8]; Available from: http://www.gabrielmetsu.org/.
40. Gergova, I. and V. Rumenov, *Landscape Images in Bulgarian Icon Painting*. 2007, Sofia: 41T Ltd. Publishers.
41. National Art Gallery Sofia. *The Crypt*. [cited 2010 April 8]; Available from: http://www.nationalartgallerybg.org/index.php?l=46.
42. Paskaleva, K., *Bulgarian Icons through the Centuries*. 1987, Sofia: Svyat Publishers.
43. Орела, М. *Влизане в Йерусалем*. 2010 [cited 2010 April 8]; Available from: http://www.flickr.com/photos/mihalorel/4503282602/.
44. Wikipedia. *National Art Gallery (Bulgaria)*. 2010 [cited 2010 April 8]; Available from: http://en.wikipedia.org/wiki/National_Art_Gallery_(Bulgaria).
45. Орела, М. *„Жътварка-30-те год.“*. [cited 2010 April 8]; Available from: http://www.flickr.com/photos/mihalorel/4268585971/.
46. Орела, М. *„Алея в тропически парк, 1932“*. [cited 2010 April 8]; Available from: http://www.flickr.com/photos/mihalorel/4269291172/.
47. Beekman, R. *Cabinet approves plan for four metropolitan museums*. Available from: http://sofiaecho.com/2010/02/11/856595_cabinet-approves-plan-for-four-metropolitan-museums.
48. *Museo Nacional del Prado, Online Gallery*. [cited 2010 March 23]; Available from: http://www.museodelprado.es/en/the-collection/online-gallery/.
49. Unknown, *Offering by Orestes and Pylades (San Ildefonso Group)*. Circa 10 BCE (Before Current Era): Museo Nacional del Prado, Madrid.
50. Museo Nacional del Prado. *Social Networks*. 2010 [cited 2010 April 12]; Available from: http://www.museodelprado.es/index.php?id=2778&L=5.
51. Wikipedia. *L. Onerva*. [cited 2010 April 10]; Available from: http://fi.wikipedia.org/wiki/L._Onerva.
52. Finnish National Gallery. *The Art Collections*. [cited 2010 April 11]; Available from: http://kokoelmat.fng.fi/wandora/w?lang=en&action=gen.
53. Wikipedia. *Caravaggio's "The Taking of Christ" (1602)*. [cited 2010 April 13]; Available from: http://en.wikipedia.org/wiki/The_Taking_of_Christ_(Caravaggio).
54. Nist´er, D. and H. Stewenius, *Scalable recognition with a vocabulary tree*, in *Conf. Computer Vision and Pattern Recognition*. 2006. p. 2161–2168.
55. Sivic, J. and A. Zisserman, *Video Google: A text retrieval approach to object matching in videos*, in *International Conference on Computer Vision*. 2003: Nice, France.
56. Jégou, H., M. Douze, and C. Schmid, *Hamming embedding and weak geometric consistency for large scale image search.*, in *European Conference on Computer Vision*, A. Zisserman, D. Forsyth, and P. Torr, Editors. 2008, Springer. p. 304–317.
57. *Bridget Riley Dialogues on Art*, ed. R. Kudielka. 1995: Thames & Hudson.
58. Cummins, M. and P. Newman, *The International Journal of Robotics Research*. 2008. 27: p. 647.

59. Cummins, M., *Probabilistic Localization and Mapping in Appearance Space*. 2009, Balliol College, Oxford.

60. Droid News Net. *App Review – PlinkArt (App, Free)*. [cited 2010 April 11]; Available from: http://www.droidnews.net/2009/12/app-review-plinkart-app-free/.

61. Droid News Net. *App Review – Google Goggles (App, Free)*. [cited 2010 April 11]; Available from: http://www.droidnews.net/2009/12/app-review-google-goggles-app-free-2/.

The anatomy of an electronic discussion list for librarians, KUTUP-L: Bibliometric and content analyses of postings

Yaşar Tonta[1]; Doğan Karabulut[2]

[1]Department of Information Management, Faculty of Letters,
Hacettepe University, 06800 Beytepe, Ankara, Turkey.
tonta@hacettepe.edu.tr
[2]Turkish Grand National Assembly Library,
06543 Bakanlıklar, Ankara, Turkey
dogank@tbmm.gov.tr

Abstract

Electronic discussion lists are widely used as a professional and scientific communication tool since late 1980s. Analysis of messages sent to discussion lists provides useful information on professional as well as scientific communication patterns. In this paper, we present the findings of a bibliometric analysis of some 20,000 messages sent to KUTUP-L, an electronic discussion list for Turkish librarians, between 1994 and 2008. We test if the distributions of messages and their authors conform to Pareto, Price and Lotka laws. We then analyze the contents of 977 messages based on a stratified sample. Findings indicate that the number of messages sent to KUTUP-L has increased over the years along with the number of authors. Two thirds (1,232) of about 1,900 list members posted at least one message to the list while the rest preferred to be so called "lurkers". Some 35 authors posted almost half (49%) the messages while 20% of the authors posted 83% of all messages. The distribution of messages to authors conform to Price ("the square root of all authors would post half the messages") and Pareto laws (so called "80/20 rule"), respectively. Of the 1,232 authors, one third (as opposed to 60% predicted by Lotka's Law) sent only one message to the list. Results of content analysis show that 40% of messages sent to the list were off-topic. Issues about or related with information management services (32%), library and information science (23%) and professional and scientific communication (19%) were discussed more often in the list. The intent analysis of the postings shows that three quarters of the messages were initiatory while the rest were reflexive. That's to say that the majority of messages posted on KUTUP-L to initiate a discussion did not seem to generate enough interest for others to reflect upon them by sending follow up messages, suggesting that professional and scientific communication taking place on KUTUP-L on certain subjects can be characterized as more of a one-way communication than a participatory one.

Keywords: KUTUP-L; electronic discussion lists; electronic publishing; professional communication; scientific communication; bibliometric analysis; content analysis

1. Introduction

The history of computer-based communication dates back to mid-1960s. Host-based mail systems were later replaced by the electronic mail (or e-mail) system of the ARPANET computer network in early 1970s. E-mail has become a "killer app" on BITNET, the predecessor of the current day Internet. LISTSERV, an electronic discussion list management software, was introduced in early 1990s and e-mail based discussion lists such as PACS-L, LIBREF, PUBLIB and WEB4LIB proliferated thereafter. Messages posted to such discussion lists contain invaluable resources for historians, social scientists, social network analysts, and bibliometricians, among others, and they can be analyzed to study professional and scientific communication patterns along with the topics discussed and the productivity of authors.

KUTUP-L, an electronic discussion list for Turkish librarians, was set up in June 1991 to share information, exchange ideas and discuss professional issues. It currently has some 1,900 registered members. Based on the analysis of messages in KUTUP-L archives, this paper aims to address the following research questions:

- Has the number of messages posted and the subjects discussed in KUTUP-L increased and proliferated over the years?
- Have professional and scientific communication patterns in KUTUP-L changed over the years in librarianship?
- Does the distribution of messages to authors (thus the authors' productivity) conform to Pareto, Price and Lotka laws?

Findings of this study will shed some light on the evolution of KUTUP-L as an electronic discussion list for the last two decades. Intent and content analysis of KUTUP-L postings will provide both quantitative and qualitative information about the level of activity as well as the types of postings, their subjects and authors.

2. Literature Review

Content analysis is a commonly used method of studying messages sent to electronic discussion lists. Wildemuth et al. [1] used content analysis to study 14 different library discussions lists and found that the relatively high percentage of messages were intended to discuss certain issues.

Content analysis of 309 messages posted at PUBYAC, a discussion list created for public librarians working in children's and young adult services, shows that the majority of postings were of reference type, indicating that the list took on the role of an information source for its subscribers [2]. PUBYAC postings were analyzed under six different categories in a different study: programs (27%), finding books (21%), collection (16%), library administration and policy (9%), professional issues (9%), and announcements (7%). Half the messages were responses to earlier requests and (37%) were inquiries while the rest (13%) were announcements and general comments. Authors of messages were generally thankful and complimented the list and its subscribers [3].

A survey of MEDLIB-L (Medical Library Association’s discussion list) users showed that about 90% of them read MEDLIB-L at work and spend less than three hours a week for this purpose. They used the list to comment on various issues and answer questions more often than to ask questions or start discussions [4]. The former types of messages are called “reflexive” ones while the latter are “initiatory” [5, 6]. Similarly, almost three quarters of messages sent to EVALTALK, a listserv for evaluation professionals, were comments/responses on requests while the rest were questions or requests, indicating that subscribers used EVALTALK discussion list as an informational tool [7]. We see the same pattern in the messages of the trombone users’ discussion list, Trombone-L and a listserv used as a journal. Some 72% of the Trombone-L messages were comments/answers and 28% questions, although percentages varied by topics discussed (One-third of the messages were off-topic.) [8]. Reflexive messages constituted 65% of listserv messages while the rest were initiatory messages [9]. More than half (51%) the messages posted to HUMANIST discussion list were made up of responses while 25% questions, 19% announcements and 5% administrative ones [10]. Some 56% of the messages sent to ABUSE-L, a discussion list on social work, were classified under “discussion” (i.e., reflexive messages) [11].

The communication patterns of authors posting at discussion lists tend to conform to some bibliometric laws such as Pareto (80% of messages get posted by 20% of all authors), Price (“the square root of all authors would post half the messages”) and Lotka (60% of authors send one message to the list while decreasing percentages of authors send more, i.e., 15% two, 6.6% three, 3.75% four and so on) [12,13,14,15]. Messages posted at two discussion lists (LINGUIST and HEL-L) seemed to conform to Lotka's Law, although the correlation was not high [16].

3. Data and Method

To address research questions, we first obtained access to KUTUP-L logs hosted by the Middle East Technical University in Ankara, Turkey. Logs archived between June 1991 and September 1994 were not available due to technical reasons. We imported the contents of all messages along with associated metadata to a spreadsheet package and cleaned the data before analysis. Descriptive statistics and bibliometric analysis are based on a total of 19,827 messages posted on KUTUP-L between 1994 and 2008. We used Pareto, Price and Lotka's laws to find out if the author productivity in KUTUP-L conforms to decreasing power laws. As given in the previous section, the first two laws are relatively easy to explain. To test if data fit Lotka's Law, we used $f(n) = C / n\alpha$ formula wherein $f(n)$ is a function of frequency, C and α are constants ($C > 0$ and $\alpha \geq 0$). Thus, the number of authors posting n messages is proportional to decreasing power law [14].

We then selected a stratified sample of 977 messages for content analysis (sample size 5%). Using Bellack's communication model, we classified each message either as "initiatory" (i.e., asking a question or initiating a discussion) or "reflexive" (i.e., answering a question or commenting on an issue) [5,6]. Based on Berman's "intent analysis", we also categorized each message according to its purpose (or "intent") under "information transfer" (IT), "information request" (IR) or "discussion of an issue" (IS) [11]. Next, we carried out content analysis to identify the subject(s) of each message using Jarvelin and Vakkari's subject classification [17, 18].

We presented the descriptive statistics and the results of bibliometric and content analysis using tables and figures. We grouped the findings of content analysis in five-year intervals to detect the changes of patterns in professional and scientific communication over the years.

4. Findings and Discussion

A total of 19,827 messages were posted on KUTUP-L between 1994 and 2008, half of which belong to the last five years (2004-2008). The average number of postings per month in recent years is about 175. Findings indicate that communication in the list has increased continuously over the years and KUTUP-L has become a major venue of communication and discussion among library and information professionals (Figure 1). The heaviest message traffic was observed in March as the Turkish Library Week gets celebrated in the last week of March every year while the list was less busy during summer months. About one third of 1,900 list members never posted a message on KUTUP-L. The gender of 1,232 unique authors contributing to the list is evenly distributed (52% female, 48% male), although males posted more (59%) messages than females.

The distribution of messages to authors conform to both Price and Pareto (80/20 rule) laws: Almost half (49%) the messages were posted by 35 out of 1,232 authors while 20% of all authors posted 83% of all messages. One third (369) of all authors posted only one message, half the percentage predicted by Lotka's Law (60%) (Table 1). More than half the authors (52.60%) posted three messages at most, constituting a mere 3% of all messages. The great majority of authors contributed to the list very little, thereby making them primarily "lurkers". Some 40% of all authors posted five or more messages to the list.

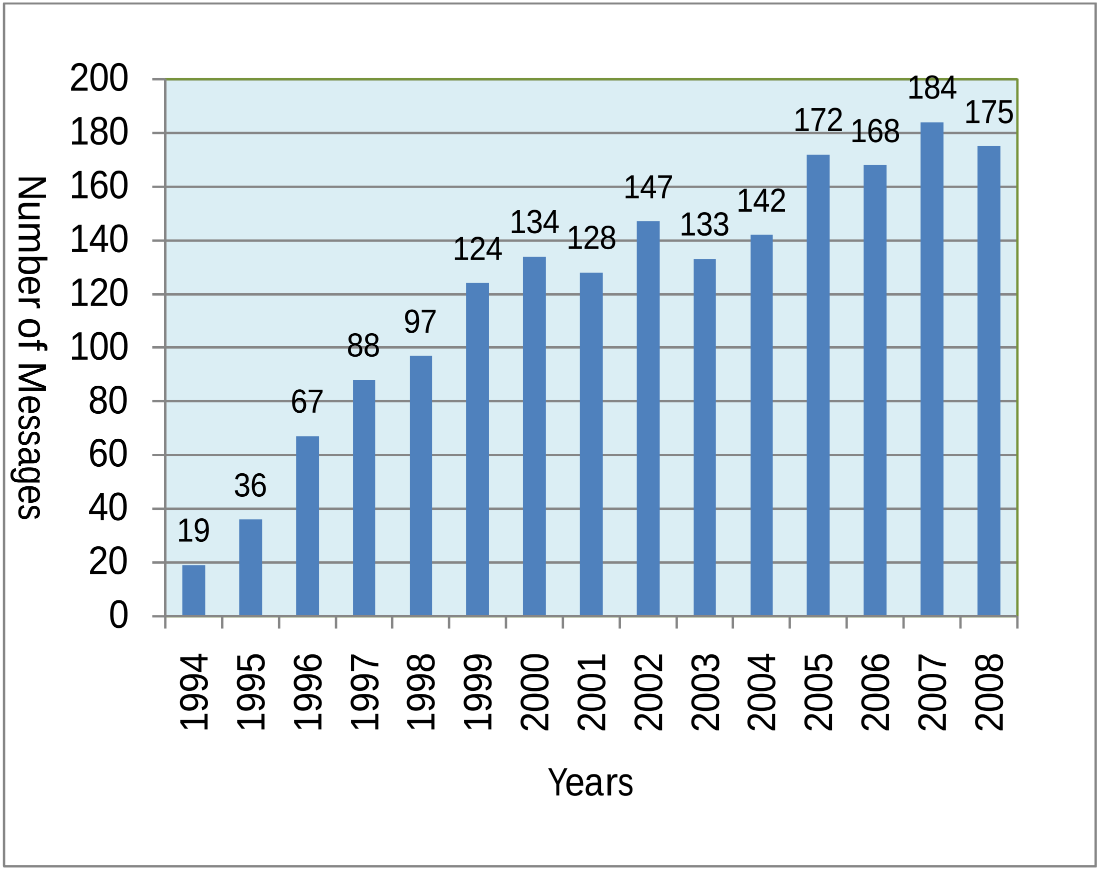


Figure 1: Average number of messages sent to KUTUP-L on a monthly basis (1994-2008)

The authors' productivity data for KUTUP-L posters seem to fit Pareto and Price laws fairly well. Yet, the distribution of messages to authors does not conform to Lotka's Law, which is due to the fact that KUTUP-L has a relatively stable base of contributors (much more than what Lotka's Law predicts) who send messages to the list from time to time. It could be that characteristics of authorship of a journal article differ from that of a post to a discussion list such as KUTUP-L.

Table 1: Test of Lotka's Law on KUTUP-L authorship data

# of messages	Expected percentage of authors according to Lotka's Law (%)	Expected number of authors according to Lotka's Law	Observed percentage of authors (%)	Observed number of authors
1	60.00	739	30.00	369
2	15.00	184	14.40	178
3	6.60	82	8.20	101
4	3.75	46	7.47	92
5	2.40	30	4.71	58
6	1.60	20	3.33	41
7	1.20	15	3.08	38
8 or more			28.71	355
Total	100.00	1,232	100.00	1,232

Note: Percentage and numbers of authors contributing more than 7 messages to KUTUP-L according to Lotka's Law are not given in the table.

Intent analysis

Three quarters (76% to be exact) of all postings were "initiatory" (i.e., asking a question or initiating a discussion) while the rest were reflexive (i.e., answering a question or commenting on earlier postings). Table 2 provides descriptive statistics, at five-year intervals, about findings of intent analysis based on a stratified sample of 977 KUTUP-L messages. The percentage of reflexive messages tended to decrease over the years, suggesting that more list members seemed to be indifferent towards KUTUP-L postings. The intention was transferring information in two thirds (67%) of all messages, followed by starting a discussion (23%) and asking for information (10%). The percentage of postings aiming to transfer information increased over the years while the percentage of postings with discussion topics decreased. Forwarded postings made up 16% of all messages, although the percentage is decreasing. The percentage of postings containing links to other web sites is on the rise (20%).

Table 2: Intent analysis of KUTUP-L postings

Type of messages	Years			Total
	1995-1999	2000-2004	2005-2008	
Initiatory	137 (63%)	279 (80%)	325 (79%)	741 (76)
Reflexive	82 (37)	70 (20)	84 (21)	236 (24)
Total	219 (100)	349 (100)	409 (100)	977 (100)

Note: Figures in brackets refer to percentages.

Content analysis

Content analysis of on-topic messages shows that about one third (32%) were related with information management services, 23% with library and information science, and 19% with professional and scientific communication (Table 3). (Each on-topic message was classified under the main topic). Topics discussed on KUTUP-L varied over the years. For instance, postings on information management and professional/scientific communication issues became more prominent in recent years while the percentage of postings on cataloging issues went down drastically (from 50% to 21%) over the years.

The percentages of reflexive postings were well over 50% for some topics (e.g., professional issues, professional training, library management and library automation), indicating that some topics drew more attention and generated more discussion on KUTUP-L. The percentage of postings intended to generate discussion has also increased and the topics of such postings were in line with those of reflexive ones, further reinforcing the willingness of KUTUP-L members to make it a more dynamic electronic discussion list.

Out of 977 KUTUP-L messages in our stratified sample, 393 (or 40% of all messages) were off-topic. Some of those messages were irrelevant while others consisted of postings of trial messages, virus warnings, announcements of social activities, deaths, and so on. The percentage of off-topic messages rose to 46% in recent years. Table 4 provides descriptive data on off-topic messages. In general, more (53%) than half of off-topic messages were irrelevant (i.e., unrelated with the purpose of the discussion list). Announcements of promotions (16%), deaths (9%), and job ads (9%) consisted of one-third of all off-topic messages.

Table 3: KUTUP-L messages by topics

Subjects	Subject 1	Subject 2	Total	%
100 Professional issues	39	4	43	7
101 Library Association's activities	30	0	30	5
The Professions total (100)	69	4	73	12
300 Publishing	18	6	24	4
400 Education in LIS	12	2	14	2
600 Analysis of LIS	1	0	1	0
701 Inter-library loan activities	59	20	79	12
702 Collections	57	6	63	10
703 Information or Reference Services	2	0	2	0
704 User education	1	0	1	0
705 Library Buildings or Facilities	12	0	12	2
706 Library Administration or Planning	10	0	10	2
707 Library Automation	22	2	24	4
708 Other Library or Information Service	14	1	15	2
LIS Service Activities total (700)	177	29	206	32
801 Cataloguing	14	3	17	3
802 Classification or Indexing	3	1	4	1
803 Information Retrieval	4	5	9	1
804 Bibliographic Databases	5	1	6	1
805 Databases	9	2	11	2
Information Retrieval total (800)	35	12	47	7
901 Information Dissemination	7	1	8	1
905 Information Use	2	0	2	0
906 Information Management	4	0	4	1
Information Seeking total (900)	13	1	14	2
1001 Scientific or Professional Publishing	10	0	10	2
1003 Other Aspects of Scientific or Professional Communication	101	12	113	17
Scientific and Professional Communication total (1000)	111	12	123	19
1100 Other LIS Aspects	148	0	148	23
Total	584	66	650	100

Table 4: Off-topic KUTUP-L messages

Types of messages	1995-99	2000-04	2005-08	Total	Avg (%)
Irrelevant messages	47 (23)	78 (37)	84 (40)	209 (100)	53
Announcements of promotions, congrats, etc.	6 (10)	18 (29)	39 (62)	63 (101)	16
Announcements of deaths, etc.	0 (0)	9 (24)	28 (76)	37 (100)	9
Job Ads	1 (3)	15 (44)	18 (53)	34 (100)	9
Trial messages	3 (21)	6 (43)	5 (36)	14 (100)	4
Messages on social activities	3 (21)	6 (43)	5 (36)	14 (100)	4
KUTUP-L	5 (42)	6 (50)	1 (8)	12 (100)	3
Virus warnings	2 (20)	7 (70)	1 (10)	10 (100)	3
Total	67	145	181	393	101

Note: Figures in brackets refer to percentages. Some totals are not equal to 100% due to rounding errors.

We calculated the productivity of 34 authors in our sample who posted six or more messages by dividing the number of off-topic messages by the total number of messages (both off- and on-topic) sent by each author. The average productivity was 60%, perfectly in line with the percentage of on-topic messages.

5. Conclusion

As an unmoderated discussion list since its inception in 1991, KUTUP-L seems to have an impact on professional lives of many Turkish librarians in that they use it as a venue to ask questions, share news and information with their colleagues, follow up current developments and discuss professional as well as social issues. The number of messages and unique authors contributing to the list has increased considerably, indicating that KUTUP-L has become a popular and dynamic discussion list.

Although a wide variety of subjects have been discussed on KUTUP-L, the percentage of reflexive messages aimed at discussing a subject or commenting on a professional issue is comparatively low (24%). A more comprehensive study on the subjects of KUTUP-L postings is in order. However, the relatively high (40%) percentage of off-topic KUTUP-L messages might have discouraged its further use, as library and information professionals may not wish to spend their precious time sifting through irrelevant postings. Screening off-topic messages before distribution may help in this respect but this would tax the list owner’s time and resources further.

Most of the KUTUP-L postings were authored by a relatively few list members. KUTUP-L authors' productivity data conform to Pareto and Price laws but not in accordance with Lotka's Law. KUTUP-L has more list members contributing two or more messages to the list.

We hope that the results of bibliometric and content analysis of KUTUP-L postings will be helpful in studying professional and scientific communication patterns of library and information professionals in a larger context.

References

[1] WILDEMUTH, B. et al. What's everybody talking about: message functions and topics on electronic lists and newsgroups in information and library science. Journal of Education for Library and Information Science, 38 (2), 1997, p. 137-156.

[2] BAR-ILAN, J; ASSOULINE, B. A content analysis of PUBYAC- A preliminary study. Information Technology and Libraries, 16, 1997, p. 165-174.

[3] EDWARDS, MM. "A content analysis of the PUBYAC discussion list". (Unpublished Master's Thesis). University of North Carolina at Chapel Hill, Chapel Hill, NC, 1999.

[4] SCHOCH, NA; SHOOSHAN, SE. Communication on a listserv for health information professionals: uses and users of MEDLIB-L. Bulletin of the Medical Library Association, 85 (1), 1997, p. 23-32. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC226219/ (April 2010).

[5] BELLACK, AA. et al. *The language of the classroom.* New York: Teachers College Press, 1966.

[6] BELLACK, AA. Methods for observing classroom behavior of teachers and students.Presented at a conference sponsored by Paedagogisches Zentrum, Berlin, November 12-15, 1968.

[7] CHRISTIE, CA; AZZAM, T. What's all the talk about? Examining EVALTALK,an evaluation listserv. American Journal of Evaluation, *25* (2), 2004, p. 219-234. Available at: doi: 10.1177/109821400402500206 (April 2010).

[8] NICKLAS, EW. "The Trombone-L e-mail discussion list : an analysis of its usersand Content." (Unpublished PhD Thesis). The University Of Missouri, 2002.

[9] PIBURN, MD; MIDDLETON, JA. Listserv as journal: computer-based reflectionin a program for pre-service mathematics and science teachers. Paper presented at the *International Conference on Science, Mathematics and Technology Education (Hanoi, Vietnam, January 6-9, 1997).*

[10] MAY, A. Automatic classification of e-mail messages by message type. Journal of the American Society for Information Science, 48, 1997, p. 32-39.

[11] BERMAN, Y. Discussion groups in the Internet as sources of information: the case of social work. *Aslib Proceedings*, 48 (2), 1996, p. 31-36. Available at: http://www.aslib.co.uk/proceedings/1996/feb/1.html (April 2010).

[12] TRUESWELL, RL. Some behavioral patterns of library users: the 80/20 rule. Wilson Library Bulletin*, 43*, 1969, p. 458-461.

[13] DE SOLLA PRICE, D. A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science, 27, 1976, p. 292-306.

[14] EGGHE, L. *Power laws in the information production process: Lotkaian informetrics*. Amsterdam: Elsevier, 2005.

[15] EGGHE, L; ROUSSEAU, R. *Introduction to informetrics: Quantitative methods inlibrary, documentation and information science*. Amsterdam: Elsevier Science Publishers, 1990. Available at: http://uhdspace.
uhasselt.be/dspace/handle/1942/587 (April 2010).

[16] KUPERMAN, V. Productivity in the Internet mailing lists: A bibliometric analysis, Journal of the American Society for Information Science and Technology, 57, 2006, p. 51-59.

[17] JARVELIN, K; Vakkari, P. Content analysis of research articles in library and information science., Library and Information Science Research*,* 12, 1990, p. 395-421.

[18] JARVELIN, K; VAKKARI, P. Evolution of library and information science 1965-1985: Content analysis of journal articles. Information Processing Management*,* 29, 1993, p. 129-144.

Constituencies of use: Representative usage scenarios in international digital library user studies, a case study on Europeana

Duncan Birrell[1]; Milena Dobreva[1]; Yurdagül Ünal[2]; Pierluigi Feliciati[3]

1 Centre for Digital Library Research (CDLR),
Information Services Division (ISD), University of Strathclyde
Livingstone Tower, 26 Richmond Street, Glasgow, G1 1XH, UK
{duncan.birrell, milena.dobreva@strath.ac.uk}
2 Department of Information Management, Hacettepe University
06800 Beytepe, Ankara, Turkey
yurdagul@hacettepe.edu.tr
3 Department of Cultural Heritage, University of Macerata
via Brunforte, 63023 - Fermo, Italy
pierluigi.feliciati@unimc.it

Abstract

Digital libraries are still being developed independently of the extensive involvement of end users, those who form their *constituencies of use*. The traditional approach to digital library development is to consult with experts or *communities of practice* in a particular field and attempt to incorporate recommendations into the interface functionality and service models, whilst user needs are often not comprehensively scoped in advance, at the development stage, or regularly consulted for the purposes of formative and summative evaluation. Recent developments in digital library design concentrate effort on the use of innovative search and browse tools, streamlined techniques for navigation and display, and the provision of personalised areas for search management and information sharing; such developments, however, remain unaligned to any thorough understanding of exactly *how* user behaviour alters depending on scenario of use, and the problems encountered by end users in task completion within different contexts. This paper reports on the deployment of usage scenarios to evaluate the Europeana digital library v1.0 prototype.

Keywords: digital libraries, user studies, tasks

1. Introduction

Digital libraries are only as good as the uses to which we can put them. Whilst user needs are signalled as a priority in the multitude of policy documents which shape online cultural heritage services, user needs are often not comprehensively scoped in advance at the development stage, or regularly consulted for the purposes of formative and summative evaluation. The traditional approach to digital library development is to consult with experts or *communities of practice* in a particular field, and incorporate advice and recommendations into the service models. In a review of the use of digitised archival collections, for example, A. Sundqvist [1] noted that "the general knowledge of user behaviour is a mixture of common sense, presumptions and prejudices" (p. 624), whilst the Institute of Museum and Library Services (IMLS) reported that "The most frequently-used needs assessment methods do not directly involve the users" (p. 2) [2]. Z. Manžuch [3] in her survey on monitoring digitisation (which summarises 11 user-related studies), showed that the most popular method deployed was the analysis of usage statistics. Digital libraries are, therefore, still being developed independently of the extensive involvement of their *constituencies of use*.

An evidence-based approach to the information behaviour of users will have tangible impact on the development of interface functionality, digital library policy, data quality and potentially on the architecture of digital libraries.[1] Such studies need to address different user groups. For example, the younger users, often referred to as "digital natives",[2] are expected to prefer enhanced *functionality* which reflects the customisable, interactive, information experience found on popular search and social networking sites; whilst, by comparison, professional users seek authoritative and trusted information *quality*.

Recent developments in digital library design have concentrated effort on the use of innovative search and browse tools, streamlined techniques for navigation and display, and the provision of personalised areas for search management and information sharing. Hence, studying the use of such digital libraries requires selecting suitable scenarios which will allow the users not only to navigate and search for data, but also deploy a range of

[1] Domains used as defined in the DELOS Digital Library Reference Model) [4]

[2] A term applied to the new generation of users who have grown up with ICT and whose patterns of knowledge creation and information sharing are largely defined by web based tools, virtual worlds and social media.

functionalities. This paper presents the experience of devising user-orientated assignments for a study of Europeana[3] undertaken in October-December 2009 by the Centre for Digital Library Research (CDLR) at the University of Strathclyde with the participation of the University of Macerata and Glasgow Caledonian University. The study aimed to gather feedback from members of the general public and younger users, since a previous web survey of Europeana users conducted in May 2009 identified these to be relatively low use consumers compared to professional in their 30s and early 40s [5].

2. Methodology and composition of the groups

Previous research on users of digital libraries has incorporated a range of methodologies including: interviews, focus groups, observations, usability testing, transaction logging, user surveys, web-based questionnaires, think aloud protocols and video data [6, 7, 8, 9, 10]. Alongside usage scenarios, the study also utilised psychological testing techniques such as eye-tracking, which had been deployed in previous studies of information behaviour (cf. [11, 12,13]), but have not, before now, been applied to assess digital libraries.

The study needed to gather demographic data which would help to analyse the profile of the participants and gather data on the tasks performed. Table 1 summarizes demographic data on the participants in the study.

The problem presented to the CDLR-led study was how to devise a cohesive methodology for the user testing of a multilingual digital library, where focus group demonstration would take place across 4 European countries with the resource being evaluated by a number of different *constituencies of use* i.e. groups composed of different cultures, linguistic communities, professions and ages. For example, 76% of study participants were between the ages of 15 (or under) to 24 years and could be categorised as belonging to the growing constituency of “digital natives”.

[3] http://www.europeana.eu/portal/. A single access point for digitised cultural heritage materials provided by various European libraries, museums, archives, galleries, and audiovisual collections. At the time of the study it was offering access to 6 million digital objects. The interface is available in 26 languages and supports both simple and advanced search, and offers additional functionalities such as a timeline and date clouds.

Table 1: Demographic characteristics of the participants in the study

	Country									
	Bulgaria		Italy		The Netherlands		UK		Total	
	N	%	N	%	N	%	N	%	N	%
Country	22	25	20	23	23	26	24	27	89	100
Gender										
Male	11	50	6	30	12	52	10	42	39	44
Female	11	50	14	70	11	48	14	58	50	56
Age										
Under 15	2	9	-	-	-	-	-	-	2	2
15-18	20	91	-	-	21	91	-	-	41	46
19-24	-	-	17	85	2	9	6	25	25	28
25-34	-	-	-	-	-	-	7	29	7	8
35-44	-	-	-	-	-	-	3	13	3	3
45-54	-	-	2	10	-	-	5	21	7	8
55-64	-	-	1	5	-	-	3	13	4	5
Profession										
At school	22	100			23	100	-	-	45	51
At College / University	-	-	20	100	-	-	5	21	25	28
Researcher	-	-	-	-	-	-	4	17	4	5
Information specialist	-	-	-	-	-	-	4	17	4	5
Manager / Administrator	-	-	-	-	-	-	3	13	3	3
Lecturer	-	-	-	-	-	-	1	4	1	1
Writer	-	-	-	-	-	-	1	4	1	1
Other	-	-	-	-	-	-	6	25	6	7

A key issue in the design of user studies is the ‘hands-on’ experience of users physically interacting with the interface and the selection of tasks for the purposes of testing and evaluation. During a study of information behaviour, users will normally have to answer specific questions, usually belonging to two categories: *navigational* and *informational* [11, 13, 14]. The selection of tasks has a central role in the user studies as they constitute the very fabric of user experience. In order to keep users engaged during the study, it is important that tasks should adequately address their likely interests. The selection of tasks is even more difficult when the study is international, addresses divergent constituencies, and must produce comparable outputs from the various groups involved for the purposes of systematic evaluation. The innovative solution for the CDLR was to devise an extended and iterative assignment which could accommodate the interests and abilities of different user groups without compromising the basic thematic structure or upsetting the essential navigational and informational elements of the tasks.

Our testing of Europeana, therefore, necessarily applied a uniform methodology to all focus groups and media labs. A standard protocol was established for the study including three questionnaires (demographic data; first impressions (since the participants were not familiar with Europeana before the study) and deeper/lasting impressions), as well as a series of key discussion points and an assignment. The thematic bond of the assignment was the creation of *a virtual portrait of the city* where each focus group was held (with the exception of the group in Fermo where the assignment was to make a virtual portrait of the city of Rome). This allowed the tasks to be both locally specific and easily translated across cultures and age groups.

The assignment requested participants to collate a PowerPoint presentation, from materials retrieved on Europeana, using a predefined set of slides which guided them to produce a virtual tour of their local city, and which also aimed to encourage use of Europeana's innovative functionality. The assignment was designed to incorporate 8 different usage scenarios: 1) finding texts on a predefined topic; 2) finding images on a predefined topic; 3) finding audio and video materials on a predefined topic; 4) finding materials presenting the same, predefined, object in different times (eg. how their city had changed over time); 5) finding materials on a specific predefined subject (like a landmark or an event or a person); 6) finding materials on a specific historical event; 7) a topic of the participants' own choice within the context of the general theme; and finally, 8) identifying the providers of digital objects who contributed the highest number of objects on a particular topic as a means of encouraging consideration of the provenance of objects; this last also asked users to identify what they found to be most useful about Europeana and to suggest areas where material may be lacking.

Deploying such scenarios requires users to formulate searches that target a range of metadata fields to retrieve a variety of material types. The approach made it possible to assess which scenarios of use are easiest to satisfy and to identify the stumbling blocks that users of the Europeana prototype might be encountering. The tasks were selected on an extended and iterative model representative of the key processes in the searching of digital libraries such as 'undirected searches' and 'monitoring a topic over time' reported in Bryan-Kinns and Blandford [15] and noted the findings of Furnas and Rauch [16] that 'one-shot query' of digital libraries, as with conventional ones, is relatively rare.

The approach is also compatible with the TIME framework for evaluating digital libraries developed by Andrew Dillon [17]. The TIME framework focuses on four elements: Task – what users want to do; Information model – what structures aid use; Manipulation of materials – how users access the components of the document; and Ergonomics of visual displays – how they affect human perception of information. The 8 scenarios of the assignment represent the Task; Europeana provides an environment in which the users can try various searches, which maps to the TIME Information model.

Triangulation of the information gathered on search strategies, the data recorded during eye-tracking sessions, and the combined responses to discussion and questionnaires, enabled researchers to analyse: 1) the user searches which show the queries used to search for objects; 2) the most extensively and frequently used (and unused) components of the interface (based on the eye tracking data); and 3) users' actual performance in relation to the specific scenarios (based on the content of presentations completed by participants).

3. Results

Table 2 below shows the results sets (at the time of the study) for Text, Image, Video and Sound materials for a simple search using the name of each city selected in the assignment (participants in Fermo chose to present a virtual portrait of Rome).

Table 2: Europeana "simple search" results for assignment cities

	Text	Images	Audio/Video
Sofia	35	668	34
Amsterdam	3,653	69,440	381
Glasgow	678	33,842	1,541
Rome/Roma	1,509/1,604	27,007/ 13,853	534/213

An inevitable consequence of adopting such a thematic approach was a marked difference in levels of digital representation (the number of digital objects available in Europeana for each location) for different geographic regions; however, the repetition of a single generic and fixed assignment throughout would have resulted in more problems, for example, with regards to the language of retrieved objects in different cultural locations.

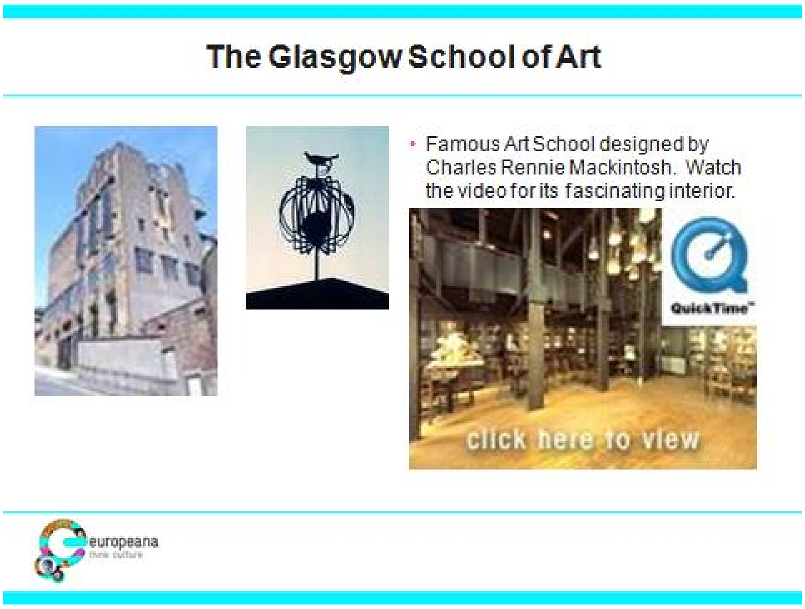
Table 3 gives the total number of presentations prepared by the various groups and the range of material types retrieved by participants in response to the task. In some instances (as in the school groups), two participants worked jointly on a single presentation, therefore, the number of completed presentations is lower than the overall number of participants in these cases. The number of slides populated with text, image or audio/visual materials is also lowest for Sofia which also reflects the relatively low ranking of Bulgarian institutions in the provision of digital collections to Europeana.

Table 3: Number of digital objects' types retrieved during task by participants

	Completed presentations	Textual resources	Image files	Audio/Video files
Sofia	15	0	0	3
Amsterdam	19	6	10	7
Glasgow	24	10	15	5
Fermo	10	5	8	0

The 8 separate tasks of the assignment required users to formulate searches that addressed differing levels of generality/specificity. From the point of view of the digital object model used in Europeana, the various tasks addressed a range of metadata fields to retrieve a variety of material types; and also involved the necessary use of various functional components. As a key aim of the study was to discover areas of difficulty for current users of the Europeana prototype, deploying a range of tasks which addressed matters of both content and functionality was deemed to be the most efficient approach.

The sample slide in Figure 1 illustrates the materials retrieved from Europeana by one Glasgow-based participant for *Scenario 5: finding materials on a specific predefined subject*, in this case the Mackintosh School of Art. The slide is representative of the difficulties encountered by a number of participants in completing the task for 3 reasons: 1) the low quality resolution of the thumbnail demonstrates the difficulty of participants in retrieving presentation quality images; 2) the insert of the Quicktime icon highlights the lack of direct access to (often subscription based) audio/visual resources experienced by the groups, and 3) users commented on a lack of textual material available to support findings.



The Glasgow School of Art

- Famous Art School designed by Charles Rennie Mackintosh. Watch the video for its fascinating interior.

Figure 1. Sample slide showing materials retrieved from Europeana by a Focus Group participant relating to the Glasgow School of Art designed by Charles Rennie Mackintosh

Table 4 below outlines the levels of performance and problems encountered across the focus groups for all 8 usage scenarios deployed.

Table 4: The 8 Usage Scenarios used for the assignment

Scenario 1	Finding texts on a predefined topic
Slide title	What do people *write* about the city of... Sofia, Amsterdam, Rome, Glasgow
General description	Participants working on assignments were tasked to identify and use reliable text resources, copy them and supply a reference to sources.
What problems were experienced?	• It was impossible to understand materials in foreign languages (experienced in Sofia, Fermo). • Maps were received as text objects (Glasgow). • In many cases texts were retrieved as digitised images and could not be used or copied easily. • Few participants added references to indicate the sources of documents.
Scenario 2	Finding images on a predefined topic
Slide titles	How do people *see*...Sofia, Amsterdam, Rome, Glasgow?
General description	Participants were tasked to identify and use relevant images, to copy the image files to their presentations and supply a reference to their

	source.
What problems were experienced?	• Images were most easy to find. • Concerns over image quality were raised, regarding the size and resolution of image files.
Scenario 3	Finding audio and/or video materials on a predefined topic
Slide title	What are the *sounds* of the city?
General description	Participants working on the assignment were asked to identify sounds which might be either typical of or unique to their city. They were expected to be able to access audio/video files and insert the resource within their presentations whilst supplying a reference to the source.
What problems were experienced?	• Big challenge to access video materials. • Audio materials easier to find and use • Generally difficult to copy such objects into presentations. • It would be helpful to have previews of material available through subscription.
Scenario 4	Finding materials presenting the same object in different times
Slide title	How has the city *changed* over time?
General description	Participants were asked to identify materials, images of landmarks etc. which could be said to represent their city in different historical periods. They were expected to be able to access and use the resource within their presentations and supply a reference to the source.
What problems were experienced?	• Users experienced difficulty in having to “guess” what digital objects were available on Europeana relating to their cultural heritage at different times – this would be made easier for users if the range of resources related to the same object were linked.
Scenario 5	Finding materials on a specific subject (like a building, place or person)
Slide titles	*Sofia as saint, princess & city; The Royal Palace on Dam Square; The Fontana dei Quattro Fiumi in Piazza Navona; The Glasgow School of Art.*
General description	Participants were asked to identify materials related to a building or landmark of popular/ iconic status within their respective cities. They could focus on its appearance or use in an historical period alongside the contemporary one. They were expected to be able to access and use the materials in presentations whilst supplying references to sources.
What problems were experienced?	• Student in Bulgaria experienced problems with polysemy of the word “Sofia”. • Although this appeared an easy task with specific objects to search for, it seems participants had difficulty in locating digital objects which matched their knowledge and expectations.

Scenario 6	Finding materials on a significant historical event
Slide titles	What happened in 1853?; Roma during the Ventennio (1924-1945); What happened in Glasgow's George Square in 1919?
General description	Participants were asked to retrieve materials relevant to a specific historic date or event. They were not restricted to what material they selected to represent the event and were encouraged to seek primary as well as secondary sources. They were expected to be able to access the materials for use in presentations and supply references.
What problems were experienced?	A general observation is that participants did not use the timeline to search for answers of these questions but rather performed general searches combining the name of their city and the year in question.
Scenario 7	Finding materials on a topic of the participants' choice within the context of the general theme
Slide title	Use this slide for your own material…
General description	Participants were invited to present materials on a subject of their own choosing. They were not restricted to what materials they could select (as long as the subject was in keeping with the thematic context of the task) as long as reference to sources on Europeana were used.
What problems were experienced?	This task redirected participants to browse mode; a low number populated the slides due mainly to lack of time for completion.
Scenario 8	Identifying the providers of digital objects who contributed the highest number of objects on a particular topic
Slide title	Europeana and... Sofia, Amsterdam, Rome, Glasgow
General description	Participants were requested to provide feedback on the institutions and partners who had supplied the most materials on their city in Europeana. Feedback was also gathered on what they considered to be the most useful aspects of the site and their recommendations for its further development.
What problems were experienced?	Generally participants did not look at the drill-down options of the search to provide the information but responses were based on their impressions of what types of materials they had retrieved during the assignment.

4. Discussion

The selection of tasks allowed for consideration of what specific difficulties were experienced in various usage scenarios. Older (more professional users), for example, complained of not being able to complete tasks due to a lack of granularity encountered in the recording of some objects in Europeana (e.g. maps being classified as texts), and cases where textual materials could only be retrieved as digitised images; whilst younger users, (who found images most easy to find) complained of not being able to freely access audio/video content in order to complete tasks, or find contemporary materials to reflect changes in their city over time. Generally, both constituencies of users expressed concerns over image quality, regarding the size and resolution of image files, stating also that it would be helpful to have previews of the materials currently available only through subscription.

The tasks of the assignment were not purely navigational but included a number of necessary navigational elements, such as use of the Europeana advanced search, time-line, results tabs and date-filters. Eye-tracking data revealed, however, that participants did not fully investigate the various drill-down options available on the search interface, and analysis of search data revealed that participants rarely used advanced search options or additional functionality but rather performed variations upon general search themes.

5. Conclusion

The CDLR-led User and Functional Testing study demonstrates that it is possible to develop large scale digital libraries, such as Europeana, with the extensive involvement of end users. In order to track the evolving requirements of both “digital natives” and other users, it is hoped that such innovative studies of digital libraries in the cultural heritage domain will continue to be conducted; a domain which is multi-cultural and multi-lingual, and whose constituencies of use seek better and more efficient forms of digital representation from their online cultural heritage institutions.

Acknowledgements

The authors would like to acknowledge the support of Emma McCulloch from CDLR (University of Strathclyde), the EDL Foundation; Jill Cousins, Adeline van den Berg, Anne Marie van Gerwen and Milena Popova from the Europeana Office. Discussions with Prof. Ian Ruthven (University of Strathclyde) and Dr. Jonathan Sykes (Glasgow Caledonian University) were particularly helpful for the study.

Notes and References

[1] SUNDQVIST, A. The use of records – a literature review. Archives & Social Studies: A Journal of Interdisciplinary Research, 1(1), 2007, p. 623-653.

[2] IMLS. Assessment of End-User Needs in IMLS-Funded Digitization Projects: Institute of Museum and Library Services, October 2003, p. 1-41. Available at http://www.imls.gov/pdf/userneedsassessment.pdf (2010).

[3] MANZUCH, Z. Monitoring digitisation: lessons from previous experiences. Journal of Documentation, 65(5), 2009, p. 768-796.

[4] CANDELA, L; et al. The DELOS Digital Library Reference Model - Foundations for Digital Libraries. Version 0.98. February 2008.

[5] EUROPEANA Online Visitor Survey: Research Report, Version 3, 2009.

[6] NORMORE, LF. Characterizing a Digital Library's Users: Steps towards a Nuanced View of the User. In: Proceedings of the American Society for Information Science and Technology, 45(1), 2009, p.1-7.

[7] BISHOP, AP.; et al. Digital Libraries: Situating Use in Changing Information Infrastructure. Journal of the American Society for Information Science and Technology, 51(4), 2000, p. 394-413.

[8] CHERRY, JM; DUFF, WM. Studying digital library users over time: a follow-up survey of Early Canadiana Online. Information Research 7(2), 2002.

[9] BLANDFORD, A; STELMASZEWSKA, H; BRYAN-KINNS, N. Use of Multiple Digital Libraries: A Case Study. In: Proceedings JCDL 2001. p. 179-188. ACM Press.

[10] TENOPIR, C.; et al. Use of Electronic Science Journals in the Undergraduate Curriculum: An Observational Study. In Proceedings of the American Society for Information Science and Technology, 41(1) p.64-71, 2004.

[11] GRANKA, LA. Eye-R: Eye-Tracking Analysis of User Behavior in Online Search. Masters Thesis, Cornell University Library Press, 2004.

[12] LORIGO, L; PAN, B; HEMBROOKE, H; JOACHIMS, T; GRANKA, L; GAY, G. The influence of task and gender on search and evaluation behavior using Google. Information Processing & Management, 42, 2006, p. 1123—1131.

[13] PAN, B; HEMBROOKE, H; JOACHIMS, T; LORIGO, L.; GAY, G.; GRANKA, L. In Google we trust: Users' decisions on rank, position, and relevance. Journal of Computer-Mediated Communication, 12(3), 2007.

[14] RELE, RS; DUCHOWSKI, AT. Using Eye Tracking to Evaluate Alternative Search Results Interfaces. In: Proceedings of the Human Factors and Ergonomics Society, p.1459—1463. Orlando, Florida, 2005

[15] BRYAN-KINNS, N; BLANDFORD, A. A survey of user studies for digital libraries, RIDL working paper, London, 2000.

[16] FURNAS, DW; RAUCH, SJ. Considerations for Information Environments and the NaviQue Workspace. *Proceedings of ACM DL* '98, pp. 79-88, 1998.

[17] DILLON, A. Evaluating on TIME: a framework for the expert evaluation of digital interface usability. *International Journal on Digital Libraries,* 2 (2/3), 1999.

ENHANCING USERS' EXPERIENCE: A CONTENT ANALYSIS OF 12 UNIVERSITY LIBRARIES FACEBOOK PROFILES

Licia Calvi[1]*; Maria Cassella*[2]*; Koos Nuijten*[1]

1 NHTV University of Applied Science:
Academy for Digital Entertainment
P.O. Box 391, 4800 DX Breda, The Netherlands
e-mail: {calvi.l; nuijten.k}@nhtv.nl;
2 Università di Torino
Via Po, 17, 10124 Torino, Italy
e-mail: maria.cassella@unito

Abstract

Facebook has become one of the most prominent tools for social networking over the last few years. Since its establishing in 2004, more and more players have made use of it: not just ordinary users willing to find their old friends and to get back into contact with them, but also, for example, more and more players from the cultural scene. These latter ones include cultural institutions willing to experiment with new ways of getting in touch with their traditional audiences but also willing to attract new audiences (like a younger audience, who is supposed to be more present on such social media); artists, who use it to create a community to share information, to promote their own creations but, more recently, also to collaborate on common project; and finally also libraries.

This paper intends to explore the use of Facebook in university libraries by making an empirical analysis of current practices. In doing so, the paper builds on the knowledge gained in a previous study on the way in which Flemish cultural institutions make use of the possibilities offered by social media to communicate with their audiences and to promote themselves [2]. The analysis on current uses we performed will help us sample existing practices and help us derive some general ideas for future best practices. And this will help libraries to better profile themselves and communicate better with their old and new audiences.

Keywords: academic libraries; social networking platforms; Facebook; Facebook academic library pages.

1. Introduction

As academic libraries strive to reposition themselves in the digital environment and try to reconfigure their role, librarians experiment the use of social tools of the Web 2.0 to advocate, promote, and raise awareness about library collections and services.

One of the most popular social networking platforms is Facebook (FB [1]). Originally developed by Mark Zuckerberg, Dustin Moskovitz and Chris Hughes in 2004 at Harvard University in order to provide Harvard students with a place in which they could keep in contact with their classmates and, most importantly, could share study-related information, Facebook "burst beyond its roots" by opening its membership to high school networks first, in 2005, and to all the net users later, in 2007. In the last few years, Facebook has globally developed into one of the most prominent tools for social networking altogether.

Since its establishing in 2004, more and more players have made use of Facebook: not just college students or ordinary users willing to find their old friends and to get back into contact with them, but also, for example, more and more players from the cultural scene have started to use Facebook. These latter ones include cultural institutions willing to experiment with new ways of getting in touch with their traditional audiences but also willing to attract new audiences (like a younger audience, who is supposed to be more present on such social media [2]); artists, who use it to create a community to share information, to promote their own creations but, more recently, also to collaborate on common project [3]; and finally also libraries.

