LET'S PUT DATA TO USE: DIGITAL SCHOLARSHIP FOR THE NEXT GENERATION

This page intentionally left blank

Let's Put Data to Use: Digital Scholarship for the Next Generation

Proceedings of the 18th International Conference on Electronic Publishing

Edited by

Panayiota Polydoratou

Alexander Technological Educational Institute of Thessaloniki, Greece

and

Milena Dobreva

University of Malta, Malta

IOS Press

Amsterdam • Berlin • Tokyo • Washington, DC

ISBN 978-1-61499-408-4 (print)
ISBN 978-1-61499-409-1 (online)
Library of Congress Control Number: 2014939899

Cover photo courtesy of Nikos Papakyriazis.

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

Distributor in the USA and Canada
IOS Press, Inc.
4502 Rachael Manor Drive
Fairfax, VA 22032
USA
fax: +1 703 323 3668
e-mail: iosbooks@iospress.com

PRINTED IN THE NETHERLANDS

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

Preface

The main theme of the 18th International Conference on Electronic Publishing (ELPUB) is the openness and use of research data as well as new and innovative publishing paradigms. Specifically, it aimed to bring together presentations and discussions that demonstrate the role of cultural heritage and service organizations in the creation, accessibility, curation and long term preservation of data. We aimed to provide a forum for discussing appraisal, citation and licensing of research data. Also, what is new with reviewing, publishing and editorial technology in a data-centric setting?

ELPUB brings together researchers and practitioners to discuss data mining, digital publishing and social networks along with their implications for scholarly communication, information services, e-learning, e-businesses, the cultural heritage sector, and other areas where electronic publishing is imperative.

ELPUB 2014 received 32 paper submissions. The peer review process resulted in the acceptance of 13 research papers and 9 posters. These papers were grouped into sessions based on the following topics: Open Access and Open Data; Know the Users Better: Researchers and Their Needs; Specialized Content for Researchers; Publishing and Access; Practical Aspects of Electronic Publishing.

The conference held 2 pre-conference workshops and one tutorial on June 18. Andreas Rauber and Kresimir Duretec (Technical University of Vienna, Austria) led the tutorial "Digital Preservation Lifecycle: from challenges to solutions". Pierre Mounier (EHESS/OpenEdition, France) and Victoria Tsoukala (National Documentation Centre, Greece) led the workshop "Non-profit Open Access ventures of significant scope in Europe" and Carla Basili (Sapienza University in Rome, Italy) led the workshop "Information Literacy in the context of scientific information".

The main program on June 19–20 features two keynotes. Herbert van de Sompel (Los Alamos National Laboratory, USA) will deliver a keynote entitled "Towards Robust Linking and Referencing for Web-Based Scholarly Communication". Mahendra Mahey (British Library Labs, UK) will deliver a keynote entitled "How the British Library's Digital Scholarship department is putting data to use for researchers through its Digital research Team and British Library Labs project".

We believe that the topics featured in the program of this year's ELPUB conference are diverse and exciting. Firstly, we would like to thank the members of the ELPUB Executive Committee who, together with the Local Advisory Committee, provided valuable advice and assistance during the entire process of the organization. Secondly, we would like to thank our colleagues in the Program Committee who helped in assuring the quality of the conference throughout the peer reviewing process. Lastly, we acknowledge the Local Organization team for making sure that all efforts materialized into a very interesting scientific event. Thank you all for helping us maintain the quality of ELPUB and deserve the trust of our authors and attendees.

We wish you all a good conference and we say farewell hoping to see you again in Malta for the next edition of the conference in 2015!

Panayiota Polydoratou and Milena Dobreva (editors)

This page intentionally left blank

Organising Chairs

General Chair: Panayiota Polydoratou – Alexander Technological Educational Institute of Thessaloniki (Greece)
Programme Chair: Milena Dobreva – University of Malta (Malta)

Executive Committee

John Smith – University of Kent at Canterbury (UK)
Peter Linde – Blekinge Institute of Technology (Sweden)
Karel Jezek – University of West Bohemia in Pilsen (Czech Republic)
Ana Alice Baptista – University of Minho (Portugal)
Sely Costa – University of Brasília (Brazil)
Jan Engelen – Catholic University of Leuven (Belgium)
Mícheál Mac an Airchinnigh – Trinity College Dublin (Ireland)
Bob Martens – Vienna University of Technology (Austria)
Leslie Chan – University of Toronto (Canada)
Susanna Mornati – CILEA (Italy)
Turid Hedlund – Swedish School of Economics and BA, Helsinki (Finland)
Yasar Tonta – Hacettepe University (Turkey)
Niklas Lavesson – Blekinge Institute of Technology (Sweden)

Organising Committee

Giorgos Christodoulou, Alexander Technological Educational Institute of Thessaloniki (Greece)
Emmanouel Garoufallou, Alexander Technological Educational Institute of Thessaloniki (Greece)
Giannis Ioannidis Alexander Technological Educational Institute of Thessaloniki (Greece)
Panayiota Polydoratou, Alexander Technological Educational Institute of Thessaloniki (Greece)
Aspasia Togia, Alexander Technological Educational Institute of Thessaloniki (Greece)

Volunteers

Theodoris Chalkidis
Dimitra Charalampidou
Vicky Georgiadou
Achilleas Gravanis
Dimitris Iliadis
Panagiota Karmpa
Artemis Lavasa
Souzana Maranga