The use of Facebook in libraries is starting to be investigated (see in [4, 5, 11, 14, 16, 17, 19]) as well as the use of other social media. Studies like the ones mentioned above focus on the tools and applications available in Facebook for librarians and make recommendations about the way libraries could benefit from using Facebook. Such applications include a Facebook Librarian [6], i.e., a virtual librarian service providing links to books and other resources; Books iRead [7], to share books with the friends in your own network; tools like the World Cat Search [8]), and several ad hoc Facebook groups [9].

This paper instead intends to explore the use of Facebook in university libraries by making an empirical analysis of current practices in 12 selected academic libraries. In doing so, the paper builds on the knowledge gained in a previous study on the way in which Flemish cultural institutions make use of the possibilities offered by social media to communicate with their audiences and to promote themselves [2]. For that analysis, a two-phase, empirical and qualitative evaluation of social media use was carried on. In a first phase, a survey was conducted on as many cultural institutions as possible in order to identify the role social media play in their current communication practices. In a second phase, the focus was narrowed down to a very specific set of institutions selected on the basis of the previous analysis for which the Facebook pages were analysed in terms of the content on each page, the updates, the degree of users' participation and the ways in which these institutions were handling users' participation, and the fidelity issue, both as it is perceived by the cultural users and as it is handled by the cultural institutions.

The results of this study show that there is a very low degree of personalisation among the cultural institutions that were analysed, although their focus and scope was intrinsically different. We noticed additionally that Facebook itself was used rather poorly, i.e., mainly to promote events or to show pictures of past events. But what was really interesting, was the fidelity issue associated with these institutions: the Facebook pages of the cultural institutions were visited by many serendipitous users, but there were very few regular and faithful ones.

With the present paper, we would like to further extend the results coming from this study and apply it to academic research libraries. The analysis on current uses we performed will help us sample existing practices and help us derive some general ideas for future best practices. And this will help libraries to better profile themselves and communicate better with their old and new audiences.

2. Facebook in academic libraries. Literature review

Since 2007, Facebook popularity is steadily advancing among colleges and universities students. Kerry estimates that 85% of undergraduates in USA have a Facebook profile [9].

Academic libraries have since then started to explore how this technology could be used in their libraries to contact and attract students, despite the fact that some very early reactions from students about the use of social networking services were not that positive at the beginning.

As a matter of fact, in an OCLC report from 2007, 6100 people aged 14-84 and 382 US library directors did not see "a role for the libraries constructing social sites and would not be very likely to contribute content" [10]. In the literature on FB in academic libraries, many librarians also express their concern about the use of social networking platforms in libraries. Charnigo and Barnett-Ellis [11], for example, found that librarians were wary about the academic purpose of FB. 54% of 126 librarians surveyed by the authors stated that it did not serve an academic purpose. 12% only was positive on this fact, and the rest were not sure. Marshall Breeding, the director of innovative technology at Vanderbilt University, wrote about the enormous opportunities of adopting Web 2.0 tools in academic libraries [12]. However, he recognized "that the very nature of Facebook works against this scenario. The natural circle of Friends centers on one's peers [...] and it is unrealistic to think that large numbers of undergraduate students would want to count librarians among their FB Friends" [12].

In many instances academic librarians adopt FB pages for their libraries but are worried about the best way to approach students. Miller and Jensen [13] advocate the aggressive "Friend and Feed" technique by which librarians "friend" as many students as possible, while Powers et al. [14} are more cautious about the practice of "friending" students. A better approach to them is to recommend mentioning one's Facebook account in library instruction sessions and reference interviews and then letting the students find that account.

In a few articles we indeed found success stories of the use of social networking platforms in academic libraries: Beth Evans [15], for example, created a "Brooklyn College MySpace page". The library then used three employees to sift through MySpace profiles to find 4,000 Brooklyn College students, faculty, and graduates. Evans invited these affiliates to be the library's friends and seven months later had approximately 2,350 friends. Evans did not mention any downsides to the Brooklyn College Library MySpace experiment and indicated that it had been well received by its audience [16]. Successful are also the results of the experiment led by Mack et al. [17], who promoted their FB library page profile for the reference service. During the fall of 2006, their librarians received 441 reference questions and 126 of these were collected through Facebook, followed by e-mail (122) and in-person consultations (112).

Studies like those mentioned above focus on the librarians' attitude and experiences with the use of FB in academic libraries, while others investigate tools and applications available in Facebook for librarians and make recommendations about the way libraries could benefit from using Facebook 5]. In 2008, Ellissa Kroski listed in her blog iLibrarian the top ten Facebook applications for libraries [18]: Books iRead to share books with the friends in your own network, LibGuides Librarian, Librarian, University of Illinois at Urbana-Champaign (UIUC) Library Catalog, del.icio.us, JSTOR, MyWikipedia, LOLCats, Slideshare, and MyFlickr. Hendrix et al. [19] also provided a different perspective to the studies of FB in academic libraries. The authors used a survey to investigate health libraries' use of the popular social network. 72 librarians responded to the survey: 12,5% (9/72) maintained a Facebook page. Libraries used FB mainly "to market the library, push out announcements to library users, post photos, provide chat reference, and have a presence in the social network" [19]. Librarians had a very positive attitude towards the future of their FB pages although its use was currently rather low.

To date, the only study focused on actual Facebook library pages use and their content is the one from Jacobson [4]. The author investigated 12 FB academic libraries using the Site Observation methodology. Results showed that FB library pages are a useful tool to market the library "and it may be valid to assert that this is currently the best use in the library realm. Whereas uses for communication from patrons or "fans", communicating library needs, and as a forum/discussion space for users may not be an ideal use" [4].

3. Scope and methodology of this study

The present paper investigated the level of use of Facebook in twelve UK research universities libraries. The scope of the authors in performing this study was:

- to assess whether FB can be an effective new tool to communicate and promote the academic library services, to outreach students, both undergraduates and graduates, to fidelize them, or whether other solutions should be preferred (i.e. a personal librarian's profile);

- to assess what the most used sections and services of a FB academic library page are;

- to highlight the potentiality of FB as a new channel to implement value-added services for students (i.e. asynchronuos reference, training courses and tutorials ….);

- to verify whether there is any positive correlation between the use of FB library pages and the number of FTE students enrolled in a university or any other possible variables (i.e. a new library building, active libraries hosting many events and exhibitions …..)

- to assess any differences in the use of FB central library pages and FB branch libraries pages.

To this end, we selected 12 UK research university libraries and classified them according to the following criteria:

1. libraries in universities with less than 10000 students
2. libraries in universties with more than 10000 students
3. branch university libraries.

This resulted in 4 libraries per category.

In order to perform some statistics (i.e., t-tests and some basic descriptives, see next section), we developed a coding instrument in line with the one used by Jacobson [4] for her analysis. Coding focused on the FB page developed by each library (number of pictures or videos, number of fans, links to social software), the kind of updates present (i.e., via blogs, newsfeeds or fans updates), the possible use of the wall (by whom, and how frequently), the presence of library applications or tools, the presence of information other than library-related one (i.e., links to external events or to possible sponsors), and whether the FB page is used for internal employees communication or announcements.

We recorded data for each library for 8 days over a period of two weeks (from 29 March 2010 to 9 April 2010), once a day, at 23 hrs, to make sure that all libraries would be already closed and that therefore no more updates from the library staff were possible.

4. Results

A quantitative analysis of the data collected with our coding instrument was processed for some descriptive statistics and correlations. In this section, we report some of the most prominent results.

First, we wanted to verify the frequency of wall use. Table 1 reports our findings: just less than 50% of the FB library pages use the wall.

Table 1: The wall is used

		Frequency	Percent	Valid percent	Cumulative percent
Valid	No	49	51,6	51,6	51,6
	Yes	46	48,4	48,4	100,0
	Total	95	100,0	100,0	

Moreover, over half of the wall postings are about a year old or older. That seems to indicate that the activity on the wall is not very well kept up to date or it means that the popularity of the platform for the library is already wearing out.

We looked at the time when postings are posted on the wall (Table 2). Table 2 shows that wall postings are usually updated by the end of the month (around 65% is done just around day 1, 2 or 30-31 of the month). Updates in-between are not really frequent.

Posting wall updates at the end of the month seems to be an explicit decision: it is bigger and more active libraries that do post at the end of the month, as to indicate that there is a communication strategy behind this choice and that the library has ad hoc staff in charge of it. Branch libraries just have postings at month edges.

Table 2: When postings are posted on the wall

		Frequency	Percent	Valid percent	Cumulative percent
Valid	1	8	17,4	17,4	17,4
	2	2	4,3	4,3	21,7
	6	6	13,0	13,0	34,8
	7	2	4,3	4,3	39,1
	8	5	10,9	10,9	50,0
	9	3	6,5	6,5	56,5
	30	1	2,2	2,2	58,7
	31	19	41,3	41,3	100,0
	Total	46	100,0	100,0	

If we look at the number of pictures in the FB library pages, we found that that libraries publish between 0 and 49 pictures: it looks as if the profiles differ quite a lot in this respect. The average number of pictures does not seem to be quite a good indicator of library behaviour on FB. By selecting only the FB pages that carry at least two pictures, then almost half of the profiles fall away. However, it seems that the profiles that remain really use those pictures (i.e., profiles seem to have either 0, 1 or 4 or more pictures incorporated). It seems that the FB profiles that carry more than two pictures really use these pictures. For the others, it seems the pictures are just an occasional and quite haphazard addition. The mean number of pictures for the FB profiles that really and intentionally use pictures is almost 26.

As for videos, most libraries use videos very limitedly. But if they put more than one video on their FB page, then they use them a lot (between 11 and 22).

The number of fans ranges between 6 and 1004. Most libraries seem to have 0 to 250 fans, and again, libraries that have a reasonable amount of fans (i.e., over 25), have very much fans (minimum 46, mean over 365). Because our data were dichotomous, it was not possible to measure the number of fan posts. Our data show that 71 out of 95 units contain updates: given the fact that not all libraries have a reasonable amount of fans, this value seems to indicate for the libraries that do have fans that these fans indeed also do posts. Although it was not possible to carry on specific demographics, it seems there are more male fans than female fans. Librarian posts are quite common (81 out of 95), instead.

As for the kind of library applications that are present on FB, we could identify that most libraries use either OPAC or database searches: it looks as if OPAC or database searches are more or less comparable services. No other applications are as successful.

5. Discussion

If we combine the results described in the previous section with a qualitative analysis of the data collected with our coding instrument, the following characteristics in the use of FB by academic libraries become more evident:

- Among the FB elements present on the library pages, the only one that is truely active is the wall.
- Wall activity however differs depening on the library size and profile: the bigger the library, the more active the wall. Branch libraries also use the wall scarcely.
- Wall activity is nevertheless still limited to an average of a couple of postings per day, in the best cases. The wall is used to post information on new libraries activities, change in the opening hours, availability of learning rooms and in one instance to promote a new collection. Very few are the postings by fans. Where the wall is active is thanks to the librarians' activity.
- Although it was almost impossible to derive some demographics of the fans for each library (since some have a high fan number), we had the impression that women are more active than men on the library walls (although in minority, see above). This statement (that can not be proved from our data) is however confirmed by Schrock [20] who claims that women are more active in social networks *tout court*.
- Most FB library pages include some library applications (e.g., OPAC or database search, JSTOR, book advice, etc.), but, again, the richest offer of library tools applies to the bigger libraries.
- The FB pages of the libraries we analysed have vey few links to other social software: Flickr, YouTube or del.ic.ous. There are no links to other external sites.
- The FB pages are not used to promote external events and only scarcely for internal communication for employees.

Although we must admit that our conclusions are based on a very small sample of FB academic libraries pages, our observations indicate that FB might be a very powerful library communication and promotion tool but at the moment its actual use is neither extensive nor advanced.

It is clear that the big libraries which are most active in cultural and learning events are also more active on FB. It is also clear that it is a librarians' task to keep the FB pages alive and that this activity might be time-consuming. According to Hendrix [19] "the time spent maintaining and updating a library Facebook page ranged "from no weekly maintenance to 120 minutes a week". Therefore, we suggest that librarians who wish to create a FB library page should consider carefully if they have enough time to dedicate to its maintenance.

Librarians should consider if it is more effective to create a FB library page or a personal librarians' profile to outreach to students. For very proactive reference librarians, for example, this might be a better strategy as FB is mainly perceived by users as a virtual personal space.

Generally speaking, for libraries to assess the best approach "to be where the users are" qualitative pre-tests and post-tests performed both on users visiting the library and on remote users might be helpful.

From these findings we can conclude that two kinds of libraries can be identified:

- those with a very active FB profile and who invest on their FB pages a lot. They have a higher number of fans, pictures and of videos which generate a higher return in terms of fans' involvement and participation.
- Those who do little either because they do not seem to appreciate the added value of having a FB page well enough or because of resource limitations.

6. Conclusions

These observations do not clear out how effective and efficient it is for libraries to develop a page on FB and if a FB page helps them achieve their goals to outreach to students. Unfortunately, we could not match these findings with a survey whereby librarians would have explained what their intentions were with opening a page on FB.

Although FB is a space made for people [10], something that our findings confirm by pointing out how the personal and the professional area seem to remain separate for FB users, we can positively conclude from these observations that FB pages help increase the communication with the students if librarians are proactive and keep the wall alive, but there is no evidence that at the moment content delivery or services delivery, i.e., reference assistance have been improved in this way.

However, the only way to profit from the added value provided by FB is to invest in it and to be rich and active: the more active, the better and higher return on investment in terms of fans' participation and involvement.

We nevertheless believe that in a few years' time social networking platforms will become more effective for academic libraries to communicate with students and to deliver them new types of services. As technology evolves, social networking platforms become more and more diffuse, pervasive, and advanced and students get used to the idea that FB might also have an institutional function/goal.

Notes and References

[1] www.facebook.com.

[2] BOOST, A. Digitale culturele communicatie: de rol van sociale media in de communicatie tussen culturele instellingen en hun doelpubliek (Digital cultural communication: the role of social media in the communication between cultural institutions and their audience). M.A. in Journalism, Lessius College, 2008/2009.

[3] Smith, S. The creative uses of Facebook as a tool for artistic collaboration. In *Proceedings of the Electronic Visualisation and the Arts (EVA) Conference.* London, 2009. Available at https://www.dora.dmu.ac.uk/handle/2086/2764?show=full (April 2010)

[4] JACOBSON, TB. Facebook as a library tool: perceived vs actual use. To be published in College & Research Libraries, the journal of the Association of College and Research Libraries. Available as preprint at http://www.ala.org/ala/mgrps/divs/acrl/publications/crljournal/preprints/crl-088.pdf (April 2010)

[5] SECKER, J. *Case Study 5: Libraries and Facebook*. London: London School of Economics and Political Science, 2008

[6] http://www.facebook.com/apps/application.php?id=3135795462&b

[7] http://www.facebook.com/apps/application.php?id=2406120893&b

[8] http://apps.facebook.com/worldcat/

[9] KERRY, B., *Are you on Facebook? A comprehensive guide*, presentation. Available at http://www.slideshare.net/KerryFSU/facebook-and-academic-libraries-presentation (April 2010).

[10] OCLC. *Sharing, privacy and trust in our networked world*. [online] 2007. Available at http://www.oclc.org/reports/sharing/default.htm (April 2010)

[11] CHARNIGO, L; BARNETT-ELLIS, P. Checking out Facebook.com: the impact of a digital trend on academic libraries. Information Technology and Libraries, 26 (1), 2007, p. 23-34

[12] BREEDING, M. Librarians face online social networks. Computer in libraries [online], 27 (8), 2007, p. 30-33. Available at http://www.librarytechnology.org/ltg-displaytext.pl?RC=12735 (April 2010)

[13] MILLER, SE; JENSEN, LA. Connecting and communicating with students on Facebook. Computers in libraries [online], 27 (8), 2007, p.18-29. Available at http://www.infotoday.com/cilmag/sep07/index.shtml (April 2010)

[14] POWERS, AC; SCHMIDT, J; HILL, C. Why can't we be friends? The MSU libraries find friends on Facebook. Mississipi Libraries. 72 (1), 2008, p. 3-5.

[15] EVANS, B. Your Space or MySpace. Library Journal, 15 October 2006. Available at http://www.libraryjournal.com/article/CA6375465.html& (April 2010)

[16] CONNELL, RS. Academic libraries, Facebook and MySpace, and student outreach: a survey of student opinion. Portal: libraries and the academy, 9 (1), 2009, p. 25-36 Available at http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v009/9.1.connell.html#f5#f5 (April 2010)

[17] MACK, D; BEHLER, A; ROBERTS, B; RIMLAND, E. Reaching students with Facebook: data and best practices. Electronic Journal of Academic and Special Librarianship [online], 2007. Available at http://southernlibrarianship.icaap.org/content/v08n02/mack_d01.html (April 2010)

[18] http://oedb.org/blogs/ilibrarian/2007/top-ten-facebook-apps-for-librarians-part-one/

[19] HENDRIX, D; CHIARELLA, D; HASMAN, L; MURPHY, S; ZAFRON, ML. Use of Facebook in academic health libraries. Journal of medical library association, 97 (1), 2009, p. 44-47. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605034/ (April 2010)

[20] SCHROCK, A. Examinig social media usage: Technology clusters and social netwrok site membership. Available at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2242/2066 (April 2010)

Use and relevance of Web 2.0 resources for researchers

Ellen Collins; Branwen Hide

Research Information Network, 96 Euston Road, London, UK
{ellen.collins, branwen.hide @rin.ac.uk}

Abstract

One of the features of the growth of Web 2.0 resources and services in recent years has been the rapid development of a range of web-based tools designed to allow researchers to generate, modify, share and redistribute information in innovative ways. There has been much discussion about the benefits of using such resources, often accompanied by an assumption, particularly from outside the academic research community, that most researchers will eventually use these tools, if they are not already doing so. However, it is not yet clear whether, why, or to what extent, researchers actually do use them. This work set out to examine the extent to which researchers use Web 2.0 tools and resources. It also set out to identify the factors that influence adoption in order to determine whether these resources are changing researchers' behaviours.

Keywords: Web 2.0, researchers, scholarly communication

1. Introduction

Over the past 15 years the World Wide Web has undergone a massive transformation from a tool for scientist at CERN to a global information source for over a billion users. The web is constantly evolving and over the past few years has entered a more social participatory phase in which the information users become the information provider by creating, sharing, and organizing content [1, 2]. Coined 'Web 2.0' these tools and resources also encourage collaboration and re-purposing of content, as well as supporting users to develop innovative ways to interact with and use these web-based platforms [3].

There is much discussion about the benefits of using these resources for educational and research purposes, with a strong belief that Web 2.0 will enable researchers to create, annotate, review, reuse and repurpose information/data. It is also believed that Web 2.0 will promote new forms of scholarly communications and drive innovation [4]. Thus, it is often assumed, particularly by those outside of the active research community, that a wide majority of researchers are using or will use these tools during the course of their research career. Currently, however, there is little evidence as to the extent to which researchers are using or intend to use these resources. In addition, there is little understanding of the factors influencing the adoption of Web 2.0 tools and resources. The evidence that does exist highlights a number of technical issues, such as the need for standardization, issues pertaining to intellectual property rights (IPR) and the problems that arise when coping with a large amount of information. In addition, some of the factors influencing the adoption of these tools are related to researchers scholarly communications practices, particularly within sub-disciplines, as well as to institutional and organizational issues such as funding and career progression mechanisms [2, 4].

Thus, this paper sets out to examine, in detail, the extent to which researchers at all stages of their career, use Web 2.0 tools and resources, the factors which influence adoption and to determine if using these resources are influencing researchers' behaviours. It also sets out to look at the implications for research practices and policy. For the purposes of this study, and in agreement with previous definitions [1], Web 2.0 encompasses web applications that facilitate interactive information sharing, interoperability, and user centred design as well as placing an increased emphasis on user-generated content. This definition is not limited to technologies, but also includes the changing ways that individuals and groups produce and communicate information [2, 4]. Therefore, there are a large number of web-based tools and resources used by researchers that fall under the term 'web 2.0'. These include generic services produced by commercial providers which are also widely used by the public at large, generic services targeted to the wider research community, services provided by publishers or librarians, and tools adapted or generated for specific research communities or worksites. During the course of this work we have categorized the Web 2.0 tools used by researchers into four distinct but overlapping groups: (1) sites for networking (e.g ResearchGate or Nature Networks), (2) sites designed for sharing information directly related to research practices and methodologies (e.g myExperiment or Methodspace), (3) resources for sharing and commenting on published outputs (e.g. Slideshare or Mendley), (4) tools for documenting and sharing experiences (e.g. blogs or wikis).

2. Methods

Several methodological techniques were used to identify the attitudes towards and patterns of adoption, of Web 2.0 tools and resources by researchers in the UK. Initially, a comprehensive survey was sent to researchers, designed to gather basic demographic information, including age, positions, gender, discipline, dissemination practices, extent to which they engage in research collaborations, use of Web 2.0 tools, and attitudes towards new technology. The survey was sent to 12,000 UK academic email addresses, with an over all response rate of 0.8%. The sample was deemed to be representative of the overall UK research population as it agreed with current Higher Education Statistics Agency [5] data on the UK research population. Researchers were not asked specifically about their use of 'Web 2.0', since many are unfamiliar with this concept. Instead, they were asked about existing scholarly communications practices and techniques, as well as attitudes towards and usage of more novel forms of scholarly communications. Focusing on specific techniques avoided problems around definition and permitted a greater degree of flexibility in analysis of the survey responses. The survey results were cross tabulated and subjected to appropriate statistical tests (Chi-squared from non-ordinal variables, Cochran-Armitage Trend Test for combinations of non-ordinal and ordinal variables and Spearman Rank Correlation for ordinal variables).

The second strand of the research used a series of semi-structured interviews (face to face and by telephone) with a stratified sample of 56 survey respondents, selected as a representative sample of all respondents. The interviews set out to explore how the researchers were making use of Web 2.0 and their perceptions of barriers and drivers to adoption. These interviews illuminated the findings of the quantitative research, and prompted further consideration of the causal relationships identified in the initial survey.

Finally, five case studies examining Web 2.0 based services were undertaken using semi-structured interviews with service developers and users. These case studies were undertaken to further investigate the adoption of web based resources and tools by researchers. The first two case studies, Nature Publishing Group (NPG) and Public Library of Science (PLoS), were chosen to illustrate how commercial and not-for-profit publishers are facilitating and enhancing access to electronic research articles. They also provide functionality for user-generated content. The third case study, Slideshare, was included to demonstrate the value of a community targeted commercial tool. The fourth case study, myExperiment, was included as an example of a researcher-generated tool that has gained support within the wider research community. The fifth case study, arts-humanities.net, highlights the growing development of publicly funded sites which support a specific research community, primarily within the UK. For each case study, several interviews were carried out with developers and users of the service.

3 Results

Use of web 2.0 tools

Table 1 presents respondents' use of a very specific sub-set of Web 2.0 tools, which focus on information sharing, rather than networking or information discovery). Respondents estimated their use of each tool separately, therefore the table does not sum to the total number of responses received.

Table 1: Use of information-sharing Web 2.0 tool

	Non-user	Occasional user	Frequent user
Write a blog	1087	155	51
Comment on other peoples' blogs	978	273	28
Contribute to a private wiki	1066	191	58
Contribute to a public wiki (e.g. Wikipedia)	1072	215	15
Add comments to online journal articles	1023	267	16
Post slides, texts, videos etc. publicly	820	382	80

The 1,282 valid responses where respondents had estimated their frequency of use of each technology were cross tabulated to create a taxonomy of overall usage:

- Frequent users (13% ;175 people) respondents, who do at least one of the activities listed in Table 1 frequently
- Occasional users (45%;589 people) individuals, who do at least one of the activities in Table 1 occasionally
- Non-users (39%;518 people) respondents who never engage in any of the web based activities indicated in Table 1.

According to this taxonomy, the majority of respondents use the specific web based tools/resources listed in Table 1 at least occasionally. As the definition of 'frequent' is weekly, which in the context of communication tools may not seem like habitual use, it is reasonable to suggest that overall use of information-sharing Web 2.0 tools is by no means intensive among researchers.

Table 2 shows usage of these tools, cross tabulated against 'stereotypical' Web 2.0 behaviours by respondents. Respondents were asked about their habits in relation to blogging, social networking and open science (sharing data and work in progress on public fora), and were determined to be users if they engaged with the tool/resource at least once a week. Unlike the frequency categorisations, the behaviours are not exclusive and it is possible for a single respondent to be a blogger, social networker, and open scientist, or none of the above. This is why the table does not sum to the total number of responses. Bloggers are a sub-set of the frequent user group, as blogging is one of the tools considered within the ranking of usage. However, not all frequent users are bloggers. Social networkers and open scientists exist within all three categories, though they remain more concentrated among users.

Table 2: Web 2.0 behaviours and use of information-sharing web 2.0 tools

	Frequent users	Occasional users	Non-users
Blogger	51	0	0
Social networker	51	80	34
Open scientist	36	24	6

Demographic characteristics of users of Web 2.0 tools

Tables 3-5 show demographic characteristics (gender, age, and career progression) cross tabulated against frequency of use. In each case, the characteristics are represented as a percentage of all users in that category, with the total number of respondents in each category (Base) shown at the bottom of each table.

Table 3: Percentage of respondents using information-sharing Web 2.0 tools by gender

	Frequent users	Occasional users	Non-users	All respondents
Female	34%	41%	52%	44%
Male	66%	59%	48%	56%
Missing	1%	0%	0%	0%
Base	175	589	518	1282

Table 4: Percentage of respondents using information-sharing Web 2.0 tools by age

	Frequent users	Occasional users	Non-users	All respondents
Under 25	2%	3%	5%	4%
25-34	26%	25%	28%	26%
35-44	34%	25%	23%	26%
45-54	19%	25%	22%	23%
55-64	16%	18%	16%	17%
Over 65	3%	4%	5%	4%
Missing	1%	0%	0%	0%
Base	175	589	518	1282

Table 5: Use of information-sharing Web 2.0 tools by position

	Frequent users	Occasional users	Non-users	All respondents
Professor	20%	21%	20%	20%
Reader	7%	9%	5%	7%
Senior lecturer	14%	18%	11%	15%
Lecturer	13%	9%	12%	11%
Research fellow	12%	13%	11%	12%
Research assistant	7%	4%	4%	4%
PhD student	19%	23%	32%	26%
Missing	7%	3%	3%	4%
Base	175	589	516	1280

Table 3 shows a clear association between being male and level of usage, which is confirmed using statistical tests (Z=5.52, p<0.001). Tests on the data shown in Table 4 suggest that a greater degree of adoption is positively associated with older age groups (rho=0.05, p=0.048). Tests on the data shown in Table 5 show that a greater degree of adoption is positively associated with more senior positions (rho=0.14, p<0.001). However, since this analysis is not multilinear, no statements can be made about the relative importance of each demographic factor. Furthermore, a relationship between variables, such as age and position, may underpin some of the observed correlations.

When considering the Web 2.0 behaviours, there are further clear demographic distinctions. Being a blogger is associated with males (p=0.07) and discipline (p=0.004), with participation more likely by those in computer science and mathematics as well as arts and humanities. Being a social networker is associated with younger age groups (Z=5.42, p>0.001), more junior positions (Z=4.64, p>0.001) and discipline (p<0.001), with participation more likely again by those in computer science and mathematics, but also by those in economics and social sciences. Being an open scientist is associated with older age groups (Z=1.70, p=0.0089), with males (P<0.001) and with discipline (p=0.009), with participation more likely by those in computer sciences and mathematics as well as arts and humanities as was seen with blogging. However, participation in open science is highly unlikely by those in the medical and physical sciences. As highlighted above in relation to the frequency categorisations, this analysis is not multilinear and there may be masking variables, such as a relationship between gender and discipline.

Attitude to Web 2.0

Table 6 cross tabulates level of usage against degree of encouragement that researchers received from various bodies, including their local research group, their department, or institution, as well as library and information services, computer support services, research funders and conference organisers. There appears to be a correlation between the level of encouragement given and the degree of adoption of web 2.0 tools. In particular, non-users of Web 2.0 tools seem to have a very low perceived level of encouragement from within their local research group. Statistical tests on the survey data also suggest that the level of collaboration is associated with degree of adoption (rho=0.26, p<0.001).

Table 6: Percentage of frequent, occasional and non-users who receive encouragement to use Web 2.0 tools

	Frequent users	Occasional users	Non-users	All respondents
Local research group encouragement	42%	23%	6%	19%
Other encouragement	31%	27%	24%	26%
No encouragement	27%	50%	70%	55%
Base	175	589	518	1282

Findings from the qualitative interviews support this overall conclusion, suggesting that high levels of local support are crucial to encouraging adoption, and that an absence of these can prevent adoption. In some cases, this may be because the researcher has no interest in changing their working practices unless they can understand why it is useful to do so:

> I do need people to recommend why I need to change to use something (Non-user)[1]

In other cases, the desire to do things differently is evident, but the researcher is unsure about how to proceed or needs encouragement to see these new practices as a priority:

> I'm enthusiastic in that I think there's a lot of potential there, but pragmatically I think there are problems still because people don't have the knowledge (…) to make use of it (Non-user)
>
> I can see other people using it and I'd like to be able [to] use it better. I really could do with having a tutorial or something, but I really don't have time to do all these things. (Occasional user)

Several respondents mentioned a lack of support from institutional IT services as a barrier to adoption.

> HEIs put [a] lot of effort into supporting innovations in teaching but little effort into supporting innovations in research' (Occasional user)
>
> 'The blog system is being run by people who we see as not technically competent enough to do it reliably (Frequent user)

Blogs were viewed by one survey respondent as a useful place to further existing connections, via research groups or other networks as they can be private spaces:

> Some of the discussions are sensitive and they want the people involved to be free to say what they want.

Table 7 cross tabulates frequency of use by attitude towards Web 2.0 tools. Very few researchers in any category have a sceptical attitude to Web 2.0, and even fewer are actively uninterested in it. However, frequent users are noticeably much more enthusiastic about Web 2.0 than respondents in other categories.

Respondents were probed more deeply on their opinions about the future of scholarly communications and the role that Web 2.0 tools might take in this. In particular, they were asked to rate the likelihood of formal peer review becoming increasingly complemented by reader-based ratings, annotations, downloads or citations and if either new types of online publication or using new kinds of media formats and content will grow in importance over the next five years. Table 8 cross tabulates frequency of use

[1] All quotes are taken from the interviews and case studies.

by rated likelihood of online supplements to peer review. Frequent users are more likely than any other group to consider online supplements to peer review to be likely. Non-users are less interested in the question, with more of them having no opinion than any other group. Overall, however, most respondents considered online supplements to peer review a likely development in their field.

Table 7: Percentage of frequent, occasional and non-users of Web 2.0 by attitude to Web 2.0

	Frequent users	Occasional users	Non-users	All respondents
Sceptical	6%	8%	10%	9%
Uninterested	1%	2%	4%	3%
Neutral	23%	49%	57%	49%
Enthusiastic	68%	38%	26%	37%
Missing	3%	3%	3%	3%
Base	175	589	518	1282

Table 8: Percentage of frequent, occasional and non-users of Web 2.0 by rated likelihood of supplement to peer review

	Frequent users	Occasional users	Non-users	All respondents
No opinion	10	17	28	20
Unlikely	23	35	33	33
Likely	65	48	38	46
Missing	2	1	1	1
Base	175	589	518	1282

Findings from the qualitative survey, however, show that some researchers remain suspicious about the value of this process:

> Things like citation rates that come out of a formal process can be tracked (…) but reader comments and ratings would be so open to abuse it’s hard to imagine that people would interpret it as valid of the paper’s worth (Non-user).

Table 9 cross tabulates frequency of use with rated likelihood of new types of online publication. All groups of users consider that new types of online publication will grow in importance in their field over the next five years.

Table 9: Percentage of frequent, occasional and non-users of Web 2.0 by rated likelihood of new types of online publication

	Frequent users	Occasional users	Non-User	All respondents
No opinion	5	10	13	11
Unlikely	13	12	13	13
Likely	81	77	73	76
Missing	2	1	1	1
Base	175	589	518	1282

Information seeking practices in relation to Web 2.0

The survey data suggested that researchers continue to place considerable emphasis on traditional forms of scholarly communication. Subscription journals, whether online or in print version, were considered the most important source of research information across all disciplines. 91% of respondents rated online journals as average or high importance, while 89% rated print journals as average or high importance. Conference presentations were also important (82% rated as average or high importance), as were proceedings (71% rated as average or high importance). 65% of respondents rated personal communications as average or high importance.

This finding is reflected in the survey interviews, where researchers placed emphasis on personal networks, and suggested that Web 2.0 tools could be valuable in enhancing their reach:

> Certainly a lot of the articles that I pick up in journals are through verbal face to face recommendations so I don't see why I wouldn't also take an online recommendation if someone in my area in a newsfeed I was to subscribe to would say that this article is important to our area, then I would take that on board and look at it (Non-user)

It is worth noting, however, that this comment is phrased as a possibility rather than a report of existing practices, suggesting that researchers may not currently make extensive use of online research recommendations.

When seeking information, researchers value services such as Google Scholar which increase the visibility of information. As one interviewee put it, the service is 'particularly useful for looking up some papers that are online but not published yet'. For most researchers, wikis and blogs were not used as a source of information, due to concerns about unreliability:

> [I] wouldn't use Wikipedia or anything like that, anything that isn't peer reviewed like that is worthless

One researcher suggested that even blogs associated with established journals were viewed with some suspicion as a source of high-quality information:

> [blogs are] not taken very seriously, even blogs based on Nature [colleagues] find it time consuming and not very credible, interesting, yes, but it's almost regarded as a piece of entertainment first and potentially useful almost serendipitously.'

However, another researcher suggested that Nature blogs had helped him to build connections within the discipline. He used the blogs:

> for searching for and about information regarding our research, with our collaborators (...) it's very useful because you get to know what other people are doing, getting to know [a] network of people. Once I saw a relevant paper written by a person in Canada, so I wrote to him to send me some of the things he was using, and within two days he sent me everything, you know. So, out of this system we are able to collaborate too, getting to know other people's work and if they are doing similar things to us, we can get in touch with them and ask questions and share ideas.

This interview data suggests that where blogs are used it is not necessarily to find information per se, but rather to connect with people or organisations that might inform a researcher's work. Web 2.0 resources, even when associated with a trusted source, have a credibility problem which prevents their widespread use as a source of research information: they have more value as a tool to increase discoverability.

Information dissemination practices in relation to web 2.0

The survey showed that, when rating methods of research dissemination, respondents continue to place considerable emphasis on traditional, peer-reviewed outputs. The exact form of these varied between disciplines, with conference proceedings rated highly in some disciplines and monographs considered more important in others. An interesting disparity arose between online and print journals: print publications were rated as very important by 70% of respondents, while only 56% rated online-only journals as very important. This may suggest that non-peer reviewed online resources such as blogs and wikis will struggle to be accepted as long as established and peer reviewed online-only journals are valued so much less than their print counterparts.

However, interview data suggests that researchers' concerns about disseminating information via web 2.0 tools were not linked to the credibility of these new media, but rather their likely impact. One non-user described novel forms of scholarly communications as a 'waste of time', and another said that 'I'd rather spend the time thinking about what I'm going to do next rather than spend it telling others what I'm doing'. Even frequent users did not necessarily begin from the assumption that the tools they were using were useful:

> People are very keen to have unconventional dissemination practices, but I think it all boils down to whether they will be valued.

In some instances, this attitude was the product of previous, unsuccessful, attempts to use Web 2.0 tools:

> The institute had a blog for two years, but we actually gave it up, because it wasn't the interactive service we thought it should be (...) nobody really commented. (Non-user)

However, frequent and occasional users did value the visibility that blogs bring to their authors.

> If it increases your profile and more people were aware of the work you did, that would be a benefit (Occasional User)
>
> There are career benefits too. Those working in the media field who are actively using these materials and are perceived to be on the 'cutting edge' are often very successful. (Frequent user)

This visibility was considered particularly useful in cases where it helped to build collaborations, share preliminary findings and increase the speed with which other researchers could see work. This reflects the use of blogs and social networks for information seeking, as outlined in the previous section.

> It is of big value to be able to communicate with academics from all over the world (Frequent User)
>
> It almost offers you a halfway house in that you can be less formal, you don't have to have completed your research project, you can talk about your research findings, as it were, and it's kind of put out there in the public space, and people can comment or interact without having to wait until your final output is a journal article that will appear in print. (Frequent User)

Within this consideration of information dissemination and Web 2.0, it is worth turning a more focused attention to the practice of open science. Open scientists formed a very small minority in the research sample, but they are particularly strong proponents of web 2.0 as a way to improve the practice of science:

> You can have a 'conversation' of more than just two-way. Other people can be watching the conversation. That's quite useful. They can contribute if they want and you can always make it private. (Open Scientist)
>
> Ultimately, it will change how people do research (...) It is about accelerating the research cycle for small pieces of research that are easily distributed. (Open Scientist)

Outside the open scientist group, some respondents were broadly supportive of the concept, without entirely understanding what it meant:

> I presume it's concerned with the production of papers and research materials that placed in some publicly accessible place. I support it, yes. (Occasional user)

However, many were concerned that the practice of open science would interfere with the established procedures which make up the so-called 'minutes of science', thereby leading to confusion:

> I do not support open science and I do not see any benefits for me. I have a negative attitude to use blogs and videos in research. Once it's finished it should be published otherwise it will be anarchy in science. (Occasional user)
>
> In our university we have a certain guideline what may or may not be put onto the blog. I have to agree that something needs to be saved and I don't want people to say: we just discovered X. (Occasional user)

This consideration of the very specific behaviour of open science reflects the overall shape of researchers' attitudes to disseminating information via Web 2.0. A small minority are actively engaging with these new publication techniques, using them to share information and hold conversations. Many are not yet engaged with the tools, and view them with some suspicion. And indeed, even the most enthusiastic proponents do not see online publication as an alternative to established peer reviewed journals.

Development and delivery of Web 2.0 services

The case studies were undertaken in order to understand more about the ways in which Web 2.0 services are developed and delivered, and users' perceptions of these services. The five case study organisations were selected to represent a range of tools, business models and disciplinary focuses.

The need to respond to a perceived gap in existing services was a stimulus for the development of many Web 2.0 tools. myExperiment, for example, was established as a tool to encourage and support scientists to share their methods as well as their data and research findings: this was felt to be an unusual practice for scientists. Similarly, SlideShare was established because the developers felt that there was no straightforward way for those in business or academia to share their presentations online. Developers seem to work in a process of continual innovation, seeking out new gaps to be filled as their service grows. In 2009, for example, Public Library of Science (PLoS) designed a strategy to explore article-level metrics of impact, in response to the perceived failings of journal impact factors. This was a logical development from its original purpose as a publisher, and later repository, for academic articles. Similarly, Nature Blogs was developed by Nature Publishing Group (NPG) in response to the observation that, while their own commenting facility for articles was not well-used, conversations about those articles were taking place elsewhere in the blogosphere.

As suggested by this latter example, innovations can also result from developments in other areas of online communication. NPG developed social bookmarking system Connotea as a scholarly version of the existing tool del.icio.us. PLoS worked in partnership with a new Google tool to develop their pre-publication service, PLoS Currents. Such innovations have varying degrees of success: Connotea, for example, is not the market leader and has not been heavily invested in. Nonetheless, NPG is investigating the ways in which it could provide useful data on the use of the tool and journal in order to inform the Group's wider management decisions. This repurposing of a 'failed' experiment illustrates the constant innovation practiced by many web 2.0 service providers.

Such innovation is often driven by developers' notions of what researchers might find useful, as with NPG's development of Connotea and Nature Blogs. However, it is sometimes undertaken through direct contact with researchers, talking to them to discover what they want from a social networking tool and using this information to build new platforms and services. Some products use relatively light touch forms of user feedback to inform their development. SlideShare, for example, places considerable value on user feedback communicated via emails and blog posts, and finds this far superior to more formal market research exercises. NPG gathers feedback from users directly via blogs, mailing lists and fora.

Others, however, use a much more intensive process when developing their services. myExperiment is an excellent example of this. During the development process, software developers were embedded in the lab with scientists so that they could understand exactly how researchers approached and carried out their work. Throughout the development stages, researchers worked with focus groups and other feedback mechanisms to regularly test their ideas, ensuring that scientists' views were taken on board. Developers were also present at introductory sessions for the scientists new to the service, who went on to become strong advocates to encourage wider adoption. This illustrates the benefits of user-centred research and development, which can ensure that the end result meets practitioners' needs.

Producing a service that researchers actually need is an important part of ensuring usage. However, as myExperiment demonstrates, it is also important to actively engage with the proposed user community, to ensure that they understand what they can gain from novel tools. SlideShare ascribes some of its success to ongoing work with other services: it has created plug-ins for Facebook, LinkedIn and PowerPoint to make its product more easily accessible. The capacity to embed presentations in a blog post has also been an important factor in SlideShare's success, as it encourages bloggers to use the service and thereby raise awareness of the service among their readers. arts-humanities.net noticed a relatively large increase in membership of their site following a concerted publicity effort involving emails, announcements and conferences within the community. Feedback suggests that once unengaged researchers are informed about the site they become very enthusiastic. This could mean that for some tools an important barrier to adoption may simply be lack of awareness.

In other instances, however, there are more complicated barriers to widespread adoption of Web 2.0 services. SlideShare users cited concerns about intellectual property rights, privacy and data protection, and had reservations about entrusting their valuable content to externally-hosted systems. This concern had been identified as a possible barrier to adoption by myExperiment, and so their service offers different levels of sharing. However, they felt that potential users needed better education about the high levels of security offered by the site, as fears about loss of intellectual property persisted. NPG noticed that their initial facility for commenting on articles was less successful than some of their competitors, such as the British Medical Journal (BMJ). They suggested that this could have been because comments at the BMJ are a formal extension of the letters to the editor, with a D.O.I. allocated to the contribution, thereby 'rewarding' the commenter.

Another barrier to widespread adoption is the number of Web 2.0 services available to researchers. Several users cited this as an issue, highlighting the amount of time required to sign up to and experiment with new technologies, and even the number of usernames and passwords that must be remembered. Another issue was the specificity of many of these web 2.0 tools, either in terms of the services they offer or their subject remit. One interviewee commented that 'Nature is very focused on certain parts of research, so doesn't allow me to follow other kinds of research'. It is not clear that individual services can do a great deal to overcome this particular barrier.

In terms of adoption, many of the case study services highlighted the different levels of engagement by researchers. SlideShare, for example, uses an 'influence pyramid' to describe its user groups: a small segment at the top who actually upload information, a larger segment who increase the visibility and usefulness of this content by commenting on it or tagging it, and the vast majority who just watch or download presentations. arts-humanities.net notes a similar hierarchy in relation to its users: it has 1,200 registered members but attracts between 6,000-7,000 unique visitors each month. Several of the case study organisations are actively targeting high-profile researchers to encourage them to use their services, in the hope that they will generate increased usage within their communities.

4. Discussion

The data collected through the survey, interviews and case studies presents a relatively detailed picture of researchers' usage of Web 2.0. Overall, it seems that researchers are not overtly hostile to new forms of scholarly communication, and that some are experimenting with these techniques. However, routine use of Web 2.0 tools does not seem to be widespread.

The cross tabulation between frequency of use of certain tools and Web 2.0 behaviour, shown in Table 2, presents some interesting results. In particular, there are a small number of open scientists sharing their data and work in progress on a regular basis, but not using blogs, wikis, comments on journal articles or slide, text and video sharing on a regular basis, prompting a question about how they are sharing this information. There is a further group, small but not insignificant, of open scientists who undertake these activities only occasionally. This suggests that the Web 2.0 tools under investigation may be seen as a convenient way to communicate when occasion demands, but not a regular and routine part of working practice. This reaffirms that it is difficult to consider researchers' use of 'Web 2.0' as a single phenomenon, since it is perfectly possible for people to be using some aspects of it frequently, while ignoring other aspects or using it only when it meets their needs.

The associations between demographic variables and frequency and type of use must be treated with a degree of caution, but remain valuable. Social networking, in comparison to blogging and open science, seems to be more important for younger and more junior researchers. This may reflect the fact that younger researchers are exploring new ways to establish their professional networks, while older and more experienced researchers have already created theirs without virtual aids. It may also be that younger researchers are more likely to use social networking tools in their personal lives, and are therefore more familiar with them and potentially more aware of their potential value in a professional context. The positive relationship between professional level and engagement with web 2.0 techniques is also interesting. This could well be prompted by a concern on behalf of more junior researchers to focus on established communication channels such as peer reviewed journals, which will have the most impact on their career and promotion prospects. More senior researchers, with an established professional reputation, may be freer to experiment with novel ways of communicating their research.

Researchers appear to be strongly influenced by their wider professional environment in their use of web 2.0 tools. The association between level of collaborative working and uptake of web 2.0 tools could be due to researchers adopting new tools and technologies to further existing collaborations. However, it is also possible that being part of a collaborative network helps researchers to discover, use and advocate new ways of working. In terms of the future of scholarly communications, researchers in all categories considered an increasing importance of new types of online publication in their field within the next five years to be likely. This is particularly interesting in light of the relatively low levels of usage of existing types of online publication such as blogs, wikis and file sharing. Researchers expect a future scenario where online publication is more important, but are not engaging with tools which could perhaps be the precursor to these new media.

Researchers' attitudes to Web 2.0 tools as a way of discovering and publishing information are decidedly mixed. The survey showed that most researchers are engaging on at least an occasional basis with Web 2.0 tools. However, the qualitative interviews suggest that only a few are using them in a systematic way as part of their investigative work. Both information seeking and dissemination are thought to benefit from Web 2.0 tools, as they improve discoverability of information and help researchers hold more effective conversations. These may be with existing members of the research team, or with other researchers in the field; sometimes the researcher would not have become aware of these people without the intervention of Web 2.0 tools. However, it is clear that online tools are valued only as a route to access high quality, trusted information: they are not seen as a source for such information. Researchers have strong reservations about the accuracy of information online. They are also dubious about the impact of Web 2.0 tools, recognising that a specific tool needs to be relatively widely used and accepted if information published on it is to have any value. Some researchers also expressed concerns that informal web-based dissemination of information could have a negative impact upon the published record of science, and felt that this could retard scientific discovery.

The case studies show that a researcher's Web 2.0 environment is somewhat complex, with a wide range of tools to choose from. Service providers are constant innovators, and often recognise the importance of engaging with researchers in order to create services that are useful. Usage of services is varied, but where services are not successful they are quickly dropped or re-purposed to better meet researchers' needs, as happened with NPG's commenting features. However, many services acknowledge that their user base is very diverse, and that it is challenging to create tools which will meet the needs of everyone. Researchers themselves do engage with these services, but retain some reservations, the principal of which is around intellectual property rights. It will be a challenge for individual service providers to overcome this concern, as evidenced by myExperiment's attempts to show researchers that their work is protected on the site.

5. Conclusion

Overall, it appears that researchers are not engaging systematically with Web 2.0 tools. They are broadly interested, and many are infrequent users of these tools. However, they do not form a core part of most researchers' working practices. Researchers value the increased visibility that Web 2.0 tools can give to research findings, but they do not hold information published via the web in equal esteem with peer reviewed journals. And while many believe that online publication tools will be increasingly important in future, relatively few are engaging with existing options such as blogs, wikis and file sharing. Web 2.0 services are rapidly evolving to attempt to meet researchers' needs, but are aware that their user base is very diverse and overwhelmed with a range of possible technological solutions to research problems.

This project suggests that any systematic changes to researchers' use of Web 2.0 tools will need to be supported by various bodies with a role in the research process, but in particular by local research groups. Web 2.0 service developers are seeking to engage high-level academics to encourage wider uptake within specific fields; there may be some virtue in examining the potential of this model more widely to encourage researchers to engage with generic Web 2.0 tools such as blogs. There are also some significant barriers to overcome. In particular, the issue of intellectual property rights and ownership of data, methods and tools must be resolved, and researchers must receive enough information to feel secure that their work is protected. There is also a credibility issue to be addressed, as researchers continue to be suspicious of information published using Web 2.0 tools, even when associated with reputable and established sources.

Future research could usefully undertake more complex multivariate analysis on the data to establish the relationship between possible causal factors for level of Web 2.0 usage. It would also be interesting to consider why, given that most researchers believe online communications will become increasingly important in future, so few of them are choosing to engage with these techniques while they are still at a relatively formative stage.

Acknowledgement

We wish to acknowledge Rob Procter from the Manchester eResearch Centre, University of Manchester, as well as Robin Williams and James Stewart from the Institute for Study of Science, Technology and Innovation, The University of Edinburgh who undertook this work on our behalf.

Notes and References

1 ANDERSON, P. What is Web 2.0? Ideas, technologies and implications for education. Available at http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf.

2 RIN. Use and relevance of web 2.0 resources for researchers. to be published May 2010, Submitted, Available at http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/use-and-relevance-web-20-researchers.

3 WARE, M. Web 2.0 and Scholarly Communication. Available at http://mrkwr.files.wordpress.com/2009/05/ware-web-2-0-and-scholarly-communication-preprint.pdf.

4 PROCTER, R; et al. Adoption and Use of web 2.0 in scholarly communications. Philosophical Transactions of the Royal Society A, Submitted

5 HESA Available at http://www.hesa.ac.uk/index.php/content/view/600/239/

What are your information needs? Three user studies about research information in the Netherlands, with an emphasis on the NARCIS portal

Arjan Hogenaar; Marga van Meel; Elly Dijk

Royal Netherlands Academy of Arts and Sciences, Research Information,
P.O. Box 19121, 1000 GC Amsterdam, The Netherlands
Arjan.Hogenaar@bureau.knaw.nl

Abstract

The NARCIS portal (www.narcis.info) provides access to science information (information about research, researchers and research institutions) and scientific information (full-text) publications and datasets. The portal is very popular, with 1.2 million users annually. NARCIS is also an important supplier of information to international services such as Google/Google Scholar, WorldWideScience.org and DRIVER. In 2009, the KNAW conducted a three-part user survey, with two online surveys and a series of semi-structured interviews. The aim was to learn more about the people who use the portal, why they use it and their ideas and wishes for improvements to the portal. Another purpose of the survey was to identify changes that could be made to improve the match between the services provided by NARCIS and the needs of existing and potential users. Surveys showed that more than half the users of NARCIS are from universities, research institutions or universities of applied science. Most searches conducted on NARCIS are for dissertations. The existence of a single gateway to different types of information is regarded as very useful. The most frequently mentioned improvement in the service would be to provide access to information from other countries as well. Respondents also mentioned the provision of *tools* for performing complex analyses of the information available via NARCIS as a worthwhile option for enhancing the service. The interviews revealed, among other things, the need for the presentation of information in context and that senior officials are often confronted with information overload. The user survey has led to a series of proposals for modifications or improvements in the service; some of them may be implemented immediately, while others will require consultation at national or international level.

Key words: user survey; questionnaires; portal; evaluation of integrated services

1. Introduction

The most important task of the Research Information (KNAW-OI) department of the Royal Netherlands Academy of Arts and Sciences (KNAW) is to help national and international users to find information about research, researchers (and their expertise), research institutions and the results of research (publications and datasets) in the Netherlands.

The Dutch Research Database (NOD)[1] is a service provided by KNAW-OI and forms the basis for its role as the national focal point for research information. Before 2005, the KNAW was involved in the development of DAREnet (network for Dutch Academic REpositories) [1], at the time the central portal for access to publications in the repositories of research institutions. Since then, KNAW-OI has been developing the national focal point for research information and research results at European level. The final result is NARCIS (National Academic Research and Collaborations Information System) [2].

NARCIS now plays a central role in searching all research-related information in the Netherlands and serves as the national showcase for researchers working in the Netherlands. Via NARCIS users have access to both the information from the Current Research Information Systems (CRISs) and the information from the Open Access Repositories (OARs).

A problem is that in the Netherlands the (OARs) and the CRISs generally fall under different organisational units of the universities: the libraries or the research administration departments. The datasets, if they are kept at all, are stored in the DANS[2] system. With all these different systems, it is inevitable that variant versions of the names of authors and researchers are in circulation. To cross-reference the different information types, unique Digital Author Identifiers (DAIs) are used. Every author or researcher is assigned a personal DAI, which creates the Academic Information Domain [3], the domain where all information relating to research is collected. Thanks to the

[1] www.researchinformation.nl

[2] Data Archiving and Networked Services (www.dans.knaw.nl)

DAI, a personal page can be compiled in NARCIS for every researcher, containing a complete overview of his or her research, publications and datasets in context, as illustrated by the example for Professor W.H.J. Meeus.[3]

NARCIS already offers users many useful functionalities such as RSS feeds, the Zotero[4] reference tool and personal pages for researchers. The portal is visited 1,200,000 times a year by researchers, policy makers, people in the media and members of the general public. The proportion of Open Access publications available is rising steadily, as Peter Suber has observed [4]. Thanks to NARCIS, these Open Access publications can be traced quickly and easily.

It is easy to discover how often NARCIS is used from the log data. In keeping with the department's tradition of conducting regular surveys, KNAW-OI conducted a user survey in 2009 to identify who the users are, where they come from and what they use NARCIS for.

2. Methodology

2.1 Analysis of IP addresses

The simplest way of discovering who is using the NARCIS portal is to check the users' IP addresses.[5] We identified the IP addresses of the 400 most frequent users (in terms of the number of NARCIS views) in January 2010 (through AW-stats[6]).

Those IP addresses were then linked to the names of institutions using IP locators Topwebhosts[7], Geobytes[8] and ip2locations.[9] In Table 1 those institutions are broken down into the following categories: University, Research institution, University of Applied Sciences, Government, Not-for-profit sector, Hospital, Business, Media and Provider.

Table 1: Share of users in each category

Category	Share of NARCIS use

[3] http://www.narcis.info/person/RecordID/PRS1237369

[4] http://www.zotero.org

[5] For reasons of privacy, no attempt has been made to connect IP addresses to individual users.

[6] http://www.ubiquityhosting.com/web-hosting/service/awstats

[7] http://www.topwebhosts.org/

[8] http://www.geobytes.com/ipLocator.htm

[9] http://www.cqcounter.com/whois/domain/ip2locations.com.html

University	36%
Research institution	4%
University of Applied Sciences	8%
Government	6%
Not for profit	2%
Hospital	1%
Business	11%
Media	1%
Provider	32%

The last category, 'provider', is a very special one: many users access NARCIS via a provider. This means that the KNAW can see which providers were used, but naturally cannot identify the individual or organisation that visited NARCIS via those providers.

2.2 Online surveys

In addition to this analysis of users on the basis of IP addresses, we also conducted two online surveys to discover more about the work environment and the professions of the users. The surveys were also designed to learn more about the types of information the users were searching for, how they rated the different functionalities in NARCIS and what developments they would like to see in the system. The two online surveys were compiled by using SurveyMonkey.[10]

The first survey was held in June 2009. It could be completed only via the NARCIS website (www.narcis.info), so that only actual portal users were aware of the survey's existence. In view of NARCIS's international character, the survey was presented in both Dutch and English. The participants were asked whether they would also be willing to take part in a follow-up survey. The survey was deliberately kept short and confined to just six questions.

The Dutch-speaking respondents who had said they were willing to participate in a follow-up study were asked to take part in another study, again compiled in SurveyMonkey, in December 2009.

As already mentioned, the two surveys were completed only by actual users of NARCIS. They left two questions unanswered:

[10] www.surveymonkey.com

- To what extent does the target group for whom NARCIS may be relevant actually use the service? Can any conclusions be drawn about non-users?
- To what extent could NARCIS be useful for non-users, given their information behaviour?

2.3 Interviews

To answer these questions, semi-structured interviews were held with 17 individuals in the final quarter of 2009. It was not known in advance whether or not they used NARCIS, but it was known that they all held senior positions in which they handled a lot of information.

The point of departure for the interviews was to identify the information ecology [5] of the interviewees, in other words what technologies they use to search for and process information.

These 17 individuals represented NARCIS's various target groups: nine researchers [humanities (3), natural sciences (3) and social sciences (3)], four policy makers, two information specialists and two journalists. This method corresponded with that used in a previous survey of needs in 2002 [6], although that earlier study related exclusively to the Dutch Research Database (NOD).

A script was written in advance to ensure that at least the following subjects were discussed with the interviewees:

- What sources of information do they use?
- How do they search for information?
- What problems do they encounter in searching for information?

Although the interviews related to the respondents' general information needs and information behaviour – in other words, their information landscape – the interviewers focused mainly on NARCIS's potential role in it.

Each interview lasted one hour. The interviews were all recorded and a report summary of each interview was produced according to a fixed format. The reports were all approved by the interviewees.

3. Results

3.1 Surveys

There were 434 respondents in the first survey, of whom 268 completed the full questionnaire; 61 respondents completed it in English.

Of the respondents, 59% were from universities, universities of applied science or research institutions. Another 15% were from the business community and 15% were from the not-for-profit or public and semi-public sector (see Figure 1).

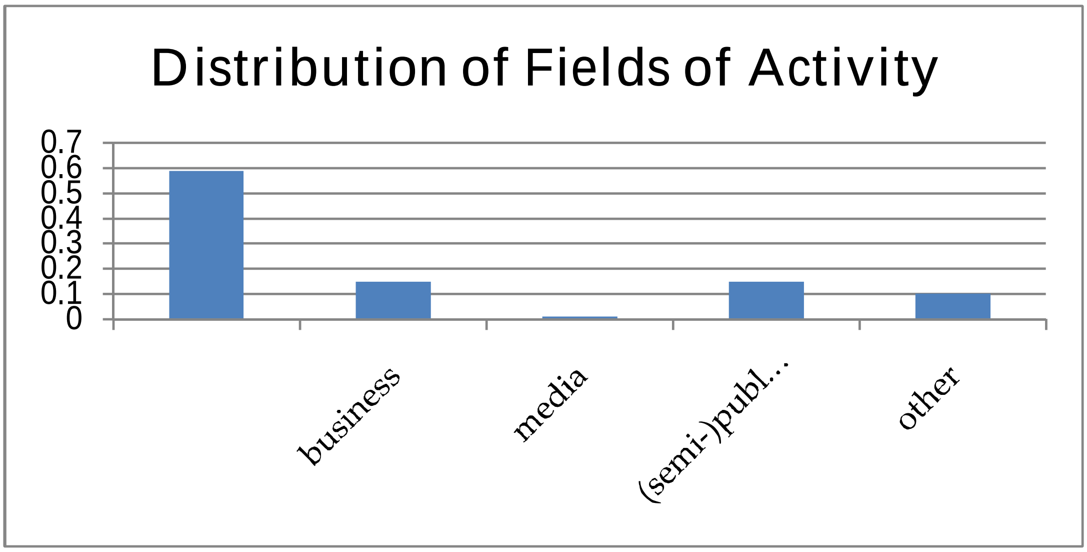


Figure 1: Distribution of fields of activity

As mentioned, it was already possible to gain an impression of the fields of activity of the users of NARCIS by identifying the names of the institutions corresponding with the IP addresses of the 400 most frequent users in January 2010. That analysis showed that at least 48% of the users were from universities, universities of applied science or research institutions, a figure that corresponds closely with findings from the survey, especially bearing in mind that quite a number of those who visit NARCIS via a provider have a position at one of those research institutions.

Thirty-six percent of the respondents described themselves as researchers and 21% as information specialists. Few described themselves as members of the other professions (e.g., policy assistant, journalist). Many respondents answered the question concerning their profession in their own words. Analysis of the information they provided suggests that almost half of those surveyed can be described as academics.

During a six-month period, 21% of the respondents use NARCIS more than 10 times, 18% use NARCIS between four and 10 times and 60% use it between one and three times. Relatively speaking, information specialists use

NARCIS most frequently (58% use NARCIS four times or more in a six-month period).

NARCIS users search mainly for dissertations, other types of publication and information about researchers. The number of searches for datasets is remarkably small (7%).

Being able to download full-text publications was mentioned as the most important feature of NARCIS by 78% of the respondents, while 60% regard the links to additional information (for example, from the description of a person to his/her publications) as important. Other important features are being able to search simultaneously in different information types as well as the presentation of an individual's entry in combination with all the relevant information about him or her.

Asked to say what they felt the most interesting development would be for NARCIS, 57% of the respondents mentioned the presentation of similar information from other countries. Other frequently mentioned suggestions for upgrading NARCIS were to make improvements in its functionality (for example, the possibility to browse) and to offer tools for complex analyses.

The follow-up survey was held among a sub-population of the respondents in the first survey, but with a similar composition. The purpose of this survey was to find out how the respondents rated the functionalities and content offered by NARCIS. For 95% of the respondents, having a single gateway to different types of information was felt to be useful or very useful.

The respondents were impressed with the option of searching on (full-text) publications and on current research. However, also this group of respondents – who search in NARCIS more frequently than the wider group of participants in the first survey – does not perform many searches on datasets.

Although the search options are highly rated (79% of respondents were satisfied or very satisfied), the respondents were not always aware of all the search functionalities in NARCIS (combining search terms; tailored RSS feeds). The respondents were most impressed by the large number of Open Access publications available through NARCIS and the overview of experts. They were also pleased with the response time.

3.2 Interviews

The interviews gave an impression of how the interviewees are using information and of what could be the potential importance of NARCIS for them. The demand for the information in NARCIS differed from one interviewee to another, so it is impossible to draw any general conclusions from the interviews. Nevertheless, a certain trend could be discerned in the interviews. The various information types to be found in NARCIS are briefly described below.

Information about individuals, organisations and current research

Information about individuals and organisations is occasionally important for researchers and non-researchers alike. Researchers use this information as background material to help them assess the value of a particular individual or organisation's publications. Non-researchers are often searching for experts in a particular field in order to gather more information.

The interviewees often have their own network and do not need to consult a database, unless they want to explore a new area or need information about less well-known individuals.

Information about current research is important to gain an early impression of work being done new fields of research.

Information sources

The interviewees use a variety of channels to gather information. The sources mentioned include those available via the Digital Library of the respondent's own university, preprints, search engines (Google, Google Scholar), personal contacts and participation at conferences and workshops in the Netherlands and elsewhere. Blogs and Twitter were also mentioned as a source of very up-to-date and opinion-forming information.

Dissertations and datasets are especially important for researchers; dissertations are a particularly valuable source of information outside the natural science sector. The same applies for other types of publication of a monographic nature. While many dissertations are nowadays available in electronic form, this is unfortunately not true of monographs in general. Nevertheless, it is shown that there is a growing demand for digital monographs [7].

In the Science-Technology-Medicine (STM) sector, the interviewees were more interested in journal articles (which also are the main elements of the dissertations in this sector).

Policy makers seek inspiration from the results of research to formulate and roll out new policy, while journalists report on that research. At most, dissertations and other scientific publications are useful to them as background information.

Most researchers and non-researchers subscribe to services alerting them to new information. One disadvantage of this method that was mentioned was that it causes information overload.

A noteworthy finding was the importance the interviewees attached to personal networks, including online networks. Some have created their own networks and they often also establish special interest groups on networks such as LinkedIn.[11] Scientific information is quickly disseminated in these networks. Trust is important in this context, which is why the digital networks are built on existing networks in real life.

Datasets

Datasets are mainly important for researchers. These datasets may consist of statistics but may in a broader sense also include, for instance, audio and video recordings. There is a certain tension between, on the one hand, the desire to write publications based on one's own raw material first, and sharing and re-analysing this material on the other.

At most, non-researchers need pre-packaged statistical information.

Context

Many researchers refer to the importance of the context of the information they find. This relates to a functionality such as links (for example from the raw data to the related publications), on the one hand, and the presentation of background information (about the author, the organisation or the research programme) relating to the information that has been found, on the other. In this way, the user can assess the merits of a particular source.

Problems in searching for and selecting information

Examples mentioned by the researchers include:

- *Quality:* it is not always easy to distinguish between information of a high quality and less valuable information.
- *Accessibility*: publications are not always available under Open Access.
- *Coverage*: a lot of material that is relevant for research and education is not available.

[11] http://www.linkedin.com

- *Context:* search engines like Google provide no information about the context.
- *Information overload:* search and alert systems are not intelligent enough, which results in a surplus of information or in irrelevant information.
- *Persistence:* researchers and documents are difficult to trace permanently on the Internet. Assigning Digital Author Identifiers and persistent identifiers to documents could solve this problem.

The non-researchers report the following problems:

- *Absence of very concise abstracts* of scientific publications
- *No free access* to some texts
- *Difficulty in finding experts* (who are needed to assess the content of news items)

Suggestions for improving the NARCIS service were made in both the surveys and the interviews. Some of the suggestions may be put into effect immediately, but some call for national or international consultation. Some of the most imaginative suggestions were:

Intelligent search and alert systems; text mining; internationalisation; permanent storage of new information types such as blogs; access to enriched publications [8, 9].

4. Discussion

The surveys have shown that a significant number of NARCIS users come from universities, research institutions or universities of applied science. They are the portal's principal target group. At the same time, it became apparent that NARCIS users are often unaware of the possibilities of the portal. For example, they are not all aware of the possibility of combining terms in a search command and do not all take advantage of the benefits of the customised RSS feeds. It is very important to display these options more clearly in NARCIS.

The interviews revealed that the interviewees first consult Google (Scholar) when searching for information. Only the biomedical specialists among them also use PubMed.

However, these interviewees are also aware of the limitations of the giant Google, the most prominent being information overload and the uncertainty about the quality of the information that is found.

The information overload can be eased by introducing the option of personalising the presentation of information in the NARCIS system by giving the greatest prominence to the information types that are most relevant to him or her.

Although search engines, and particularly Google, are popular, the interviewees did say that they would like information to be presented by subject. To present information by subject (for example, on the topic of historical sciences), a service depends on the metadata that is supplied.

Past experiments by the KNAW with tools for automatic categorisation suggest that it does not lead to acceptable results in a multidisciplinary database. Thematic presentation might be possible with Web 2.0 facilities (along the lines of Flickr), with users applying their own tags to information objects.

The NARCIS information is already highly accessible in Google. Google often shows users information from NARCIS without their realising it. Google can therefore be regarded as a supplementary source of access to NARCIS. The benefit for the user of searching directly in NARCIS is the availability of additional functionalities that Google does not offer. A public relations effort is needed to inform users of these extra options in NARCIS.

This user survey was confined to the reaction of human users. However, a service like NARCIS is also for non-human users. For example, NARCIS provides crucial information to services that operate at European and global level (DRIVER [12] ; Scientific Commons [13] ; Google (Scholar); WorldWideScience.org[14]). It is clear, for example, that the co-ordination between NARCIS and Google works well from the fact that the website www.narcis.info has the high page ranking of '8'[15], a scale that is awarded only to one percent of the websites displayed by Google.

According to the interviews, NARCIS is not adequately promoted or publicised. None of the senior officials interviewed use the service, which is not surprising since until 2010 there had never been a publicity campaign for NARCIS. The launch of a new version of NARCIS in March 2010 is now being used to bring NARCIS to the attention of a wider public.

The campaign will stress the key role that NARCIS plays in the Dutch national information landscape. By using Digital Author Identifiers (DAIs) and showing relationships between types of information, NARCIS is the leading site for searching for and finding scientific information in context.

[12] http://search.driver.research-infrastructures.eu/

[13] http://www.scientificcommons.org/

[14] http://www.Worldwidescience.org/

[15] http://www.thegooglepagerank.com

However, broader applications of identifiers are possible, particularly identifiers for persons, who may act as a researcher, as an author or even as the subject of a study. A number of interviewees suggested enriching the identifiers with a definition of the various roles as a way of improving the system.

The NARCIS concept is unique in the Netherlands, and even in Europe. There are no other services that provide a combination of scientific information (publications and datasets) and science information (information about researchers, research, research institutions).

The integrated supply of so many types of information automatically creates a desire for more complex text-mining tools, which can display clusters of researchers or publications. Naturally, that implies that users must be able to visualise the results of these analyses.

5. Conclusions

The user survey in 2009 proved very useful. The surveys gave an impression of the backgrounds of the NARCIS users, of the NARCIS functionalities they appreciated and of possible improvements. The interviews provided an understanding of the information needs of persons who use information a lot but are not yet familiar with NARCIS.

The surveys reveal that more than half of the NARCIS users come from universities, scientific institutions or universities of applied science. Most searches in NARCIS are for dissertations.

The most valuable functionalities are the option of downloading publications, the links from individuals to publications and the ability to search simultaneously for different information types. The existence of a single gateway to different types of information is regarded as very useful.

The possible upgrade that was most frequently mentioned is access to information from other countries. Providing tools to perform complex analyses on the material accessed via NARCIS was also mentioned as a worthwhile option. The conclusion to be drawn from these responses is that there is room for a service like NARCIS alongside a ‘one size fits all’ search engine like Google. NARCIS can already largely meet the wishes of the interviewees in its current form. With just a few minor modifications – such as the introduction of the possibility of browsing through the information – NARCIS will also be far better equipped to their needs.

The value of NARCIS can be further increased by intensifying the publicity about the portal and by continuing to improve the service.

Acknowledgements

KNAW-OI is very grateful to Repke de Vries and Henk Voorbij for their assistance in preparing and carrying out the user survey and processing the results.

Notes and References

[1] WAAIJERS, L. The DARE Chronicle: Open Access to Research Results and Teaching Materials in the Netherlands. Ariadne 53, October 2007. Available at: http://www.ariadne.ac.uk/issue53/waaijers/ (March 2010).

[2] DIJK, E; BAARS, CHR.; HOGENAAR, A.; MEEL, M. VAN. NARCIS: The Gateway to Dutch Scientific Information. Paper presented at ELPUB Conference 2006, Bansko, Bulgaria. Available at: http://elpub.scix.net/cgi-bin/works/Show?233_elpub2006 (March 2010).

[3] BAARS, CHR; DIJK, E; HOGENAAR,A; Meel, M. VAN. Creating an Academic Information Domain: a Dutch example. Paper presented at EuroCRIS 2008, Maribor, Slovenia, 2008. pp. 77-78. Available at: http://www.eurocris.org/fileadmin/Upload/Events/Conferences/CRIS2008/Papers/cris2008_Baars_Dijk.pdf (March 2010).

[4] SUBER, P. Open Access in 2009. SPARC Open Access Newsletter, issue #141, 2010. Available at: http://www.earlham.edu/~peters/fos/newsletter/01-02-10.htm#2009 (March 2010).

[5] NARDI, B.O.; O'DAY,V.L. Information Ecology : Using Technology with Heart. First Monday, 4(5), May 1999. Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/672/582 (March 2010).

[6] KOOPMANS, N.I. 'What is your question? The need for research information from the perspective of different user groups' in W. Adamczak and A. Nase (eds). *Gaining Insight from Research Information: Proceedings of the 6th International Conference on Current Research Information Systems*, *University of Kassel, August 29-31, 2002* (Kassel) pp.

183-192, 2002. Available at: http://www.upress.uni-kassel.de/online/frei/978-3-933146-84-7.volltext.frei.pdf

[7] ADEMA, J; RUTTEN, P. Digital Monographs in the Humanities and Social Sciences: Report on User Needs. OAPEN report 2010. Available at: http://www.oapen.org/images/D315%20User%20Needs%20Report.pdf.

[8] Enriched publications consist of a traditional publication, linked to information that is used in the writing of the publication (a dataset, for example) and to related information that has appeared since the publication (citations; publications based on it; commentary).

[9] HOGENAAR, A. Enhancing Scientific Communication through Aggregated Publications Environments. Ariadne 61, October 2009. Available at: http://www.ariadne.ac.uk/issue61/hogenaar/ (March 2010).

An effective and automated publishing process to improve user interface style guides

Martin Lugmayr; Johann Schrammel; Cornelia Gerdenitsch; Manfred Tscheligi

CURE – Center for Usability Research and Engineering
Modecenterstrasse 17/Objekt 2, Vienna, Austria
{lugmayr, schrammel, gerdenitsch, tscheligi}@cure.at

Abstract

Style guides have become an important and common way to improve and standardise development of user interfaces. However, there are several well-known problems on using style guides. Having these problems in mind we present an effective and automated publishing process. First we will introduce a role-based approach to model style guides. After that we will focus on the steps of the publishing process and describe them in detail with their outputs. By that we want to focus on the practical and theoretical advantages of our methodology and their limitations. In summary, this paper will describe in detail how mentioned techniques and components work together and how we build up a useful publishing process for adaptable and usable style guides.

Keywords: user interface style guide; DITA maps; user centred design; usability

1. Introduction

User interface style guides are a central and important element for developing graphical user interfaces (GUIs). With their aid it is possible to guarantee consistency (e.g. menu guidance, Look and Feel) between different applications [2], to provide a high quality human-computer interaction [5] and to simplify interdisciplinary and multinational collaboration in developing GUIs [4].

There are several typical problems that occur in developing, implementing and using style guides that are related to the commonly used traditional publishing process. Scientific literature and practical experiences illustrate that the reason for many problems is related to the preparation and representation of information in a style guide. Wilson [6] reports issues on updating, bad usability, insufficient indexes and others. Similar problems are formulated by [4] and affect for example style of preparing materials, media (paper-based vs. online) and the complexity in practice of style guides. Those authors also pointed out that people who use style guides want to get the information they look for structured and fast.

Reasons for those problems that exist in using style guides are e.g. formulated by [6]. The author lists thirteen reasons that can cause problems in using style guides. Five main reasons concerning the design and implementation of the style guide are formulated as follows:

- Extent of the style guide: Although guides should be very easy to understand, complete style guides became very big.
- Possibilities of updating: The question is how to distribute updates to the style guide. Some try to put the guide online, but in this case you still need to alert people to do changes.
- Bad usability: Users often do not understand the guides and also have to look for the information they need very long.
- Insufficient indexing: There are not enough index terms (which are additionally not sufficient) and there is also a poor use of cross-referencing.
- Too much formulated text: There is too much formulated text instead of integrating screenshots or bullet points.

Other problematic areas are the complexity of the style guide and that style guides are laborious to use [4].

In this paper we want to present a new style guide approach that gives a solution for many of those problems. We present style guides based on DITA maps. First we want to describe the role-based approach, the publishing process and the technical implementation of the guide. Afterwards we provide a demonstration of our methodology. In the discussion section we work out positive and negative aspects of our work.

2. Methodology

Taking above mentioned problems into account, the aim of our research was to establish a new method to develop, publish and maintain style guides.

To achieve our goals we use two different approaches. First we want to reduce complexity of the style guide by introducing role-based style guides.

Second, our aim was to improve and automate the update process by using DITA.

Within our publishing methodology typical problems can be avoided, and as a consequence we expect an increase in the acceptance of style guides in the practical context.

The next step is to evaluate our approach in user trials and interviews with real users of the developed style guide.

2.1. The role-based approach

The main idea of the role-based approach is to split up the information into small logical units that are maintained centrally, to determine the relevance of these information units for the different user groups and output formats and to produce tailored documents automatically. Based on the modelled scaffold and with the help of automated transformation processes, specific documents for specific users and publication formats are generated. Consequently we can make sure that different user groups get just the information they need and we can avoid providing irrelevant and unnecessary content.

Regarding the implementation of our particular style guide, the following three roles were defined to cover the different needs and requirements of the different user groups:

1. User Interface Designer
2. Developer
3. Library-Developer

Corresponding to these three roles we split up the content of the whole style guide and assigned each logical unit to at least one role. Due to the fact that we designed the roles according to an extensive analysis of internal processes and division of work of our project partner, other roles will probably be more feasible for other applications.

Using the tagged content for each role, a tailored style guide can be generated automatically. Thus, each user will get the required information for his/her role. Nevertheless, each user is given a chance to explore the whole style guide and not only the filtered parts. This will provide an overview of all topics additionally.

2.2. The publishing process

In addition to the aim of reducing complexity of the style guide, our second approach is to avoid the mentioned typical problems that occur when using a conventional production and publishing process for style guides. The technical implementation focuses especially on an efficient publishing process, which supports the user in getting the required information.

Contrary to conventional style guides which are monolithic systems, we used a modular concept. Therefore, we split the style guide and extracted hundreds of small topics from chapters, sections and subsections. Consequently, a topic is our basic information unit.

2.3. Technical Implementation

According to our idea of structured content, we required technology which provides both a reliable and automated publishing process and the ability to handle tagged content units.

DITA (Darwin Information Typing Architecture) achieves these requirements. Embedded Ant and Batch scripts for automation as well as a topic-based structure provide the required functionality. Thus, we designed the publishing process of our style guide based on DITA specifications.

The mentioned topics are stored in XML-files and build the basis for further data processing in DITA. DITA Open Toolkit[1] is an open source implementation of DITA and thus the main component of our publishing process. DITA Open Toolkit – an XML based framework – is used for generating, distributing and reusing technical information and we will show how the capabilities of DITA Open Toolkit fit for our aims of structure, changeability and automated publishing process.

[1] http://dita-ot.sourceforge.net/

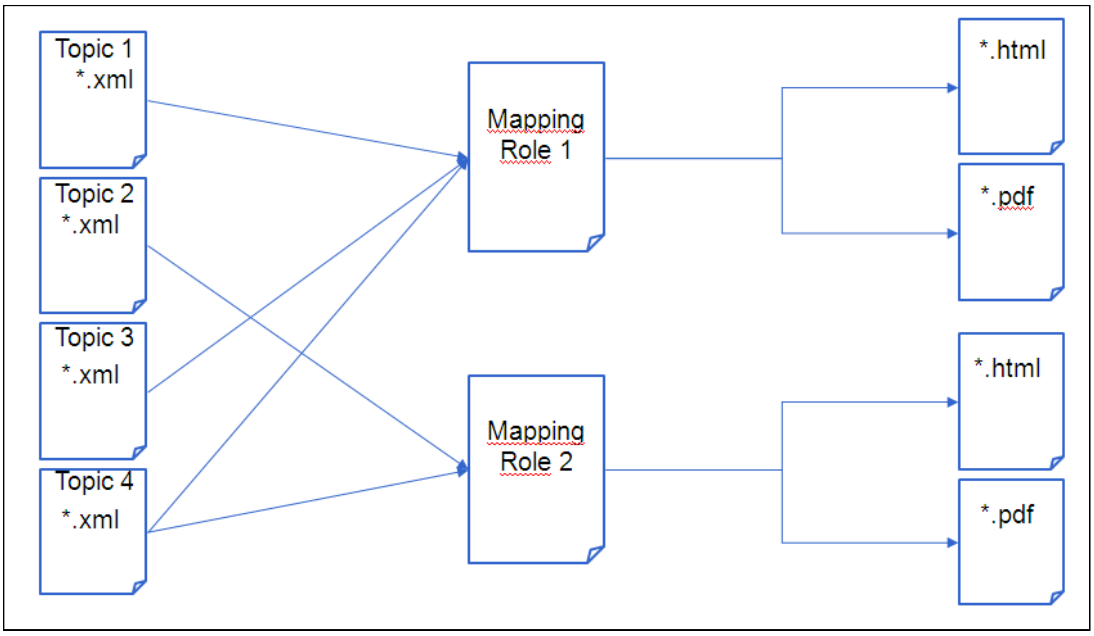


Figure 1: Schematic representation of the automated publishing process

DITA maps are a main concept in DITA and are mainly used to define the structure of a style guide document. In DITA maps all necessary topics are listed hierarchically by using a reference to the corresponding XML files, as shown in Figure 1. According to the structure in the DITA map, DITA Open Toolkit creates table of contents and structured documents automatically.

As DITA Open Toolkit implements the DITA specifications, we adopt all the useful characteristics of DITA like various output formats (HTML, PDF, MS HTML Help) from a single XML repository.

Due to the fact that DITA Open Toolkit is based on Java, XML and XSLT, we got platform-independency on the top.

The role-based approach is managed in DITA Open Toolkit by using XML attributes. Basically, each topic is labelled by an XML attribute that defines for which role it is important. On using different values for different roles it is possible to filter out the unnecessary topics. Having integrated this functionality in our publishing process, we make sure that each user gets the information s/he needs, according to the role-based approach.

Corresponding to the aim of an automated publishing process we used Ant and Batch scripts for automation. At this point of process it is defined which output formats should be created. In the background, DITA Open Toolkit uses different XSL transformations for different output formats (e.g. HTML, PDF, MS HTML Help Files, RTF). Ant scripts enable an easy and fast publishing process, which means that it is possible to build all defined output formats for all defined roles just by a double click.

A WYSIWIG XML Editor is also an essential part of our publishing process. So, defined administrators get the possibility to apply necessary changes to the style guide. Providing such update mechanisms is very important for acceptance of style guides [6]. According to this, we also introduced a SVN (Subversion[2]) repository for interoperability and traceability of changes.

2.3.1. Steps of the publishing process

To give an overview of how the publishing process is used, we describe it in the following step by step. Central aim was also to keep the update process as simple as possible.

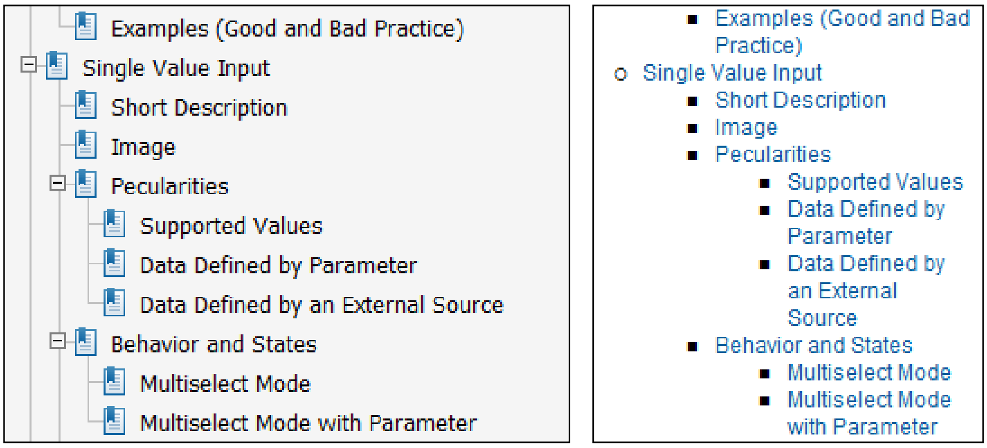
1) Check content out of version control (Subversion): At first it is necessary to fetch the content from a central version control repository, due to tracking and security considerations.
2) Changing content in WYSIWIG XML Editor: Defined “content administrators” can insert, update or delete content. They have to assign the topics to one or more roles. That means they have to be familiar with underlying role approach and have to decide which content relates to whom.
3) Check content in to version control: Changed topics have to be stored back to version control repository, so, more than one person can administer the content of the style guide.
4) Start batch process: Re-building of the style guide and its various files can be done regularly or – if required – by executing a batch file.
5) HTML, PDF, Help file for defined roles will be created automatically: The user can always work with latest version of the style guide.

3. Live demonstration

In the following we present some output examples and details of the style guide we developed to demonstrate the advantages of our approach of the publishing process.

[2] http://subversion.tigris.org/

As discussed before, we implemented an automated publishing process which allows content distribution to various output formats for different user groups (roles). Figure 2 shows at the left side a snippet of the section overview of the PDF style guide and at the right side the HTML equivalent. As you can see, there is no difference in content. Just the appearance differs slightly due to technology constraints.

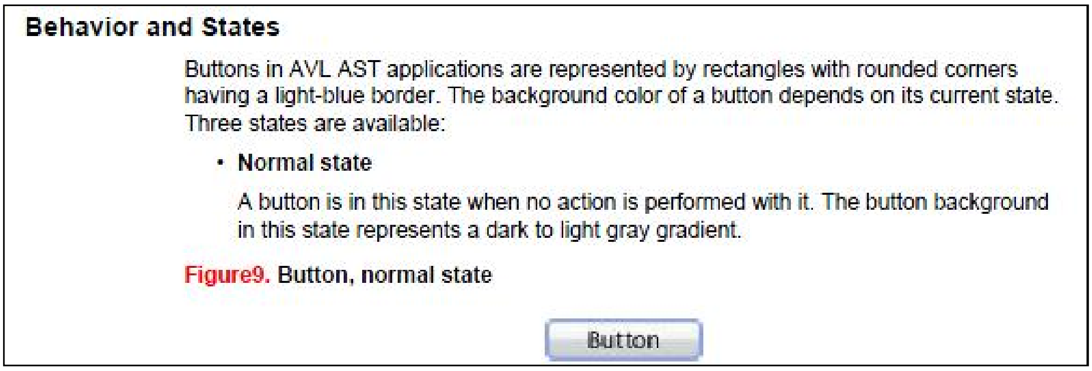


- Examples (Good and Bad Practice)
- Single Value Input
  - Short Description
  - Image
  - Peculiarities
    - Supported Values
    - Data Defined by Parameter
    - Data Defined by an External Source
  - Behavior and States
    - Multiselect Mode
    - Multiselect Mode with Parameter

- Examples (Good and Bad Practice)
- Single Value Input
  - Short Description
  - Image
  - Peculiarities
    - Supported Values
    - Data Defined by Parameter
    - Data Defined by an External Source
  - Behavior and States
    - Multiselect Mode
    - Multiselect Mode with Parameter

Figure 2: Section overview of content in PDF (left) and HTML (right)

Of course, not only the structure but also the content has to be the same in different output formats. Figure 3 and Figure 4 demonstrate the advantage of storing information in XML. Due to this, it is possible to generate various different output formats automatically without additional efforts. As a result, we can provide the same content for the same role in e.g. HTML, PDF, MS HTML Help. Thus, users of the style guide can switch between e.g. HTML and PDF without loss of orientation.



Behavior and States

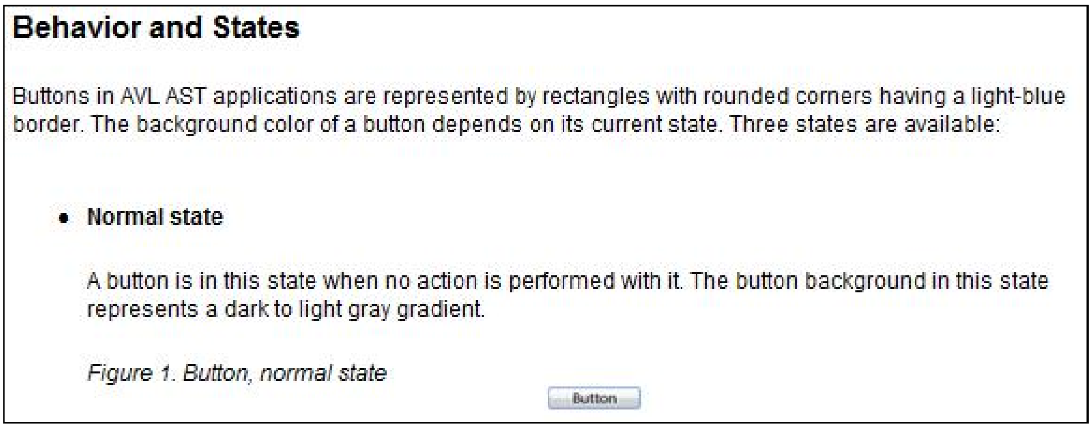
Buttons in AVL AST applications are represented by rectangles with rounded corners having a light-blue border. The background color of a button depends on its current state. Three states are available:

- **Normal state**

  A button is in this state when no action is performed with it. The button background in this state represents a dark to light gray gradient.

Figure9. Button, normal state

Figure 3: Content presented in PDF



Behavior and States

Buttons in AVL AST applications are represented by rectangles with rounded corners having a light-blue border. The background color of a button depends on its current state. Three states are available:

- **Normal state**

  A button is in this state when no action is performed with it. The button background in this state represents a dark to light gray gradient.

Figure 1. Button, normal state

Figure 4: Content presented in HTML

According to the role-based approach we modelled the information about e.g. “Buttons” differently for the role “library developer” and “developer”. As you can see in Figure 5, “library developer” will get information about the “Behavior and States”, because this role is responsible to implement new buttons (or other user interface elements) in a consistent way.

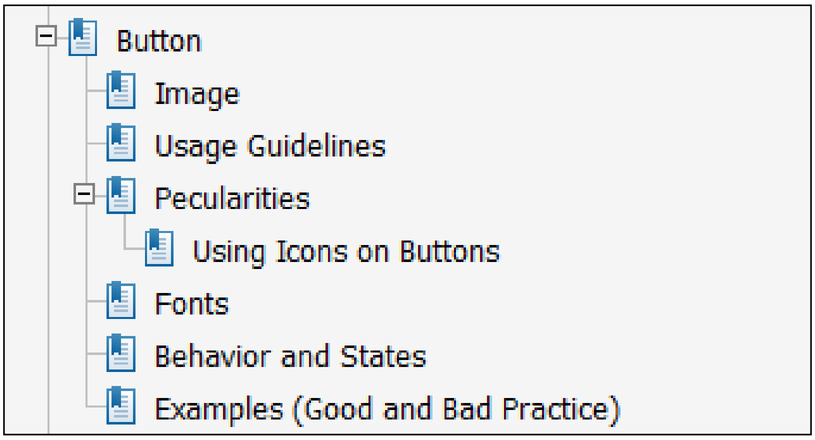
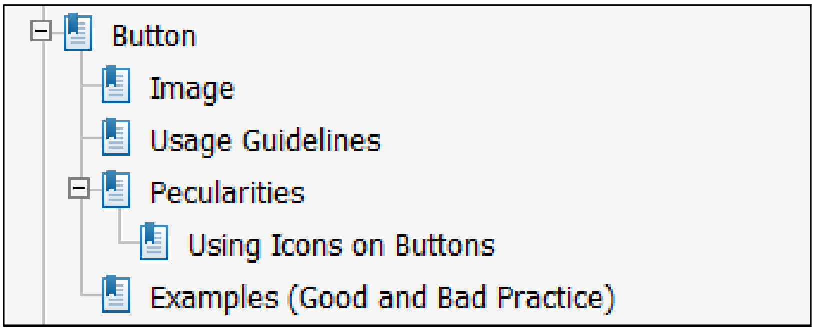


Figure 5: Section “Button” for role “library developer”

As shown in Figure 6, we defined the role “developer” as not in need of the subsection “Fonts” and “Behavior and States” of buttons. The reason for this is that it’s not relevant for a developer how the button will appear, because s/he has just to use it and implement the functionality. Due to this the developer gets information about “Usage Guidelines” and “Examples”.



- Button
  - Image
  - Usage Guidelines
  - Pecularities
    - Using Icons on Buttons
  - Examples (Good and Bad Practice)

Figure 6: Section “Button” for role “developer”

4. Discussion

Developing and implementing a style guide based on DITA maps will lead to several advantages.

- Firs, it is possible to actively integrate the user in the development process of the style guide, which will consequently lead to a more flexible and easier service of the style guide.
- Second, the consistency of the style guide will be increased and for that it is much easier to compare different style guides or versions of style guides.
- Third, we want to argue that through the integration of DITA maps regular services will be done easily and fast, because it is possible to quickly react on new requirements and update styles. DITA maps allow a direct connection on online-publishing channels and for that changes came fast to the specific user and without expensive distribution costs.

Summarizing our arguments, using DITA maps can deal with traditional problems in implementing style guides.

As mentioned before, DITA maps allow integrating the user in the developing process and increasing the commitment of the end-user. For further development we suggest specifying the introduced DITA maps in a more individual way and adapt them to specific user groups. For that we suggest formulating a scaffold, where it is possible to use it for contents of specific issues (e.g. developing an e-commerce portal). We assume to individualize the style guides as much as possible to provide the appropriate information.

We suggest integrating users in the development process as they can give valuable feedback and stimulations in developing the styles (e.g. through discussion boards). This will lead to a higher commitment on the style guide and allows adaptive improvements of the style guide within the development circle.

Further research and development should also concentrate on the output format of the DITA Open Toolkit. An opportunity in this context would be to provide code snippets as well as to formulate existing patterns and design. We want to point out that this will lead to an optimized designing process of the user interface.

5. Conclusions

Finally, development and initial setup of a style guide using both the role-based approach and the automated publishing process with DITA are more time-consuming than writing a common style guide. But in the long run there are several advantages which have to be taken into account also.

The most important advantage is the updating process of this approach, as changing content doesn’t lead to extensive re-design and expensive distribution costs. Thus, the content of the style guide can be held up-to-date easily and the user gets more current information.

In addition, the role-based approach provides more relevant information to the user, because just the content that is most informative and useful for a particular role is presented. But, of course, the quality of relevance of content depends on an extensive analysis of required roles at the beginning of the style guide development process.

Both together improve style guides essentially and avoid the common problems like too large style guides and missing updating possibilities.

Notes and References

[1] Gale, S. A Collaborative Approach to Developing Style Guides. Conference proceedings on Human factors in Computing Systems, April 13 - 18, 1996, Vancouver Canada. ACM Press, pp. 362-367. 1996

[2] Henninger, S. Creating organization-specific usability guidelines. In CHI '97 Extended Abstracts on Human Factors in Computing Systems: Looking To the Future, Atlanta, Georgia, March 22 – 27. 1997

[3] Quesenbery, W. Building a better style guide. Proceedings of UPA 2001.

[4] Schemenauer, P.J., Pawlick, C. Evaluating Guidelines for Writing User Interface Text. SIGDOC'07, October 22-24. 2007

[5] Willems, P., Verlinden, J., Troost, P.J. Towards a Framework of Methods on UI Style Guides. CHI 2000, 1-6 April, p.131. 2000

[6] Wilson, C. E., STC Usability SIG Newsletter: Usability Interface, Vol 7, No. 4, April 2001 (http://www.stcsig.org/usability/newsletter/0104-style.html) (January 2010).

Analysis of e-book use: The case of ebrary

Umut Al; İrem Soydal; Yaşar Tonta

Department of Information Management, Faculty of Letters,
Hacettepe University, 06800 Beytepe, Ankara, Turkey
{umutal, soydal, tonta}@hacettepe.edu.tr

Abstract

The interest in the use of electronic media in scholarly communication is growing. Nowadays, libraries reserve much larger budgets for electronic information sources as users tend to get access to the full-texts of journal articles and books online. The effective management of library collections is only possible through studies identifying user needs as well as studies of usage analysis showing how much of what is being used in these collections. Although e-books are a significant part of library collections, studies on e-book use are scarce. In this paper, we have analyzed about half a million use data of ebrary e-book database by the users of Hacettepe University Libraries within a four-year period (2006-2009). We obtained COUNTER-compliant use data identifying, among other things, book title, publisher, and publication year for each transaction to analyze the use. We identified the most frequently used e-book titles by Hacettepe University ebrary users in each Library of Congress (LC) subject class. E-books on Medicine (R) were used most often, followed by books on Education (L) and Language and literature (P). A small number of e-books in each subject class satisfied half the demand, while an overwhelming majority of e-book titles were never used. Findings of this study can be used to develop an e-book collection management policy and understand its implications for consortial licensing of e-book packages.

Keywords: e-book use; collection management; collection development; ebrary; Hacettepe University Libraries

1. Introduction

Libraries continue to develop their own collection management policies that suit their users' information needs. Limited library budgets put the consortium-type collaboration efforts on the agenda. Therefore, libraries must not only understand the needs of their potential users but also be aware of the needs of the users of other consortium members. Developing effective collection management policies and executing them in a consortial environment requires careful work as well as polished negotiation skills.

Databases provide instant access to information. Full-text databases are more intensely used than bibliographic ones as they provide direct access to the sources. Full-text electronic books (e-books) are now available via several database vendors or aggregators, and they are becoming an important part of library collections. The increase in the variety of e-book packages forces libraries to be selective as their diminishing budgets are not enough to cope with the growth of databases. Libraries have to investigate the trends and choices of their users to be able to build an effective policy for collection development and management. Therefore, it is important to address the following questions: Who are the actual users of these databases? Do tendencies on using e-books differ across the subjects? Are current e-books requested more often by the users?

This paper addresses some of these questions by analyzing the usage of e-book titles in the ebrary database by the users of Hacettepe University Libraries. It identifies the most frequently used e-book titles and tries to shed some light on the non-use of titles in some subjects. Findings can be used to improve the micro-management of e-books collections and understand its implications on a wider scale for library consortia.

2. Literature Review

Generally, an "e-book" is defined as a digital version of a traditional print book designed to be read on a personal computer or an e-book reader [1]. Related studies explain some advantages of e-books such as adjustable font size and font face, easy access to the content, multimedia display capabilities, no cost for paper, ink, binding, wrapping, postage, or transport, no physical space requirements, on-demand availability, and searchability within a book [2, 3, 4]. Typically, e-books are cheaper than hard copy versions.

Despite the many advantages of e-books, some studies showed that their usage can be very low [5, 6, 7, 8]. Most users are unaware of the existence of e-books in library collections. Although they are willing to discover and use e-books more effectively, user unfriendly interfaces or usability problems tend to hinder their further use. Some studies comparing usage statistics of e-books to that of their print counterparts concluded that e-books are used more often [9] while others did just the opposite [10]. Usage patterns should be taken into consideration in interpreting these somewhat conflicting findings.

E-book use is hard to measure when compared to printed ones. For instance, the use of printed books can be measured by the number of loans or in-library use whereas e-book use can only be measured by access statistics. Access to e-books, on the other hand, can be defined in different ways such as print, view or download. Access statistics provided by e-book vendors differ in this respect. Hence, a standard presentation of parameters is not available for e-books. It can be difficult to compare e-book usage across different packages since there is almost no consistency in usage statistics between vendors [11]. Therefore, most of the e-book usage studies in the literature focused on e-books provided by the same vendor rather than making comparisons across different vendors [12]. COUNTER (Counting Online Usage of Networked Electronic Information Resources) is an international initiative to set practical world-wide standards for the recording and reporting of vendor-generated usage statistics in a consistent and compatible way [13, 14].

Libraries benefit from usage statistics of e-books in collection development despite the lack of standardized reports. Statistics of e-book use show the demand for e-books as well as give some important clues about the most popular subject categories for e-books. For instance, according to the usage statistics of the netLibrary, e-books on computer science and engineering (39%), business, economics, and management (17%), arts and humanities (14%), and natural sciences and mathematics (13%) were the most demanded ones [11]. Safley [11] emphasized in the same study that the use of ebrary has increased 190%. In a different study almost 19% of the library's e-book titles had been accessed at least once [12]. The results of the Springer's e-book use study showed that chapters from e-books on chemistry and materials science were downloaded most often [15]. The study also pointed out that the growth of sales of Springer's printed books is commensurate with the increase in Springer's e-book usage, and that the use of Springer's e-books has increased 60% in Turkey between 2007 and 2008. The study concludes that the e-book use was on the rise and printed books were not being "cannibalized" by e-books. E-books were even driving print book sales especially in the countries which have a large e-book penetration [15].

There are many surveys on who use e-books, how and why they use them. For instance, more than 60% of netLibrary users in a university library said they preferred print books over electronic ones [16]. They generally read a chapter or a few pages instead of reading the whole e-book. In a different study, special libraries were the largest users of e-books (15% to 60%) compared to academic (5%) and public (2%) libraries [17]. Users of special libraries appreciated the remote use and they preferred the convenience of e-books. Academic librarians thought that e-book usage was better for browsing and reference work [17]. Studies also show that users are unaware of the existence of e-books. More than 70% of students of an academic library were familiar with the term "e-book", yet almost 60% of them did not use e-books and did not know that e-books were available in their university library [18]. (See also [19].)

3. Method

This study analyzes nearly half a million e-book "sections" (a section is defined as a chapter or an entry [20]) requested from ebrary, an e-book database, by the users of Hacettepe University Libraries within a four-year period (2006-2009). Hacettepe University has currently some 28,000 students and 3,500 academic staff and Hacettepe University Libraries offer a rich collection of both printed and electronic sources of information to its users including more than 70 databases [21, 22]. The University added ebrary database to its collection in 2006 as a member of the Consortium of Anatolian University Libraries (ANKOS). ebrary has been Hacettepe's most frequently used and the largest e-book database containing more than 45,000 e-book titles in several disciplines. We obtained COUNTER-compliant use statistics of Hacettepe University Libraries from ebrary. We then looked into the usage of e-books classified under different Library of Congress (LC) Classification System subjects. The term "usage" in this study is defined as the number of times each e-book is requested from the ebrary e-book database.

This paper addresses the following research questions:

- Which e-books are requested most frequently from the ebrary collection?
- What are the subjects of the most frequently requested e-books?
- Does the usage of e-books differ across the subjects?

- Does the usage of e-books change by year?

Finding answers to these research questions would empower Hacettepe University Libraries to better manage its precious resources, negotiate better deals with e-book database vendors and aggregators, and align its collection management policy with that of the Consortium.

4. Findings and Discussion

The total number of books in ebrary database is 45,147. Users of Hacettepe University Libraries requested a total of 499,841 sections from 12,826 different books in ebrary collection between the years 2006 and 2009, indicating that just over a quarter (28%) of the overall ebrary collection was used. Figure 1 provides the number of unique book titles requested from ebrary collection between 2006 and 2009 (4,213, 4,548, 4,343, 5,072, respectively), which at any given year constitutes only a fraction (about 10%) of the total number of books available in the ebrary collection (45,147). Although the number of requested books has increased in 2009, the number of sections requested by the users of Hacettepe University has been decreasing in the last three years.

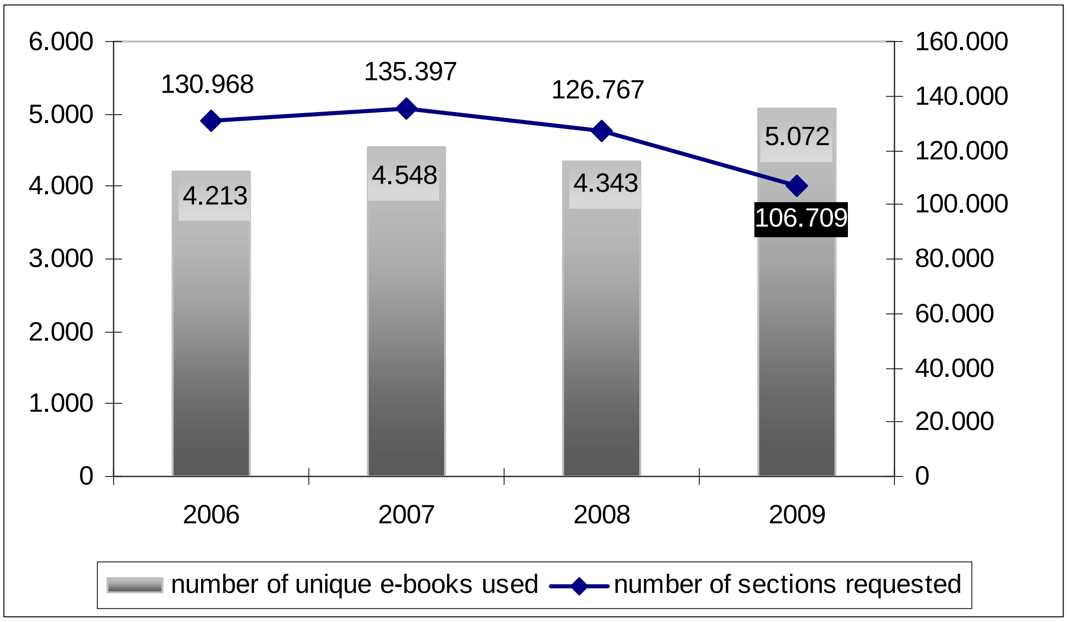


Figure 1: Number of unique e-books used and sections requested therefrom (2006-2009)

The distribution of 43,422 e-books and sections requested therefrom under LC Classes is given in Table 1 (no LC subject classification information was available for 1,725 titles). The number of e-books under each LC class ranges between 40 (A: General works) and 10,021 (H: Social sciences), median being 1,142. Books classified under Social Sciences constitute 23% of all books in the ebrary e-book database, followed by Language and literature (P) (12%) and Science (Q) (11%) (column 3 of Table 1). Books classified under General works, Naval science (V), Library science (Z), on the other hand, make up less than 1% of the overall ebrary collection.

Of 43,422 of e-book titles available in the ebrary database, 12,826 (or 29.5%) were used at least once. The number and percentage of e-book titles used under each LC subject classification (excluding 1,484 titles with no LC Class numbers) is given in columns 4 and 5 of Table 1. On the average, 74% of 43,422 e-book titles were not used at all. The proportion of use of e-books under each LC subject varied, although none was over 40%. For instance, only 39.8% of e-book titles on Medicine (R) were ever consulted, followed by e-books on Education (L), General works and Library science (*circa* 35% each). About 90% of e-book titles on American history, Military science (U) and Naval science were not used at all. The use of e-books seems to be closely related with the disciplines studied at Hacettepe University, which has one of the top medical schools in the country along with a Faculty of Education, but has no military/naval school or a department on American history (F).

The total number of sections requested from 12,826 e-books was 499,841. The distribution of 429,049 sections under each LC subject (as frequencies and percentages) is given in columns 6 and 7 of Table 1 (no LC subject classification information was available for 70,792 sections). The number of sections requested from e-books under each LC subject ranges between 90 (Naval science) and 78,157 (Medicine), median being 3,841. More than 51% of all sections requested came from e-books on Medicine (18.2%), Social sciences (16.8%) and Language (16.5%). E-books under 11 LC subjects satisfied less than 5% of all requests. Sections from e-books on Medicine, Education, Language and literature and Science have been requested more heavily. For instance, the number of e-books available on Medicine and Philosophy, psychology, religion (B) are almost the same (each constitutes 8.4% of all e-books). Yet, more than twice as many sections were requested from books on Medicine (18% of all section requests) compared to that of Philosophy, psychology, religion. The number of sections requested from all but four subjects (Medicine, Education, Language and literature and Science) was not commensurate with the number of e-books available under those LC subject classes.

Table 1: Number of books and sections requested from ebrary database (under Library of Congress subject classes)

LC Class	# of books (a)	*% of total books (43,422) (b)	# of requested books (c)	**% of books requested (c / a)	# of sections requested (d)	***% of all sections requested (d / 429,049)	d / c
General works (A)	40	0.1	14	35.0	269	0.1	19.2
Philosophy, psycho., religion (B)	3,667	8.4	866	23.6	36,051	8.4	41.6
Auxiliary sciences of history (C)	161	0.4	53	32.9	998	0.2	18.8
History (D)	2,319	5.3	507	21.9	17,297	4.0	34.1
American history (E)	1,101	2.5	123	11.2	3,288	0.8	26.7
American history (F)	614	1.4	52	8.5	684	0.2	13.2
Geography, anthro., recreation (G)	1,142	2.6	324	28.4	11,296	2.6	34.9
Social sciences (H)	10,021	23.1	2,271	22.7	72,162	16.8	31.8
Political science (J)	1,639	3.8	379	23.1	10,844	2.5	28.6
Law (K)	1,195	2.8	151	12.6	2,161	0.5	14.3
Education (L)	1,727	4.0	612	35.4	24,383	5.7	39.8
Music (M)	659	1.5	147	22.3	3,581	0.8	24.4
Fine arts (N)	405	0.9	105	25.9	3,159	0.7	30.1
Language and literature (P)	5,393	12.4	1,819	33.7	70,615	16.5	38.8
Science (Q)	4,962	11.4	1,472	29.7	59,032	13.8	40.1
Medicine (R)	3,675	8.5	1,461	39.8	78,157	18.2	53.5
Agriculture (S)	557	1.3	110	19.7	3,841	0.9	34.9
Technology (T)	3,391	7.8	744	21.9	28,752	6.7	38.6
Military science (U)	467	1.1	49	10.5	985	0.2	20.1
Naval science (V)	71	0.2	8	11.3	90	0.0	11.3
Library science (Z)	216	0.5	75	34.7	1,404	0.3	18.7
Total/Average	43,422	100.0	11,342	26.1	429,049	****99.9	37.8

*Ratio of the *# of books* in LC classes to the *total # of books* **Ratio of the *# of requested books* to the *total # of books* under corresponding LC Class

***Ratio of the # *of requested sections under* corresponding LC class to the total # of requested sections ****Percentages may not equal 100% due to rounding

The last column of Table 1 provides the average number of sections requested from books under each LC subject. Medicine came first with an average of 54 sections per e-book used, followed by Philosophy, psychology, religion (42 sections), Language and literature, and Education (40 sections each). Users studying Medicine seem to attach greater importance to electronic information sources [23]. Monographs in electronic form appear to be in high demand in Language and literature as well as in Education [24]. E-books used in Naval science, American history and Law (K) were consulted much less often (average of 11, 13 and 14 sections per book, respectively). E-books under these LC classes can be considered as prime candidates to be excluded from the collection.

Figure 2 shows the cumulative percentage of 12,826 e-books in the ebrary database satisfying Hacettepe users' demand. About 10% of books satisfied 63% of the total requests and 20% did 78% of all requests, conforming to Trueswell's well-known 80/20 rule [25]. When the entire collection of ebrary database with 45,147 books is taken into account, the concentration of requests on a relatively few e-books is even more remarkable: 10% of books satisfied 90% of the total demand.

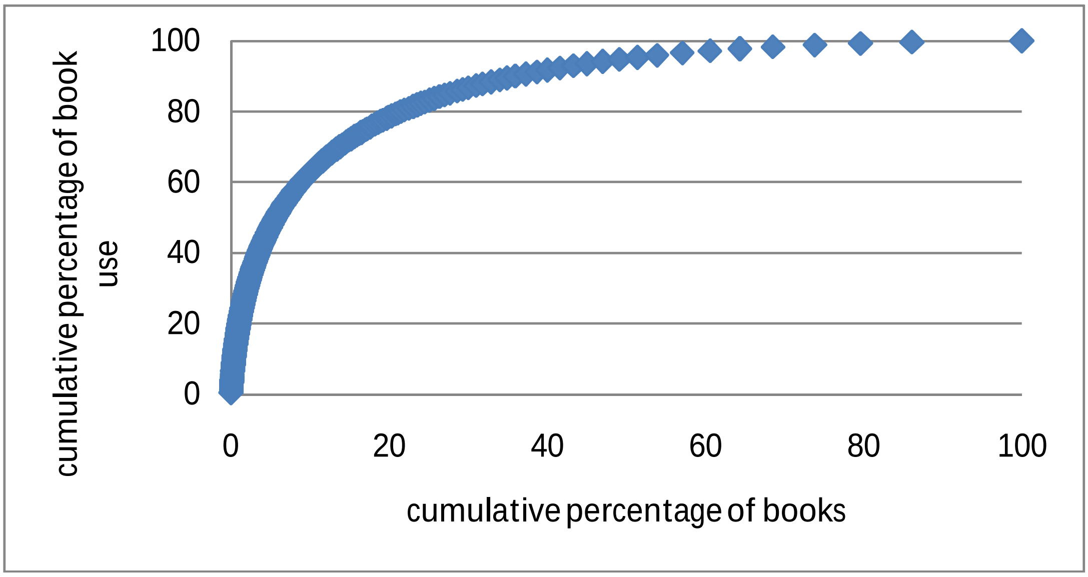


Figure 2: Usage of e-books in ebrary database

We obtained similar results for e-books used under each LC subject. A small number of books in each LC subject class consistently satisfied one third or half of all requests (Table 2). For instance, in Social sciences, a subject with the highest number of books (10,021), 60 books satisfied 33% of all requests while 142 books did 50%. Similarly, 7 books each in General works and American history, 8 books each in Auxiliary sciences of history (C), American history (E) and Fine arts (N), 6 books in Agriculture (S), 2 books in Military science, and 5 books in Naval science satisfied half the demand. In general, a few books satisfied the overwhelming majority of demand while the great majority of books were not used at all. It should be kept in mind that the proportion of books satisfying 33%, 50%, 67% and 100% of the demand in each LC subject would be much lower if we used the total number of books available under each LC subject instead of total number of books used (as we did in Table 2). For instance, 5% of out of all e-books *used* to satisfy 40% of the total demand in Language and literature subject class constitute only 1% of all e-books *available* under that class.

Table 2: Number of books satisfying demand

LC Class	33%	50%	67%	100%	# of books
A	5	7	10	14	40
B	20	44	95	866	3,667
C	5	8	12	53	161
D	16	36	68	507	2,319
E	3	8	17	123	1101
F	4	7	12	52	614
G	9	19	37	324	1,142
H	60	142	287	2,271	10,021
J	10	22	47	379	1,639
K	8	16	30	151	1,195
L	21	49	97	612	1,727
M	48	76	102	147	659
N	3	8	20	105	405
P	30	84	197	1,819	5,393
Q	34	81	182	1,472	4,962
R	48	105	197	1,461	3,675
S	3	6	11	110	557
T	17	38	81	744	3,391
U	2	2	5	49	467
V	3	5	6	8	71
Z	5	12	20	75	216

No discernable pattern of yearly variations was detected in the use of e-books under different LC subject classes, although this may change in the long run. Subjects of e-books requested in different years were similar.

We also tested if the distribution of books to requests under each LC subject fits the Price Law, which states that the square root of all books would satisfy half the demand [26]. Books classified under General works and Library science were in accordance with the Price Law to some extent. The rest did not fit the Price Law. For LC subject classes of Social sciences, Education, Music (M), Language and literature, Science and Medicine, the number of books that satisfied 50% of all requests in respective fields were higher than that predicted by Price Law whereas the number of books satisfying half the demand for subject classes of Philosophy, psychology, religion, Auxiliary sciences of history, History (D), American history (E & F), Geography, anthropology, recreation (G), Political science (J), Law, Fine arts, Agriculture, Technology (T), Military science, and Naval science were lower than that predicted by Price Law. It could be that relatively small number of requests for e-books under most LC subject classes was not enough to test the validity of a power law such as Price Law.

Table 3: The most frequently used 20 ebrary books

Book	LC	N
International Encyclopedia of Ergonomics and Human Factors*	T	2,194
5 Steps to a 5 on the Advanced Placement Examinations: Calculus	Q	2,147
Beginner's Guide to Structural Equation Modeling*	Q	1,436
Beginning Programming	Q	1,411
Character Strengths and Virtues: A Handbook and Classification	B	1,372
Routledge Critical Dictionary of Semiotics and Linguistics	P	1,346
Psychology of Humor: A Reference Guide & Annotated Bibliography	P	1,308
On That Point! An Introduction to Parliamentary Debate	P	1,246
Harrison's Manual of Medicine (16th Edition)	R	1,224
Jacques Derrida and the Humanities: A Critical Reader	B	1,202
Speaking, Listening and Understanding: Debate for Non-Native English Speakers	P	1,186
Talking Gender and Sexuality	P	1,170
Routledge Reader in Politics and Performance	J	1,169
Speech Acts in Literature	P	1,164
Provocations to Reading: J. Hillis Miller and the Democracy to Come	P	1,150
Argument and Audience	P	1,106
Advanced Mathematical Thinking	Q	1,094
Theoretical Aspects of Heterogeneous Catalysis	Q	1,080
Discourse	P	1,067
SPSS for Intermediate Statistics: Use and Interpretation*	H	1,032
Total		26,104

Note: Titles with asterisks (*) did not have LC subject classes assigned to them. They were classified by the authors.

The most frequently requested 20 e-book titles satisfied 5% of the total requests (Table 3). Almost half (9) of 20 book titles came from Language and literature, 5 from Science, 2 from Philosophy, psychology, religion, and 1 each from Political science, Medicine, Technology, and Social sciences.

5. Conclusion

ebrary has been the most heavily used e-book database at Hacettepe University Libraries since 2006. Although the number of subscribed books in ebrary database was increasing since then, the number of sections requested has been decreasing in recent years.

The use of e-books under different LC subject classes differs tremendously. There was almost 20-fold difference between the average number of total requests in Medicine and that in American history. Books classified under Naval science, Law and Military science were used very infrequently. Relatively lower rates of use of e-books in these subjects are probably due to the fact that Hacettepe has no military and naval schools and the Faculty of Law has recently been founded. Hence, a few Hacettepe researchers appear to study in these fields. Hacettepe University Libraries have to promote e-books more intensely to increase their use.

The use of e-books exhibits a Bradfordian distribution in that relatively few titles satisfied the majority of the requests, perfectly in line with the findings of similar studies. The fact that the majority of e-book titles in some subject areas were not used at all suggests that libraries should review their collection management policies. It could be that a potential user base for e-books in certain subjects may be lacking. Or, the existence of licensed e-book packages such as ebrary may need further promotion within the campus to increase the awareness. In any case, the unit cost of using a title from a licensed e-book database could be quite expensive if such e-book packages are not carefully selected according to collection development policies and user needs of respective libraries. The lack of consortial collection development and management policies complicates the issue of cost further.

The following recommendations can be offered:

- The usage levels of e-book databases licensed by universities through consortia have to be measured.
- Findings of in-depth usage analysis studies should be taken into account when negotiating deals with e-book suppliers for license agreements or renewals thereof.

- In addition to usage statistics, feedback should be gathered from users through questionnaires and interviews to find out why they use or do not use e-books.
- The percentage of non-used e-book titles should be figured into the license renewal terms to get discounts or additional titles.

Findings of this study can be used by libraries to improve their e-book collection management policies. Further studies on cost-benefit analysis of e-book use, comparison of e-book use in different libraries and its impact on library consortia are needed.

Acknowledgements

We are grateful to Ms. Ayşen Küyük, Director of Hacettepe University Libraries, for making their ebrary usage data available to us for analysis and use.

Notes and References

[1] REITZ, JM. *ODLIS: online dictionary for library and information science*. 2007. Available at http://lu.com/odlis/odlis_e.cfm (4 April 2010).

[2] BROWN, GJ. Beyond print: reading digitally. Library Hi Tech, 19(4), 2001, p. 390-399.

[3] GUNTER, B. Electronic books: a survey of users in the UK. Aslib Proceedings: New Information Perspectives, 57(6), 2005, p. 513-522.

[4] RAO, SS. Electronic books: a review and evaluation. Library Hi Tech, 21(1), 2003, p. 85-93.

[5] ABDULLAH, N; GIBB, F. A survey of e-book awareness and usage amongst students in an academic library. In: *Proceedings of International Conference of Multidisciplinary Information Sciences and Technologies, Merida, 25-28 October*, 2006. Available at http://strathprints.strath.ac.uk/2280/1/strathprints002280.pdf (4 April 2010).

[6] ANURADHA, KT; USHA, HS. Use of e-books in an academic and research environment: A case study from the Indian Institute of Science. Program: electronic library and information systems, 40(1), 2006, p. 48-62.

[7] BENNETT, L; LANDONI, M. E-books in academic libraries. The Electronic Library, 23(1), 2005, p. 9-16.

[8] ISMAIL, R; ZAINAB, AN. The pattern of e-book use amongst undergraduates in Malaysia: a case of to know is to use. Malaysian Journal of Library & Information Science, 10(2), 2005, p. 1-23.

[9] LITTMAN, J; CONNAWAY, LS. A circulation analysis of print books and e-books in an academic research library. Library Resources & Technical Services, 48(4), 2004, p. 256-262.

[10] CHRISTIANSON, M; AUCOIN, M. Electronic or print books: Which are used? Library Collections, Acquisitions, & Technical Services, 29(1), 2005, p. 71-81.

[11] SAFLEY, E. Demand for e-books in an academic library. Journal of Library Administration, 45(3-4), 2006, p. 445-457.

[12] SPRAGUE, N; HUNTER, B. Assessing e-books: Taking a closer look at e-book statistics. Library Collections, Acquisitions, & Technical Services, 32(3-4), 2009, p. 150-157.

[13] COUNTER. About COUNTER. 2009. Available at http://www.projectcounter.org/about.html (4 April 2010).

[14] SHEPHERD, PT. COUNTER: towards reliable vendor usage statistics. VINE: The Journal of Information and Knowledge Management Systems, 34(4), 2004, p. 184-189.

[15] VAN DER VELDE, W; ERNST, O. The future of eBooks? Will print disappear? An end-user perspective. Library Hi Tech, 27(4), 2009, p. 570-593.

[16] LEVINE-CLARK, M. Electronic book usage: A survey at the University of Denver. portal: Libraries and the Academy, 6(3), 2006, p. 285–299.

[17] BLUMMER, B. E-books revisited. Internet Reference Services Quarterly, 11(2), 2006, p. 1-13

[18] ABDULLAH, N; GIBB, F. Students' attitudes towards e-books in a Scottish higher education institute: part 1. Library Review, 57(8), 2008, p. 593-605.

[19] GREGORY, CL. "But I Want a Real Book" an investigation of undergraduates' usage and attitudes toward electronic books. Reference & User Services Quarterly, 47(3), 2008, p. 266–273.

[20] CONYERS, A. Usage statistics and online behaviour. In *The E-resources management handbook*, v.1, 2006, p.17-27. Available at http://uksg.metapress.com/media/5g048jxytg5xxxxtwc2l/contributions/5/5/0/u/550u5m1aku3aqjvk.pdf (4 April 2010).

[21] HACETTEPE UNIVERSITY. General overview. 2010. Available at http://www.hacettepe.edu.tr/english/ortak/universite/genel.php (4 April 2010).

[22] HACETTEPE UNIVERSITY LIBRARIES. Hacettepe University Libraries. 2009. Available at http://www.library.hacettepe.edu.tr (4 April 2010).

[23] TENOPIR, C. *Library resources: An overview and analysis of recent research studies*. Washington, D.C.: Council on Library and Information Resources. 2003. Available at http://www.clir.org/pubs/reports/pub120/pub120.pdf (4 April 2010).

[24] THOMPSON, JW. The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. Libri, 52(3), 2002, p. 121-136.

[25] TRUESWELL, RL. Some behavioral patterns of library users: The 80/20 rule. Wilson Library Bulletin, 43 (5), 1969, p. 458–461.

[26] EGGHE, L; ROUSSEAU, R. *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier. 1990. Available at http://hdl.handle.net/1942/587 (4 April 2010).

Digital content convergence: Intellectual property rights and the problems of preservation: A US perspective

John N. Gathegi

School of Library and Information Science, University of South Florida
4202 E. Fowler Ave, CIS1040, Tampa, FL 33620
jgathegi@cas.usf.edu

Abstract

One of the issues that this conference explores is the continuing phenomenon of convergence of communication, caused in part by the convergence of media and digital content. In this paper, we will review some of the intellectual property challenges that loom in this environment, with an emphasis on the situation in the Unites States. We shall discuss some of the peculiar features inherent in digital content that exacerbate the intellectual property problem, such as non-permanence, multiple, heterogeneous. We shall examine a couple of cases that illustrate some of the problems in this area. We shall then conclude with the problem of intellectual property and the multiple goals of digital content collections.

Keywords: digital rights, intellectual property, digital content

1. Introduction

The literature continues to indicate a continuing phenomenon of convergence of communication, caused in part by the convergence of media and digital content. We will review in this paper some of the intellectual property challenges that loom in this environment, with an emphasis on the situation in the Unites States. Some of the peculiar features inherent in digital content that exacerbate the intellectual property problem, such as non-permanence, multiple, and heterogeneous media, will be discussed. We shall examine a couple of cases that illustrate some of the problems in this area. We shall then conclude with the problem of intellectual property and the multiple goals of digital content collections.

2. The problem of non-permanency of dynamic content

Digital content has, unlike its print counterpart, some unique features that present challenges in both development and management, especially from a legal perspective. We shall examine three such features in this paper, the proposition that digital content (1) is dynamic, (2) may suffer issues of non-permanence, and (3) may have more than one media format.

Digital content is dynamic. As the need arises, items are constantly added and corrections and modifications made to specific files and databases. This means that a file may change from day to day. The problem then emerges on how to preserve digital content and vouch for its integrity.

Preservation efforts face many legal problems. The primary problem is how to ensure the non-infringement of copyright, by avoiding unauthorized exercise of the authors' exclusive rights, as well as determining what content is protected by copyright, to facilitate access to content as well as consent from copyright holders. A persistent question is whether the digital content manager still has the necessary rights in the e-content. Also, issues of privacy and confidentiality may be raised by the dynamism and non-permanence of digital content, as may ethical issues in health and personal data [1].

Despite perceptions to the contrary, "digital information is in fact fragile and at risk." Changes in technology can render some digital files corrupt and unreadable [2]. The longer the time frame required for future access, the more the uncertainty with information preservation. Challenges include changes in format, data definitions, and metadata content [1]. The format problem is exacerbated by the fact that many formats are proprietary and continue to evolve into more complex versions with newer features and functions, sometimes 'orphaning' earlier versions [3]. Legal access problems can occur when a proprietary owner contractually limits access or goes out of business [4].

One way of handling format changes in digital preservation is migration of data, both in terms of software and hardware. This will sometimes involve re-arranging structural and data elements sequence [3]. Two copyright problems arise. First, the act of migration usually will involve copying of the information, which may be an infringement of the author's exclusive reproduction right. Also, the re-arrangement of the structural and data elements may trigger the trampling of another right: the author's exclusive right to make derivative copies. Permission to migrate may have to be sought from the copyright holder. Other issues include, in the case of the

United States, whether a file conversion would be a violation of the Digital Millennium Copyright Act, and whether, for evidentiary purposes, a migrated file is the same as the original file.

3. Multi-media content and its complexity

As well as being dynamic and raising the problem of non-permanency, digital content may also contain a mixture of different media formats, including text, sound, graphics, video, and a variety of other file formats.

Good examples of multimedia digital content are electronic books, or e-books. An e-book could have, for example, an article about a country, a video about parts of the country, and a sound file of examples of music from the country.

E-books are usually in proprietary devices, but may also be accessible through a central server. An owner of an e-book collection has many of the similar features to a publisher of any other digital content in terms of the susceptibility of the content to be easily copied. Digital rights management technology is used to control access to e-book content that is copyright protected, to preserve the copyright owner's exclusive rights. There are, however, e-books available free of copyright protection that a digital content manager can link to from the digital collection [5]. E-book aggregators, such as netLibrary (a division of OCLC), provide access to a digital library's e-content on a 24/7 basis by negotiating intellectual property rights with publishers to provide access to content hosted on their servers. Aggregators usually provide their own digital rights management technology, thus easing legal issues for the digital content manager [6].

"Stocking" or publishing e-books in a digital collection requires that the digital collection manager understand the access limitations that come with the digital rights management, and the different pricing models. These models can range from outright purchases (much like print versions) to limited time and number of persons per access, and may also come with use restrictions that define practices such as printing, downloads, and amount of content that can be accessed.

Legal issues in this area are complicated by the fact that some media formats are covered by rules specific to the media (e.g. sound files). Also, conversion of media from one format to another may trigger copyright infringement (e.g. conversion of text into audio formats). More so than text works, dates on which a sound recording was first fixed determine the nature of the legal protection available in the United States. For example, no federal copyright protection was available to sound recordings prior to February 15, 1972, but the Sound Recording Amendment Act of 1971, rectified the situation by providing federal copyright protection to works recorded or fixed after that date [7]. Pre-1972 works, however, may be protected by state criminal law statutes or common law, against unfair competition or misappropriation, until February 15, 2067 [8].

Page thumbnails and document icons

Other newer versions of familiar formats, such as document icons and page thumbnails, present new legal issues.

Document icons are small visual representations of documents [9]. Icons can include information about a document format or genre (e.g., pdf document, web page or folder). Page thumbnails, on the other hand, are small images of a page usually in reduced resolution, that can be enlarged by a reader for viewing.

In discussing thumbnails, two rights that are exclusive to the copyright holder are implicated. Because they make copies of the images they crawl, search engines may violate the author's exclusive right to make reproductions of a work [10]. Also, because the thumbnails are shown to the users, search engines may also violate the author's exclusive right to public display [10]. However, the use of thumbnails may rely, as we see in the Kelly v. Arriba [11] and Perfect 10 cases below, on one of the exemptions to the author's exclusive rights: fair use.

In Kelly v. Arriba Soft Corp., a photographer whose copyrighted images were displayed by a visual search engine operator on the operator's web site and those that it had licensed sued the operator. The operator had built its database by copying images from web sites and reducing these images into "thumbnails" that could be enlarged by clicking on the thumbnail. The lower court ruled the operator's use of the thumbnails fair use, as the character and purpose of its use was "significantly transformative and the use did not harm the market for or value of [the photographer's] works" [11]. The 9th Circuit affirmed the lower court's ruling that the display of thumbnails was fair use.

In Perfect 10 v. Google [12] it was a website operator's turn to sue Internet search engines. Perfect 10 published adult photographs in both a magazine and a web site, and had expended considerable resources to the development of the brand name for the magazine and web site. Google and Amazon, search engine operators, have an image search function that retrieves thumbnail images in response to a textual search string query. Some of the images so retrieved came from Perfect 10's website, and it sued both

Google and Amazon. The district judge, when considering plaintiff Perfect 10's motion for an injunction against Google, put the issue in a perfect context:

> The principal two-part issue in this case arises out of the increasingly recurring conflict between intellectual property rights on the one hand and the dazzling capacity of internet technology to assemble, organize, store, access, and display intellectual property "content" on the other hand. That issue, in a nutshell, is: does a search engine infringe copyrighted images when it displays them on an "image search" function in the form of "thumbnails…?" [12]

The district court was of the view that Perfect 10 was likely to succeed in its claim that the display of thumbnails was a direct infringement by Google of its copyrighted images, and issued a preliminary injunction from creating and displaying Perfect 10's images. The district court distinguished Arriba's use of thumbnails in Kelly, in that Perfect 10's market for downloading reduced-size adult thumbnails into cell-phones was superseded by Google's use of Perfect 10's thumbnails. However, the 9th Circuit later ruled that the thumbnails were fair use because they did not detract from the economic value of the images, and thus Google could continue displaying Perfect 10 thumbnails that came up following a search [13].

4. Multiple, heterogeneous content: the legal complexities

A related feature to the complexity of digital content discussed above is the multiple, heterogeneous nature of digital collections. This feature can usefully be analyzed in two parts: the different types of digital collections, and the different goals of digital collections. Digital collections often have content of different types designed to meet a variety of goals.

The Internet, for example, can be viewed as one giant digital collection; a sort of a "meta-collection." Individuals, libraries, and other organizations often take subsets of the Internet to form specific collections. This is generally done through book-marking and linking. While there have been no legal challenges yet to bookmarking, linking has generated some legal issues, less so in the United States than in some European countries.

Likewise, libraries and other organizations may have, as part of their digital collections, commercial databases. The major issues here involve copyright protection and licensing issues. Data sets, on the other hand, may be viewed in the same category as commercial databases, with less emphasis on copyright and more emphasis on licensing issues.

One type of content, collective works and compilations requires a more extended discussion here. Section 101 of the U.S. copyright code defines a compilation as “a work formed by the collection and assembling of preexisting materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship. The term “compilation” includes collective works.” [14].

Just to make sure it is understood that compilations fall under the subject matter of copyright specified in section 102, section 103 explicitly declares that compilations are indeed included, but points out that protection extends only to what the author has contributed, and not to the underlying or pre-existing material. Nor does it extend to preexisting material that has been used unlawfully [15].

Collective works and compilations may or may not have common characteristics. In a collective work, individual components are generally independent copyrightable works, while compilations may include material that is not necessarily copyrightable [16]. Separate contributions to a collective work can have copyright protection that is distinct from copyright in the collective work as a whole. Owning a copyright in a collective work entitles the copyright owner to “only the privilege of reproducing and distributing the contribution as part of that particular collective work, any revision of that collective work, and any later collective work in the same series” [17]. The court in New York Times Co., Inc. v. Tasini [18] explored the question of whether a copyright owner in a collective work who republished all or a part of the compilation in an electronic database could prevail against an assertion of copyright infringement from the author of a contribution in the compilation. The case involved freelance writers who had sued a newspaper publisher for making their articles available in electronic databases. The newspaper asserted a privilege offered by section 201(c) of Title 17. Under section 201(c):

> Copyright in each separate contribution to a collective work is distinct from copyright in the collective work as a whole, and vests initially in the author of the contribution. In the absence of an express transfer of the copyright or of any rights under it, the owner of copyright in the collective work is presumed to have acquired only the privilege of reproducing and distributing the contribution as part of that particular collective work, any revision of that collective work, and any later collective work in the same series. [17]

The Supreme Court focused on the perception of a user of the articles as presented in the database, and rejected the newspaper's reliance on the section 201(c) privilege. The privilege, however, continues to be available given the right circumstances. A second circuit court, for example, affirmed the granting of summary judgment to the National Geographic publisher who had made a searchable digital collection of past issues of the magazine (dubbing it the Complete National Geographic), against freelance authors and photographers who had sued the magazine for the use of their work in this new medium [19]. The Court relied on its holding the fact that the original context of the magazine was present in the new Complete National Geographic, and that the digital work was a new version of the National Geographic Magazine. The database at issue in Tasini, on the other hand, did not allow users to view the underlying works in their original context.

Divergent goals

As well as being heterogeneous and having multiple formats, digital collections also have different goals. One of the goals of digitization, as mentioned above, is preservation. Some institutions have a legal privilege to preserve. Section 108 of the U.S. Copyright Code, for example, addresses the need for preservation and conservation [20]. Legal issues that are likely to arise here not only include copyright, but also evidence.

The issue of non-permanence that we discussed above acquires critical importance when it comes to maintaining documents for legal evidentiary purposes. In December 2006, changes were put into effect in the U.S. Federal Rules of Civil Procedure that institute a new category of evidence: the Electronically Stored Information (ESI), which is designed to work within the existing rules of production of "documents." Under the rules, a party must provide to other parties: "a copy--or a description by category and location--of all documents, electronically stored information, and tangible things that the disclosing party has in its possession, custody, or control and may use to support its claims or defenses…" [21]. While the rule does not specify the version of the electronically stored information that should be produced, Rule 26(f) does oblige the parties to conference and "…discuss any issues about preserving discoverable information" as well as "any issues about disclosure or discovery of electronically stored information, including the form or forms in which it should be produced."

A closely related issue to non-permanence for evidence is authenticity. Digital information can be vulnerable to tampering or corruption. Depending on the nature of the collection, authentication methods such as digital signatures, version control, and encryption may be necessary [22].

Finally, we cannot conclude without mentioning something about access. Access and preservation are much intertwined. There could be any number of reasons for seeking access, including for entertainment, research, or safeguarding culture. Copyright is always an issue, as, for example, not all copyrightable works have the same protection duration. Different publications are covered under different copyright protection terms, depending on when they were created or published. However, the issues most likely to rise are those of licensing for access. By access, we are also here referring to use. Unfortunately, this issue is for the moment outside the scope of this paper.

Notes and References

[1] ANDERSON, WL. Some Challenges and Issues in Managing, and Preserving Access to, Long-Lived Collections of Digital Scientific and Technical Data. Data Science Journal 3, 2004, p. 191-201. Available at http://www.jstage.jst.go.jp/article/dsj/3/0/191/_pdf (June 29, 2009).

[2] ADAMS, G. Partners Go Dutch to Preserve the Minutes of Science. Research Information, September/October, 2004. Available at http://www.researchinformation.info/risepoct04archiving.html (March 10, 2010).

[3] ARMS, CR. Keeping Memory Alive: Practices for Preserving Digital Content at the National Digital Library Program of the Library of Congress, RLG DigiNews, 4(3) 2000, Available at http://www.rlg.org/preserv/diginews/diginews4-3.html#feature1

[4] BARNES, I. *The Preservation of Word Processing Documents.* Canberra: Australian Partnership for Sustainable Repositories, 2006. Available at http://www.apsr.edu.au/publications/word_processing_preservation.pdf (June 29, 2009).

[5] ESPOSITO, J. The Processed Book, First Monday, 8 (3) 2003 (Updated 2005). Available at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1038/959 (March 25, 2010)

[6] GARROD, P. Ebooks in UK Libraries: Where Are We Now? Ariadne, no. 37 (2003). Available at http://www.ariadne.ac.uk/issue37/garrod/

[7] BESEK, J.M. *Copyright Issues Relevant to Digital Preservation and Dissemination of Pre-1972 Commercial Sound Recordings*. Washington,

DC: Council on Library and Information Resources and Library of Congress, 2005. p. 16-17.

[8] 17 USC 301(c) provides: “With respect to sound recordings fixed before February 15, 1972, any rights or remedies under the common law or statutes of any State shall not be annulled or limited by [Title 17] until February 15, 2067.”

[9] JANSSEN, WC. *Document Icons and Page Thumbnails: Issues in Construction of Document Thumbnails for Page-Image Digital Libraries*. Palo Alto, CA: Palo Alto Research Center, 2004

[10] 17 USC 107

[11] 280 F.3d 934 (9th Cir. 2002)

[12] 416 F. Supp. 2d 828 (C.D. Cal. 2006)

[13] PERFECT 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007).

[14] 17 USC 101

[15] 17 USC 102, 103

[16] ROY Export Co. Establishment of Vaduz, Liechtenstein, Black Inc., A. G. v. Columbia Broadcasting System, Inc., 503 F. Supp. 1137 (S.D. N.Y. 1980), judgment aff'd, 672 F.2d 1095 (2d Cir. 1982)

[17] 17 USC 201(c)

[18] NEW York Times Co., Inc. v. Tasini, 533 U.S. 483

[19] FAULKNER v. National Geographic Enterprises Inc., 409 F.3d 26 (2d Cir. 2005)

[20] 17 USC 108

[21] FEDERAL Rules of Civil Procedure 26(a)(1)(A)(ii)

[22] GERMAIN, CM. Legal Information Management in a Global and Digital Age: Revolution and Tradition, International Journal of Legal Information, 35 (1) 2007. p.134-163, 156-157

An adaptable domain-specific dissemination infrastructure for enhancing the visibility of complementary and thematically related research information

Engin Sagbas;[1] York Sure[1, 2]

[1] GESIS – Leibniz Institute for the Social Sciences, Information Processes in the Social Sciences (IPS), Bonn, Germany
[2] University of Koblenz-Landau, Institute for Computer Science, Information Systems and Semantic Web (ISWeb), Koblenz, Germany
{engin.sagbas, york.sure}@gesis.org

Abstract

We introduce an adaptable domain-specific infrastructure for dissemination of heterogeneous outcomes (e.g. publications) from thematic complementary and related projects. Our aim is to enhance the visibility of thematically related research information and to face obstacles from both sides: needs from information users and information providers. Users are confronted with finding sources for relevant information, handling with heterogeneous information display, varying information granularity on different sources, extracting and compiling the information found whereas information providers have costs for implementing and maintaining such an infrastructure from scratch, limit or omit coupling with different related sources and offer information partly in an interconnected manner. The contributions of this paper include a model closely related to the CERIF standard and a technical infrastructure ready to reuse to set up a research information system for a new research topic. We created a reference portal on the topic “Governance in the EU”.

Keywords: dissemination infrastructure; information retrieval; research information; CERIF

1. Introduction

Information visibility of complementary and related information on the web is an important claim from a user's perspective. For example, collaborative research in large projects and complementary research by other related projects across national and international research institutes have the problem not to be adequately visible for those who are not familiar with the related projects. Transparency is hampered, e.g. about the produced outcomes in a research field, established research structures and connections to other related projects. Furthermore, there are different target groups with different information needs like researchers seeking new, relevant papers; project coordinators and managers looking for project specific documents; and the general public interested in new developments in specific research topics. Users with these different needs usually start finding the relevant information by using different search engines or available specific project information systems. It is a very tedious and time consuming task for a user to find and use several relevant information systems and websites. The success of finding the requested information across several sources is uncertain. In addition, the results found are heterogeneous, i.e. they mostly have a different kind of information display and granularity. Besides, if the information is not directly interconnected to related sources, the access to relevant complementary resources is hampered. For the information provider there is a challenge to build such a project dissemination infrastructure usually from scratch that gathers these information needs from the user.

2. Challenges

We identify challenges from two sides: On the one side, the information consumer needs, and, on the other side, the information provider needs. The information consumer side usually consists of individuals participating in the projects, users of the projects' results like external researchers, policy makers, and the interested public. They are confronted with the following obstacles:

- In project information systems like CORDIS (European Commission's Research Information System) [1], information is currently available at the level of the individual research projects. Persons interested in individual project outcomes like conferences and publications are required to visit the websites of all projects dealing with the topic of interest. The visibility of the projects and of their collective contribution to the realization of the Framework

Programme priorities and European Research Area [2] is therefore rather limited.

- Users who will not know in advance which type of information or service is to be expected from each project website, are forced to find and visit all project websites including those not relevant to them.
- Due to the lack of interconnections to complementary and related information across project boundaries, users will usually have to visit multiple websites for further information needs.
- By visiting each website, users have to learn the sites' structure, how to find and access relevant information on each project website, and finally compile themselves the heterogeneous materials with varying qualities found on different sites [3]. Mostly, it is a time consuming and inconvenient task for users.
- Biased by the above problems, users miss the big picture for relevant and related information.

In contrast, the information provider needs are characterized with the following common situations:

- Spending time and financial resources for implementing the project dissemination activities for each project resulting in several websites with similar infrastructures which are usually project-specific and isolated. Therefore, they are not coupled with the complementary information from other information providers.
- Across all projects, different dissemination and sustainability strategies beyond the projects' duration will make it difficult to ensure the availability of project results in the long run yielding information websites that are not maintained or no more visible [4].
- The lack of a topic-oriented research infrastructure for dissemination of complementary project outcomes can lead to an unnecessary duplication of work on the provider side, and an increased effort for finding and accessing relevant information on the user side.

3. Approach

Our contributions are making thematically connected research activities visible at a single place together with their results, giving users integrated access to currently distributed resources at a common level of quality of service; to provide an adaptable technical infrastructure for information providers facilitating dissemination of heterogeneous outcomes from thematic complementary projects targeted to different audiences; to integrate and compile heterogeneous data from different sources providing quality data for the purpose of analyzing, visualizing and reusing by other services; to provide a collaborative infrastructure connecting interested and active researchers; to reuse the complete information infrastructure for a new domain reducing costs for acquisition and implementing; and to facilitate sustainability of project outcomes after the project's end.

The main pillar of work carried out focused on the development of a technical dissemination infrastructure which covers all entities relevant in the context of research information (i.e. actors, activities and results) at a very detailed level and at the same time interconnected them both within the context of an individual project and across project boundaries.

3.1 The Conceptual Model

In the first phase of the EU project IConnectEU [5], the different outcomes produced by eight complementary projects and the target audience of these outcomes were analyzed and a core model was defined for documenting these outcomes together with information about participating institutes and persons at a very detailed level. The core model is closely related to the Common European Research Information Format (CERIF), which was funded by the European Commission and is maintained by euroCRIS [6], a professional organization dedicated to improvement of research information availability since the release of CERIF2000 [7]. Compatibility to CERIF, in specific to its exchange format CERIF-XML [8], supports reusing research information across institutional and geographic boundaries.

The CERIF model is built around three core entities of research information and three result entities. Core entities are project, person and organizational unit. The result entities consist of publication, patent, and product. These entities are connected with typed links which represent the semantic relationships between these entities expressing, e.g. the members of a project, the affiliation of a person, project outcomes, authors of publications, and persons with specific project roles like coordinator, to name a few examples.

These entities are reflected in the information architecture, where semantic annotations, i.e. attributes, were used to describe these entities. They have been partly expanded in regard to the attribute set defined in CERIF. This not only includes additional information on e.g. project work packages, data collections or scientific methods, but also includes geographic location and coverage of all entities, target groups of activities and results.

We specified the basic requirements in a core model that includes all relevant entities, with their describing attributes and the relationships between them. We modeled the research information context comprising of the following core entities: project, institute (institutional participants involved in a research project), person (doing research and affiliated with an institute), and research results (including project research outcomes like publications, events, research data or other produced results).

Figure 1 depicts the conceptual model. Each of these entities (displayed in oval form) has its own set of mandatory and optional attributes, which adequately describe the single entity. Attributes are grouped in formal attributes, specific attributes, attributes for content describing and indexing, and finally, semantic relations for interconnecting entities.

For example, attributes for describing projects are divided in:

- formal attributes: title, acronym, begin, end, funding, URL etc.,
- specific attributes: funding agency, thematic priority, instrument etc.,
- content describing attributes: summary, research area, objectives, work packages etc.,
- content indexing attributes: geographic coverage, thesaurus keywords, free keywords etc., and
- semantic relations: linking semantically the project entities to institutes, persons, and results and interconnecting all entities in the core model to express relationships between two entities, e.g. “Mike” (person) is involved in “Apollo” (project). Other relation type attributes can be defined and used like “coordinator of”, “author of”, or “cooperate with”.

The result entity is subdivided in different research outcome entities produced by all projects. Especially publications are the most prominent example for representing research outcomes. Events like conferences or workshops, produced results like research data, and other generated project resources are also covered project outcomes. Each result entity has its own describing attribute set and is interconnected within the research context.

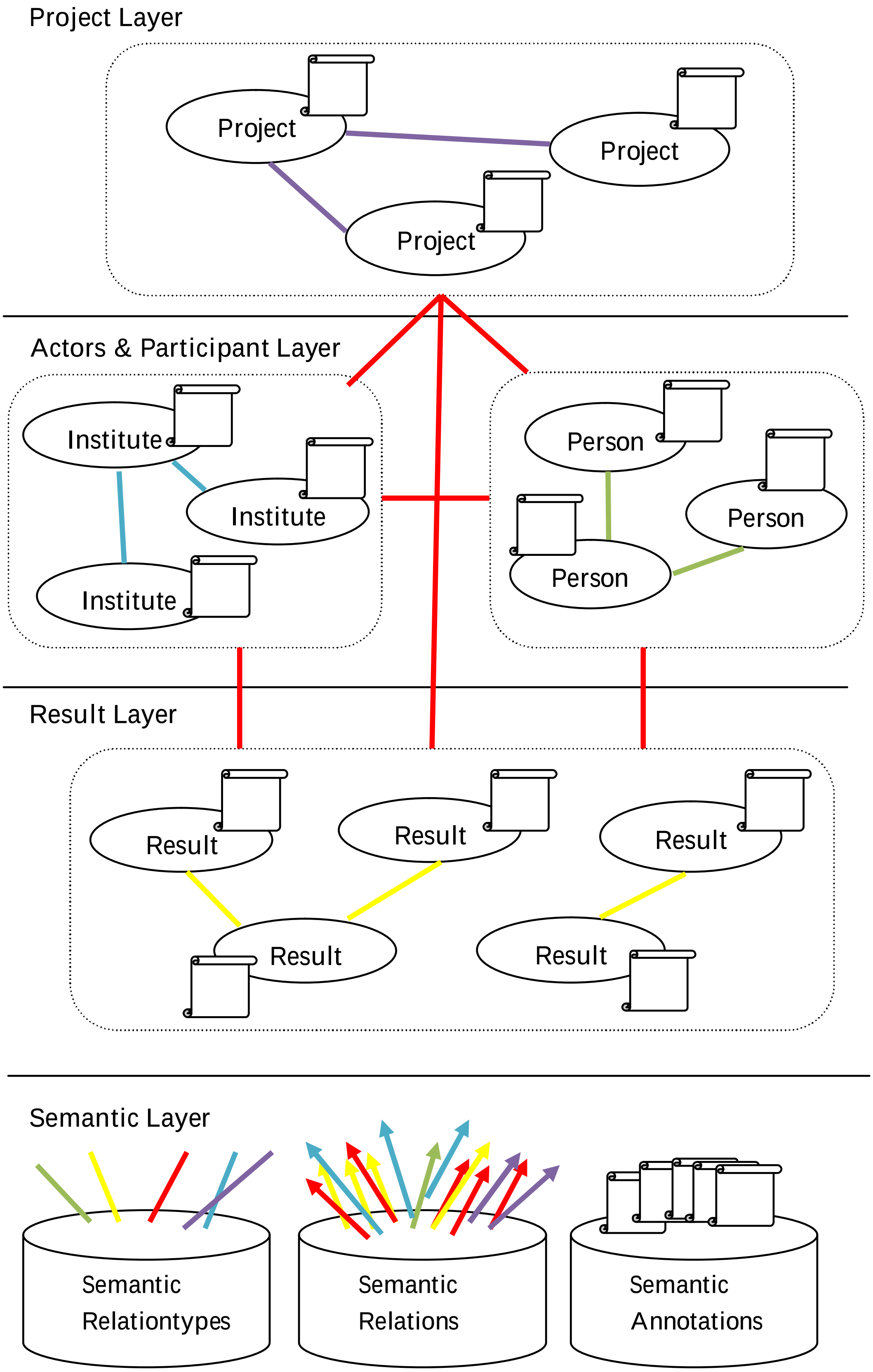


Figure 1: Conceptual model in the context of research information

3.2 The Architecture Model

The overall architecture model is shown in Figure 2. The model considers the different information needs and implements the conceptual model described in the last section. The middleware consists of two main parts. A content management system manages the editorial static contents. The dynamical contents representing the conceptual model in Figure 1 are realized by a cataloguing system. This combination allows exploiting the synergy effects of both specialized systems so that information providers can both build complex portal structures combined with functionality of a cataloguing system providing dynamic contents.

We use for the technical infrastructure of the IConnectEU reference portal as cataloguing system DBClear [9], which is developed in a project funded by the German Research Foundation (DFG). Due to its flexibility, DBClear has successfully been adapted to several use cases where a web-based cataloguing system was needed to collect and map information, e.g. in the FP5 project "MORESS - Mapping of Research in European Social Sciences and Humanities", in around 10 Digital Libraries in Germany, and recently in "SSOAR – Social Science Open Access Repository"[10], which was also funded by the DFG. As content management system we adopted Typo3 [11]. Both software packages are open source.

Our infrastructure is flexible in regard to adapting it to a new domain. Since the conceptual model developed is generic for research information, it is not restricted to a particular research discipline. All kinds of research topics in different research domains can in principle be covered by the model, e.g. research topics in social sciences, life sciences, or natural sciences. Reusing the conceptual model can significantly reduce the effort to set up a specific topic-oriented research information system.

Customizing and extensions might be useful to adapt the model to new emerging needs, e.g. for particular requirements of a new domain. DBClear has two main strengths. The flexibility both in defining and editing semantic annotations and flexibility in adapting the information display view for the user interface, both feasible during and after system implementation. For example, we could define a new relation type called "cooperate-with" and then annotate persons who cooperate with each other (semantic relations). This would represent a cooperation network of persons. There are no limitations, i.e. we can add or adapt all semantic annotations and semantic relation types for the given entities in Figure 1.

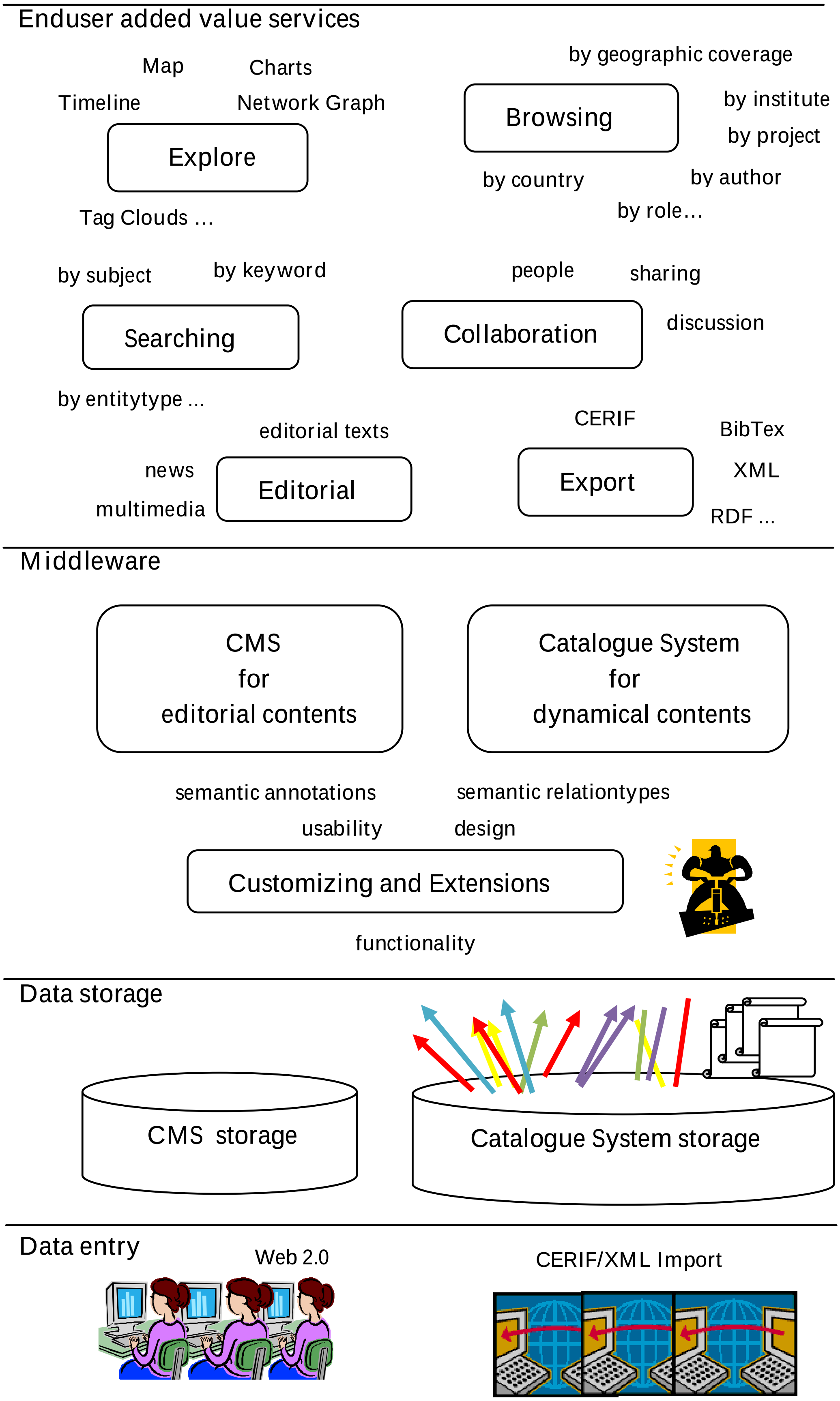


Figure 2: Overall architecture model

3.3 Enduser added value services

Building a useful research information system from a user's perspective requires attention for different user needs. Especially, added value services could be a crucial incentive using it. Figure 2 lists a few valuable services. For example, browsing frequently requested research information in predefined categorized views facilitates quick access to the requested information, e.g. persons can be browsed by project members, by institute members, by project roles, by authors or by another person view adopting a relation type to a person. Further services like new visualization types (e.g. map, charts) or useful export formats for publications like BibTex can be realized.

3.4 Data entry and Data gathering

Each information system with its varying functionality and services rely on the availability of quality and up-to-date data. Thus, an issue is where to get these data. There are approaches for automatically harvesting and collecting relevant data, e.g. by web mining, but we focus here on the idea of Web 2.0. This has the advantage that a human collaboration led to better results of gathering quality data by exploiting the social intelligence [12]. We focus on researchers who would provide their own data and integrate it with the others. Another associated issue is, how we can get the user's data without forcing them to use multiple places for data entering. In this case, we can use standards for reusing research data. The CERIF standard adopting by research information systems (CRIS) makes research information across different systems available using the CERIF exchange format. Export formats from proprietary systems are imported by using XML and CSV formats. The flexibility of DBClear to define new mapping templates for different data export formats by using XSLT makes large amounts of data reusable from other quality controlled systems.

4. Related Work

The IConnectEU model is closely related to the CERIF model which is regularly updated. The current release is CERIF 2008 [13]. IConnectEU is consistently structured according to the CERIF core entities: project, institute, person, and results. There is a mapping defined for the intersection between the IConnectEU and the CERIF entities. Due to the specific use case of IConnectEU, not all entities from CERIF are relevant or used. Otherwise there are some entities and attributes defined in IConnectEU but not present in

CERIF. If required, the conceptual model can be evolved or adapted to the specific discipline or emerging new requirements.

Relying on the CERIF standard ensures that the data can be reused by other third-party research information systems. For example, the IST World project [14] adopts the CERIF standard so that the IConnectEU data can be further reused, e.g. to apply advanced technologies for visualization like a competence diagram. The IST World project complements IConnectEU since its focus is on analyzing research competencies across European countries [15]. IConnectEU's primary focus is to provide a complete model for covering typically research information combined with a ready for reuse software infrastructure applicable for any research topic.

There are other portals like sowiport.de [16] (one of the largest information portals for the social sciences in Germany) or vascoda.de [17] (an interdisciplinary portal for scientific information in Germany) which provide a broad range of information from multiple integrated databases. Science gateways like WorldWideScience.org [18] enable federated searching of national and international scientific databases and portals. In contrast to these portals, IConnectEU has a narrow thematic focus on a research topic within a discipline connecting only thematically related projects, e.g. projects with research on the topic "EU Governance". In this sense, it is a lightweight thematically focused information system not intended to be a literature database with millions of entries from different areas. Especially, IConnectEU covers the project context with research information (e.g. persons, institutes, results) in a semantically interconnected manner which is not or partly provided in this form by the mentioned portals. To sum up, IConnectEU can contribute topic oriented research information and data to larger (multi-) disciplinary portals and profit from accessibility from those portals gaining new users who have a special research focus.

Search engines like Google might be useful but their retrieval effectiveness [19] is limited in the context of finding complementary research information. Due to the fact that IConnectEU is thematically focused and the data is handpicked and quality controlled by the partners, all search results are thematically relevant and related, which result in a high retrieval quality.

5. Conclusion and future work

We addressed common needs of both information users and information providers. From the user's perspective there are obstacles to find and retrieve relevant and related research information on the web. These include finding the relevant websites and sources, confronting with the heterogeneity on different sources like different information display and granularity, and extracting as well as compiling the collected information from the web. In contrast, obstacles for information providers are costs for implementing and maintaining a complex dissemination infrastructure, coupling with other related sources, and providing interconnected information to thematically related information. Another issue is sustainability of information beyond the projects' duration.

We introduced the benefits of the IConnectEU infrastructure allowing topic-oriented organization of complementary research information and outcomes. The typical research information, e.g. from projects, institutes and persons to publications, conferences and other results are covered and interlinked semantically. Research networks can be represented, i.e. linking researchers and institutes across projects and countries. Besides, detailed information and extensive metadata for a large number of information entities are provided. The data is compatible with the CERIF standard for research information promoted by the European Commission. Reusing data by other services for the purpose of analyzing and visualizing adds several new dimensions to geographical analysis, e.g. mobility of researchers, development of collaboration networks, and inclusion of regions in European funded research.

IConnectEU strengths are based on an adaptable infrastructure. Using it for dissemination in other research topics will significantly reduce the provider's costs. The software developed in IConnectEU is made freely available as open source software to third parties.

Future work includes the Web 2.0 approach to get research information for the data entry process directly from the involved persons. This requires a collaborative infrastructure that eases the data gathering process since everyone would be responsible for maintaining his/her own part of contribution. Our approach for dealing with the issue will include:

- Providing an incentive by establishing new added value services like new visualization and exploration services for data, e.g. map visualizations allowing geographical analysis.
- Bridging to social networks for special target groups like researchers using XING, and making an incentive to join the platform [20].

We will extent the current dissemination platform to a collaborative infrastructure where users maintain their own data and collaborate together, e.g. discussions on a topic. Further, we prove use cases for connecting to social network users by using the open social standard [21], which is also part of future work.

Notes and References

[1] CORDIS: http://cordis.europa.eu/

[2] European Commission Research http://ec.europa.eu/research/

[3] P. Oliveira, F. Rodrigues, P. Henriques, und H. Galhardas, A taxonomy of data quality problems, *Proceedings of 2nd International Workshop on Data and Information Quality,* 2005, p. 219-233.

[4] K.H. Lee, O. Slattery, R. Lu, X. Tang, und V. McCrary, The state of the art and practice in digital preservation, *Journal of Research-National Institute of Standards and Technology*, vol. 107, 2002, p. 93-106.

[5] IConnectEU: http://www.iconnecteu.org/

[6] euroCRIS: http://www.eurocris.org/

[7] A. Asserson, K.G. Jeffery, und A. Lopatenko, CERIF: Past, Present and Future: an Overview. Gaining Insight from Research Information, *6th International Conference on Current Reseach Information Systems*, 2002, p. 29-31.

[8] B. Jörg, O. Krast, K.G. Jeffery, und G. van Grootel, CERIF 2008–1.1 XML: Data Exchange Format Specification*. euroCRIS*, March, 2010.

[9] H. Hellweg, B. Hermes, M. Stempfhuber, W. Enderle, und T. Fischer, DBClear: A Generic System for Clearinghouses, *Gaining Insight from Research Information*, 2002, p. 131.

[10] SSOAR: http://www.ssoar.info/en/

[11] Typo3: http://typo3.org/

[12] J.F. Jensen, User-generated Content – a Mega-trend in the New Media Landscape, *Interactive TV: Shared Experience, TICSP Adjunct Proceedings of EuroITV2007*, 2007, p. 29–30.

[13] B. Jörg, K.G. Jeffery, A. Asserson, und G. van Grootel, CERIF 2008–1.1 Full Data Model: Introduction and Specification*. euroCRIS*, March, 2010.

[14] IST World: http://www.ist-world.org/

[15] B. Jörg, J. Ferle, H. Uszkoreit, und M. Jermol, Analyzing European Research Competencies in IST: Results from a European SSA Project, Bošnjak&Stempfhuber, 2008.

[16] Sowiport: http://sowiport.de/

[17] Vascoda: http://vascoda.de/

[18] WorldWideScience: http://worldwidescience.org/

[19] D. Lewandowski, The retrieval effectiveness of web search engines: considering results descriptions, *Journal of Documentation*, vol. 64, 2008, s. 915-937.

[20] L. Leung, User-Generated Content on the Internet: An Examination of Gratifications, Civic Engagement, and Psychological Empowerment, *New Media & Society*, 2009.

[21] OpenSocial: http://www.opensocial.org/

Translation of XML documents into logic programs

Martin Zima; Karel Jezek

Department of Computer Science & Engineering
University of West Bohemia in Pilsen
Universitni 8, 306 14 Pilsen, Czech Republic
{zima, jezek_ka}@kiv.zcu.cz

Abstract

The semantic web is supposed to become a characteristic phenomenon of the worldwide web in the next decade. One of the basic semantic web tools is the XML language. The aim of this paper is to provide information on how web documents written in the XML language can be rewritten into logic forms expressed as Prolog/Datalog programs. The XML language constitutes the basis of many semantic web languages and information in XML documents is usually retrieved with the help of procedural language called XQuery. Retrieving based on logic formulas gives us the chance to take advantage of deduction and this way to gain new, originally hidden information.

Keywords: semantic web; logic programming; XML; Datalog language.

1. Introduction

An interesting and topical research issue is the use of logic rules (logic program) to evaluate a query about XML documents [1]. It provides an option to combine XML technology with the inference capabilities of logic programming. Logic programming allows us to evaluate the queries that require computation of a transitive closure of relations. This means that we can query such information that is not explicitly included in the documents. In other words, we are able to draw deductive conclusions concerning the facts contained in the documents and in this way to find new, but originally hidden facts. Suppose, for example, that we have an XML collection of scientific articles. The logic program allows a query such as: *find the names of all co-authors of the given author, including co-authors of found co-authors, etc.* It means this query evaluates the transitive closure of the relation co-author. There are many similar tasks having something to do with transitivity of relations, e.g. looking for a path from one place to some destination, searching for owners of a given company (as frequently real final owners are hidden behind the companies that transitively own other companies), etc. All compound queries have the form of logic formulas. Therefore it is natural to use the logic query languages and, consequently, transform XML documents into expressions written in the logic language. This task is particularly interesting for us, as a few years ago we implemented an experimental deductive system [2], translating the Datalog language into the PL/SQL procedural language [3], which is used in the DBMS Oracle.

The use of Datalog instead of Prolog has some pragmatic reasons too. A database-oriented system with a professional database management system is able to process large data collections within a reasonable time. This is particularly important in the case of web documents processing.

2. Logic programming and Datalog

To be able to explain the method of translation, let us briefly introduce some principles of logic programming and its form used in the language Datalog [4]. Datalog is a slightly modified version of the primary logic language Prolog [5] and is tailored to database processing.

The logic program consists of a set of facts, a set of logic rules and a query. On the basis of the logic program execution, a set of new facts can be inferred and delivered as the query result. Every logic program can contain constants and variables. The names of variables begin with upper case letters. The exceptions are anonymous variables, i.e. variables whose values we are not interested in. They are marked with an underscore. Names of predicates begin with lower case letters and predicates are distinguished by the number of their arguments as well. Facts have the common form

```
predicate_name(list of constant arguments).
```

Rules have the common form `head :- body.` The symbol “`:-`” means “if” and expresses an implication between the truth of the body and head predicates. The head is a predicate name, the arguments of which are mostly variables. Such variables are evaluated during the program execution and in case the head predicate is TRUE, they are returned as the result of the rule processing. The rule body consists of predicates whose arguments have to contain all variables from the rule head. The head predicate becomes TRUE if there exist such values of variables in the logic program that the values of all predicates in the rule body are TRUE too. If such values of variables do not exist, the rule is evaluated as FALSE. The query consists of a predicate whose arguments are variables or constants. The deductive system tries to find values of query variables which are derivable from existing program facts with the possible use of program rules. The query succeeds if such values exist, otherwise, the query predicate has the value FALSE and the query answer is NO.

Our experimental deductive system implements an extended version of the Datalog language. The current extensions include relational operators, assignment and not operation for negation.

3. Translation of XML documents

Advantage of logic programming for XML documents querying and processing has inspired other researchers too. To our knowledge, a similar issue was described by Jesús M. Almendros-Jiménez [6]. He proposed a possible solution to the problem of converting XML documents to a logic program written in Prolog. His solution uses a list structure which describes data of the XML document and represents the result of the XPath query [7]. The rules of the logic program define the structure of the XML document (the way the elements are nested within other elements). He introduces specific functions with a different number of arguments to specify the XML documents structure, but the process of functions evaluations is missing. The resulting logic program contains all data of the input XML document in a set of facts. The structure of the XML document is fixed in logical rules. This means that each XML document is transformed into a different set of rules.

On the contrary the method proposed by us generates a logic program which consists of universal rules. That is, two different logic programs (results of the transformation of two different XML documents) contain the same logic rules. Differences between the structures of various XML documents are captured by the facts. This technique also eliminates the need to work with lists that our implementation of Datalog still lacks.

3.1. Construction of the set of facts

To show our method of generating a set of logic facts, we need to choose some example of an input XML document. Such suitable candidate is e.g. books.xml, a modified XML document describing a library content, which was adopted from [6]. The text of the XML document, accompanied by a number of lines, is shown in Fig. 1.

```
 1 <?xml version="1.0" ?>
 2 <bookshelf>
 3   <book year="2003">
 4     <author>Abiteboul</author>
 5     <author>Buneman</author>
 6     <author>Suciu</author>
 7     <title>Data on the Web</title>
 8     <review>A fine book.</review>
 9   </book>
10   <book year="2002">
11     <author>Buneman</author>
12     <title>XML in Scotland</title>
13     <review>The best ever!</review>
14   </book>
15 </bookshelf>
```

Figure 1: XML document books.xml

As [6] shows, the information "*Buneman is the author of the first book*" describe this fact:

```
author('Buneman', [2, 1, 1], 3, 'books.xml').
```

- The first argument is the value of the element <author>.
- The second argument defines the XML document structure: Number 2 means: the element is the second one inside another element. The first 1 means that the element <book> is the first one inside another element <bookshelf>. The second 1 stands for the element <bookshelf> which is the root element of the document.
- The remaining arguments do not require further explanation.

Before explaining the above-presented transformations, let us modify the sample XML document so it is consistent with the standards of the Semantic Web. An adapted version of the XML document is shown in Fig. 2. The above-mentioned information *"Buneman is the author of the first book“*, will be written into a triple of auxiliary logic facts that define the predicate xml. In this way we simultaneously eliminate the need to use the data structure list.

```
xml(5, 'dc:creator', 2, 3).
xml(5, 'bk:book', 1, 2).
xml(5, ′bk:bookshelf′, 1, 1).
```

The predicate `xml` mostly shows only the structure of the document. The first argument of the `xml` predicate holds the row number of the input XML document. This value will be the same for all logic facts defining the predicate `xml`, i.e. facts which describe a specific occurrence of the element written in a given row of the XML document. The other three arguments define the path in the XML document, i.e. the path from the given element to the root element of the document. To be specific, on line 5 there is recorded the element `<dc:creator>`, whose parental element is `<bk:book>`. The element `<bk:book>` contains a total of 5 children, of which the second child element is `<dc:creator>` located on line 5. The element `<bk:book>` has a parental element `<bk:bookshelf>`. This element contains 2 children (books). The element `<dc:creator>` from line 5 is contained in the first element `<bk:book>`. The last of the three facts says that the described `<bk:bookshelf>` element is the root element of the document.

```
 1 <?xml version="1.0" ?>
 2 <bk:bookshelf
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:bk="http://example.org/books/">
 3   <bk:book year="2003">
 4     <dc:creator>Abiteboul</dc:creator>
 5     <dc:creator>Buneman</dc:creator>
 6     <dc:creator>Suciu</dc:creator>
 7     <dc:title>Data on the Web</dc:title>
 8     <bk:review>A fine book.</bk:review>
 9   </bk:book>
10   <bk:book year="2002">
11     <dc:creator>Buneman</dc:creator>
12     <dc:title>XML in Scotland</dc:title>
13     <bk:review>The best ever!</bk:review>
14   </bk:book>
```

Figure 2: Adapted XML document

```
xml(2, 'bk:bookshelf', 1, 1).
xml(3, 'bk:book', 1, 2).
xml(3, 'bk:bookshelf', 1, 1).
xml(4, 'dc:creator', 1, 3).
xml(4, 'bk:book', 1, 2).
xml(4, 'bk:bookshelf', 1, 1).
xml(5, 'dc:creator', 2, 3).
xml(5, 'bk:book', 1, 2).
xml(5, 'bk:bookshelf', 1, 1).
xml(6, 'dc:creator', 3, 3).
xml(6, 'bk:book', 1, 2).
xml(6, 'bk:bookshelf', 1, 1).
xml(7, 'dc:title', 4, 3).
xml(7, 'bk:book', 1, 2).
xml(7, 'bk:bookshelf', 1, 1).
xml(8, 'bk:review', 5, 3).
xml(8, 'bk:book', 1, 2).
xml(8, 'bk:bookshelf', 1, 1).
xml(10, 'bk:book', 2, 2).
xml(10, 'bk:bookshelf', 1, 1).
xml(11, 'dc:creator', 1, 3).
xml(11, 'bk:book', 2, 2).
xml(11, 'bk:bookshelf', 1, 1).
xml(12, 'dc:title', 4, 3).
xml(12, 'bk:book', 2, 2).
xml(12, 'bk:bookshelf', 1, 1).
xml(13, 'bk:review', 5, 3).
xml(13, 'bk:book', 2, 2).
xml(13, 'bk:bookshelf', 1, 1).

data(4, 'dc:creator', 'Abiteboul').
data(5, 'dc:creator', 'Buneman').
data(6, 'dc:creator', 'Suciu').
data(7, 'dc:title', 'Data on the Web').
data(8, 'bk:review', 'A fine book.').
data(11, 'dc:creator', 'Buneman').
data(12, 'dc:title', 'XML in Scotland').
data(13, 'bk:review', 'The best ever!').

attribute(3, 'bk:book', 'year', '2003').
attribute(10, 'bk:book', 'year', '2002').
```

Figure 3: Generated facts

The data itself (names and attributes of elements, text content, etc.) has to be saved through other predicates. Let us introduce for this purpose a piece of predicate `data` which will contain the line number, the element name that contains the relevant text information and the value, i.e. the text content of this element. The fact that *"Buneman is the author of the first book"* is expressed as follows:

```
data(5, ′dc:creator′, ′Buneman′).
```

Attributes and values will also be recorded in the form of facts. We can introduce the predicate `attribute` containing the line number, the element name by which an attribute is defined, the attribute name and finally its value. For example, as our document defines only one attribute `year` within the element <`bk:book`>, the corresponding fact will have the form:

```
attribute(3, ′bk:book′, ′year′, ′2003′).
```

Fig. 3 shows the full set of facts defining the predicates `xml`, `data` and `attribute`, which were obtained by translation from the XML document in Fig. 2.

3.2. Universal rules

The logic facts defining the predicate `xml` hold the structure of the input XML document. Therefore, it is possible to determine which element is a part (descendant) of another element. As this process is recursive we can define an auxiliary predicate `intersection`. This predicate is defined by two rules, which look for the intersection of sets of facts determining the predicate `xml`. Each set describes an element from one line of the XML document. The set of logic facts from Fig. 3 defines 11 various sets in total. To demonstrate, we selected two sets which describe the elements listed in lines 3 and 6.

```
xml(3, 'bk:book', 1, 2).
xml(3, 'bk:bookshelf', 1, 1).

xml(6, 'dc:creator', 3, 3).
xml(6, 'bk:book', 1, 2).
xml(6, 'bk:bookshelf', 1, 1).
```

At first sight it is clear that both sets have three common arguments: `'bk:bookshelf', 1, 1`. They describe the root element of the XML document, which must be specified in any set of facts defining the predicate `xml`. The form of rules defining the predicate `intersection` is as follows:

```
intersection(Line1, Line2, Element, N, 1) :-
  xml(Line1, Element, N, 1),
  xml(Line2, Element, N, 1),
  Line1 < Line2.
```

The above-mentioned sets also have another three common arguments: `'bk:book', 1, 2`. The value 2 defines the level of nesting of the element, the value 1 indicates the root element. To ensure that these two facts will also be included in the intersection, the intersection must include the fact defined at a lower level. This results in the following recursive rule:

```
intersection(Line1, Line2, Element, N, P2) :-
  xml(Line1, Element, N, P2),
  xml(Line2, Element, N, P2),
  P1 := P2 – 1,
  intersection(Line1, Line2, _, _, P1).
```

The pairs of line numbers, which are the results of an evaluation of the predicate `intersection`, do not guarantee so far that on these XML lines there are written two immediately nesting elements, e.g., that on `Line2` there is written such element, whose parent is written on `Line1`. This condition applies only in the following case. With `Line1` there is associated such a set which is identical to the intersection of both sets and the second set, associated with `Line2`, contains one more fact. This condition is true for elements on lines 3 and 6 of the XML document, but it is false in the case of elements on line 4 and 7. The following rule describes this condition.

```
child_lines(Line1, Line2) :-
  intersection(Line1, Line2, _, _, P1),
  P2 := P1 + 1,
  not xml(Line1, _, _, P2),
  xml(Line2, _, _, P2),
  P3 := P2 + 1,
  not xml(Line2, _, _, P3).
```

XML documents often contain deeply nested elements. We call nested elements the descendants of a surrounding element. The nesting has to be verified. The following recursive predicate will do this activity.

```
descendant_lines(Line1, Line2) :-
  child_lines(Line1, Line2).

descendant_lines(Line1, Line3) :-
  child_lines(Line1, Line2),
  descendant(Line2, Line3).
```

To make our list of universal rules complete, we have to add a rule detecting which element is on the specified line. This rule must be used in case of elements without any attributes, e.g. the element <`bookshelf`> (see Fig. 1). The rule `element_line` looks for the specified number of such line element (in the set of predicates `xml`) which has the greatest value of the last argument, i.e. the level of nesting.

```
element_line(Line, Element) :-
  xml(Line, Element, _, P1),
  P2 := P1 + 1,
  not xml(Line, _, _, P2).
```

All listed and described logic rules are universal. If we use the proposed transformation on two different XML documents, the resulting logic programs will contain different set of facts, but the logic rules will be the same.

4. Queries

The logic program is complete if it contains a query we want to evaluate. Datalog has the basic form of a query:

```
?- predicate(list of arguments).
```

For the formulation of a query, it is usually necessary to define additional predicates in the form of one or more logic rules. This approach is used for all queries listed below.

The first query looks for the names of all authors participating in books published in 2003. The rule defining the predicate `authors_2003` and its corresponding query are as follows:

```
authors_2003(Author) :-
  attribute(Line1, 'bk:book', 'bk:year', '2003'),
  data(Line2, 'dc:creator', Author),
  descendant_lines(Line1, Line2).

?- authors(Author).
```

The rule looks for all lines (see variable Line1) where books issued in 2003 are recorded and all lines (see variable Line2) where all authors existing in the relevant facts of the predicate data are recorded. The predicate descendant_lines searches only pairs values of variables Line1 and Line2, which satisfy the condition that the element written on Line2 is a descendant of the element written on Line1. So the predicate searches only for the names of the authors of those books that were issued in 2003.

The last query is more complicated, working with several auxiliary rules. It searches for all co-authors of a given author (e.g. Abiteboul), including all co-authors of the searched co-authors, etc. This means it is looking for the transitive closure of a co-authorship relation (for the connected component of the co-authorship graph). The core of the recursive evaluation is given in a simplified form:

```
coauthors(New_coauthor) :-
  coauthors(Coauthor),
  search_book(Coauthor, Book),
  serarch_new_coautor(Book, New_coauthor).

?- coauthors(Coauthor).
```

The predicate coauthors will bind the variable Coauthor to the name of the previously found co-author. The predicate search_books finds out such books in which the Coauthor participated with other co-authors. The predicate search_new_coauthor evaluates the names of these co-authors (the value of the variable New_coauthor). This recursive evaluation stops when no other co-authors are found.

5. Conclusions

This paper shows one possible transformation of an arbitrary XML document into a logic program. The advantage of our transformation is the use of universal rules. These rules are the same in all generated logic programs. The programs differ only in facts.

The shortcoming of the proposed procedure is the assumption that the elements in XML documents will not have mixed content. This means that the element will contain either text or other nested elements. For example, the element `<review>A <em>fine</em> book.</review>` has a mixed content. It contains both text as well as a nested element `<em>`. Such information cannot be recorded into facts. If the proposed procedure transforms the RDF, RDFS or OWL document written in XML syntax, the elements with mixed content are not occurred.

In the future, we suppose to extend the set of universal rules and add rules which simplify the queries formulation.

Acknowledgements

This work was partly supported by Ministry of Education, Youth and Sports of the Czech Republic – the project 2C06009 *Knowledge base tools for natural language communication with semantic web*.

Notes and References

[1] W3 CONSORTIUM. *Extensible Markup Language (XML) 1.0 (Fifth Edition)*, November 2008, `http://www.w3.org/TR/xml/`.

[2] ZIMA, M. *Experimental Deductive Database System with Uncertainty [in Czech]*, Ph.D. Thesis, University of West Bohemia in Pilsen, 2002.

[3] ORACLE CORPORATION. *Oracle Database PL/SQL User's Guide and Reference 10g Release 2 (10.2)*, June 2005, Available at `http://www.oracle.com/`.

[4] CERI, S; GOTTLOB, G; TANCA, T. *What you always wanted to know about Datalog (and never dared to ask)*, IEEE Transactions on Knowledge and Data Engineering 1(1), March 1989, p. 146-66.

[5] STERLING, L; SAPIRO, E. *The Art of Prolog, Second Edition: Advanced Programming Techniques (Logic Programming)*, The MIT Press, March 1994.

[6] ALMENDROS-JIMENEZ, J.M. *An RDF Query Language based on Logic Programming*, Electronic Notes in Theoretical Science, 200, 2008, p. 67-85.

[7] W3 CONSORTIUM. *XML Path Language (XPath) 2.0*, January 2007, `http://www.w3.org/TR/xpath20/`.

[8] W3 CONSORTIUM. *RDF/XML Syntax Specification (Revisited)*, February 2004, `http://www.w3.org/TR/rdf-syntax-grammar/`.

Geo information extraction and processing from travel narratives

Rocío Abascal-Mena; Erick López-Ornelas

Universidad Autónoma Metropolitana – Cuajimalpa
Departamento de Tecnologías de la Información
Avenida Constituyentes 1054, 4° piso, Col. Lomas Altas,
Delegación Miguel Hidalgo, C.P. 11950, México, D.F., México
{mabascal, elopez@correo.cua.uam.mx}

Abstract

Travel narratives published in electronic formats can be very important especially to the tourism community because of the great amount of knowledge that can be extracted. However, the low exploitation of these documents opens a new area of opportunity to the computing community. In this way, this article explores new ways to visualize travel narratives in a map in order to take advantage of experiences of individuals to recommend and describe travel places. Our approach is based on the use of a Geoparsing Web Service to extract geographic coordinates from travel narratives. Once geographic coordinates are extracted by using eXtensible Markup Language (XML) we draw the geo-positions and link the documents into a map image in order to visualize textual information.

Keywords: electronic publishing; knowledge representation; mapping services; Geoparsing Web Service; social network.

1. Introduction

The growing predominance of adding information in social web sites presents new challenges to the computing community in terms of thinking of new ways to extract, analyze and process the information contained. Nowadays, there are interesting works around the capacity to extract and map geotagged photos [1], coming from social web sites like Flickr, but there is a lack in the extraction of geographic concepts coming from unstructured text documents [2].

The text documents that we will use in this paper are the travel narratives. These documents represent the observations and experiences of individuals who visited foreign countries or places and constitute a special category of primary source for people who want to visit these places.

These narratives provide information about a foreign society or culture that the traveller guides do not provide. This way, we have realized that information coming from travel narratives as opposed to travel guides, allow, from a personal point of view, the exchange of experiences, and recommendations by way of letters, diaries or chronicles. These narratives can be used to know, for example, what museums have to be visited in a determined place and what are the most interesting things to do inside (from the point of view of the writer). So, in our work we decided to investigate how to link, automatically, travel narratives containing personal descriptions of visited places on a map and to explore the different web mapping services existing in this specific application.

Web mapping services applications like Google Maps, Google Earth, NASA World Wind or Flickr Map are a new way to organize the world's information geographically. Mapping services have been used in many areas including weather forecast, tourism and asset management. They provide geospatial visualization of information so the users can analyze, plan and take decisions based on geographic location. They help users understand the relationship between data and geographic location. All mapping applications provide an intuitive mapping interface with detailed street and aerial imagery data embedded. In addition, map controls can be embedded in the product to give users full control over map navigation. The primary goal behind its rapid acceptance as an Internet mapping viewer is the ability to customize the map to fit application-specific needs.

The article is structured as follows. Some background of Geoparsing Web Services (GWS) is provided in Section 2. Section 3 describes the methodology followed to identify visual elements in travel stories. Experimental results are described in Section 4 while some conclusions and further work are shown in Section 5.

2. Geoparsing Web Services background

Traditional Information Extraction (IE) has involved manual processing in the form of rules or tagging training examples where the user is required to specify the potential relationships of interest [3].

Geoparsing is the process of recognizing geographic context [4]. The first step involves extracting geographic entities from texts and distinguishing them from other entities such as names of people or organizations, and events. In natural language processing this is referred to as Named Entity Recognition (NER) and is central to other text processing applications such as information extraction (IE) and information retrieval (IR).

Geoparsing is most frequently used to automatically analyze collections of text documents. There are a number of commercial products with a geoparsing capability. Companies like MetaCarta[1] extract information about place and time, while others like Digital Reasoning[2] (GeoLocator), Lockheed Martin (AeroText)[3], and SRA (NetOwl)[4] extract places along with other entities, such as persons, organizations, time, money, etc. To process the large volumes of data, these systems rely on automated techniques optimized for speed.

These geoparsing systems are not perfect. Identifying and disambiguating place names in text are difficult and vulnerable to the vagaries of language. Just identifying which words are associated with place names can be a challenge. The geoparsing software must not only understand the words, but whether the words that form a name actually refer to a place. The software must understand that “Paris” in “Paris, France” refers to an urban area; in “Paris Creek” refers to a stream; in “Paris Hilton” refers to a person or to a hotel; and in “Paris Match” refers to a magazine name.

Once a place name has been identified, disambiguation remains a challenge. For example, there are over 2,100 names in the National Geospatial-Intelligence Agency which exactly match San Antonio. Sometimes, without being the author of a document, it is simply not possible to identify, with any confidence, the place to which a name refers.

Given these difficulties, it is understandable that automated geoparsing software will miss some place names, identify non-place text as place names, and sometimes identify the location of place incorrectly if multiple choices are possible.

Rather than focus on analyzing collections of documents, some other approaches focuses on the individual document, allowing authors to efficiently ensure that the place names are identified correctly and are discoverable by other users. Just as map documents go through a review and

[1] http://www.metacarta.com/

[2] http://www.digitalreasoning.com/

[3] http://www.lockheedmartin.com/products/AeroText/index.html

[4] http://www.sra.com/netowl/

validation process, this approach allows authors to confirm that the places in their documents are correctly identified and located at the time of writing. One example is GeoDoc[5] where the user has to identify and tag the place names manually, the application starts by automatically extracting place names and highlighting them on the display.

Current work on query processing for retrieving geographic information on the Web has also been done by Chen et al [5]. Their approach requires a combination of text and spatial data techniques for usage in geographic web search engines. A query to such an engine consists of keywords and the geographic area the user is interested in (i.e., query footprint).

2.1 The Yahoo! Placemaker Web Service

Yahoo! Placemaker[6] is a geotagging web service that provides third-party developers the means to enrich their applications or Web sites with geographic information. The service is able to identify, disambiguate, and extract place names from unstructured and semi-structured documents. It is also capable of using the place references in a document, together with a pre-determined set of rules, to discover the geographic scope that best encompasses its contents. Thus, given a textual document, Yahoo! Placemaker returns unique *Where-on-Earth Identifiers* (WOEIDs) for each of the named places and scopes. Through these identifiers, one can use the Yahoo! GeoPlanet[7] Web service to access hierarchical information (i.e., containing regions) or spatial information (i.e. centroids and bounding boxes).

There are two flavours of document scopes in Placemaker, namely the geographic scope and the administrative scope. The geographic scope is the place that best describes the document. The administrative scope is also the place that best describes the document, but is of an administrative type (i.e., Continent, Country, State, County, Local Administrative Area, Town, or Suburb). Since the reference document collection that we used for our experiments only contains documents assigned to administrative regions, we limited our cross-method comparison to using Placemaker's administrative scopes.

Placemaker is a commercial product and not many details are available regarding its functioning. However, some information about the service is available in the Web site, together with its documentation. For instance, the

[5] http://geodoc.stottlerhenke.com/geodoc/

[6] http://developer.yahoo.com/geo/placemaker/

[7] http://developer.yahoo.com/geo/geoplanet/

Web site claims that when the service encounters a structured address, it will not perform street level geocoding but will instead disambiguate the reference to the smallest bounding named place known, frequently a postal code or neighbourhood. The Web site also claims that besides place names, the service also understands geography-rich tags, such as the W3C Basic Geo Vocabulary and HTML micro-formats. However, no details about the rules that are used in the scope assignment process are given in the documentation for the service.

The Placemaker Web service accepts plain text as input, returning an eXtensible Markup Language (XML) document with the results. The service has an input parameter that allows users to provide the title of the document separately from the rest of the textual contents, weighting the title text as more representative. In our experiments we used the Web service as a black-box to assign scopes to the Web documents, using the option that weights the title text as more important than the rest.

3. Methodology

Our approach is based on the use of a Geoparsing Web Service (GWS) which enriches content with geographic metadata by extracting places from unstructured texts, the travel narratives. Geoparsing offers the ability to turn text documents into geospatial databases. This process is done in two steps: 1) entity extraction and 2) disambiguation, which is also known as grounding or geotagging. Geospatial entity extraction uses natural language processing to identify place names in text, while disambiguation associates a name with its correct location.

In order to access the GWS we have used the Yahoo! Placemaker, which is a GWS that provides third-party developers the means to geo-enrich content at scale. The service identifies, disambiguates, and extracts places from unstructured and structured textual content: web pages, RSS (and Atom) feeds, news articles, blog posts, etc. It is an open API that assists developers in creating local and location-aware applications and datasets. Placemaker is a geo-enrichment service that assists developers in determining the whereness of unstructured and atomic content, making the Internet more location-aware.

To access the GWS we have used the Yahoo Query Language (YQL) which is a query influenced by the Structured Query Language (SQL) but diverges from it as it provides specialized methods to query, filter, and join data across web services. The process is shown in figure 1.

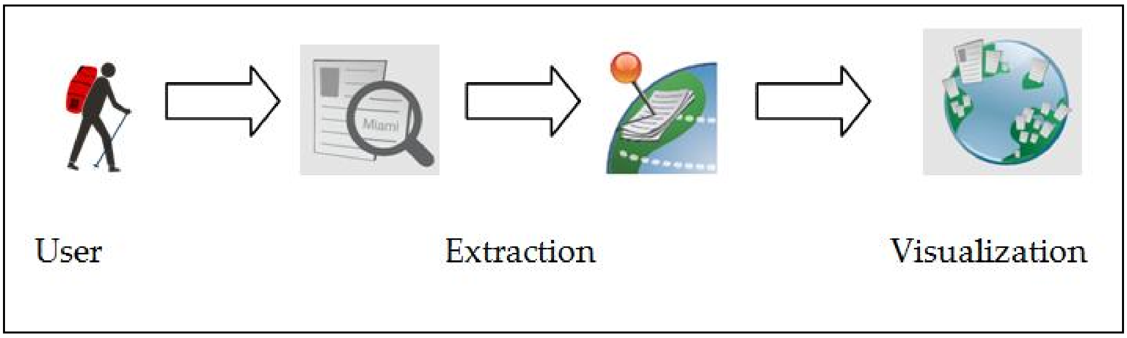


Figure 1. Geoparsing and visualization process.

4. Experimental results

On the web we find many websites containing travel stories. But, to present our experimental results we have chosen, in an aleatory way, only one of these sites. We selected "The Adventure Prone Site" (http://www.adventureprone.com/). From this site, we have extracted all the URLs containing travel narratives by using the web spider called Robot V1 (http://www.semantic-knowledge.com/). This spider is designed to collect websites and extract texts from Internet, following the links from a starting Web page to other pages, until the process is finished. Thus, we have obtained pages coming from this site. To analyze and extract the geographic aspects of these pages we have implemented a system able to communicate to the GWS by using YQL. For each URL extracted we have obtained the geographic coordinates. But, in order to present our results in ELPUB conference, we decided to choose only one travel story: "A Tale of Ten Cities" (http://www.adventureprone.com/travel/stories/cities.html). Analyzing this page, we have obtained 48 places with the next geographic elements (name, latitude, longitude):

1. Budapest, Budapest, HU (47.5062, 19.0648)
2. San Francisco, CA, US (37.7792, -122.42)
3. France (46.7107, 1.71819)
4. Silicon Valley, CA, US (37.3953, -122.053)
5. Italian Town, AL, US (33.1183, -87.1)
6. Danube, HU (46.3298, 18.9073)
7. Poland (51.9189, 19.1343)
8. Naples, Campania, IT (40.8399, 14.2519)
9. Gary, Midi-Pyrénées, FR (43.6959, 1.91942)
10. Greece (39.0724, 21.8456)

11. Prague, Hlavni mesto Praha, CZ (50.0791, 14.4332)
12. Armenia (40.0662, 45.0399)
13. Acropolis, Athens, Attiki, GR (37.9714, 23.7238)
14. England, GB (52.8836, -1.97406)
15. Mont Blanc, Bossons, Rhône-Alpes, FR (45.8359, 6.86211)
16. Berlin, Bundesland Berlin, DE (52.5161, 13.377)
17. Pantheon, Rome, Lazio, IT (41.8987, 12.4769)
18. Colosseum, Rome, Lazio, IT (41.8902, 12.4929)
19. Montpellier, Languedoc-Roussillon, FR (43.6109, 3.87609)
20. Rome, Lazio, IT (41.9031, 12.4958)
21. Italy (42.5038, 12.5735)
22. Spain (39.895, -2.98868)
23. Venice, Veneto, IT (45.4383, 12.3185)
24. Bolivia (-16.2883, -63.5494)
25. Manarola, Liguria, IT (44.1075, 9.73006)
26. Chamonix-Mont-Blanc, Rhône-Alpes, FR (45.9249, 6.87193)
27. Florence, Tuscany, IT (43.7824, 11.255)
28. Pisa, Toscana, IT (43.71, 10.3995)
29. London, England, GB (51.5063, -0.12714)
30. Vesuvius, Torre del Greco, Campania, IT (40.8, 14.4)
31. Newquay, England, GB (50.4158, -5.07558)
32. Monastiraki, Athens, Attiki, GR (37.9782, 23.7268)
33. St Peter's Basilica, Vatican City, VA (41.9022, 12.4547)
34. Vatican Museums, Vatican City, VA (41.9069, 12.454)
35. St Peter's Square, Vatican City, VA (41.9023, 12.4576)
36. Athens, Attiki, GR (37.9762, 23.7364)
37. United Kingdom (54.3141, -2.23001)
38. St. Columb Major, England, GB (50.4325, -4.93688)
39. Vatican City (41.9038, 12.4525)
40. Morocco (31.8154, -7.067)
41. Buda, Budapest, Budapest, HU (47.5131, 19.0241)
42. Africa (2.07079, 15.8005)
43. Trevi Fountain, Rome, Lazio, IT (41.9009, 12.4833)
44. Cairo, Al Qahirah, EG (30.0837, 31.2554)
45. Olympia, Olimbia, Dytiki Ellada, GR (37.6405, 21.6281)
46. Europe (52.9762, 7.85784)
47. London Stansted International Airport, Takeley, England, GB 51.8894, 0.26256)
48. Delphi, Dhelfoi, Sterea Ellada, GR (38.472, 22.4746)

In the previous results presented we find places like “San Francisco, CA, US” which is compared in the story even if it is not part of the trip: *“We would go out to dinner every night, and seeing as things were so "cheap" compared to San Francisco, spend US$50/head on a fantastic meal with great wines which would have cost twice as much at home.”* This is an example of further work that must be done in order to extract only the pertinent places according to the context.

With the geographic coordinates obtained we display each position on an Equirectangular projection of the Earth. For example, for the 48 coordinates extracted we have obtained the projection show in Figure 2.

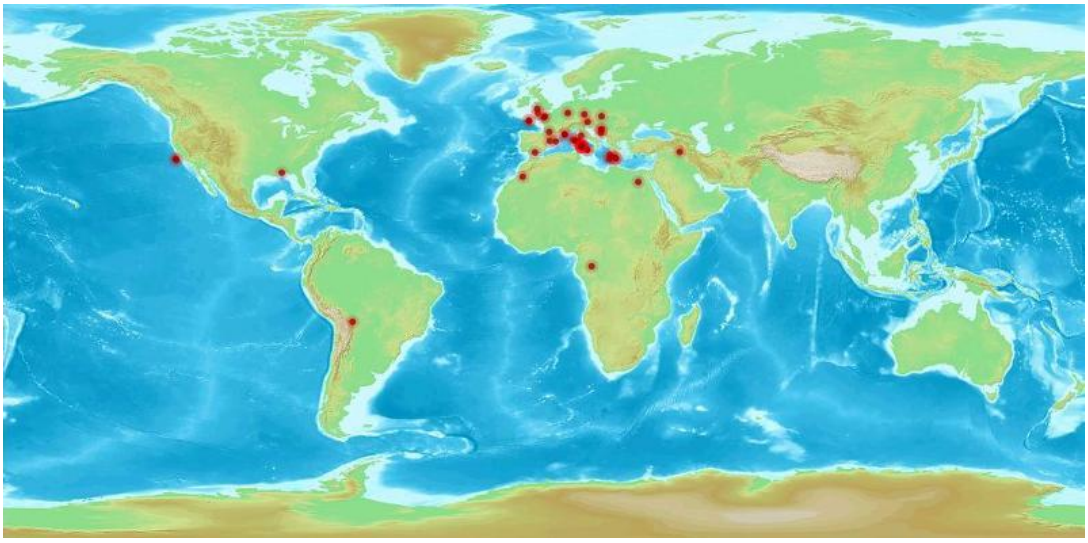


Figure 2: Representation of 48 coordinates extracted from a previous travel narrative analyzed

To present in a clear way the results obtained, we have selected only the coordinates belonging to Europe to show them in an Equirectangular Europe projection. In this way, Figure 3 presents the places visited in Europe during the tour written in the analyzed travel story.

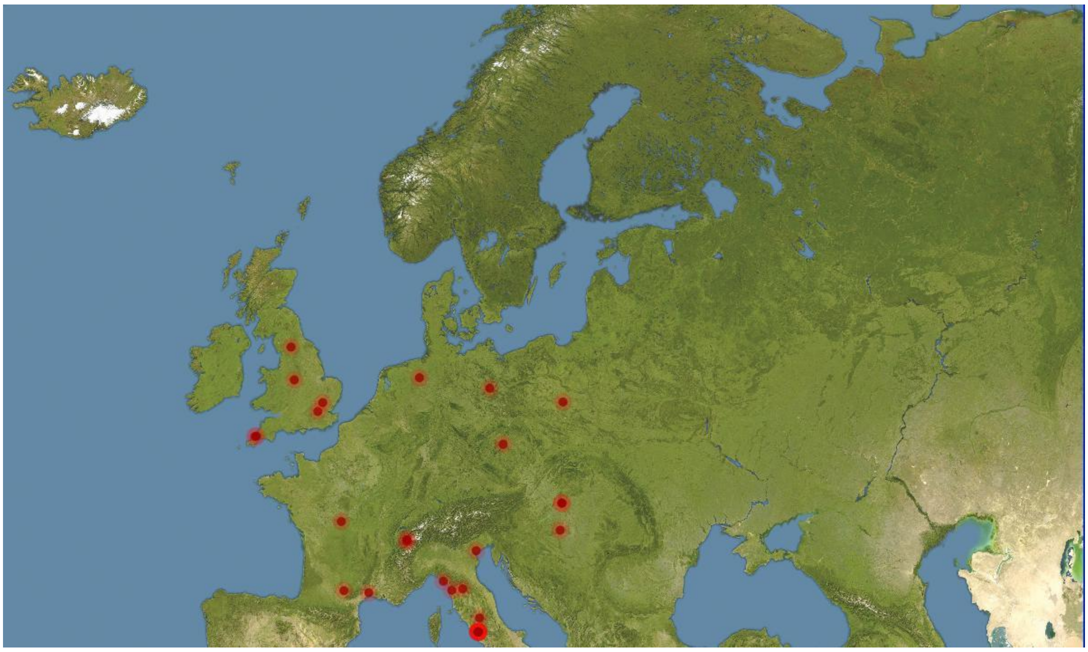


Figure 3: Representation of extracted cities corresponding to Europe

Each of the places linked to the map contains its respective paragraph where the place is mentioned.

5. Conclusions and further work

This article explores the possibilities given by a Geoparsing Web Service in order to extract and contextualize unstructured text documents. Our work presents a first exploration of travel narratives in order to map them into an Earth projection and visualize personal descriptions of the marked places. Further work will be in the next axes:

(1) Identify other non-structured documents to verify and check the Geoparsing extraction like news, historical documents or collections of documents like Wikipedia [6].

(2) Compare different Geoparsing methods (Metacarta, NetOWL, GeoLocator) in order to identify the best extraction process with the different applications.

(3) Having the extracted spatial position of the travel narrative we will have to find new ways of extracting some spatial knowledge like GeoProfiling and find new geo-visualization tools.

(4) The contextualization and disambiguation of the places named in each travel narrative sometimes is not very clear, so a work to contextualize and extract pertinent information of the original document with the use of ontologies is required [7]. Also, the use of

information extracted from a large encyclopedic collection and the Web could be a solution to explore [8].

(5) Compare concepts, coming from different documents, such as the hierarchical nature of geographic space and the topological relationships between geographic objects in order to represent relationships between different documents [9].

(6) Geo-temporal criteria are important for filtering, grouping and prioritizing information resources [10]. In this way, the capacity to automatically link travel stories on a time scale like the proposal done in articles of the Wikipedia shown by Bhole et al. [11] is an approach to be explored.

(7) Classify documents according to their implicit location relevance [12].

Results presented here show that it is possible to automatically extract places from unstructured texts in order to visualize them and provide other kind of services to the users. In this way, we are working in order to provide to the users with the capability to manipulate textual travel narratives by clicking places of interest or even having the capacity to visualize visited places according to time.

References

[1] CRANDALL, D; et al. Mapping the World's Photos. *WWW 2009 Madrid. Track: Social Networks and Web 2.0.* 2009. Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.1382&rep=rep1&type=pdf (April 2010).

[2] ABASCAL-MENA, R; LÓPEZ-ORNELAS, E. Structured Data Analysis of Travel Narratives by Using a Natural Language Processing Tool. *IADIS International Conference WWW/Internet 2009,* Rome, Italy. pp. 242-246. 2009.

[3] BANKO, M; CAFARELLA, M.J; SODERLAND, S; BROADHEAD, M; ETZIONI, O. Open Information Extraction from the Web. *Proceedings Twentieth International Joint Conference on Artificial Intelligence 2007*, pp. 2670–2676. 2007.

[4] LARSON, R.R. Geographic Information Retrieval and Spatial Browsing. *Smith, L., Gluck, M. (eds.) University of Illinois GIS and Libraries: Patrons, Maps and Spatial Information*, pp. 81–124, 1996.

[5] CHEN, Y; SUEL, T; MARKOWETZ, A. Efficient Query Processing in Geographic Web Search Engines. *Proceedings SIGMOD 2006*, pp. 277–

288. 2006. Available at http://portal.acm.org/citation.cfm?id=1142505 (April 2010).

[6] WITMER, J; KALITA, J. Extracting Geospatial Entities from Wikipedia. *IEEE International Conference on Semantic Computing.* 2009. Available at http://www.computer.org/portal/web/csdl/doi/10.1109/ICSC.2009.62 (April 2010).

[7] ZUBIZARRETA, A; FUENTE, P; CANTERA, J. M; ARIAS, M; CABRERO, J; GARCÍA, G; LLAMAS, C; VEGAS, J. Extracting Geographic Context from the Web: GeoReferencing in MyMoSe. *Proceedings of the 31th European Conference on IR Research on Advances in information Retrieval.* 2009. Available at http://www.springerlink.com/content/u364332137807568/ (April 2010).

[8] CUCERZAN, S. Large-scale named entity disambiguation based on Wikipedia data. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pp. 708-716. Prague, June 2007. Available at http://acl.ldc.upenn.edu/D/D07/D07-1074.pdf (April 2010).

[9] BRISABOA, N. R; LUACES, M. R; PLACES, Á. S; SECO, D. Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index. *Geoinformatica* 14, 3 (Jul. 2010), pp. 307-331. 2010. Available at http://www.springerlink.com/content/2553w445515231q1/ (April 2010).

[10] MARTINS, B; MANGUINHAS, H; BORBINHA, J. Extracting and Exploring the Geo-Temporal Semantics of Textual Resources. *IEEE ICSC*, pp. 1–9. 2008. Available at http://www.computer.org/portal/web/csdl/doi/10.1109/ICSC.2008.86 (April 2010).

[11] BHOLE, A; FORTUNA, B; GROBELNIK, M; MLADENIC, D. Extracting Named Entities and Relating Them over Time Based on Wikipedia. *Informática*, 31(2007) pp. 463-468. 2007. Available at http://www.informatica.si/PDF/31-4/12_Bhole-Extracting%20Names.pdf (April 2010).

[12] ANASTACIO, I; MARTINS, B; CALADO, P. Classifying Documents According to Location Relevance. *Proceedings of the 14th Portuguese Conference on Artificial intelligence: Progress in Artificial intelligence*. Lecture Notes in Computer Science. pp. 598-609. 2009. Available at http://www.springerlink.com/content/d551604105m05g07/ (April 2010).

Semantic enrichment for 3D documents: Techniques and open problems

Torsten Ullrich[1]; Volker Settgast[1]; René Berndt[2]

1 Fraunhofer Austria Research GmbH, Visual Computing
Inffeldgasse 16c, A-8010 Graz, Austria,
{torsten.ullrich, volker.settgast}@fraunhofer.at
2 Institut f. ComputerGraphik und WissensVisualisierung,
Technische Universität Graz, Inffeldgasse 16, A-8010 Graz, Austria,
r.berndt@cgv.tugraz.at

Abstract

With increasing knowledge the process of knowledge management and engineering becomes more and more important. Enriching documents by using markup techniques and by supporting semantic annotations is a major technique for knowledge management. This invaluable information is of extreme importance in the context of civil engineering, product life cycle management, virtual archival storage, and preservation. In these fields of applications annotation techniques for 3D documents are a vital part. They provide semantic information that makes up the basis for digital library services: retrieval, indexing, archival, and searching. Furthermore, metadata are of significant importance as they set the stage for data re-use and they provide documentation of data sources and quality, which is vital for every engineering department. Using metadata helps the user to understand data. Additional information allows focusing on key elements of data that help to determine the data's fitness for a particular use and may provide consistency in terminology. In this paper we give an overview on state-of-the-art annotation techniques focussed on 3D data.

Keywords: annotation techniques; semantic enrichment; geometry processing

1. Introduction

In 1998 William J. Clinton announced at the 150th Anniversary of the American Association for the Advancement of Science that “the store of human knowledge doubles every five years”. With increasing knowledge the process of knowledge management and engineering becomes more and more important. Enriching documents by using markup techniques and by supporting semantic annotations is a major technique for knowledge management. It allows an expert to establish an interrelationship between a document, its content and its context.

Annotations made by groups or individuals in the context of teamwork or individual work allow to capture contextual information, which can improve and support cooperative knowledge management policies; i.e. annotations can be considered under the perspective of documentation. In fact, tracking the changes and focal points of annotations implies tracing the underlying reasoning process.

This invaluable information is of extreme importance in the context of civil engineering, product life cycle management, virtual archival storage, and preservation. In these fields of applications annotation techniques for 3D documents are a vital part. In this paper we give an overview on state-of-the-art annotation techniques focussed on 3D data.

2. Terms and Definitions

Different documentation standards and annotation processes are used in various fields of applications. Unfortunately, each branch of science has slightly different definitions of bibliographical terms. To clarify these terms and to avoid misunderstandings and misconceptions we present the relevant definitions of terms used in this article.

A document is any object, “preserved or recorded, intended to represent, to reconstruct, or to demonstrate a physical or conceptual phenomenon”. This definition has first been verbalized by Suzanne Briet in her manifesto on the nature of Documentation: Qu’est-ce que la documentation? [1]. In Michael K. Buckland’s article “What is a document?” various document definitions are given and compared to each other [2]. As we will concentrate on 3D data sets, the “physical or conceptual phenomenon” will always be a three-dimensional phenomenon.

A distinct, separate subpart of a document is called entity. Other authors refer to a subpart as segment. Metadata about documents or parts of documents are defined as “structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities.” The American Library Association formalized this definition in its Task Force on Metadata Summary Report [3]. According to this definition metadata is always structured. Unstructured, encoded data, such as comments and free texts, are hereinafter called annotations. As metadata is always structured, it can be specified in a formal, explicit way: an ontology is a “formal, explicit specification of a shared conceptualisation”. It provides a shared vocabulary, which can be used to model a domain; i.e. the type of objects and/or concepts that exist, and their properties and relations. Tom Gruber established this definition in his article “A translation approach to portable ontology specifications” [4]. The connections between a document and its metadata or annotations are called markup instructions. They provide local or global reference points in a document.

In the context of computer-aided design, reconstruction, and archival storage, a document is very often the result of a process chain. Data describing a single processing step or a document’s process chain is termed paradata.

3. Classification

Metadata and annotations – semantic information in general – can be classified in several ways. Depending on the field of application, they can be classified according to the following criteria.

3.1 Document data type

Semantic information enriches a document. As documents can be grouped according to their type, these categories can be transferred to metadata and annotations as well. We concentrate on 3D data in this article, which can be subdivided further into different kinds of 3D representations.

Boundary Representation. Boundary and surface representations are the most common kind of data representations in computer graphics. These representations comprehend lists of triangles, polygonal meshes, spline surfaces, and subdivision surfaces, etc. to name a few. Several annotation systems allow users to leave text messages on the surface of a 3D model [5] or to draw annotations on the surface and in free space of a virtual scene [6]. The main field of applications is architectural model annotation. Adobe embeds 3D models into PDF documents [7] and combines this technique with its annotation system.

Point Clouds. In a digital documentation process, the input data set is very often a simple point cloud – measured points in space without any additional structure – recorded by e.g. a laser scanner. Although many tools exist (for example: geomagic[1], Leica Cyclone[2]) to annotate and markup point clouds, the automatic documentation of the recording process has many gaps.

Volume Data. Computer tomography and similar acquisition techniques generate volumes of measured points consisting of many layers of high-resolution 2D images. This is the predominant acquisition technique in biomedical sciences, in which documentation, markup and annotation has always been put into practice. Consequently, many established annotation and markup systems exist; for example the volume data annotation tool called VANO by Peng et al. [8] and the Annot3D Project by Balling et al. [9].

Miscellaneous. Besides these “main data types” numerous data representations specialized in its field of application are available. As annotation is a key activity of data analysis, many visualization systems offer annotation capabilities [10], [11].

3.2 Scale of Semantic Information

Semantic information can be added for the entire data set or only for a fragment of the object. For some metadata like “author” it can be sufficient to mark the entire document. But 3D data creation is often a collaborative task with many people working on one complex object. For comments and detailed descriptions a specific place within the 3D data set is needed. Also to communicate suggestions for improvement during an evaluation process it is necessary to make comments on some parts of an object but independent from the entities of the object. This can be done by defining an anchor, like a point, a surface or a volume in addition to the actual data set. It is essential that an anchor can be defined independently from the object. In this way it is also possible to annotate something which is missing in a specific place. Often also the viewer’s parameters are stored together with the added information to make it easier to read.

[1] http://www.geomagic.com

[2] http://leica.loyola.com

3.3 Type of Semantic Information

The “Metadata Encoding & Transmission Standard”[3] defines the following types of metadata and annotation:

Descriptive Information. Descriptive information describe the content of an object and comprehend amongst others the Dublin Core metadata set [12].

Administrative Metadata. Administrative metadata provide information regarding how a document was created and stored, as well as intellectual property rights, information regarding the provenance of a document, transformation/conversion information, etc.

Structural Metadata. A structural map and structural links outline a hierarchical structure for a digital library object and embed a document into a context.

3.4 Type of creation

The creation of semantic enrichment of 3D documents fall basically in two categories: manual or automatic. Most of the metadata (especially administrative and descriptive metadata) can be generated automatically, but depending on the domain certain fields need support from an expert. Especially categorizing 3D models can be a difficult task for automatic indexing, e.g. classify buildings according to architectural theory or genres (e.g. the Getty Art & Architecture Thesaurus (AAT)).

Annotations, in terms of free text like comments and remarks, are usually entered manually (the translation of a comment using an automatic translation service would be an example for automatic created annotations). While the manual processing of structured metadata is done by experts, comments or remarks can also come from non-expert users. This method (social tagging) has become very popular within Web 2.0.

3.5 Data organization

The data organization is an important aspect thinking of the sustainability of the annotation. There are two basic concepts how programs can store annotations.

The truth is in the file. The metadata and annotations are stored within the original documents. EXIF or XMP are good examples for that strategy. The main drawback is that the file format must support the possibility to add such arbitrary data. While modern 3D formats like Collada offer this functionality (e.g. the extra tag), others do not. For these a sidecar file can be an appropriate location for storing the annotation. Putting the original document and the sidecar file(s) in a container (portmanteau) is a very popular technique. In

[3] http://www.loc.gov/standards/mets/

most cases the container is a simple zip archive, but with a different and unique extension (examples are the Open Document Format (ODF) or Microsoft Office Open XML).

The truth is in the database. In this concept the metadata and annotations are stored within a database system. This guarantees that the user will access up-to-date data, but he needs to retrieve the associated annotations separately. In addition the application must be able to access the database.

3.6 Information comprehensiveness

Semantic enrichment can be further classified by the comprehensiveness of the information. The amount of comprehensiveness can vary from low to high in any gradation. An example for a low comprehensiveness would be the Dublin Core metadata set [13]. It allows 15 properties to be added as semantic information. In contrast, for example the CIDOC Conceptual Reference Model (CRM) is a scheme with a high comprehensiveness. It is a framework for the definition of relationship networks of semantic information in the context of cultural heritage [14]. CIDOC CRM offers a formal ontology with 90 object classes and 148 properties (in version 5.0.1) to describe all possible kinds of relations between objects.

4. Standards and File Formats

An important aspect of semantic enrichment is to agree on standards. For semantic information it is more a question of organizing documents in a standardized way. In the area of 3D data representations an important aspect is the file format and its ability to support semantic enrichment. Many concepts for encoding semantic information can be applied to 3D data but only a few 3D data formats support semantic markup.

4.1 Semantic Information

The definition and storage of semantic information develops mainly around textual documents. Established standard schemes like Dublin Core and CIDOC CRM (see Section 3) can also be used for 3D content. The Moving Picture Experts Group (MPEG) defined a standard for the annotation of media resources. MPEG-7 is a scheme to describe metadata for audio and visual contents. 3D models however are not part of the scheme. In [15] Bilsco et al. propose an extension to the standard for 3D data.

4.2 Three-Dimensional Content

For 3D data there is not one single standard format like for example JPEG can be seen as standard for digital photos. Many different formats with a large variation of shape descriptions and features are in practical use. For semantic information on 3D data it is even harder to identify one or a few standard formats because most of the commonly used 3D formats do rarely support semantic enrichment.

On the one hand it is possible to extend existing 3D file formats to support metadata and annotations. Especially XML based formats are well suited to be extended while still being readable by existing applications. Some extensions have been proposed for Extensible 3D (X3D) and for Collada. On the other hand there are document formats which have been extended to support 3D content, like PDF 3D.

Collada. The XML-based Collada format was initiated by Sony Entertainment mainly to establish a standard way of data exchange between different creation tools for 3D content. It is hosted by the Khronos group as an open standard. The Collada description allows storing metadata like title, author, revision etc. not only on a global scale but also for parts of the scene like nodes and geometry. Custom extensions to the format are possible as part of the XML scheme. The Collada format can be found in Google Warehouse and the Google Earth application. Metadata like the location of the 3D data on the virtual earth however is stored in separate files.

PDF 3D. Since version 1.6 the Adobe PDF format supports the inclusion of 3D content. For presentations the complete set of data can be exported to PDF 3D. 3D data has to be converted to the Universal 3D (U3D) format or to the Product Representation Compact (PRC) format to be used inside a PDF file. PDF 3D allows to store annotations (logically) separated from the 3D data of the annotated object. Typically the exported 3D data is modified to get a smaller file size. But only if the model is stored without lossy compression it is possible to extract the original data and use the PDF 3D as an exchange format. The PRC format supports product manufacturing information and allows to use the geometry data to be used as input for computer aided manufacturing. U3D is mainly used as a visualization and publication format. Also the PDF format allows adding additional annotations to the model and even annotating the annotations. 3D PDF has the potential to become a standard for 3D models with annotations. The viewer application is widely spread and PDF documents are the quasi standard for textual documents.

The export of 3D content containing metadata or annotations to another file format often leads to information loss. To minimize data loss some semantics, for example labels and measurements, can be integrated into the

3D data as geometry. But afterwards depending on the format it may be hardly possible to distinguish between the metadata and the geometry.

Some software companies offer solutions for their own product lines. For examples Dassault Systems introduced an XML-based format called 3D XML. It is supported by all of their applications as an exchange format. There are many of those solutions but none of them is a standard.

5. Examples

To illustrate semantic information processing, we present results from three on-going research projects: CityFIT, PROBADO and 3D-COFORM[4].

Currently the state-of-the-art for automatically generated city models are basically just extruded ground polygons with roofs. Still missing are detailed 3D models of facades. The goal of the CityFIT project is to reconstruct these facades automatically using the example of Graz, Austria. The main idea of CityFIT [16] is to turn the general implicit architectural knowledge about facades into explicit knowledge, based on both architectural theory and empirical evidence. To achieve this, it combines inductive reasoning with statistical inference. In this example semantic knowledge is encoded in a facade library in form of algorithms. Each execution of such an algorithm generates a (part of) 3D model. During the reconstruction process, the textured point cloud is matched by algorithm instances of the facade library and its instantiation parameters are added as metadata to the point cloud. Therefore, this reconstruction process identifies architectural patterns and enriches the data set semantically.

This metadata is essential in digital libraries. Hence, considering content-based retrieval tasks, multimedia documents are not analyzed and indexed sufficiently. To facilitate content-based retrieval and browsing, it is necessary to introduce recent techniques for multimedia document processing into the workflow of nowadays digital libraries. The PROBADO-framework will integrate different types of content-repositories – each one specialized for a specific multimedia domain – into one seamless system, and will add features available in text-based digital libraries (such as automatic annotation, full-text retrieval, or recommender services) to non-textual documents [17].

[4] http://www.3d-coform.eu/

The third example is a project in the context of cultural heritage (CH). The context of cultural heritage distinguishes itself by model complexity (“masterpiece of human creative genius”[5]), model size (archaeological excavation on the scale of kilometers with a richness of detail on the scale of millimeters), and imperfection (natural wear and tear effects). In this context, the interplay of content and metadata as well as paradata is extremely complex and difficult to model. The aim of the 3D-COFORM project is to establish 3D documentation as an affordable, practical and effective mechanism for long term documentation of tangible cultural heritage. In order to make this happen the consortium is highly conscious that both the state of the art in 3D digitization and the practical aspects of deployment in the sector must be addressed.

6. Recommendations and Open Problems

In this section we discuss some good practices (and also problems) for handling semantic information. Special attention is drawn to aspects which influence the sustainability of the semantic information (e.g. integrity of information, long-term preservation).

Techniques as presented in [18] offer a great flexibility in terms of annotating 3D formats which do not offer a built-in mechanism. One problem of this approach is that modifying the model afterwards can break the integrity of the semantic information. Any 3D authoring operation might change the geometry in a way that the referenced part of the model either no longer exists or has changed its meaning. A possible solution to detect such conflicts is to add a checksum to the annotations. But without deep knowledge of the semantic data structure it is impossible to preserve the semantic information through a processing pipeline. While for special domains a number of complex and expensive product lifecycle management (PLM) solutions exists, other domains, e.g. cultural heritage still lack such a common infrastructure.

For the issue of long-term preservation the use of open standards is a very essential point. This addresses both the 3D file format and the format for the semantic information (if not already included in the 3D format). Especially the next version of Adobe’s PDF standard for long-term archiving (PDF/A-2) will play a major role. PDF/A-2 as PDF/A-1[19] will use the ”truth is in the file”

[5] http://whc.unesco.org/en/criteria

strategy. This guaranties that all information will be available even in an offline scenario. One drawback is that PDF 3D relies on either U3D or PRC. This requires in most cases a conversion step, which inevitably leads to loss of information. Because of the wide distribution of the Adobe Reader, more and more domains start to use PDF 3D as a format for their visualizations. Some examples are described in [20].

Which approach fits best depends on the project constraints. In many cases the 3D file format cannot easily be replaced by another; even worse, most tools still build on their own proprietary formats.

References

[1] S. Briet, Qu'est-ce que la documentation? EDIT -´editions documentaires industrielles et techniques, 1951.

[2] M.K. Buckland, "What is a "document"?," Journal of American Society of Information Science, vol. 48, no. 9, pp. 804–809, 1997.

[3] ALCTS Committee on Cataloging: Description and Access. "Task Force on Metadata -Summary Report," American Library Association, vol. 4, pp. 1–16, 1999.

[4] T.R. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition -Special issue: Current issues in knowledge modeling, vol. 5, no. 2, pp. 199–220, 1993.

[5] T. Jung, M.D. Gross, and E. Y.-L. Do, "Annotating and Sketching on 3D Web models," Proceedings of the International Conference on Intelligent User Interfaces, vol. 7, pp. 95–102, 2002.

[6] K. Osman, F. Malric, and S. Shirmohammadi, "A 3D Annotation Interface Using DIVINE Visual Display," Proceeding of the IEEE International Workshop on Haptic Audio Visual Environments and their Applications, vol. 5, pp. 5–9, 2006.

[7] I. Adobe Systems, "3D Annotations Tutorial," Adobe Solutions Network, vol. 7, pp. 1–16, 2005.

[8] H. Peng, F. Long, and E. W. Myers, "VANO: a volume-object image annotation system," Bioinformatics, vol. 25, no. 5, pp. 695–697, 2009.

[9] J. S. Balling, "Visualization, Annotation, and Exploration of Three Dimensional Datasets Using an Existing 3D Rendering System," BioEngineering Senior Design, vol. 1, pp. 1–6, 2004.

[10] M. M. Loughlin and J. F. Hughes, “An annotation system for 3D fluid flow visualization,” Proceedings of the IEEE Conference on Visualization, vol. 5, pp. 273 – 279, 94.

[11] L. Offen and D. W. Fellner, “BioBrowser – Visualization of and Access to Macro-Molecular Structures,” Mathematics and Visualization – Visualization in Medicine and Life Sciences, vol. 5, pp. 257–273, 2008.

[12] S. Sugimoto, T. Baker, and S. L. Weibel, “Dublin Core: Process and Principles,” Lecture Notes in Computer Science – Digital Libraries: People, Knowledge, and Technology, vol. 2555/2010, pp. 25–35, 2002.

[13] “Dublin Core Metadata Initiative.” http://dublincore.org/, 1995.

[14] C. C. S. I. Group, Definition of the CIDOC Conceptual Reference Model. ICOM/CIDOC Documentation Standards Group, 2003.

[15] I. M. Bilasco, J. Gensel, M. Villanove-Oliver, and H. Martin, “An MPEG-7 framework enhancing the reuse of 3D models,” Proceedings of the International Conference on 3D web technology, vol. 11, pp. 65–74, 2006.

[16] B. Hohmann, U. Krispel, S. Havemann, and D. W. Fellner, “Cityfit: High-Quality Urban Reconstructions by Fitting Shape Grammars to Images and Derived Textured Point Clouds,” Proceedings of the ISPRS International Workshop 3D-ARCH, vol. 3, pp. 61–68, 2009.

[17] R. Berndt, H. Krottmaier, S. Havemann, and T. Schreck, “The PROBADO-Framework: Content-Based Queries for non-textual Documents,” Proceeding of the Conference on Electronic Publishing (ELPUB), vol. 13, pp. 485–500, 2009.

[18] S. Havemann, V. Settgast, R. Berndt, and Ø. Eide, “The Arrigo Showcase Reloaded – towards a sustainable link between 3D and semantics,” Proceedings of the 9th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST), vol. 9, pp. 125–132, 2008.

[19] International Organization for Standardization (ISO), “ISO 19005-1:2005 (Document management – Electronic document file format for long-term preservation),” 2005.

[20] M. Strobl, R. Berndt, V. Settgast, S. Havemann, and D. W. Fellner, “Publishing 3D Content as PDF in Cultural Heritage,” Proceedings of the 10th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage (VAST), vol. 6, pp. 117–124, 2009.

Costs and benefits of alternative scholarly publishing models: Lessons and developments

John W. Houghton

Centre for Strategic Economic Studies, Victoria University
Melbourne, Australia
E-mail: John.Houghton@vu.edu.au

Abstract

The JISC study *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits*, released early in 2009, focused on three alternative models for scholarly publishing: subscription or toll access publishing, open access publishing using the author-pays model, and self-archiving. The research approach involved a combination of process mapping, activity costing and macro-economic modelling. Since its release, there have been six follow-on studies applying elements of the same basic methodology. This paper describes the research approach and explores some of the major issues arising and lessons learned from this ongoing research. Drawing on experience from a number of studies and countries, it attempts to distil and summarise the key research issues and policy messages arising.

Keywords: Scholarly publishing; economics of publishing; cost-benefit analysis; open access.

1. Introduction

The JISC report *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits*[1] was greeted with praise in some quarters and opposition from others. During 2009, there were a number of follow-on studies. These included national studies in The Netherlands and Denmark, and a three-country comparison that explored the impacts of alternative scholarly publishing models for one of the larger (United Kingdom), a mid-sized (Netherlands) and one of the smaller European countries (Denmark). During the first half of 2010, there have been three further projects, two of which are still underway. The first focuses on Germany, and brings the

German National Licensing Program (NLP) into the mix of alternative scholarly communication models. The second, conducted by Alma Swan of Key Perspectives, focuses on the United Kingdom, using the JISC EI-ASPM on-line cost model to examine the cost implications of alternative scholarly publishing models for a sample of UK universities.[2] The third significantly extends one aspect of the underlying method used in the original study to explore the possible return on investment implications of the proposed *Federal Public Research Access Act* (FRPAA) in the United States. This paper explores some of the major issues arising and lessons learned from this ongoing research. Drawing on experience from a number of studies and countries, it attempts to distil and summarise the key research issues and policy messages arising.

2. The JISC EI-ASPM study

The JISC study *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits*[3] focused on three alternative models for scholarly publishing: subscription or toll access publishing, open access publishing using the author-pays model, and self-archiving. Because self-archiving, of itself, does not constitute formal publication, analysis focused on two publishing models in which self-archiving is supplemented by the peer review and production activities necessary for formal publishing, namely: 'Green OA' self-archiving operating in parallel with subscription publishing, and the 'deconstructed' or 'overlay journals' model in which self-archiving provides the foundation for overlay journals and services.[4] Hence, each of the publishing models explored includes all of the key functions of formal scholarly publishing, including peer review and quality control.

The approach taken to the JISC EI-ASPM study involved a combination of process mapping, activity costing and macro-economic modelling, and the research process involved four main steps.

Process mapping

Björk (2007) developed a formal model of the scholarly communication lifecycle, based on the IDEF0 process modelling method which is often used in business process re-engineering.[5] Björk's central focus was the single publication (primarily the journal article), how it is written, edited, printed, distributed, archived, retrieved and read, and how eventually its reading may affect practice. To provide a solid foundation for our analysis, we developed and extended Björk's model to include five core scholarly communication lifecycle activities, namely: (*i*) fund research and research communication; (*ii*) perform research and communicate the results; (*iii*) publish scientific and scholarly works; (*iv*) facilitate dissemination, retrieval and preservation; and (*v*) study publications and apply the knowledge. Each of these activities is further subdivided into a detailed description of the activities, inputs, outputs, controls and supporting mechanisms involved, creating a lifecycle process model with some 53 diagrams and 190 activities.[6]

Activity costing

This formal process model provided the foundation for detailed activity costing, using a spreadsheet-based cost model that included all of the elements in the lifecycle model, as well as the base data necessary for the study (*i.e.* relating to the UK and UK higher education). The costings relied primarily on existing sources, and collating activity cost information from a wide-ranging literature on scholarly communication and publishing.[7] Where necessary, these sources were supplemented by informal consultation with experts in the field. For the UK national and higher education data, we relied on national and international sources on R&D expenditure and personnel by activity and sector, expenditure and employment trends. Detailed data on higher education were sourced from such agencies as SCONUL and HESA. The resulting activity cost model included more than two thousand data elements.

Macro-economic modelling

To capture the impacts of alternative scholarly publishing models on returns to R&D expenditure, we developed a modified Solow-Swan model. The standard Solow-Swan approach makes some key simplifying assumptions, including that: all R&D generates knowledge that is useful in economic or social terms (*the efficiency of R&D*); and that all knowledge is equally accessible to all entities that could make productive use of it (*the accessibility of knowledge*). Addressing the fact that these assumptions are not realistic we introduced *accessibility* and *efficiency* into the standard model as negative or friction variables, to reflect the fact that there are limits and barriers to access and to the efficiency of production and usefulness of knowledge. Then we explored the impact on returns to R&D of changes in accessibility and efficiency.[8]

A stepwise approach

There were four main steps in the research process. In the first, we produced a detailed costing of all of the activities identified in the scholarly communication lifecycle model, focusing on areas where there were likely to be activity and, therefore, cost differences between the alternative publishing models. In the second, we summed the costs of the three publishing models through the main phases of the scholarly communication lifecycle, so we could explore potential system-wide cost differences between the alternative publishing models. In the third of the three major research steps, we used the modified Solow-Swan model to estimate the impact of changes in accessibility and efficiency on returns to R&D. The final step was to compare costs and benefits, for which we used the three elements outlined: (*i*) the direct costs associated with each of the models, (*ii*) the associated indirect system-wide costs and cost savings, and (*iii*) the benefits accruing from increases in returns to R&D resulting from increases in accessibility and efficiency. Because the returns to R&D lag expenditure and accrue over a number of years, the cost-benefit comparisons were made over a 20 year transitional period.

Findings and conclusions

Our analysis of the potential benefits of more open access to research findings suggested that open access could have substantial net benefits in the longer term, and while net benefits may be lower during a transitional period, they are likely to be positive for both open access publishing and overlay alternatives (*Gold OA*) and for parallel subscription publishing and self-archiving (*Green OA*).

For example, during a transitional period of 20 years we estimated that, in an all open access world:

- The combined cost savings and benefits from increased returns to R&D resulting from open access publishing all journal articles produced in the UK’s universities using an author-pays system (*Gold OA*) might be around 3 times the costs;
- The combined cost savings and benefits from open access self-archiving in parallel with subscription publishing (*Green OA*) might be around 7 times the costs; and
- The combined cost savings and benefits from an alternative open access self-archiving system with overlay production and review services (*overlay journals*) might be around 4 times the costs.

While the benefits from unilateral national adoption of open access alternatives would be lower, they would be substantial – ranging from 2 to 4 times the costs.

3. Responses to the JISC EI-APSM study

Responses to the JISC report have been polarised. While recognising the inherent limitations in such modelling, academic and professional commentary has been generally positive. A detailed peer review of the report undertaken by Professor Danny Quah, Head of Economics at The London School of Economics, provides an example of the academic and professional reception of the work. He concluded:

> *The report addresses an important and difficult problem, and is clearly the result of a lot of very careful thinking about the issues. The methodology is sound and the analysis is extremely detailed and transparent. The multi-stage model of production that is used is complex, and does require calibration according to a large number of parameters, many of which are necessarily estimates, where possible taken from published sources or the wider literature. If demonstrably better estimates become available then these could improve that calibration still further. The report represents the best evidence so far on the questions it addresses.*[9]

Initial comments from some publishers' representatives, including The Publishers' Association, the Association of Learned and Professional Society Publishers and the International Association of *stm* Publishers, focused on the modelling assumptions and calibration, while implicitly accepting the methodology and underlying analysis. Ware and Mabe (2009) summarised the critique, noting that:

> "*[The Houghton Report] underestimated the efficiencies of the current subscription system and the levels of access enjoyed by UK researchers. Many of the savings hypothesized would depend on the rest of the world adopting author-pays or self-archiving models. The calculated savings would remain hypothetical unless translated into job losses… Critics also argue that Houghton* et al*. underestimated the costs of switching to an author-pays model because they underestimated the true costs of publishing an article only, and because additional costs such as the infrastructure required to manage the many small publication charges were not included.*"[10]

Although referring to critics, Ware and Mabe (2009) failed to cite a single source. Nevertheless, JISC (2009) released a response addressing the criticisms soon after the release of the report and a response to Ware and Mabe was posted on the Liblicense-L list.[11]

Later in 2009, the International Association of *stm* Publishers commissioned Steven Hall to provide a critique of the JISC report, which resulted in a paper and presentation at the Berlin7 Conference. Hall's analysis rested on such claims as:

> "*The fact is, researchers today have immediate access to the vast majority of the scientific articles that they could need for their research.*"
>
> "*The fact is, the report's authors have failed to show that there is any real gap between the access that researchers have today to the scientific literature that they need and that which they might have under an open access model.*"[12]

Unfortunately, of course, the fact is that there is widespread evidence that such claims are baseless. Much of the evidence is cited in the JISC report, although some important studies have been published since which confirm, yet again, that access gaps remain a major concern. Ware (2009) reported that 73% of UK small firms experienced difficulties accessing articles and just 2% reported having all the access they needed.[13] As the Research Information Network recently noted:

> "*…access to research information content issues must be addressed if the UK research community is to operate effectively, producing high-quality research that has a wider social and economic impact.*"[14]

A response to Hall's critique is available from the Berlin7 Conference website.[15]

Much of the critique rests on the assertion that one should choose different variables. However, the project website has included an on-line version of the underlying cost model since the report's release, which allows anyone to experiment with alternative values for the major parameters (http://www.cfses.com/EI-ASPM/). Consequently, the critique rather misses the mark, as anyone could test different parameters and publish the results along with their evidence for the choice of alternative values. Moreover, our own sensitivity testing suggested that the bottom-line answer does not change for any plausible values that we have tried.

In the absence of any serious critique of the approach, subsequent studies have focused on its further development, refinement and application.

4. The three-country studies: a comparison

During 2009, the same basic approach as that used in the JISC study was applied in the Netherlands[16] and Denmark[17] with a view to exploring the potential impacts of alternative publishing models in a mid-sized and a smaller European country, as well as one of the larger European countries. For the purposes of presenting a summary of the three country studies in Brussels, Knowledge Exchange facilitated a workshop and released summary report.[18]

In exploring the potential impacts of alternative publishing models in the three countries, differences in the modelling *per se* were kept to a minimum, although some minor adjustment of the basic model to fit different national circumstances was necessary. Nevertheless, there are a number of factors that can affect the benefit/cost estimates for different countries. As modelled, these included such things as: the number and size of universities and research institutions; the implied number of institutional and other repositories, each with substantial fixed costs and relatively low variable costs; the ratios of publicly funded and higher education research spending to gross national expenditure on R&D; historical and projected rates of growth of R&D spending by sector; relative national and sectoral publication productivity; historical and projected growth in publication output; and the mix of publication types.

There are also inherent data limitations that varied somewhat between the countries. For example, in addition to cost differences between the countries, there were minor differences in the methods used to estimate full cost for researcher activities. In the UK, we used the official higher education costing methodology for full economic costing of research (TRAC fEC), for the Netherlands we used an averaged GERD/FTE researchers triangulating with a variation of a full cost model from the University of Amsterdam, and in Denmark we used a simple HERD/FTE researchers calculation. Minor differences between these methods relate primarily to the inclusion (exclusion) of the technicians counted among research personnel into (from) overheads. In addition, some UK R&D data related to 2006, whereas data for the Netherlands and Denmark were all from 2007.

Despite these influences, the different national studies produced very similar results and exhibited broadly similar patterns within the results. The cost-benefits of the open access 'author-pays' publishing model were similar across the three countries. In terms of estimated cost-benefits over a transitional period of 20 years – open access publishing all articles produced in universities in 2007 would have produced benefits of 2 to 3 times the costs in all cases, but showed benefits of 5 to 6 times costs in the simulated alternative 'steady state' model for unilateral national open access, and benefits of around 7 times the costs in an all open access world.

One observable difference related to scale and the impacts of unilateral national adoption of open access, with the benefits of worldwide adoption being relatively larger for smaller countries as they produce a smaller share of the world's journal articles. However, the most obvious difference between the results related to the 'Green OA' self-archiving and repositories model, which did not look quite as good in the Netherlands as in the UK and nothing like as good as it did in Denmark. This is due to the implied number of repositories, each with operational overheads. As modelled, the number of institutional repositories required in each country related to the number of institutions and their operational overheads were shared across the number of articles produced and archived. For example, under the modelled assumptions, for 2007 outputs, the Netherlands' 86 higher education institutional repositories might have housed around 26,000 articles (an average of 302 each from that year), the UK's 168 higher education institutional repositories might have housed around 100,000 articles (an average of 595 each from that year), and Denmark's 8 universities' repositories might have housed around 14,000 articles (an average of 1,750 each from that year). As modelled, these differences materially affected the implied per article cost of self-archiving. Of course, had we used a averaged per article lifecycle costing, these differences would not have been apparent.

Notwithstanding these differences, the modelling suggested that open access alternatives would be likely to be more cost-effective in a wide range of countries (large and small), with 'Gold OA' or author-pays publishing, the deconstructed or overlay journals model of self-archiving with overlay production and review services, and 'Green OA' self-archiving in parallel with subscription publishing progressively more cost-effective.

5. Germany: incorporating the NLP

As a part of a much larger ongoing project, funded by DfG in Germany, we have been working with colleagues at Goethe Universität in Frankfurt on a study that brings the German National Licensing Program (NLP) into the mix of alternative models, and compares the NLP with the subscription and open access alternatives.

The German NLP provides enhanced access for researchers in Germany through an extended form of consortial purchasing and licensing. While it centralises a number of activities in the lifecycle process relating to facilitating dissemination, retrieval and preservation (*e.g.* negotiation and licensing), the NLP does not fundamentally change the activities performed. Since the scholarly communication lifecycle process model focuses on activities without pre-judging which actors undertake them, incorporating the NLP has not necessitated changes to the underlying process model.

Nevertheless, the German NLP does impact a number of the five main activity areas in the scholarly communication lifecycle.

- *Fund research and research communication:* The NLP has little or no impact on the activities performed by research funders, with the exception of DfG which funds it, so no impacts were included in the modelling.
- *Perform research and communicate the results:* With the exception of time saving related to permissions and research reporting, the NLP facilitates much the same potential time saving as open access alternatives for German researchers, but scaled to the share of worldwide journal articles/titles that are encompassed by the NLP.
- *Publish scientific and scholarly works:* While it could be seen as a new sales strategy for publishers, the NLP has little or no impact on publisher costs except for possible minor savings on marketing, the operation of servers and user support. As these activities are still performed for content lying outside the NLP and the rest of the world outside Germany these impacts were excluded.
- *Facilitate dissemination, retrieval and preservation:* The NLP leads to research library savings in handling, support and purchasing and negotiation activities, scaled to the number of titles in the NLP. The counter-factual to the NLP is not readily knowable as we cannot know if the NLP content would have been subscribed to without the NLP. Hence we explored per title impacts, then multiplied by the number of titles accessible through subscriptions and through the NLP (combined). It is assumed that the NLP reduces non-negotiation and licensing subscription-related library activities by 50% (*i.e.* 50% of the non-negotiation and licensing subscription-related activity is handled centrally under the NLP and 50% is still done by the institutional research library).
- *Study publications and apply the knowledge:* The impacts of the NLP on accessibility and efficiency were modelled as follows:

- In relation to accessibility, the NLP leads to (*i*) a marginal increase in returns to German R&D through an increase in German access, which would be very small and was not included; and (*ii*) no increase in access to German research outside Germany, as it is published in the same way; and
- In relation to efficiency, the NLP's impacts are less than those of open access as it has no impact on the speed of publication and facilitates domestic collaboration only. Hence the efficiency impacts were scaled.

There is one important difference between the comparisons undertaken in the German study and those that preceded it. Subscription and open access publishing perform very different roles. To the limits of affordability, subscription access seeks to provide an institution's or country's researchers with access to the worldwide research literature; whereas open access seeks to provide worldwide access to an institution's or country's research output. These are very different things, but to compare cost-effectiveness it is necessary to compare like with like. Consequently, the JISC EI-ASPM study compared the costs associated with publishing, handling and accessing UK article output under different models. In contrast, the German study compares the costs of operating within alternative models. This does not compare the cost of using alternative models to achieve a comparable task, rather it compares the cost implications of the alternative models for a particular actor or actors (in this case for Germany).

Modelling the impacts of an increase in *accessibility* and *efficiency* resulting from more open access on returns to R&D over a 20 year period and then comparing costs and benefits, we found that the benefits of open access publishing models were likely to substantially outweigh the costs. The German National Licensing Program (NLP) returned the highest benefit/cost ratio during a transitional period and the second highest (to 'Green OA' self-archiving) in a simulated steady state alternative scenario. Whether 'Green OA' self-archiving in parallel with subscriptions is a sustainable model over the longer term is uncertain. So too is the potential for developments in open access or other scholarly publishing business models to significantly change the relative cost-benefit of the NLP over time.

Currently, the project focus is on ascertaining what impact the NLP may have on: (*i*) the take up of open access alternatives in Germany (*e.g.* by improving access for German researchers, does it reduce awareness of and pressures for open access, or does it enhance awareness of the importance of access?), and (*ii*) levels and patterns of use of the content available (*e.g.* does the NLP materially affect usage patterns, perhaps increasing usage, and, if so, how does usage under the NLP compare with that under open access?).

6. Impacts on UK universities

Alma Swan, of Key Perspectives, has recently completed another follow-on study for JISC in which she applied the on-line cost model produced as a part of the original JISC EI-ASPM study to an examination of the cost and benefit implications of alternative publishing models for a small sample of UK universities.[19] As in the German study, Swan compared the costs of operating within alternative models, in this case for a sample of universities, by setting the cost of publishing UK articles under alternative publishing models against the costs of subscription to that share of worldwide articles currently subscribed to. Again, this does not compare the cost of using alternative models to achieve a comparable task, rather it compares the cost implications of the alternative models for a particular actor or actors (in this case a sample of UK universities).

Swan found that:

- There are potential economic savings for universities from open access. Economic savings accrue to universities according to the detail of how each operates its library services and its repository, and the level of research intensiveness of the institution.
- Moving to open access as the basis for disseminating research outputs can bring economic and academic benefits for all universities, though the most research-intensive universities may face additional costs under some conditions.
- If universities continue to pay for subscription-based journals while simultaneously making their outputs freely available through their repositories, as they currently do, they are likely to make savings. Savings accrue from increased efficiencies in the research and library handling processes.
- If universities switch from the current subscription-based system to publishing all their articles in open access journals that charge an article-processing fee, there would be savings for all universities with the article-processing fee at GBP 700 per article or less.

Swan showed how universities can compare the impacts of alternative publishing models for themselves, and that by looking at whole-of-system costs we can start to question the simplistic arguments that suggest that in research-intensive universities author-pays fees may be higher than current subscription expenditures. While that may be true in some cases, it is apparent from this study that the potential savings in research time, library handling costs, etc. that could arise from more open access would also be greatest in the more research-intensive universities.

7. The United States: incorporating the FRPAA

Early in 2010, the Scholarly Publishing and Academic Resources Coalition (SPARC) supported a feasibility study that sought to outline one possible approach to measuring the impacts of the proposed US *Federal Research Public Access Act* (FRPAA) on returns to public investment in R&D.[20] The aim of the study was to define and scope the data collection requirements and further model developments necessary for a robust estimate of the likely impacts of the proposed FRPAA archiving mandate.

The project involved a major shift from previous studies in that its focus was on the modified Solow-Swan model, rather than the scholarly communication lifecycle model or the associated activity cost model. That focus enabled further development and refinement of the modified model, particularly in relation to the most appropriate lag and distribution over time of returns to R&D, the most appropriate depreciation rate for the underlying stock of R&D knowledge arising from federally funded R&D, and metrics to measure potential changes in accessibility and efficiency.

As noted, the standard approach makes some key simplifying assumptions, including that all R&D generates knowledge that is useful in economic or social terms (*the efficiency of R&D*), and that all knowledge is equally accessible to all entities that could make productive use of it (*the accessibility of knowledge*). These assumptions are not realistic. In the real world, there are limits and barriers to access and limits to the usefulness of knowledge. So, we introduced *accessibility* and *efficiency* into the standard model as negative or friction variables, and then looked at the impact on returns to R&D of reducing the friction by increasing *accessibility* and *efficiency.*

To operationalise the model it was necessary to establish values for the *accessibility* and *efficiency* parameters, as well as a number of other parameters, such as rates of return to R&D and of depreciation of the underlying stock of research knowledge. To establish plausible base case values for these parameters we drew on the extensive literature on returns to R&D.[21]

The other piece of the puzzle is the input data required for the modelling. These include the implied archiving costs, the volume of federally funded research outputs (journal articles), and the levels of federal research funding and expenditure trends. For the purposes of preliminary analysis we used publicly available sources and published estimates. Data relating to federal research funding, activities and outputs were taken from the National Science Board *Science and Engineering Indicators 2010*,[22] and we explored three sources for archiving costs: the LIFE^2 Project lifecycle costs,[23] and submission equivalent costings from arXiv[24] and NIH.[25] In order to enable anyone to use alternative values for the various parameters, test sensitivities and explore the issues for themselves, we created a simplified model in MS Excel format.[26]

Preliminary modelling suggests that over a transitional period of 30 years, the potential *incremental* benefits of the proposed FRPAA archiving mandate for all federally funded R&D might be worth around 3 times the estimated cost using the higher end LIFE^2 lifecycle costing, 6 times the cost using the NIH costing and 17 times the cost using the arXiv costing. Perhaps two-thirds of these benefits would accrue within the US, with the remainder spilling over to other countries. Hence, the US national benefits might be of the order of 2 to 11 times the costs.

Exploring sensitivities in the model in order to identify major sensitivities and, thereby, prioritise areas for further data collection and model development, we found that the benefits exceed the costs over a wide range of values. Indeed, it is difficult to imagine any plausible values for the input data and model parameters that would lead to a fundamentally different answer.

8. Summary and conclusions

With the exception of the US project, the studies have combined three main approaches: a scholarly communication lifecycle or process re-engineering model; a spreadsheet-based activity cost model; and a modified macro-economic model into which we introduced accessibility and efficiency as negative variables. The studies sought to map activities throughout the scholarly communication lifecycle, then attach costs to each of the activities to explore the direct activity cost differences and indirect system-wide cost differences between publishing models. The modified Solow-Swan model was used to explore the impact of increased accessibility on returns to R&D. In the final step, the costs are set against system-wide cost savings and benefits in the form of increases in returns to R&D to estimate cost-benefits. While there are many limitations to such modeling, the results were not particularly sensitive. We found that the benefits of more open access exceeded the costs across a wide range of values, and it is difficult to imagine any plausible values for the main variables that would result in a fundamentally different answer.

Given the sometimes heated debate on open access and the detail and complexity of the research methods used, it is perhaps not surprising that reactions have been polarized and somewhat piecemeal. While recognising the inherent limitations in such modelling, academic and professional commentary has been generally positive. Comments from some publishers' representatives have been critical, focusing on the details of the modelling assumptions and calibration. Much of the critical comment has been based on unsubstantiated claims, misunderstandings and falsehoods, and we have provided a response to each of the commentators. We have also provided on-line versions of the models to enable anyone to explore the impacts of using their own preferred values for the main variables.

Given that there has been no substantive critique of the work and that the results appear to hold across a wide range of values and a range of countries, the evidence would seem to suggest that more open access could have substantial net benefits in the longer term, and while net benefits may be smaller during a transitional period, they are likely to be positive for both open access publishing and overlay alternatives (*Gold OA*) and for parallel subscription publishing and self-archiving (*Green OA*). At the institutional level, Swan (2010) has shown that the benefits would be likely to outweigh the costs for all but the most research-intensive of universities.[27]

Given the capacity to enhance access at very little cost, self-archiving alternatives appear to be the more cost-effective – although whether self-archiving in parallel with subscriptions is a sustainable model over the longer term is debateable. Similarly, the German NLP provides enhanced access for researchers in Germany through an extended form of consortial purchasing and licensing, which by centralising certain library subscription-related activities is effectively negative cost (*i.e.* the centralised costs are less than the de-centralised cost savings). Hence, it too provides a highly cost-effective avenue for enhanced access. The downside risks of such a program include the potential for developments in open access or other scholarly publishing business models to significantly erode the relative cost-benefit of the NLP over time, and the potential impact the NLP may have on slowing the take up of open access alternatives in Germany (*e.g.* by improving access for German researchers, does it reduce awareness of and pressures for open access, or does it enhance awareness of the importance of access?).

Nevertheless, the evidence would suggest that archiving policies and mandates, be they at the national, sectoral, funder or institutional levels, can enhance accessibility and improve efficiency at relatively little cost and with no immediate disruptive change to scholarly publishing practices and traditions. As such, archiving mandates provide an obvious focus for policy and implementation activities in the immediate term, while more fundamental changes to scholarly communication practices evolve over time.

Notes and References

[1] HOUGHTON, J.W., RASMUSSEN, B., SHEEHAN, P.J., and OPPENHEIM, C., MORRIS, A., CREASER, C., GREENWOOD, H., SUMMERS, M. and GOURLAY, A. *Economic Implications of Alternative Scholarly Publishing Models: Exploring the Costs and Benefits*, Bristol and London: The Joint Information Systems Committee (JISC), 2009. Available http://www.jisc.ac.uk/publications/publications/economicpublishingmodelsfinalreport (December 2009).

[2] SWAN, A. *Modelling scholarly communication options: costs and benefits for universities*, Bristol and London: The Joint Information Systems Committee (JISC), 2010. Available http://ie-repository.jisc.ac.uk/442/ (February 2010).

[3] HOUGHTON, J.W., RASMUSSEN, B., SHEEHAN, P.J., and OPPENHEIM, C., MORRIS, A., CREASER, C., GREENWOOD, H., SUMMERS, M. and GOURLAY, A. *Economic Implications of Alternative Scholarly Publishing Models: Exploring the Costs and Benefits*, Bristol and London: The Joint Information Systems Committee (JISC), 2009. Available http://www.jisc.ac.uk/publications/publications/economicpublishingmodelsfinalreport (December 2009).

[4] SMITH, J.W.T. 'The Deconstructed Journal, a new model for Academic Publishing,' *Learned Publishing* 12(2), 1999, pp79-91.; SMITH, J.W.T. Open Access Publishing Models: Reinventing Journal Publishing, *Research Information*, May-June 2005.; VAN DE SOMPEL, H. *et al.* 'Rethinking Scholarly Communication: Building the system that scholars deserve,' *D-Lib Magazine* 10(9) September 2004.; SIMBOLI, B. *Subscription subsidized open access and the crisis in scholarly communication*, Lehigh University, 2005.; HOUGHTON, J.W.

'Economics of Publishing and the Future of Scholarly Communication' in Eds. GORMAN, G.E. & ROWLAND, F. *International Year Book of Library and Information Management 2004-2005: Scholarly Publishing in an Electronic Era*, London: Facet Publishing, 2005.

[5] BJÖRK, B-C. 'A model of scientific communication as a global distributed information system,' *Information Research* 12(2) paper 307, 2007.

[6] Details of the entire model in 'browseable' form can be found at http://www.cfses.com/EI-ASPM/SCLCM-V7/

[7] TENOPIR, C. and KING, D.W. *Towards Electronic Journals: Realities for Scientists, Librarians and Publishers,* Washington D.C.: Special Libraries Association, 2000.; TENOPIR, C. and KING, D.W. 'Reading behavior and electronic journals,' *Learned Publishing* 15(4), 2002, pp259-265.; TENOPIR, C. and KING, D.W. 'Perceptions of value and value beyond perceptions: measuring the quality and value of journal article readings,' *Serials* 20(3), 2007, pp199-207. Available http://www.uksg.org/serials (March 2009).; KING, D.W. The cost of journal publishing: a literature review and commentary, *Learned Publishing 20(2)*, April 2007, pp. 85-106.; HALLIDAY, L. and OPPENHEIM, C. *Economic Models of the Digital Library*, eLib, United Kingdom Office of Library and Information Networking, 1999.; FRIEDLANDER, A. and BESSETTE, R.S. *The Implications of Information Technology for Scientific Journal Publishing: A Literature Review*, Washington DC: National Science Foundation, 2003.; OECD. *Digital Broadband Content: Scientific Publishing*, Paris: OECD, 2005. Available http://www.oecd.org/dataoecd/42/12/35393145.pdf (February 2010); EUROPEAN COMMISSION. *Study on the economic and technical evolution of the scientific publication markets in Europe*, Brussels: European Commission, 2006.; EPS, *et al. UK scholarly journals: 2006 baseline report – An evidence-based analysis of data concerning scholarly journal publishing*, London: Research Information Network, Research Councils UK and Department of Trade and Industry, 2006. Available at http://www.rin.ac.uk/data-scholarly-journals (December 2009).; BJÖRK, B-C. 'A model of scientific communication as a global distributed information system,' *Information Research* 12(2) paper 307, 2007.; CEPA. *Activities, costs and funding flows in the scholarly communications system in the UK*, London: Research Information Network (RIN), 2008.; CLARKE, R. 'The cost profiles of alternative approaches to journal publishing,' *First Monday* 12(12), December 2007.

[8] HOUGHTON, J.W. and SHEEHAN, P. 'Estimating the potential impacts of open access to research findings,' *Economic Analysis and Policy* 39(1), 2009. Available http://eap-journal.com/ (January 2010).; HOUGHTON, J.W., RASMUSSEN, B., SHEEHAN, P.J., and OPPENHEIM, C., MORRIS, A., CREASER, C., GREENWOOD, H., SUMMERS, M. and GOURLAY, A. *Economic Implications of Alternative Scholarly Publishing Models: Exploring the Costs and Benefits*, Bristol and London: The Joint Information Systems Committee (JISC), 2009. Available http://www.jisc.ac.uk/publications/publications/economicpublishingmodelsfinalreport (December 2009).

[9] JISC. *JISC's response to comments from publishers' representative groups*, 2009. Available http://www.jisc.ac.uk/media/documents/publications/responseoneiaspmreport.pdf (March 2010).

[10] WARE, M. and MABE, M. *The STM Report: An overview of scientific and scholarly journal publishing*, Oxford: STM Publishers Association, 2009. Available http://www.stm-assoc.org/news.php?id=255&PHPSESSID=3c5575d0663c0e04a4600d7f04afe91f (December 2009).

[11] JISC. *JISC's response to comments from publishers' representative groups*, 2009. Available http://www.jisc.ac.uk/media/documents/publications/responseoneiaspmreport.pdf (March 2010).; Houghton, J.W. *Re: Growth for STM publishers in 2008*, 2010. Available http://www.library.yale.edu/~llicense/ListArchives/0910/msg00056.html (March 2010).

[12] HALL, S. 'Widening access to research information: collaborative efforts towards transitions in scholarly communications', Paper presented at The Berlin7 Conference, Paris, December 2009. Available http://www.berlin7.org/IMG/pdf/hall.pdf (February 2010).

[13] WARE, M. *Access by UK small and medium-sized enterprises to professional and academic literature,* Bristol: Publishing Research Consortium, 2009. Available http://www.publishingresearch.net/SMEaccess.htm (December 2009).

[14] RIN. *Overcoming barriers: access to research information content*, London: Research Information Network, 2009. Available http://www.rin.ac.uk/system/files/attachments/Sarah/Overcoming-barriers-report-Dec09_0.pdf (December 2009).

[15] HOUGHTON, J.W. and OPPENHEIM, C. *Widening access to research information: A response*, The Berlin7 Conference, Paris, 2010. Available http://www.berlin7.org/spip.php?article57 (February 2010).

[16] HOUGHTON, J.W., DE JONGE, J. and VAN OPLOO, M. *Costs and Benefits of Research Communication: The Dutch Situation,* Utrecht: SURFfoudation, 2009. Available http://www.surffoundation.nl/wiki/display/economicstudyOA/Home (December 2009).

[17] HOUGHTON, J.W. *Costs and Benefits of Alternative Publishing Models*: *Denmark*, Copenhagen: DEFF, 2009. Available http://www.knowledge-exchange.info/Admin/Public/DWSDownload.aspx?File=%2fFiles%2fFiler%2fdownloads%2fDK_Costs_and_benefits_of_alternative_publishing_models.pdf (December 2009).

[18] HOUGHTON, J.W. *Open Access: What are the economic benefits? A comparison of the United Kingdom, Netherlands and Denmark*, Brussels: Knowledge Exchange, 2009. Available http://knowledge-exchange.info/Default.aspx?ID=316 (December 2009).

[19] SWAN, A. *Modelling scholarly communication options: costs and benefits for universities*, Bristol and London: The Joint Information Systems Committee (JISC), 2010. Available http://ie-repository.jisc.ac.uk/442/ (February 2010).

[20] HOUGHTON, J.W., RASMUSSEN, B. and SHEEHAN, P.J. *Economic and Social Returns on Investment in Open Archiving Publicly Funded Research Outputs*, Washington DC: SPARC, 2010 (Forthcoming).

[21] SALTER, A.J. and MARTIN, B.R. 'The economic benefits of publicly funded basic research: a critical review,' *Research Policy* 30(3), 2001, pp509-532.; MARTIN, B.R. and TANG, P. *The benefits of publicly funded research*, SWEPS Paper No. 161, Brighton: Science Policy Research Unit, 2007. Available http://www.sussex.ac.uk/spru (December 2009).; SVEIKAUSKAS, L. *R&D and Productivity Growth: A Review of the Literature,* US Washington DC.: Bureau of Labor Statistics Working Paper 408, BLS, 2007. Available www.bls.gov/osmr/pdf/ec070070.pdf (February 2010).; HALL, B.H., MAIRESSE, J. and MOHNEN, P. *Measuring the returns to R&D*, NBER Working Paper 15622, Cambridge MA: NBER, 2009.

[22] NATIONAL SCIENCE BOARD (NSB). *Science and Engineering Indicators 2010*, Arlington, VA.: National Science Foundation (NSB 10-01), 2010. Available http://www.nsf.gov/statistics/ (January 2010).

[23] AYRIS, P. *et al*. *The LIFE^2 Final Report*, London & Bristol: The Joint Information System Committee (JISC), 2008. Available http://www.life.ac.uk/2/documentation.shtml (January 2010).

[24] ARXIV.ORG. *arXiv Business Model White Paper*, Cornell University, 2010. Available http://arxiv.org/help/support/whitepaper (January 2010).

[25] NATIONAL INSTITUTES OF HEALTH (NIH). *Analysis of Comments and Implementation of the NIH public access policy*, Washington DC.: NIH, 2008. Available http://publicaccess.nih.gov/analysis_of_comments_nih_public_access_policy.pdf (February 2010).

[26] See http://www.cfses.com/SPARC-FRPAA/.

[27] SWAN, A. *Modelling scholarly communication options: costs and benefits for universities*, Bristol and London: The Joint Information Systems Committee (JISC), 2010. Available http://ie-repository.jisc.ac.uk/442/ (February 2010).

The Open Access Landscape 2009

Bo-Christer Björk[1], Patrik Welling[1], Peter Majlender[1], Turid Hedlund[1], Mikael Laakso[1], Gudni Gudnasson[2].

[1] HANKEN School of Economics, Helsinki, Finland
[2] Innovation Center Iceland, Reykjavik, Iceland

Abstract

The Internet has technically facilitated making scientific results available to a much wider readership than ever before, both via electronic subscriptions but also for free in the spirit of Open Source licensing of software and the knowledge sharing of Wikipedia. This emerging openness has important implications for better impact of published research in general and for bridging the digital divide between the researchers of the leading universities and the developing nations.

A central question many policymakers ask is how common Open Access is today and how fast the share of OA is increasing. What proportion of journal articles are OA and to what extent do researchers post OA copies in repositories? Accurate answers to such questions would be very valuable for instance for research funders and for university administrators. The purpose of the study reported on in this paper is to provide answers to this type of questions.

Keywords: Open access publishing, prevalence of open access articles

1. Earlier studies

Although some estimates of OA prevalence have been published over the last few year there is a clear need for rigorously conducted studies. Also the share is constantly changing and thus studies need to be up-to-date. So far the volume of OA has been studied for instance in the following ways.

- For gold OA publishing it has been easy to compare the number of OA journals listed in the DOAJ index to the total number of active peer reviewed scholarly journals listed in the Ulrich's periodicals directory.
- For green OA there are directories listing repositories and statistics of how many documents these contain.
- For particular limited disciplines it is possible to take the content in a few leading journals and check the availability of OA copies via googling
- For larger masses of articles the availability of full text versions OA can be checked by web crawling robots (cf. ex, [1]) that are fed by article titles from indexing services such as Web of Science.

All these methods suffer from limitations. On average OA journals publish far fewer articles per annum than subscription based one [2] and thus the share of OA articles in the total global article volume is much lower than the share of titles. Secondly the criteria for inclusion in DOAJ and Ullrich's might differ, so that the number of journals may not be directly comparable.

Counting the number of documents in repositories may tell a lot about the growth of the repositories, but the numbers cannot usually easily distinguish between copies of articles published elsewhere and a wide range of other materials (thesis, working papers, research data, teaching material etc).

2. Research methods and Results

The proportion of all peer reviewed scholarly journal articles, which are available openly on the web without any restrictions (Open Access), was empirically studied. A sample was constructed using articles from 2008 obtained from a citation indexing service (Scopus) which covered approximately 1,2 million articles, estimated to represent around 80 % of the whole peer reviewed article stock of that year. The sample, which all in all included 1773 titles, was stratified over 9 disciplines so that roughly equal numbers of titles were included in each of the sub-samples. The research method consisted of using a web search engine in order to find free full text copies of the articles. A team of researchers shared the workload, aided by a spreadsheet tool linked to the search engine.

The detailed results will be presented at the conference itself. We have submitted manuscripts to journals, which prohibit publishing them in the conference proceedings (which are openly available) at this stage. Readers of this abstract can check for the results later by googling using the authors' names.

References

[1] Hajjem, C., Harnad, S., Gingras, Y. (2005). Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact. IEEE Data Engineering Bulletin, 28(4): 39-47. Available at: http://eprints.ecs.soton.ac.uk/11688/

[2] Björk, Bo-Christer, Roos, Annikki, Lauri, Mari 2009 Scientific journal publishing – yearly volume and open access availability, Information Research, 14(1) paper 391. http://InformationR.net/ir/14-1/paper391.html]

Mapping the structure and evolution of electronic publishing as a research field using co-citation analysis

Yaşar Tonta; Güleda Düzyol

Department of Information Management, Faculty of Letters,
Hacettepe University06800, Beytepe, Ankara, Turkey
{tonta, gduzyol}@hacettepe.edu.tr

Abstract

Electronic publishing can be defined as making full-texts of journal articles and books available through the network. Although e-publishing has been in existence for over 30 years in various forms such as CD-ROMs, it owes much of its current level of development to the Internet and the Web. This paper attempts to chart the evolution of e-publishing as a research field over the last 31 years using CiteSpace, an information visualization tool. It maps the intellectual structure of e-publishing based on 493 articles that appeared in professional literature on the subject between 1979 and 2009. Document co-citation and author co-citation patterns and patterns of noun phrases and keywords of papers on e-publishing are visualized through a number of co-citation maps. Maps show the major research strands and hot topics in e-publishing such as "open access" and would improve our understanding of the e-publishing as a research field.

Keywords: electronic publishing; information visualization; CiteSpace

1. Introduction

Scientific papers and publications reflect the rapid growth of human knowledge. Studying citations in research papers describes the development of science and explains the starting point and intellectual bases of the scientific research [1]. Bibliometrics uses citation data to trace the growth of published literature and study the patterns of publications and specific scientific developments within a field [1, 2]. Co-citation analysis can be used to study various aspects of scientific networks and to map structures of scholarly research in a certain field [2, 3, 4]. It identifies how often "two documents are . . . co-cited when they both appear in the reference list of a third document" [5]. Author co-citation analysis (ACA) is used to find out the number of times "that selected author pairs are cited together in articles, regardless of which of their works are cited" and it tries to "identify influential authors and display their interrelationships from the citation record" [6]. Co-word analysis, on the other hand, is based on the co-occurrence frequency of pairs of words or phrases" and "used to discover linkages among subjects in a research field and thus to trace the development of science" [7]. "Åström found a good correspondence between maps based on author-co-citation analysis and on co-occurrence of descriptors" [8]. Such relationships between citations and words reveal networks of documents, authors and words, respectively [9, 10, 11].

Studying networks has been an established research topic in information science and other disciplines. A network consists of nodes (i.e., articles, words or authors) and links (to other articles, words or authors). "Each node in the network represents a reference cited by records in the retrieved dataset" [12]. The size of a node and its label is proportional to the frequency of citations. Colors on a node (so called "rings") correspond to the time slice in which citations were made. The thicker the ring for a certain color, the more citations the paper received from that time slice [13]. The lines between these circles represent co-citations. The width and length of links are proportional to the co-citation coefficient. Colors of links indicate the first appearance of those links [4]. Thicker lines and closer nodes indicate that the pairs are co-cited more frequently and thus more similar [2].

Social network analysis (SNA) used in creating co-citation maps is based on graph theory. SNA offers several measures such as density and centrality to study the characteristics of a network and conceptualize it [14]. The "density" of a network is defined as the number of actual links between nodes divided by the number of possible links and represents the connectedness of the graph [15]. The "centrality" of a network, on the other hand, measures relationships between nodes in terms of degree, closeness and betweenness. Central nodes are more important in a network [16]. Degree centrality is the number of direct relationships that a node has. Betweenness centrality is an indicator of a node's ability to make connections to other nodes in a network while closeness centrality measures how quickly a node can access more nodes in a network [17].

Highly cited, and thus important, articles in a co-citation network form "landmark" nodes. Articles that have many connections to other articles are called "hubs". The "pivot" nodes, on the other hand, connect different sub-networks in a co-citation network through playing a brokerage role [18]. Scientific networks tend to change over time in various ways. Moderate as well as dramatic changes may be observed [4]. As a scientific field matures, new nodes and links get added to the network while some of the existing ones get merged with other nodes or would disappear altogether.

This paper aims to assess the evolution of e-publishing as a research field using scientific visualization techniques. Tracing its historical development between 1979 and 2009, we carried out a domain analysis of the e-publishing field so as to see how it is that the intellectual structure of e-publishing has changed over time. In addition to providing descriptive statistics on e-publishing, we addressed the following research questions:

- What are the prominent articles in the field of e-publishing?
- What major areas of e-publishing exist and how are they interlinked?
- Which authors are major knowledge producers?
- Is there an evolving area in e-publishing as a research field?

We used the CiteSpace software (http://cluster.ischool.drexel.edu/~cchen/citespace/) to explore the research fronts in e-publishing field and addressed the research questions by means of co-citation analysis and scientific information visualization tools.

2. Methodology

We performed a topical search on Thomson Reuters' Web of Science (WoS) online database to identify papers on e-publishing that appeared in the literature between 1979 and 2009 [19]. We used the terms "electronic publishing", "e-publishing" and "digital publishing" for topical searches. A total of 1,182 papers were identified. Some 689 contributions other than journal articles (book reviews, editorials and other document types) were excluded. The full bibliographic records (including authors, titles, abstracts and reference lists) of the remaining 493 journal articles were downloaded along with a total of 1,895 citations that they received.

We used CiteSpace to analyze and visualize co-citation networks. Developed by Dr. Chaomi Chen, CiteSpace facilitates the analysis of emerging trends in a knowledge domain [4]. CiteSpace is part of the developing field of "knowledge domain visualization" aimed at creating a picture of how science grows and evolves over time [18]. "Compared with earlier visualizations, the new methods in CiteSpace have improved the clarity and interpretability of visualizations" [16]. CiteSpace supports collaboration networks of co-authors, institutions and countries, document co-citation networks, concept networks of noun phrases and keywords, and hybrid networks that consist of multiple types of nodes and links [9]. CiteSpace reduces the number of links that must be shown and weights the remaining ones, thereby preserving the network's basic structure.

We analyzed the data using two-year time slices, making altogether 16 slices for the entire period of 1979-2009. In each time slice, a co-citation network was constructed based on the co-citation instances made by the top 30 most cited records published in the corresponding time interval and the threshold values.

3. Findings and Discussion

Table 1 provides descriptive statistics on papers on electronic publishing that appeared in professional literature between 1979 and 2009. During this period, a total of 493 papers with "electronic publishing" or "e-publishing" or "digital publishing" in their topics were published and they were cited 1,895 times. On the average, 16 papers appeared on electronic publishing annually (SD = 11) and they received 61 citations (SD = 72).

Table 1: Number of articles and citations on electronic publishing (1979-2009)

Year	# of articles	# of times cited	Year	# of articles	# of times cited
1979	1	12	1995	32	238
1980	0	0	1996	30	111
1981	0	0	1997	34	62
1982	4	0	1998	36	88
1983	4	2	1999	29	216
1984	7	43	2000	42	220
1985	11	26	2001	28	94
1986	9	13	2002	21	127
1987	11	11	2003	22	145
1988	5	10	2004	16	50
1989	10	11	2005	15	45
1990	5	0	2006	17	37
1991	12	11	2007	15	12
1992	14	189	2008	21	25
1993	9	14	2009	10	4
1994	23	79			
Total			Total	493	1,895

While the number of papers and citations thereto were not high between 1979 and 1993 (average of 7 papers and 23 citations per year), they have increased considerably between 1994 and 2000 (average of 32 papers and 145 citations per year). This is probably due to the fact that the number of Internet and Web users proliferated in early 1990s when the Internet became available outside the academia, thereby increasing both the number of e-publishing activities and papers engendered therefrom. The increase has slowed down after the year 2000 (average of 18 papers and 60 citations per year), which can perhaps be explained by the appearance of more specific papers on e-publishing indexed under more specific keywords.

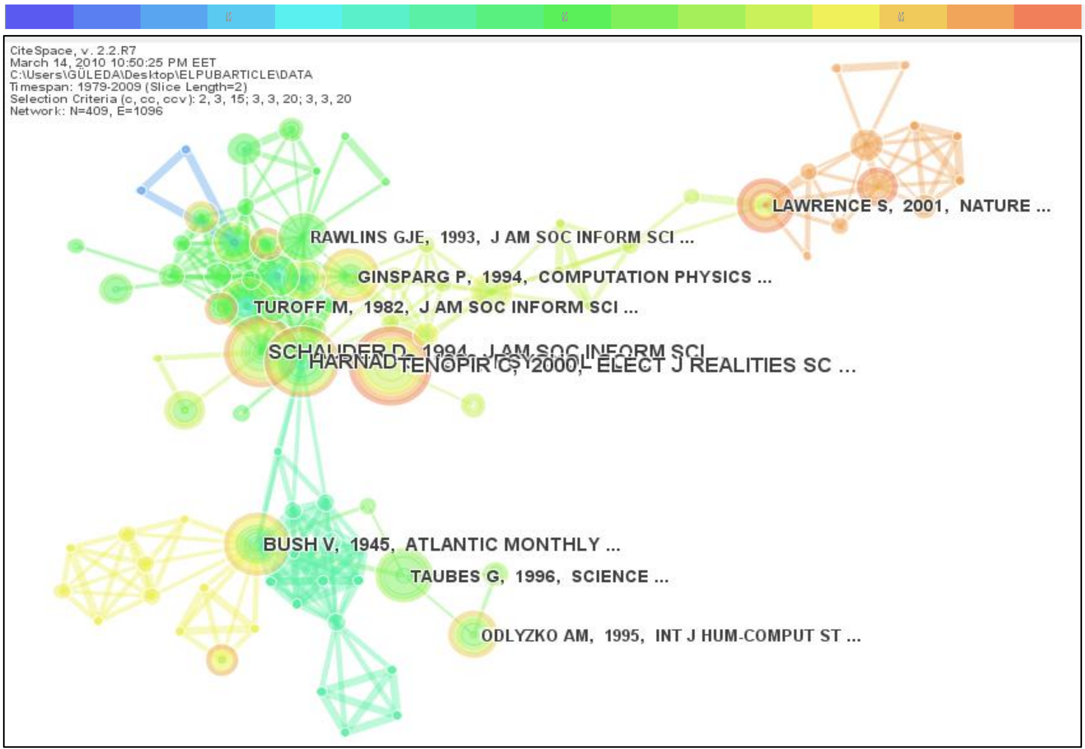
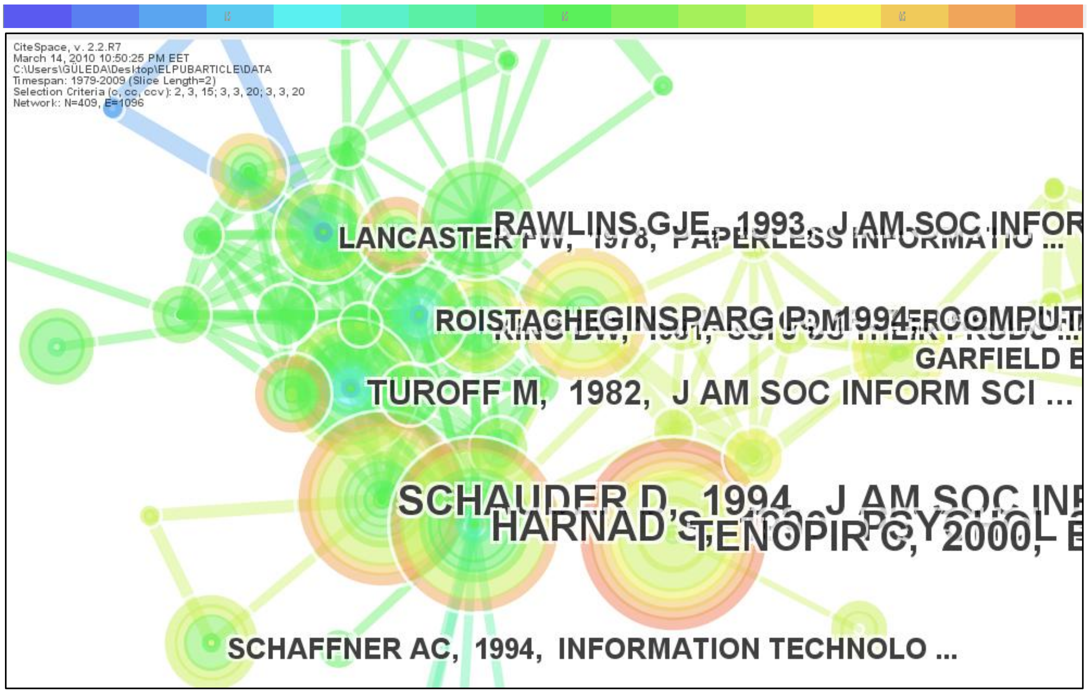

CiteSpace, v. 2.2.R7
March 14, 2010 10:50:25 PM EET
C:\Users\GÜLEDA\Desktop\ELPUBARTICLE\DATA
Timespan: 1979-2009 (Slice Length=2)
Selection Criteria (c, cc, ccv): 2, 3, 15; 3, 3, 20; 3, 3, 20
Network: N=409, E=1096
LAWRENCE S, 2001, NATURE ...
RAWLINS GJE, 1993, J AM SOC INFORM SCI ...
GINSPARG P, 1994, COMPUTATION PHYSICS ...
TUROFF M, 1982, J AM SOC INFORM SCI ...
SCHAUDER D, 1994, J AM SOC INFORM SCI
HARNAD
TENOPIR C, 2000, ELECT J REALITIES SC ...
BUSH V, 1945, ATLANTIC MONTHLY ...
TAUBES G, 1996, SCIENCE ...
ODLYZKO AM, 1995, INT J HUM-COMPUT ST ...

Fig. 1: A document co-citation network of electronic publishing (1979-2009)

Figure 1 shows a document co-citation network derived from the citing behavior of authors writing on e-publishing. This network is the result of merging 15 two-year and 1 one-year (2009) document co-citation networks generated by the WoS dataset (1979-2009). It consists of 409 papers that have been cited by two or more e-publishing articles and 1,096 co-citation links. Each co-citation link represents at least three co-citations. Citations made in earlier years are shown in blue links, mid-range years in green and yellow, and recent years in orange. The colors of co-citation links depict the earliest year in which the connection between two documents was made for the first time. For example it is quite possible that papers published in the 1980s were not co-cited until 1990s [20].

Structurally strategic nodes can easily be identified in Figure 1[21]. Tenopir and King's book (2000) on electronic journals appears to be the most prominent source as it was cited the most. Journal articles on e-publishing by Harnad ("Scholarly skywriting", 1990), Schauder ("Electronic publishing of professional articles", 1994) and Ginsparg ("First steps towards electronic research communication", 1994) were the second most highly cited articles in the network. These four sources started to get cited soon after publication and still continue to be cited today, as the outer orange rings indicate.

In Figure 1, we see three distinct clusters in the network. (These three clusters are shown in detail in Figures 2, 3 and 4). We have already pointed out the strategic nodes of Tenopir-2000, Harnad-1990, Schauder-1994 and Ginsparg-1994 at the middle of the network. Figure 2 shows the middle and upper left-hand cluster in Figure 1 in more detail. Figure 2 comprises papers with mainly green links, indicating that this cluster was formed between 1991 and 2002. Papers in this cluster (e.g., Tenopir-2000, Harnad-1990, Schauder-1994, Ginsparg-1994) have also been cited after 2002. Rawlins-1993, Turoff-1982, and Lancaster-1978 have not been cited after 2003. Tenopir-2000, Harnad-1990 and Ginsparg-1994 provide connection with the recently formed upper right-hand part of the network (see Fig. 3). To put it differently, they were cited by papers in this cluster whose centroid is represented by Lawrence's seminal letter ("Online or Invisible?") that appeared in the journal *Nature* in 2001. The linkage between the two clusters was formed in 2001-2002 time slice, which roughly corresponds to the rise of open access debate in early 2000s. The debate was (and, to some extent, still is) centred on the potential impact of e-publishing through open access e-journals in terms of use and citations exemplified in Antelman-2004 ("Do open access articles have a greater research impact?") and Kurtz-2005 ("The effect of use and access on citations"), for example.



CiteSpace, v. 2.2.R7
March 14, 2010 10:50:25 PM EET
C:\Users\GÜLEDA\Desktop\ELPUBARTICLE\DATA
Timespan: 1979-2009 (Slice Length=2)
Selection Criteria (c, cc, ccv): 2, 3, 15; 3, 3, 20; 3, 3, 20
Network: N=409, E=1096

LANCASTER FW, 1978, PAPERLESS INFORMATIO ...
RAWLINS GJE, 1993, J AM SOC INFOR
ROISTACHE
GINSPARG P, 1994, COMPUT
KING DW, 1981,
GARFIELD E
TUROFF M, 1982, J AM SOC INFORM SCI ...
SCHAUDER D, 1994, J AM SOC INF
HARNAD S, 1990, PSYCHOL
TENOPIR C, 2000, E
SCHAFFNER AC, 1994, INFORMATION TECHNOLO ...

Fig. 2: The middle and upper left-hand part of network in Figure 1 in detail

The cluster in the upper right-hand part of the network seems to have been formed recently, as the prevalent orange and red rings indicate. This part shows the most recent active area of e-publishing field. Sources in that cluster were cited mostly after 2005. This part of the network shown in detail in Figure 3 represents an evolving thread and contains highly cited articles by Lawrence-2001, Antelman-2004, Kurtz-2005, Miller-2004, Odlyzko-2002, Swan-2005, and Jones-2006. The first paper published by Lawrence-2001, was first cited in 2001 and heavily cited after 2005, whereas the Kurtz-2005 paper was not cited between 2007 and 2009.

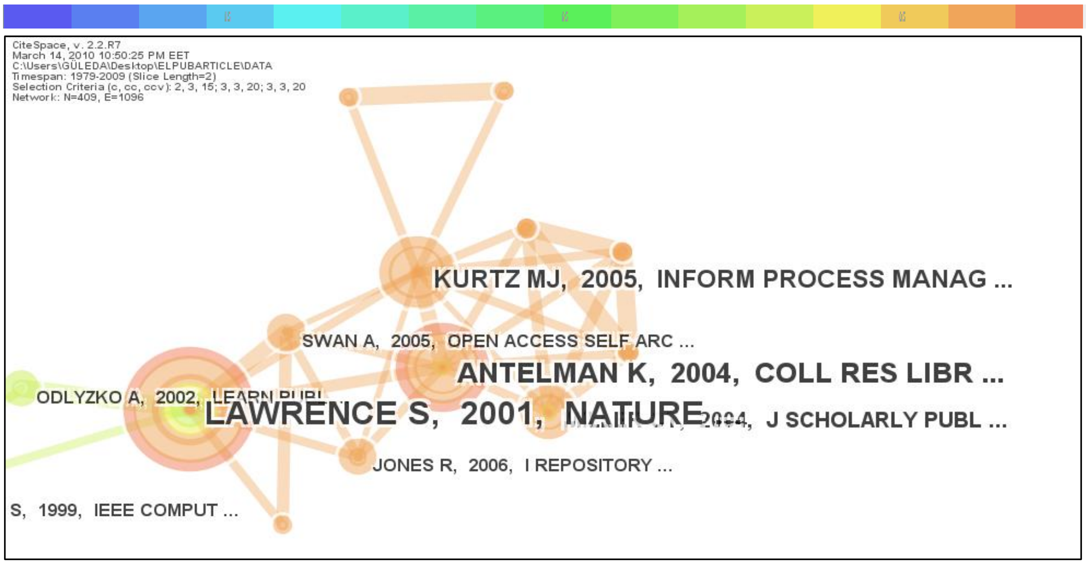

CiteSpace, v. 2.2.R7
March 14, 2010 10:50:25 PM EET
C:\Users\GÜLEDA\Desktop\ELPUBARTICLE\DATA
Timespan: 1979-2009 (Slice Length=2)
Selection Criteria (c, cc, ccv): 2, 3, 15; 3, 3, 20; 3, 3, 20
Network: N=409, E=1096
KURTZ MJ, 2005, INFORM PROCESS MANAG ...
SWAN A, 2005, OPEN ACCESS SELF ARC ...
ANTELMAN K, 2004, COLL RES LIBR ...
ODLYZKO A, 2002, LEARN PUBL
LAWRENCE S, 2001, NATURE
2004, J SCHOLARLY PUBL ...
JONES R, 2006, I REPOSITORY ...
S, 1999, IEEE COMPUT ...

Figure 3: The upper right-hand part of the network in Figure 1 in detail.

Figure 4 shows the lower part of the network that was formed starting from 1988. Note that the seminal article by Vannevar Bush ("As we may think", 1945) is one of the nodes connecting two clusters in the network and continued to be cited until 2006. Also, papers by Odlyzko and Negroponte have been cited up until recent years. We can see a dense cluster in yellow on the left-hand side next to the Bush's 1945 paper. This cluster was formed in 2003-2004 time slice and contains papers by Negroponte-1995, Ormes-2001, Crawford-2000, Hawkins-2000, and Sottong-2001.

We also carried out a network analysis of authors contributing to e-publishing literature (author co-citation analysis) (Fig. 5). The network contains 340 authors cited by the e-publishing dataset and 1091 co-citation links. The largest connected component of this network is densely connected and therefore it is difficult to identify sub-networks, even if they exist. (See Fig. 6 for the blow up of the densest part of Fig. 5.) Increasing the threshold value does not help much in this respect, as "meaningful pairwise associations are broken" and related authors "appear in different components" [5].

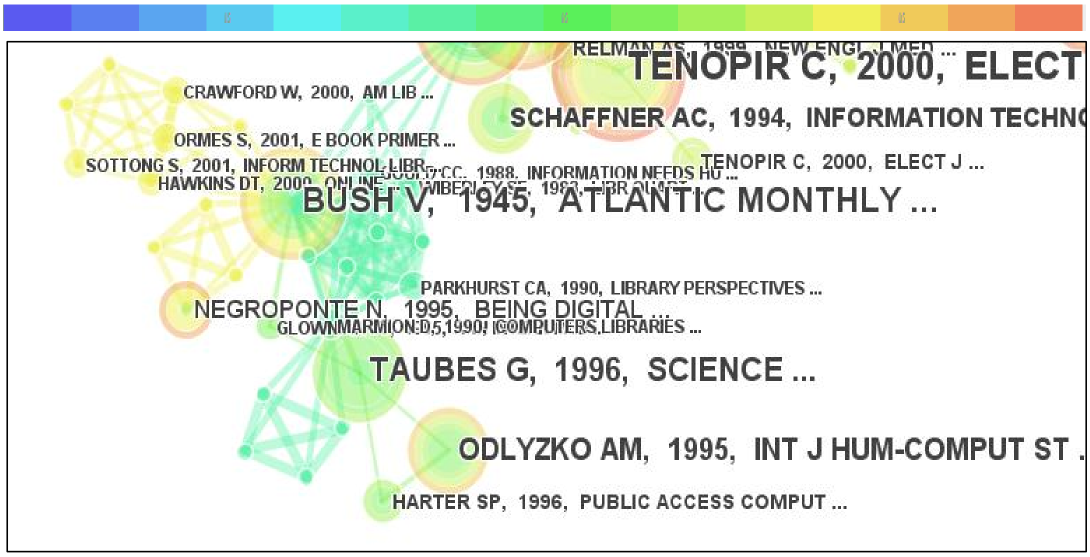

RELMAN AS, 1999, NEW ENGL J MED ...
TENOPIR C, 2000, ELECT
CRAWFORD W, 2000, AM LIB ...
SCHAFFNER AC, 1994, INFORMATION TECHN
ORMES S, 2001, E BOOK PRIMER ...
SOTTONG S, 2001, INFORM TECHNOL LIBR...
TENOPIR C, 2000, ELECT J ...
HAWKINS DT, 2000, ONLINE ...
BUSH V, 1945, ATLANTIC MONTHLY ...
PARKHURST CA, 1990, LIBRARY PERSPECTIVES ...
NEGROPONTE N, 1995, BEING DIGITAL ...
TAUBES G, 1996, SCIENCE ...
ODLYZKO AM, 1995, INT J HUM-COMPUT ST
HARTER SP, 1996, PUBLIC ACCESS COMPUT ...

Figure 4: The lower part of the network in Figure 1 in detail

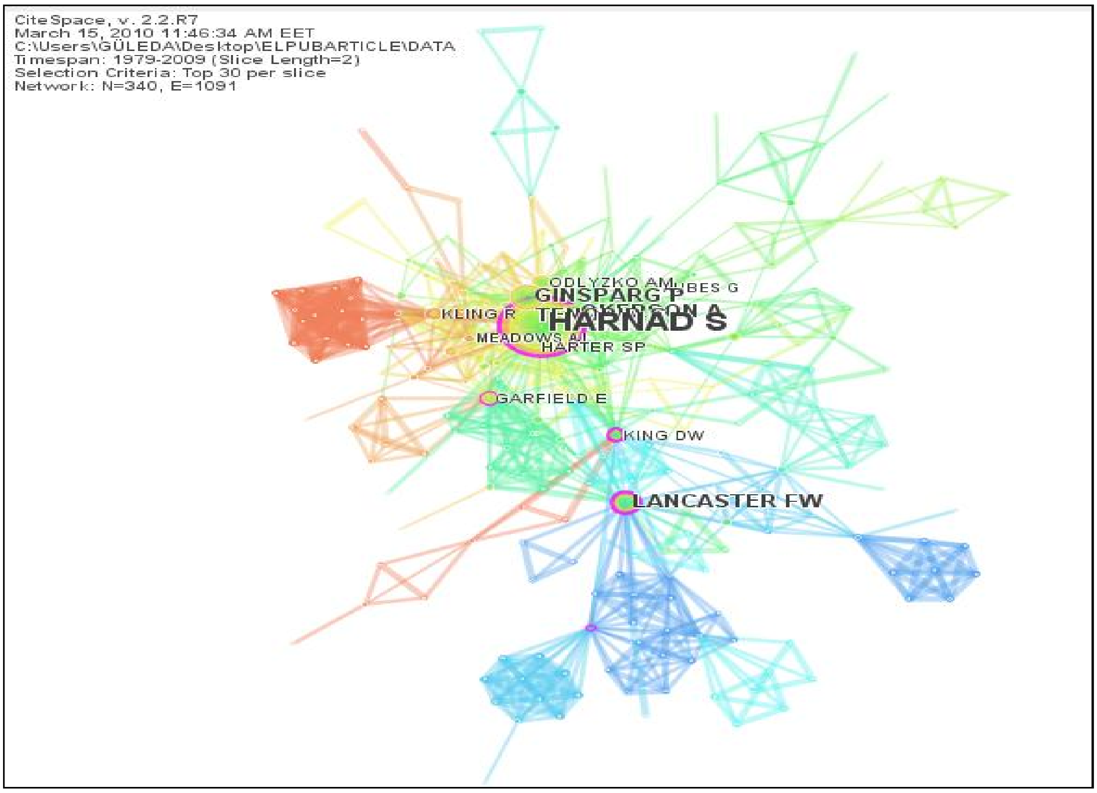

CiteSpace, v. 2.2.R7
March 15, 2010 11:46:34 AM EET
C:\Users\GÜLEDA\Desktop\ELPUBARTICLE\DATA
Timespan: 1979-2009 (Slice Length=2)
Selection Criteria: Top 30 per slice
Network: N=340, E=1091
ODLYZKO AM
GINSPARG P
KLING R
HARNAD S
MEADOWS AJ
HARTER SP
GARFIELD E
KING DW
LANCASTER FW

Figure 5: An author co-citation network (1979-2009), including 340 cited authors and 1091 co-citation links.

The size of a node is proportional to the number of e-publishing articles one has published. The colors of tree-rings indicate the temporal patterns of an author. For example, Harnad has the largest citation circle. On the other hand, the author co-citation map conveys additional information about how these authors have been cited. The nodes of Harnad and Lancaster have purple rings, indicating that they are pivotal nodes in the network with the highest betweenness centrality. In other words, they are strategically important in pulling other nodes together [20]. The same can be said, to a lesser extent, for the nodes of Garfield and King. The citation tree-ring of Harnad shows thick layers of green-orange rings, indicating that the majority of citations to Harnad were received in recent years (e.g., 2000s). The open access expert Stefan Harnad, the founder of arXiv Paul Ginsparg and Ann Okerson of Yale University Libraries are usually co-cited.

The prominent nodes are dominated by green citation rings (see Fig. 6). This pattern suggests that these authors frequently published e-publishing papers in the green time slices, which corresponds to the 1990s and first years of 2000s. The outermost authorship tree-rings of most of the authors are orange, suggesting that many of these authors continue to publish papers that continue to be cited. The names of those authors can be seen in Figure 7 along with linkages among them.

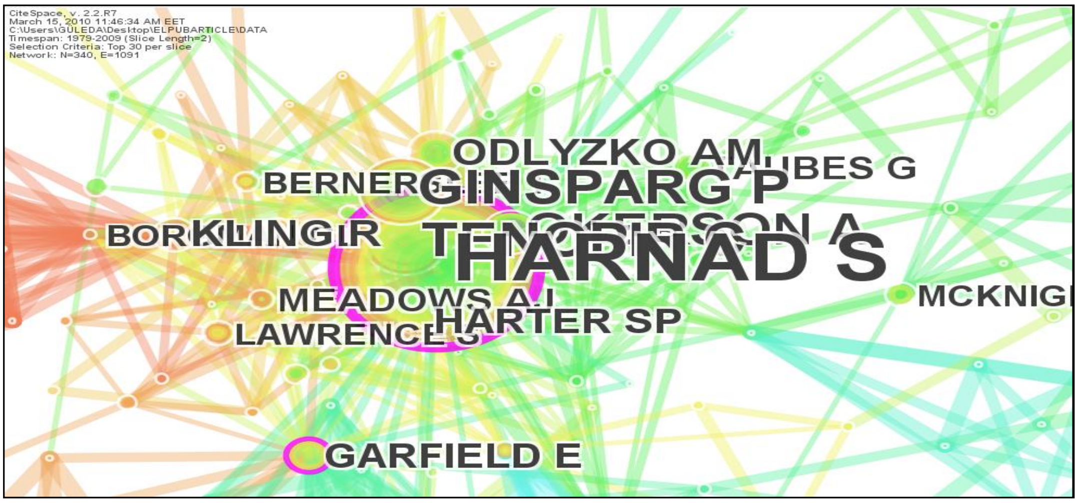


Fig. 6: The densest part of the author co-citation network (1979-2009) in Fig. 5.

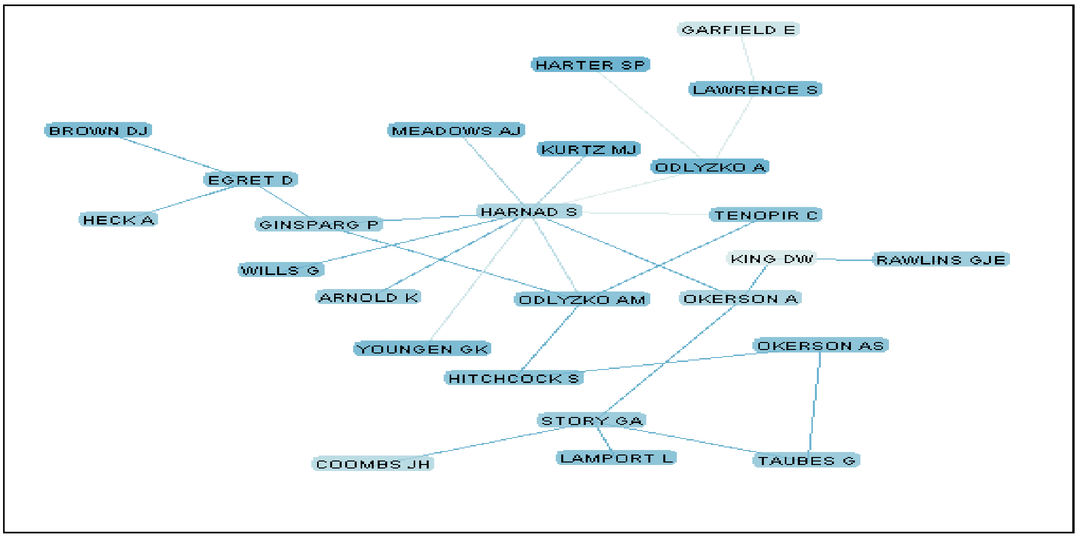


Fig. 7: The largest connected component of the e-publishing authorship network

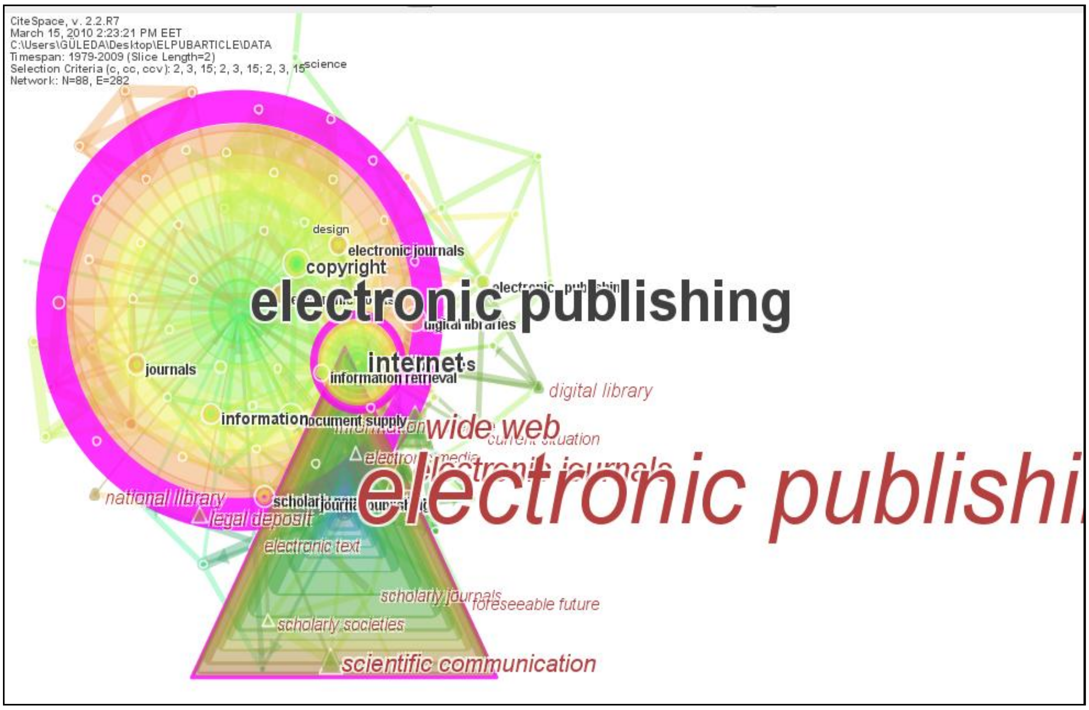


Fig. 8: A hybrid networks of keywords (shown as circles with black labels) and noun phrases (shown as triangles with dark red labels) (1979-2009)

Figure 8 shows a hybrid network of keywords as circles and noun phrases as triangles, extracted from titles and abstracts of papers. A noun phrase consists of a noun and adjective(s). Pivotal nodes are shown with purple rings (e.g., electronic publishing, internet).

Figure 9 draws a minimum spanning tree using the hybrid network of keywords and noun phrases in Figure 8. Keywords represent more general topics whereas noun phrases represent microscopic analysis. So, the hybrid map of keywords and noun phrases is expected to reveal concrete connections at different granularity levels [20]. The inclusion of the map is to provide an overall orientation of the conceptual structure of papers on e-publishing.

This map includes hubs of electronic publishing, internet and research funding. Internet and research funding are interconnected with electronic publishing. The hub of electronic publishing is connected to other keywords or noun phrases such as open access, copyright and electronic books; the hub of internet is connected to electronic books, information, information retrieval, information technology and so on. Concept maps can be useful to identify specific terms that are closely related with the field of e-publishing [20].

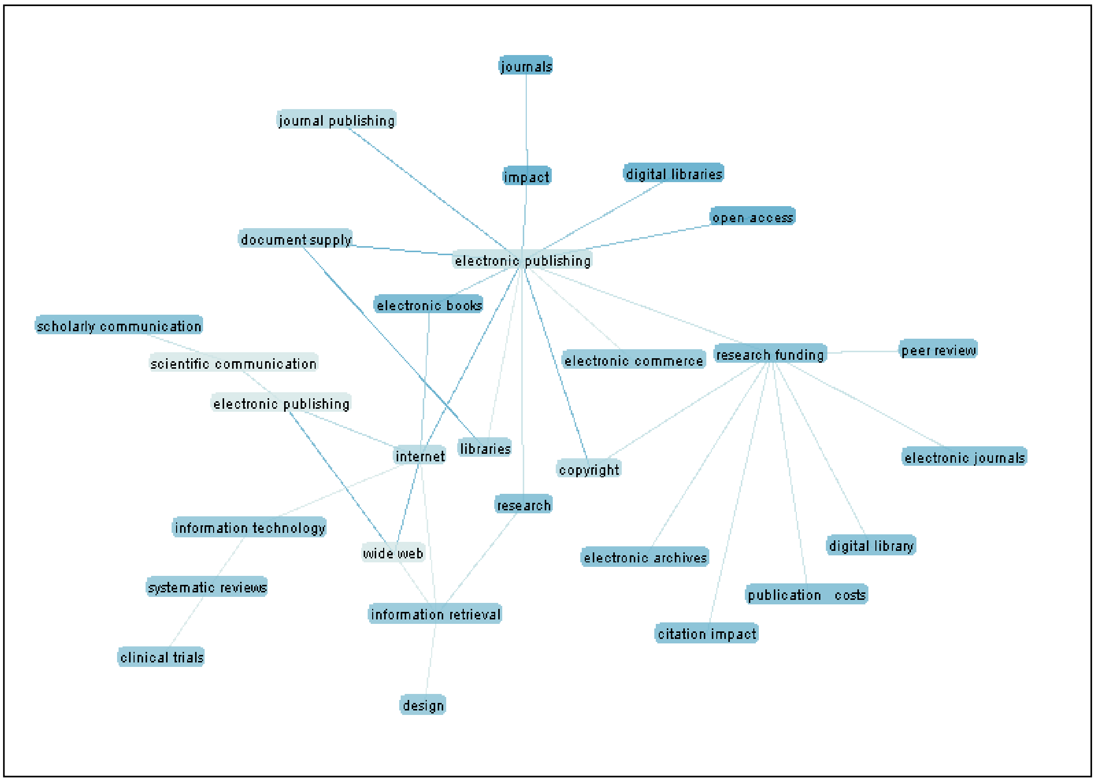


Fig. 9: A concept map of keywords assigned by authors to their own papers and noun phrases extracted from titles and abstracts of papers. Citespace thresholds: 3,3,15; 3,3,20; 3,3,20

4. Conclusion

We have analyzed the structure and evolution of electronic publishing field through articles published between 1979 and 2009 using co-citation networks derived from CiteSpace. Findings of our study show that e-publishing is an emerging research field. The three most prominent sources in e-publishing field are Tenopir and King's book ("Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers", 2000), Harnad's article ("Scholarly skywriting and the prepublication continuum of scientific inquiry", 1990) and Schauder's article ("Electronic Publishing of Professional Articles: Attitudes of Academics and Implications for the Scholarly Communication Industry", 1994). There is a recently formed part of the network that represents "open access". The open access evangelist Stefan Harnad seems to be the most influential author. Open access, e-journals, e-books, digital libraries are among the major research tracks in e-publishing as indicated in the hybrid map of keywords and noun phrases. Findings of this study can be used to identify landmark papers along with their impact in terms of providing different perspectives and engendering new research areas.

Notes and References

[1] JIANHUA, H; et al. The information visualization analysis of the study in international S&T policy. Proceedings of the Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting, Berlin, 2008. Available at http://www.collnet.de/Berlin-2008/HouJianhuaWIS2008iva.pdf (January 2010)

[2] ESTABROOKS, CA; et al. The intellectual structure and substance of knowledge utilization field: A longitudinal author co-citations analysis, 1945 to 2004. Implementation Science, 3, 2008, p. 49. Available at http://www.implementationscience.com/content/3/1/49 (January 2010)

[3] MOED, HF. *Citation analysis in research evaluation*. Netherlands: Springer, 2005

[4] CHEN, C. Searching for intellectual turning points: Progressive knowledge domain visualization. Proceedings of the National Academy of Sciences of the USA (PNAS), 101 (Suppl.(1)), 2004, p. 5303-5310. Available at http://www.pnas.org.content/101/suppl.1/ 5303.full.pdf (January 2010).

[5] EGGHE, L; ROUSSEAU, R. *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier Science Publishers, 1990. p. 239. Available at http://uhdspace.uhasselt.be/dspace/handle/1942/587 (April 2010).

[6] WHITE, HD; MCCAIN, KW. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. Journal of the American Society for Information Science, 49, 1998, p. 327-355. p. 327.

[7] HE, Q. Knowledge discovery through co-word analysis. Library Trends, 48 (1): 133-159, 1999. p. 133.

[8] ÅSTRÖM, F. Visualizing Library and Information Science concept spaces through keyword and citation based maps and clusters. In: Bruce, Fidel, Ingwersen & Vakkari (Eds). Emerging frameworks and methods: Proceedings of the fourth international conference on conceptions of Library and Information Science (CoLIS4), 2002, p. 185-197. Greenwood Village: Libraries Unlimited. (as cited in http://www.db.dk/bh/Core%20Concepts%20in%20LIS/articles%20a-z/coword_analysis.htm).

[9] CHEN, C; et al. Visual analysis of scientific discoveries and knowledge diffusion. Proceedings of the 12th International Conference on

Scientometrics and Informetrics (ISSI 2009). July 14-17, 2009. Rio de Janeiro, Brazil. Available at http://cluster.cis.drexel.edu/~cchen/papers/2009/issi2009/issi2009.pdf (January 2010).

[10] PETERSON, I. Mapping scientific frontiers. Science News Online, 165 (11), 2004. Available at http://cluster.cis.drexel.edu/~cchen/citespace/doc/mathtrek.pdf (January 2010).

[11] CHEN, C; et al. Making sense of the evolution of a scientific domain: A visual analytic study of the Sloan Digital Sky Survey research. Scientometrics, in press (DOI 10.1007/s11192-009-0123-x). Available at http://www.springerlink.com/content/46661328402643l0/fulltext.pdf (April 2010).

[12] CHEN, C; et al. Towards an explanatory and computational theory of scientifiv discovery. Journal of Informetrics, 3 (3), 2009, p. 191-209. Available at http://arxiv.org/ftp/arxiv/papers/0904/0904.1439.pdf (April 2010).

[13] CHEN, C; et al. The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. Journal of the American Society for Information Science and Technology, in press (DOI: 10.1002/asi.21309). Available at http://www3.interscience.wiley.com/cgi-bin/fulltext/123324662/PDFSTART (April 2010).

[14] OTTE, E; ROUSSEAU, R. Social network analysis: A powerful strategy, also for the information sciences. Journal of Information Science, 28, 2002, p. 443–455.

[15] SCOTT, J. *Social network analysis: A handbook* (second ed.). Thousand Oaks, CA: Sage Publications, 2000.

[16] CHEN, C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57 (3), 2006, p. 359-377. Available at http://cluster.cis.drexel.edu/~cchen/citespace/doc/jasist2006.pdf (January 2010).

[17] Sentinel Visualizer. Social network analysis (SNA), 2009. Available at http://www.fmsasg.com/SocialNetworkAnalysis/ (January 2010).

[18] DELL, H. Mapping intellectual milestones. BioMedNet: Special Report. Available at: http://cluster.cis.drexel.edu/~cchen/citespace/doc/biomednet.pdf (January 2010)

[19] One of the reviewers commented that papers published in proceedings of ELPUB conferences should be included in our sample. However, this was not possible in that metadata, references and citations of ELPUB papers were not readily available and they would have to be pre-processed in order for them to be entered into CiteSpace software.

[20] CHEN, C; et al. The thematic and citation landscape of Data and Knowledge Engineering (1985-2007). Data and Knowledge Engineering, 67, 2008, p. 234-259.

[21] Details for all papers in Figures 1-4 are given in Appendix.

Appendix: Papers depicted in the network clusters and mentioned in the text

Antelman, K. (2004). Do open access articles have a greater research impact? College and Research Libraries, 65 (5): 372-382.

Bush, V. (1945). As we may think. The Atlantic Monthly, 176 (1): 101-108.

Crawford, W. (2000). Nine Models, One Name: Untangling the e-book Muddle. American Libraries, 31, 56-9.

Harnad, S. (1990) . Scholarly skywriting and the prepublication continuum of scientific inquiry. Psychological Science, 1, 342–344.

Jones, R., Andrew, T., & MacColl, J. (2006). The Institutional Repository. Oxford: Chandos.

Ginsparg, P. (1994). First Steps Towards Electronic Research Communication. Computers in Physics, 8 (4): 390-396.

Hawkins, D.T. (2000). Electronic books: a major publishing revolution (part 1). Online, 24 (4): 14-28.

Kurtz, M.J..et al. (2005). The effect of use and access on citations. Information Processing & Management, 41, 1395-1402.

Lancaster, F.W. (1978). Toward Paperless Information Systems. Orlando, FL: Academic Press.

Lawrence, S. (2001). Online or invisible? Nature, 411 (6837): 521.

Miller C.T., & Harris, J. C. (2004). Scholarly Journal Publication: Conflicting Agendas for Scholars, Publishers, and Institutions. Journal of Scholarly Publishing, 35 (2): 73-91.

Negroponte, N. (1995). Being Digital. New York: Alfred A. Knopf.

Odlyzko, A.M. (1995). Tragic loss or good riddance? The impending demise of traditional scholarly journals. International Journal of Human-Computer Studies, 42, 71–122.

Ormes, S. (2001). An e-book primer, available at: www.ukoln.ac.uk/public/earl/issuepapers/ebook.htm

Rawlins, G.J.E. (1993). Publishing over the Next Decade. Journal of the American Society for Information Science, 44 (8): 474-479.

Schauder, D. (1994). Electronic Publishing of Professional Articles: Attitudes of Academics and Implications for the Scholarly Communication Industry. Journal of the American Society for Information Science, 45, 73-100.

Sottong, S. (2001). E-book technology: Waiting for the false pretender. Information Technology and Libraries, 20 (2): 72-80.

Swan, A. (2005) Open access self-archiving: An Introduction. Technical Report UNSPECIFIED, JISC, HEFCE.

Tenopir, C., & King, D.W. (2000). *Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers*. Washington, D.C.: Special Libraries Association.

Turoff, M., & Hiltz, S.R. (1982). The Electronic Journal: A Progress Report. Journal of the American Society for Information Science, 33 (4): 195-202.

Electronically published scientific information in technical university libraries

Kate-Riin Kont

Tallinn University of Technology Library: Acquisitions Department
Akadeemia 1, 12618 Tallinn, Estonia
kont@lib.ttu.ee

Abstract

The use of electronic information resources is growing rapidly. The actual science information is electronic as a rule - practically all the journals of engineering and natural science have electronic versions and a certain number of them are available only electronically. Electronic scientific information in technical universities is the basis for research and development, degree study and professional specialty, to a certain extent. It is widely agreed by producers and purchasers of information that the use of electronic resources should be measured in a more consistent way. Librarians want to understand better how the information they buy from a variety of sources is being used; publishers want to know how the information products they disseminate are being accessed. Findigs of this study suggest that the financial opportunities of technical university libraries in the four neighboring countries - Swedish Royal Institute of Technology, Helsinki University of Technology, Tallinn University of Technology Library, and Scientific Library of Riga Technical University (henceforth referred to as KTHL, HUTL, TUTL and RTUL respectively) - to spend resources on electronic publications are very different.

Keywords: university libraries; digital libraries; electronic scholarly communication; library services; performance measurement.

1. Introduction

Libraries in the Nordic European countries (Denmark, Finland, Iceland, Norway and Sweden) began the process of digital library building in 1980s with the implementation of computerized library catalogues [1]. The main purpose of the *Nordic Council for Scientific Information and Research Libraries*

(NORDINFO), founded in 1977 (closed in 2004) was to promote Nordic cooperation within the field of scientific information and documentation, principally in connection with the research library systems.

The Nordic Electronic Research Library is a concept which is based on national developments within the research library sector in each of the five Nordic countries. The goals of the Nordic Electronic Research Library are to make scientific and technical information easily available in all the Nordic region [2].

In contrast to the situation in the Nordic countries, the development of the information system of Eastern Europe libraries is weakly included in the national program for the development of the information society. According to Virkus [3], academic libraries in the former communist countries of Baltic states have experienced a period of rapid and profound change during the last decade, in connection with the transformation in the political and economic structures, changes in territorial and administrative situations, as well as with the rapid development of information and communication technologies.

The purpose of the present poster is to analyze the essential data, details of the use of e-resources and the cost of electronic scientific information as well as the cost of the most important performance indicators related to the increasing usage and acquisition of electronic scientific information of the leading technical university libraries in Sweden, Finland, Estonia and Latvia. These university libraries are also members of the IATUL (International Association of Scientific and Technological University Libraries). The choice of the period 2004-2008 is justified by the fact that during that time the libraries underwent a substantial increase in e-services as well as in expenditures on electronic scientific information.

2. Methodology

The data used in this paper is based on the analysis of relevant literature. The details of the size, cost and usage of the collections of university libraries, based on the annual reports of these libraries (in the case of the HUTL and TUTL) as well as on the questionnaires sent to directors of libraries (in case of the KTHL and RTUL), are analyzed.

3. Findings

The numbers in Figure 1 indicate that students constitute only one part of the readership of the technical university libraries, while a considerable part of the readership (for example ca 30% in TUTL) is formed by other target groups (lecturers, scientists and other interested groups). Therefore, the role of technical university libraries is much broader, offering services to different users.

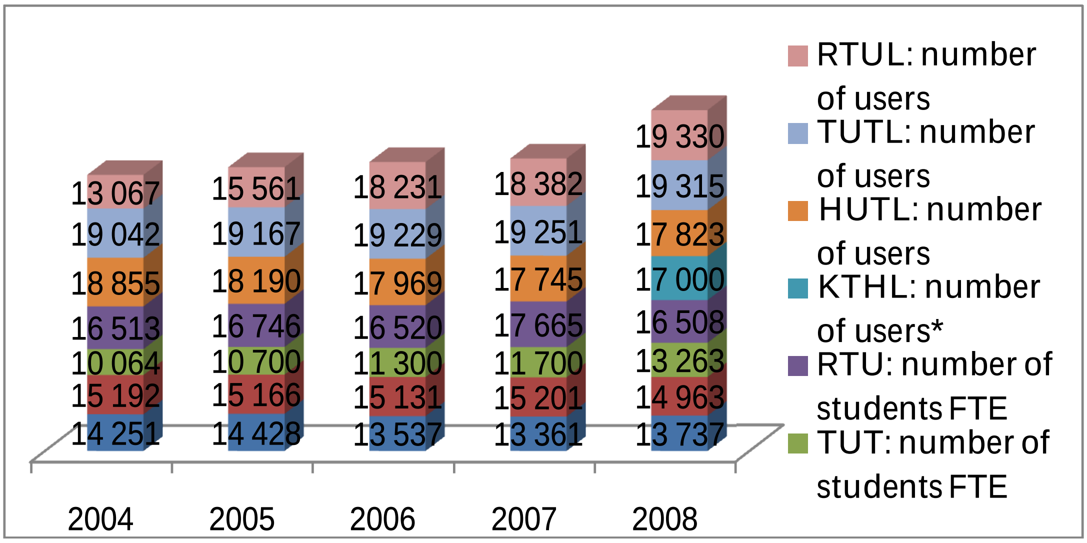


Figure 1: The number of students FTE/the number of registered users

Data given in Figure 2 shows that the number of licensed databases has grown during the last three years in all libraries. Since 2006, HUTL began to distinguish between 66 separate CSA (Cambridge Scientific Abstracs) database, hence the sudden increase in the number of databases. The number of licensed databases in RTUL was not calculated until 2006, but the number of databases is still extremely small compared to other libraries.

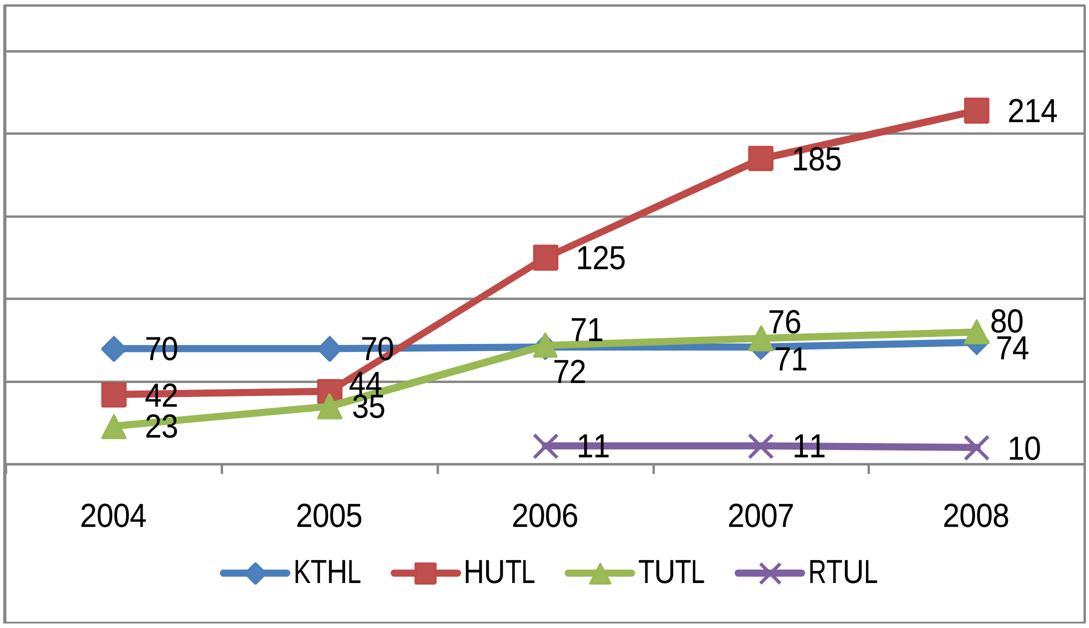


Figure 2: Electronic collection: number of licensed databases

Table 1 lists the comparison of collections on physical carriers (includes for example books and periodicals on paper, but also CDs, VHSs, DVDs etc.) and electronic collections (number of the titles of e-books and e-periodicals) of technical university libraries. These data are collected according to the standard the ISO 2798:2006 [4]. In 2004-2008, the number of collections on physical carriers was stable in all libraries. Of the considerably increased number of e-publications in HUT Library in 2006, around 240,000 are various digitized historical books from different disciplines published in the UK and US in the 15th to18th centuries and made available via different databases. Unfortunately it is not possible to compare side by side the number of electronic collections between libraries in 2004-2008. The reason for this is that RTUL does not reflect these numbers in statistics.

Table 1. Collections on physical carriers / electronic collection: the number of e-publication titles (e-books + e-periodicals)

Library	2004	2005	2006	2007	2008
KTHL	864 661/ 626 274	865 723/ 633 412	877 572/ 638 348	833 379/ 638 667	837 770/ 665 392
HUTL	237 087/ 10 999	241 104/ 13 068	240 800/ 260 228	240 875/ 311 547	234 894/ 326 151
TUTL	718 536/ 31 000	723 136/ 37 800	723 906/ 43 800	733 4867/ 55 000	723 630/ 69 474
RTUL	2 333 910/*	2 301 858/*	2 205 044/*	2 084 972/*	1 961 419/*

*Records were not considered

Table 2 compares traditional loans and downloaded electronic content units in the technical university libraries. Home lending, on-site-loans, loans through the self-rental machine and renewals (but not in-house usage) are taken into consideration in the case of traditional library loans. *Content downloaded* is defined as content unit (full-text article or abstract), that is successfully requested from a database, electronic serial or library digital collection.

Table 2. The usage of the collections: traditional library usage: loans/ electronic library usage: content units downloaded

Library	2004	2005	2006	2007	2008
KTHL	107 563/ 482 729	105 953/ 728 787	96 305/ 720 443	84 031/ 772 317	82 140/ 914 318
HUTL	181 557/ 245 046	253 264/ 255 642	271 545/ 264 895	256 447/ 291 849	241 760/ 353 627
TUTL	183 246/ 136 244	193 497/ 418 538	193 518/ 545 804	193 960/ 684 623	193 545/ 436 788
RTUL	752 243/*	756 730/*	670 780/ 95 000	630 261/ 74 213	352 680/ 229 754

*Records were not considered

Figure 2 indicates a big difference in the acquisitions costs of the libraries, due to which the libraries have very different financial means to spend on electronic publications, unfortunately to the disadvantage of TUTL and RTUL.

The proportion of the expense of e-documents in the acquisition costs is considered an important performance indicator, which is included in official statistics since 2006, but has been recorded by libraries even earlier. The spending on electronic collections – purchased access to databases and acquired licenses – has been the largest in KTHL - 69% of acquisition costs in 2004, 73% in 2005, 72% in 2006, 78% in 2007 and 89% in 2008, followed by HUTL - 80% of acquisition costs in 2004, 84% in 2005, 87% in 2006, 88% in 2007 and 90% in 2008. In TUTL, the spending on electronic collections increased from 32% of acquisition costs in 2004, to 36% in 2005, 39% in 2006, 54% in 2007 and dropped to 36% in 2008, while in RTUL the spending on electronic collections has not increased, being 15% of acquisition costs in 2004, 6% in 2005, 17% in 2006, 14% in 2007 and 13% in 2008.

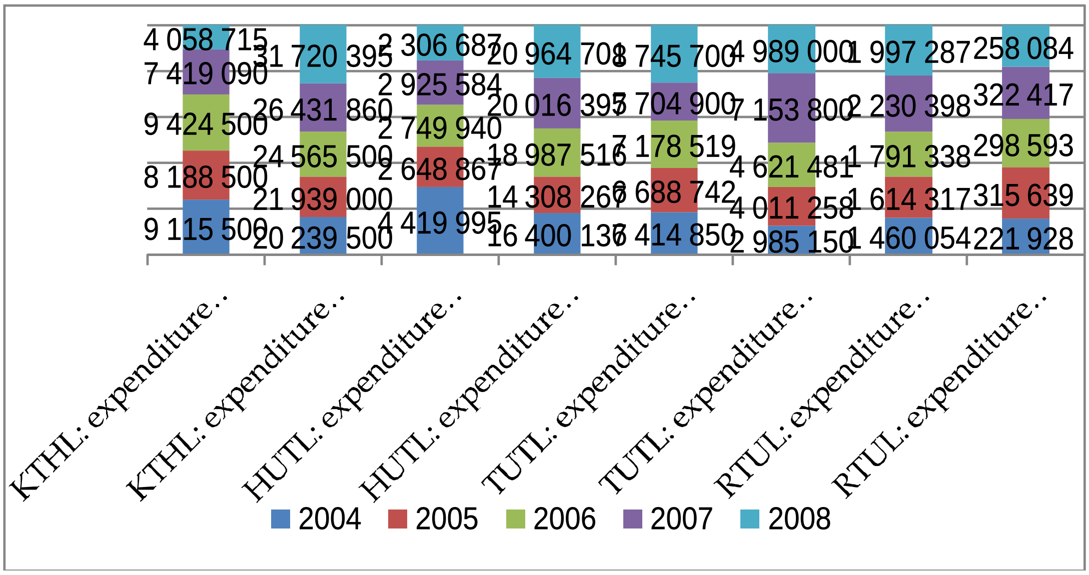


Figure 2. Acquisition costs: expenditure on print materials / expenditure on electronic materials (EEK) [1 EUR=15,6466 EEK]

The two most interested stakeholder groups in the case of university libraries are the population the library is set up to serve and the institution to which it belongs. The institution, especially if it provides funding, will see university library quality on another scale i.e., the library is good if it helps to shorten studying time, produces graduates that quickly find a job, supports research in an effective way, helps to raise the image of the institution, and if it is cost-effective overall. The last issue will often be the most important when resources are scarce [5]. To measure this, university libraries are using a performance indicator given in Table 4 – acquisition costs per student.

Table 4. Acquisition costs per student: expenditure on print materials/ expenditure on purchased access to databases, e-publications (EEK).

Library	2004	2005	2006	2007	2008
KTHL	640/1420	568/1521	696/1815	555/1978	295/2309
HUTL	29/1079	175/942	182/1255	192/1317	154/1401
TUTL	637/297	625/375	624/402	488/611	659/376
RTUL	88/13	96/57	108/18	126/18	121/16

Note: The values are calculated as follows: expenditure on print materials, expenditure on purchase of e-documents, databases / number of students.

Acquisition costs of electronic publications per student have steadily increased in KTHL and HUTL since 2006. In TUTL this cost was the highest in 2006 (402 EEK) and in 2007 (611 EEK). Acquisition costs of electronic publications per student in RTUL have been very low as well as acquisition costs of print materials per student per year.

A number of cost indicators in library work are based on the relationship between a certain statistical indicator and the operating expenditures of the library [6].

Table 5. Cost per loan and cost per contents downloaded

Library	2004	2005	2006	2007	2008
KTHL	932/ 208	919/ 134	1017/ 136	1064/ 116	1075/ 94
HUTL	376/ 279	256/ 253	260/ 267	296/ 260	330/ 226
TUTL	93/ 131	101/ 47	114/ 41	138/ 39	150/ 67
RTUL	9/*	12/*	16/ 110	26/ 218	58/ 89

*Number of content units downloaded was not recorded. The values are calculated as follows: operating expenditure/number of loans, number of contents unit downloaded

The objective of the indicator of the cost per traditional loan is to establish a relation between the number of loans and the cost of providing all services of the library, based on this can be estimated the overall efficiency of the service, especially in the university libraries, where loans are the dominant service. The objective of the indicator cost per content unit downloaded is to assess the contractual cost of an electronic resource related to the number of content units downloaded. A lower value indicates cost efficiency for electronic resources [5]. In addition, the cost indicators of regular loans have become considerably more expensive throughout years when compared to the usage of the electronic library.

4. Conclusions and Discussion

Since Estonia and Latvia joined the EU in May 1, 2004 – all foreign electronic publishers changed their pricing policy towards our countries. The Baltic region is no longer a region of transition and therefore many current benefits (for example discounts preferences for developing countries) have disappeared. The usage license fees for electronic resources and prices of printed books and journals continue to rise. However, expanding the choice of electronic scientific information in the Baltic countries cannot be done without additional financing at the national level.

The analysis of the most important cost indicators shows that the main cost indicator of the electronic library - the cost of the downloaded e-content unit– has become cheaper than traditional loans to the library, which affirms that the costs on the electronic library of the university library –e-resources, are well worth making due to smaller cost indicators.

Perhaps a future suggestion would be the establishment of a consortium of the libraries of universities of technology in the Nordic and Baltic countries. The need for certain specific and expensive databases would well justify that wish.

Acknowledgements

This study was supported by European Social Fund's Doctoral Studies and Internationalisation Programme DoRa.

References

[1] FENDIN, M-L. Digital libraries in the Nordic countries: with practical examples for the creation and development of 'libraries without walls' from the Nordic Africa Institute Library and other libraries within the Nordic countries. Paper presented in Conference of Electronic Publishing and Dissemination, from 1st to 2nd September in Africa, Senegal, 2004. Available at: http://www.codesria.org/Links/conferences/el_publ/fendin.pdf [December 2009].

[2] HANNESDOTTIR, S.K. The Nordic Electronic Research Library in different dimensions. *Library Consortium Management: An International Journal*. Vol. 2, no.5/6, 2001, p. 122-131.

[3] VIRKUS, S. Academic Libraries in Eastern Europe. In: Encyclopedia of Library and Information Science. London: Taylor & Francis, 2005, p. 11-21.

[4] ISO 2789:2006 *Information and documentation. International library statistics.* International Organization for Standardization.

[5] POLL, R. Performance, Processes and Costs: Managing Service Quality with the Balanced Scorecard. *Library Trends*, 2001, Vol. 49, no. 4, p. 709-717.

[6] ISO 11620:2008 *Information and documentation*. *Library performance indicators*. International Organization for Standardization.

Exploratory study of quality control mechanism for academic papers in the Internet era: A case study of Sciencepaper Online in China

Cantao Zhong [1]*; Meng Wan* [2]

1 Institute of Advanced Technology, Peking University
Beijing, 100871, China
ctzhong@pku.edu.cn;
2 Centre for Sci-tech Development of the Ministry of Education
Beijing, 100080, China
wanmeng@cutech.edu.cn

Abstract

Information and communication technologies such as the Internet can both challenge traditional ways and open opportunities for solving existent problems of present academic quality assurance system. Sciencepaper Online in China (CSPO) has adopted a sophisticated mechanism for quality controlling, which can be represented by "Publish Online First, Author Selected Peer-Review Later". Using a five-star rating system for quality labelling, each reviewed paper will be assigned a one to five star grade label, corresponding with Poor, Fair, Good, Very Good, and Excellent. With this system, CSPO innovatively solves the conflict between rapid publication and quality assurance. This paper investigates this unique quality mechanism with the aim to understand its operation more thoroughly and evaluate its value to the scientific communication community.

Keywords: Sciencepaper Online; quality labelling; quality control; academic paper

1. General Background

New technologies such as the Internet enable new publication models for academic papers. The ways scientists share and use research results are changing rapidly, fundamentally and irreversibly [1]. Information and communication technologies can both challenge traditional ways and open remedies for existent problems of present academic quality assurance system. New forms of ex-ante and of ex-post quality control may partly replace and partly amend peer review. Open peer review, online commenting, rating, access counts and use tracking are also potential contributors [2].

Current development of ’open’ movements, including Open-access, Open-data, and Open-science has evolved from only coping with the serials crisis into reflection and re-engineering of the entire scholarly communication processes. Scholarly publishing mainly comprises four functions: registration to establish intellectual priority, certification to certify the quality/validity of the research, awareness to assure accessibility of research and archiving to preserve research for future use. Convergence of technologies enables new business models for scholarly communication. A variety of business models can be explored with the four core functions disentangled or recombined [3].

Traditional academic journals are switching or have switched to Internet platforms to facilitate the reviewing and editing process, thus to shorten the time for papers to reach their readers. But, fundamentally, most online journals are simply digital editions of their print analogs [1]. They still use the traditional subscription-based business model without overcoming all the inherent access-limiting drawbacks of traditional journals.

Golden OA journals eliminate the access barrier. But most of them still use traditional forms of peer review. Some have begun to use innovative new forms that take advantage of the new medium and the interactive network joining scholars to one another. One example is PLoS’s light-touch peer review. The truly radical thing about PLoS ONE is that it has redefined the nature of peer review. Using this ‘light touch’ refereeing process, the only criterion for publication is that a paper is methodologically sound. So the time for a paper to appear on the web or PubMed repositories will be shortened into two to four weeks. Although there are many debates about this light-touch refereeing method, PLoS ONE has made a success with this disruptive business model. Peter Suber said that ‘removing access barriers and reforming peer review are independent projects’ [4], which is right in concept level. However, the case of PLoS ONE indicates that they can be combined to bring more benefits to the academic arena. Leo Waaijers even suggests in Ariadne that funders should fund research on alternative "non-proprietary peer review" services [5].

2. Introduction of Sciencepaper Online

There are also pioneers in the scientific communication area in China, and Sciencepaper Online (CSPO) is the most aggressive one, which integrates the concepts and implementations of preprint repository, open access journals, and coalition of institutional repositories into one platform. CSPO is sponsored and operated by the Center for Science and Technology Development (CSTD) of the Ministry of Education (MOE) with strong financial support from the government.

CSPO offers an online publication platform for new and innovative ideas to the academic community with the aim to facilitate fast exchanges and instant adoption of new academic achievements. It also serves as an important platform for the optimization of academic environments and for the improvement of academic behaviors [6].

CSPO will accept and publish any papers that meet its basic format requirements for online publication in one week after the submission, no publishing fee needed. It bypasses traditional publication procedures such as ex-ante peer reviews, revisions, editing and printing. CSPO does not hold any copyright of any paper, the copyright still belongs to the author, and it allows and encourages authors to submit the paper to other professional journals.

CSPO has begun operation formally since August, 2003. By the end of March 2010, it has more than 200,000 registered (which is free) users, with daily IP visits about 10,000. The number of total papers published is about 41,000, and this number increases steadily about 1,000 per month. (See Fig. 1)

The operator, CSTD, also plays as a funder and a government agency, which greatly helps the operation of CSPO. For example, it requires that all projects funded by the Doctoral Program of MOE must publish 1-2 original papers on Sciencepaper Online. In addition, as a funder and government agency, it has an expert database, including all the doctoral advisors in China. This provides strong support for its peer review and quality control system.

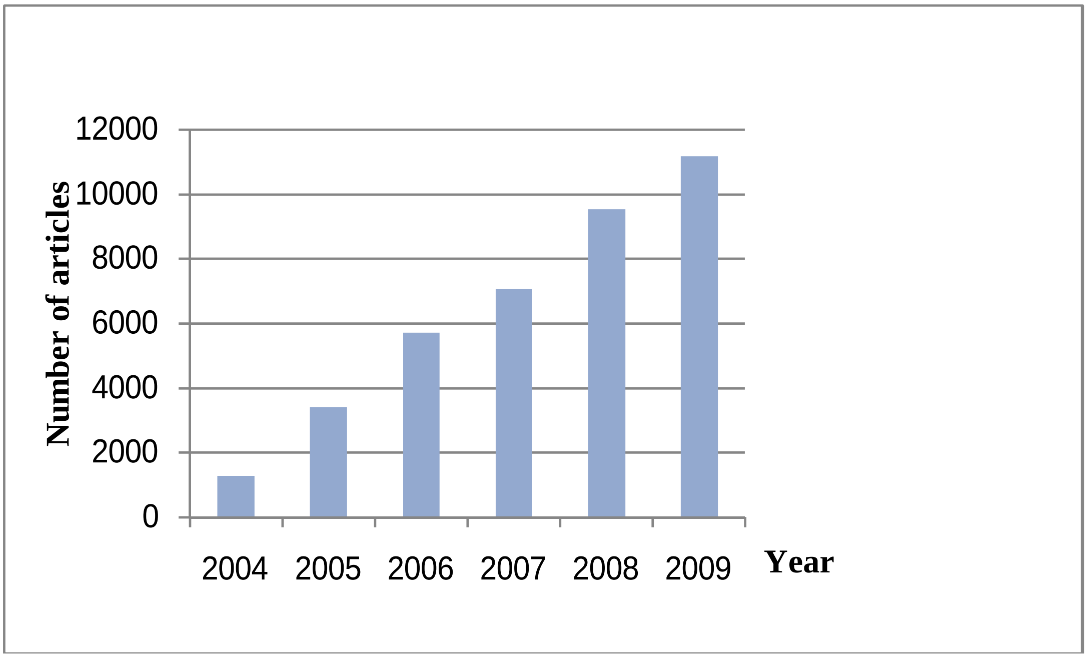


Figure 1: Annual published articles on Sciencepaper Online

3. Unique quality mechanism

Sciencepaper Online offers a fast and real-time platform for exchanging new academic ideas and disseminating new academic achievements. In this aspect, Sciencepaper Online acts much like a preprint repository, or archive. In addition, CSPO also provides an opportunity to publish scientific content that would not be accepted in a traditional journal.

However, quality control mechanism is essential for academic papers to have real value to the scientific community. So CSPO adopts a sophisticated mechanism for quality controlling, which can be represented by "Publish Online First, Author Selected Peer-Review Later" since Oct. 2005.

When a paper is submitted to the site, its author can select whether they want formal peer review services (free). The paper, waiting for peer review upon its author's request, will appear first on the site with a "waiting for review label" along its side. Reviewers will evaluate the paper around several aspects, including Title, Chinese and English abstracts, scientific innovativity and originality, rationality of the research plan, methodology of data processing and bibliometric references (See Table 1).

Table 1: Review Criteria

Criteria	Options				Note
Title	Very Good	Good	Fair	Poor	Appropriate with content
Abstract (Chinese)	Very Good	Good	Fair	Poor	Terse and Concise
Abstract (English)	Very Good	Good	Fair	Poor	Terse and Concise
Scientific Innovativitiy	Very High	High	Some	No	For review articles, this will be academic value.
Research Plan	Very Good	Good	Fair	Poor	For review articles, this will be coverage of references.
Data processing and Reasoning	Very Good	Good	Fair	Poor	For review articles, this will be logic deduction and proof.
Written expression	Very Good	Good	Fair	Poor	Clear and formal
References	Very Good	Good	Fair	Poor	Comprehensive and concise
Overall Review Suggestion	Suggest to pub. with priority	Agree to pub.	Need minor mod.	Much mod. needed	Not pub.
Specific suggestions	(No less than 50 words)				

mod. = modification pub.=publish

CSPO uses a five-star rating system for quality labelling. According the final comprehensive peer evaluation results generated from the reviewing process, each reviewed paper will be assigned a one to five star grade label, corresponding to Poor, Fair, Good, Very Good, and Excellent, respectively. The final reviewing commentary will also appear alongside the paper without the name of the reviewer to ensure that reviewers can criticize openly without the danger that the author would know the originator and be resentful [2]. Furthermore, CSPO also permits registered users to comment on all published papers, encouraging academic criticism and discussion with the aim to evaluate the real academic value of a paper more objectively.

4. Data analysis and discussion

The "Author-selected peer-review after publishing online" method started from October 2005 as an effort to make papers published on CSPO to be accepted by the academic community. Appropriately-arranged peer-review can encourage academic communication, including critics and discussions around sumitted papers. By the end of March 2010, CSPO has published about 41,000 originated articles, among which about 88 percent has selected peer-review. The high ratio reflects the fact that most authors are serious when submitting their articles to CSPO. They want to demonstrate academic values of their articles.

For all peer-reviewed articles, the proportion of different star levels, are 13, 34, 19, 18, and 16 percent for one to five stars, respectively. Because only articles with 3 stars or higher level have the the opportunity to be recognized as eligible academic papers, so we can consider the rejection rate of CSPO is about 36 percent, which is not very high. Now, only 35 universities and/or research institutes consider papers published on CSPO as eligible academic papers for tenure, promotion, and/or graduation purpose. So, only a little fraction of high-rating papers can be formally treated and brought actual impact to their authors.

CSTD, the operator of CSPO, started to publish a traditional journal *Sciencepaper Online* (ISSN 1673-7180) since August 2006. The *Sciencepaper Online* journal is different from CSPO, but those best articles on CSPO, will be arranged to be re-published in this journal with priority. Since last year, CSTD started another journal, the *Sciencepaper Online Collections*, which only selects excellent articles from CSPO.

Due to time and other constraints, complete citation data for CSPO is not obtained. However, we found citation data for *Sciencepaper Online* journal from one of the most widely used database –China National Knowledge Infrastructure (CNKI). There are only 74 articles with non-zero citations for this journal, and the largest number of citations for a single article is only 5.

These data indicate that the quality of articles published on *Sciencepaper Online* journal and on CSPO site need to be improved. Of course, the lack of an OAI-compliant interface may also be one of the reasons for its low impact, because readers or academics can't use popular search engines, i.e., Google Scholar, to find most articles on CSPO.

As said above, China Sciencepaper Online is sponsored and operated by a government affiliate, the Center for Science and Technology Development (CSTD) of the Ministry of Education (MOE). Because its inherent non-commercial and non-profit nature, CSPO does not charge authors for publishing, and even pay a little fee to reviewers. This makes it very different from other new publishers in the western world, such as PLoS. For example, PLoS ONE uses light-touch peer-review, mainly to cope with its sustainability problem, in other words, to attract more authors to publish in it and earn more author-paid fees.

Another reason for its low impact may arise from its wide discipline coverage. Academics usually read a few journals focusing on their disciplines, or use search engines for initial literature investigation in their study. Coverage being too-wide and the lack of indexing by search engines means fewer readers and users, resulting in low academic impact.

5. Conclusion

Through its “Publish Online First, Author Selected Peer-Review Later” method and five-star quality labelling system, Sciencepaper Online innovatively solves the conflict between rapid publication and quality assurance. Since quality control is so important for academic papers, further in-depth investigation about this unique quality mechanism and its long-term impact will tell more about the nature and changes of scientific communication in the Internet era.

Notes and References

[1] R. K. Johnson, "Will Research Sharing Keep Pace with the Internet?," *J. Neurosci.,* vol. 26, 2006, pp. 9349-9351, September 13.

[2] M. Nentwich, "Quality control in academic publishing: challenges in the age of cyberscience," *Poiesis & Praxis: International Journal of Technology Assessment and Ethics of Science,* vol. 3, 2005, pp. 181-198.

[3] R. Crow, "The Case for Institutional Repositories: A SPARC Position Paper," ed. Washington, D.C.: Scholarly Publication and Academic Resources Coalition, 2006.

[4] P. Suber. (2007, 1-28). *Open Access Overview*. Available: http://www.earlham.edu/~peters/fos/overview.htm (March 2010)

[5] L. Waaijers, "Publish and cherish with non-proprietary peer review Systems," *Ariadne,* no. 59, 2009. Available: http://www.ariadne.ac.uk/issue59/waaijers/ (April 2010)

[6] Sciencepaper Online. (2003, 1-28). *Introduction about Sciencepaper Online*. Available: http://www.paper.edu.cn/en/aboutus_zaixian.php (March 2010)

Sophie 2.0 - a platform for reading and writing of interactive multimedia books in a networked environment

Kalin Georgiev[1]*; Miloslav Sredkov*[2]

1 Faculty of Mathematics and Informatics, Sofia University,
5 J. Boucher str., Sofia, Bulgaria
kalin@fmi.uni-sofia.bg
2 Astea Solutios AD, 20 Lozentz str., Sofia, Bulgaria
milo@asteasolutions.com

Abstract

Sophie is software for reading and writing networked multimedia electronic books. It has been designed to let those without professional design skills create and distribute complex documents that take full advantage of new media and the Internet. Sophie brings together all of the pieces of media-rich writing. In addition, Sophie fosters collaboration, allows instant reader feedback, and encourages interactivity. Sophie lets users create communities around projects; with Sophie, "books" become "places" where people meet. In addition to its powerful capabilities for combining various media formats and interactivity, Sophie Server, a significant part of the Sophie platform, allows authors to collaborate – working on the same content simultaneously in real time or offline, and later integrate their changes with the work of others when an Internet connection becomes available. Sophie also offers integrated reader communication capabilities allowing readers to ask questions and comment on specific sections of the book.

Keywords: e-book; e-publishing; collaboration; Sophie; rich media;

1. Introduction

The digital, networked nature of the World Wide Web provides significant opportunities for the dissemination of electronic content, opportunities not available for conventional printed content. There are several widely used technologies that today facilitate the delivery of media intensive, interactive content over the Web. These technologies include Flash, HTML, and PDF [1].

Collectively, current digital content 'carriers' offer the following features, providing the fundamental advantages that are propelling the acceptance of electronic content:

- Support for various asset types (text, images, audio, video, and others) [2]
- Support for interactive features and scripting
- Advanced layout of content organized as building blocks

The widespread availability and features of existing software tools for writing and publishing is also a major factor for many authors. Every modern authoring tool provides at least some of the following advantages:

- Ability to author, manage, and publish rich multimedia books without prior technical training (ease of use)
- Availability of facilities for the publishing and review process
- Support for author collaboration and reader feedback [3]
- Low price (or free)

There are certainly other criteria that make one authoring tool preferable over another. However, we have limited our list to the set of factors that the project team has identified as those authors take into consideration when choosing an authoring tool.

Each technology taken alone will provide many of the desired capabilities, but none of them provides all of them in a single package. For example, Flash supports various asset types and text flow (using the recently developed Text Layout Framework [4]). However, it requires technical training. It is not intended to make content *per se*, but rather animations and applications. There are many easy to use authoring tools for HTML, and it does support video assets [5]. However, advanced text flows around other assets are not supported, nor are some matrix transformations over media (rotation of images, for example).

We will not continue comparing existing technologies that enable authoring of electronic content since a detailed list is not the purpose of this paper. Rather, it is to make several simple points:

- There are too many authoring tools and the choice authors have to make is not trivial. Often, by choosing a particular technology, authors sacrifice the ability to make use of some of the potential capabilities of electronic content because of technology limitations
- There are too many competing concepts in the world of authoring electronic content. For example, there is a conceptual abyss between making Flash content and making HTML 5 content
- Creating quality interactivity elements in electronic content requires professional experience, tools and training

- Integration of multiple sources of content and reuse of content created with different technologies is complicated and often impossible
- Free and easy to use tools for building advanced, rich content are rare
- Publishing, author collaboration, and reader feedback facilities are often beyond the reach of technically untrained authors.

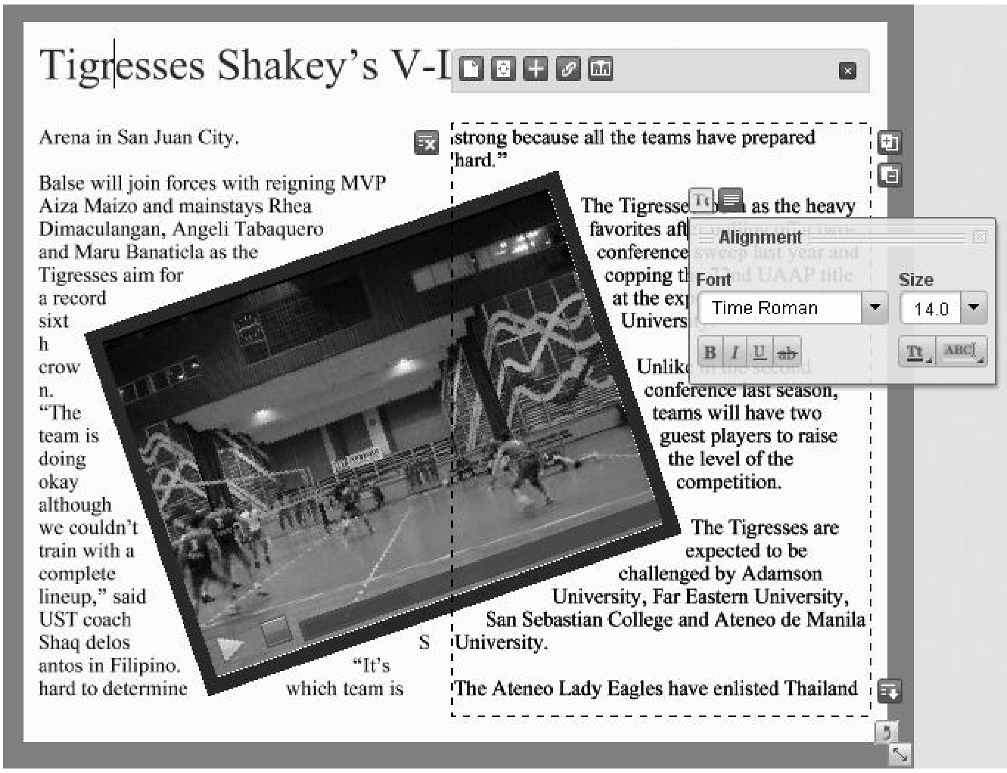
2. What is Sophie?

Sophie is a software package for writing and reading interactive books in a network environment. Its aim is to support authors of all levels in the world of electronic content by addressing the challenges discussed above. In addition, Sophie makes several new conceptual approaches possible for creating interactive content without prior technical training. Sophie also provides a platform for real time author collaboration and gathering reader feedback.

3. Static Sophie Content

Sophie content incorporates a wide variety of asset types such as text, images, sound, videos, and more. Sophie goes even further by enabling the incorporation of PDF documents, Flash, and, more exotically, HTML content. In addition to reusing existing documents and assets in other formats, Sophie books are embeddable one into the other, allowing the authoring process to be decomposed by creating smaller, reusable content components.

Text is of special importance for Sophie as Sophie is software for creating books. In addition to most of the text styling features authors have come to expect in rich text editors, Sophie provides dynamic text flow. Text wraps around other shapes in the document (images, videos, and other text areas, for instance) and is flowed into a sequence of independent rectangular shapes called *a text chain*. Manual chaining of text blocks allows authors to build various types of standard layouts like the simple book layout, or the multi-column layout with images that is the standard for newspapers. Turning the automatic chaining option on makes Sophie automatically generate new pages whenever typed or pasted text exceeds the limits of the manually created text chain.



Tigresses Shakey's V-L

Arena in San Juan City.

Balse will join forces with reigning MVP Aiza Maizo and mainstays Rhea Dimaculangan, Angeli Tabaquero and Maru Banaticla as the Tigresses aim for a record sixth crown.

"The team is doing okay although we couldn't train with a complete lineup," said UST coach Shaq delos Santos in Filipino. "It's hard to determine which team is strong because all the teams have prepared hard."

The Tigresses ... as the heavy favorites aft... conference ... copping t... at the ex... Univers...

Unlik... conference last season, teams will have two guest players to raise the level of the competition.

The Tigresses are expected to be challenged by Adamson University, Far Eastern University, San Sebastian College and Ateneo de Manila University.

The Ateneo Lady Eagles have enlisted Thailand

Alignment

Font: Time Roman

Size: 14.0

Figure 1: Halo buttons and HUDs conceptualize available settings and tools

Supporting a wide variety of asset formats is a significant advantage; however, in most software tools, this often leads to bloating of the user interface. Many different asset types usually mean many different tools and properties, accessible through many different palettes, dialog boxes, and inspectors. Sophie addresses these challenges by proposing an innovative universal approach to the user interface for manipulating any content type. Sophie introduces halo buttons and HUDs (Head-up Displays) which contextualize the access to available operations over the specific assets.

4. Dynamic Features for Sophie Content

Content displayed on a computer not only allows incorporation of predefined dynamic behavior (such as effects, transitions, and time-based behavior), but also enables interactivity. Sophie's timeline feature allows authors to synchronize dynamically changing properties of page elements in a Sophie book with time. Effects such as "start the video at the n-th second, while at the same time making the page title turn white" are achieved by simply clicking the desired moment on the timeline and setting the attributes of the elements to be changed. Further, the time line allows synchronization of audio, video, images, and transcripts through the same intuitive interface.

Sophie provides a set of triggers that fire when a certain event occurs within page elements, when an element is clicked, for example. Authors can associate triggers with an action from a predefined set, such as hide or show an asset, play or stop a video. Advanced links can be built by defining dynamic behavior of content elements to be triggered by user input. Possible actions include changing the current page to any other page, which makes possible the implementation of non linear reading.

Sophie's scripting system lets authors write JavaScript code that accesses the Book's DOM model to implement complex interactivity or batch operations.

5. Publishing with Sophie

No matter how impressive the content itself is, publishing and distribution of interactive, rich media content is commonly problematic to authors. Electronic content is, for example, often published on web sites in PDF format.

Sophie 2.0 consists of three building blocks - the Sophie Author, the Sophie Reader and the Sophie Server. Sophie Author seamlessly connects to Sophie Server, allowing authors to publish their books on the Server with a single click. Books become accessible to readers through the Sophie Server's web interface. Readers read books using Sophie Reader. Sophie Reader is capable of running in a web page (as an applet), which allows Readers to read without installing any software on their computers and authors to integrate Sophie books into web pages. Additionally, Sophie Server supports the full history of book editions.

Sophie introduces the concept of 'book extras', additions to the main book content. An example of a book extra is the use of Sophie annotations that allow authors and readers to create additional notes to book sections, paragraphs, or individual assets. The novel aspect of Sophie's book extras is that they can be distributed separately from the book. In this way there may be several differing sets of book annotations made available to different target groups of readers.

6. Collaboration and Feedback

Sophie Server provides real time collaboration features for authors connected to the server. If multiple authors are connected to Sophie Server and working on the same book, changes made by any author are reflected to the local copies of all other authors. Authors will also be able to work offline, without a connection to Sophie Server, and their work will be automatically integrated with the work of others when a connection becomes available.

Through the use of a Sophie Feature called “comment frames”, authors are capable of requesting live feedback from readers. When reading a book, readers are able to input their comments or suggestions in the comment frames making their input available to other readers and the author in real time. Comment frames can be associated with specific book segments, which allow focusing on gathered feedback on specific parts of the book.

7. Sophie as a Platform

In addition to being open source, Sophie 2.0 is being developed with extensibility in mind. Java is chosen for implementing most of Sophie 2.0 because of its high popularity and rich availability of tools. Sophie 2.0 runs on Java SE 5 and higher. Modularization support is built on top of OSGi. Each of the products (Author, Reader, and Server) is implemented as subsets of the modules of the platform. Extension points for additional content types, additional user interface elements, and others are defined to allow extra features to be added without the need of modifying the existing source code. Third party module developers can take advantage of the underlying implementations for resources, text layout, graphic scenes, collaboration and other fundamental Sophie libraries. The file formats are open, documented and XML based. The source code is documented and follows strict conventions.

8. Applications and target users

Sophie 2.0 has been developed by academics and focused at academia. Sophie is being evaluated by professors creating content to be used for their classes and other academic activities. However, the broad set of capabilities Sophie provides opens it to various other target groups, such as self publishing authors, traditional publishers, artists, designers, and others. Sophie is under continuous development, and new features and improvements are being introduced that enhance Sophie's publishing capabilities as well as its content authoring features.

Acknowledgments

Sophie 2.0 is a project developed by the University of Southern California's School of Multimedia Literacy and Astea Solutions AD of Sofia, Bulgaria, under a grant from The Andrew W. Mellon Foundation. Sophie 2.0 is distributed under the Educational Community License (ECL). The project is a continuation of the work on Sophie 1.0, initiated in 2006 by Robert Stein, founder of the Institute for the Future of the Book. Software development has been done in close collaboration with the Laboratory of Interactive Multimedia at the Faculty of Mathematics and Informatics of Sofia University.

This report is supported by the SmartBook project, subsidized by the Bulgarian National Science Fund, under Grant D002-111/15.12.2008 and the EC FP7 SISTER Project, Grant Agreement Number 205030.

Notes and References

[1] WIKIPEDIA. *Comparison of e-book formats*. Available at http://en.wikipedia.org/wiki/Comparison_of_e-book_formats

[2] WIKIPEDIA. *Multimedia*. Available at. http://en.wikipedia.org/wiki/Multimedia

[3] NEW MEDIA CONSORTINUM AND THE EDUCAUSE LEARNING INTIATIVE*. The 2010 Horizon Report*. Available at http://wp.nmc.org/horizon2010/

[4] ADOBE LABS*. The Text Layout Framework*. Available at http://labs.adobe.com/technologies/textlayout/

[5] W3C. *HTML 5 reference*. Available at http://dev.w3.org/html5/html-author/

E-books: Finally there?

Jan Engelen

Katholieke Universiteit Leuven, ESAT-SCDocarch
Kasteelpark Arenberg 10 box 2442, B-3001 Leuven (Belgium)
jan.engelen@esat.kuleuven.be

Abstract

Widespread distribution of electronic book readers, commonly called “e-readers” seems to have taken off seriously over the last year. Although the first e-readers popped up almost twenty years ago, last year’s market appearance of a completely reworked Amazon Kindle 2, a new series of Sony PRS readers and several Bookeen & Cybook devices made e-books quite popular. All of them can store thousands of books, are extremely light weight and very mince. However, many of them present problems for persons with low vision or blindness. We will discuss briefly the situation and possible solutions.

Keywords: electronic books, e-books, Kindle, print impaired persons

1. Introduction

One of the major recent e-book reader improvements is linked to the display technology: so called e-ink pages [1] are using battery power only during a text change; in between, their consumption is almost nil. Battery life for these devices now routinely extends to one or two weeks. However, e-ink screens do not produce light: one needs ambient lighting on the screen to see the text.

E-books (as opposed to e-book readers) have yet to gain global distribution. Amazon provides a huge amount of e-books, but only in their own (Kindle) format and up to now only in English. Publishers in other languages often prefer the more open ePub format (e.g. BE & NL).

Furthermore, not all authors have endorsed the concept of electronic publishing. J.K Rowling, author of the Harry Potter series, has stated that there will be no e-versions of her books . . . [2]

E-books can easily be obtained from many internet bookstores. Two bottlenecks do remain, however: their price (still often at almost the same level as printed copies) and a copyright protection mechanism that hinders changing the electronic format of a book (e.g. when a person buys another brand of e-reader, previously bought e-books become unusable).

E-books and audiobooks have been discussed at previous ELPUB conferences [3]. Therefore this contribution focuses on recent developments only.

The International Digital Publishing Forum (IDPF) provides impressive statistics on the booming US e-book industry (cf. Fig. 1). [4]

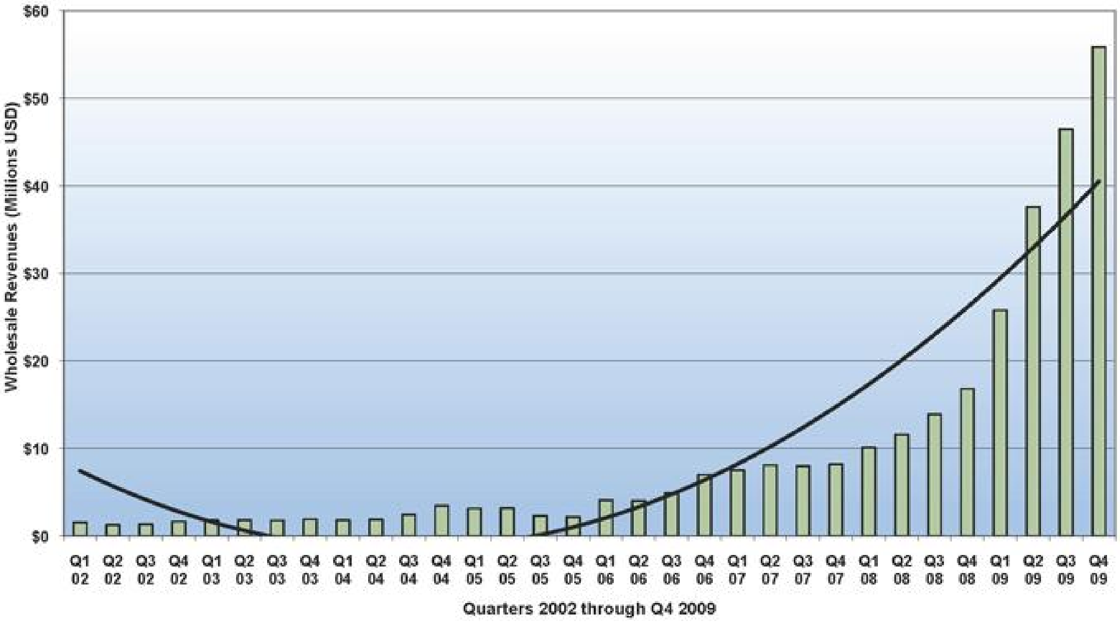


Figure 1. US book industry (2002-2009)

2. E-book formats

Current e-books are made to several different standards with their own advantages and disadvantages, which will briefly be described below.

Besides popular formats such as HTML, RTF and PDF that can be read by all e-book readers, following major e-book standards are commonly used [5]:

- Amazon Kindle has his own format, loosely based on an older French Mobipocket format
- Sony uses the ePub format (Open standard for e-books)
- the Cybooks use an Adobe protected PDF format

Except for public domain books an encryption feature (Digital rights management) is used on all e-books.

3. E-book distribution

Evidently most e-books simply can be downloaded from the Internet.

As Internet was not yet routinely available when travelling (and therefore buying books on the move was not possible), Amazon decided to distribute the Kindle books through Whispernet which is based on mobile network technology (3G, EDGE…). The cost of the data connection is covered by Amazon (and constitutes part of the e-book price). Up till last year it was impossible to download books within Europe where Amazon's Whispernet was not available. Since then agreements with EU mobile providers have been made.

4. Using e-book readers as audiobook or talking book devices

Starting with the Kindle-2 [6] in 2009 a text-to-speech module has been incorporated in an e-reader device. This feature permitted to listen to the text and was very much appreciated by print impaired persons (persons with low vision, blindness, dyslexia or a severe motor handicap). Unfortunately this possibility was turned off soon after by Amazon unless the author had explicitly agreed with it or if public domain books (=old) are read. In practice the ban on text-to-speech output was almost general.

This led to several US cases in court (e.g. National Federation of the Blind *vs* the Arizona State University that planned to provide university course material in Kindle format only).

Another bottleneck is commanding the e-reader itself. Up to now no auditive feedback was produced when choosing commands on the device. But at the end of 2010 a new Kindle will become available in which the speech output can be used for accessing the menu functions and an extra large font will be added so that the device is more usable for persons with low vision. However nothing apparently will change for the text-to-speech locking for most books…

5. Daisy format e-books (and audio-books)

For use within the community of print impaired persons an e-book and audiobook standard has been developed almost 15 years ago. The Daisy format not only links the text to the audio version of the same book but permits also an extensive tree-like navigation through the document. This results in easy jumping to parts of the book, including up to the sixth level in a table of contents. Furthermore Daisy permits to produce several kinds of e-books such as text-only, audio-only, text & audio, text & images etc [7]..

Daisy is promoted by the Daisy consortium and their standards are nowadays recognised by international bodies [8].

Daisy books technically consist of a collection of computer files that can be stored on a CD (very popular as it can be read with a portable CD player) but also on any other computer medium including SD-cards (popular for pocket size readers) and simply via the internet.

Despite a decennium of efforts the Daisy standard is still not in use outside the field of print impaired users. To make it more popular several open-source software solutions have been developed. So it is possible to produce a Daisy talking book directly from within Microsoft Word [9]. Within the European Aegis project three add-ons for OpenOffice.org (file extension: *.odt) have been developed at K.U.Leuven [10]:

- an *odt* to Daisy convertor
- an *odt* to Daisy talking book convertor
- an *odt* to Braille convertor – still under development

The Daisy Consortium itself focuses on

- DAISY Online Delivery Protocol: this is basically a definition of SOAP messages permitting easy web-based content provision [11]
- Daisy version 4.0: this standard will permit an easier transfer of e-books and talking books to the ePub format (mentioned in section 2)
- copyright protection for Daisy books.
  Up to now such protection was deemed unnecessary as special equipment or software was needed to read a Daisy book.

6. Conclusions

It can be stated that the market of e-books and e-readers finally has taken off. Although the phenomenon in general terms still remains a byproduct of standard and traditional book publishing, new applications e.g. for print impaired persons seem to be growing. But a tough copyright hurdle is still to be taken before e-books routinely also will become audio or talking books.

Acknowledgements

The author would like to acknowledge the help of the Aegis and Kamelego projects in preparing this contribution. Furthermore the support of the Kath. Univ. Leuven is highly appreciated.

Notes and References

[1] Details at: http://en.wikipedia.org/wiki/E_Ink

[2] This statement is based on fear for piracy. Ironically enough, all Harry Potter books have been turned (illegally) into e-books within hours of the release of the printed version. More on: http://www.bradsreader.com/2009/01/jk-rowling-harry-potter-ebooks-and-the-definition-of-irony/

[3] ENGELEN, J. Marketing Issues related to Commercial and Specialised Audiobooks, including Digital Daily Newspapers, ELPUB2009. *Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies - Proceedings of the 13th International Conference on Electronic Publishing* held in Milano, Italy 10-12 June 2009 / Edited by: Susanna Mornati and Turid Hedlund. ISBN 978-88-6134-326-6, 2009, pp. 621-624; ENGELEN, J. A Rapidly Growing Electronic Publishing Trend: Audiobooks for Leisure and Education, (Electronic Publishing conference - ELPUB-2008, Toronto, June 26-27, 2008), published in *Open Scholarship: Authority, Community and Sustainability in the Age of Web 2.0*, Edited by Leslie Chan & Susana Mornati, ISBN 978-0-7727-6315-0. Both papers are available electronically from elpub.scix.net.

[4] More trend figures and numbers can be found at: http://www.idpf.org/doc_library/industrystats.htm

[5] http://en.wikipedia.org/wiki/Comparison_of_e-book_formats

[6] Details at: http://en.wikipedia.org/wiki/Amazon_Kindle

[7] http://www.daisy.org/daisy-technology

[8] Daisy 3.0 is in fact an ANSI standard, "ANSI/NISO Z39.86 Specifications for the Digital Talking Book"

[9] Details on: http://www.daisy.org/project/save-as-daisy-microsoft

[10] Details on: http://www.daisy.org/project/save-as-daisy-openoffice.org

[11] Details on: http://www.daisy.org/projects/daisy-online-delivery/drafts/20100402/do-spec-20100402.html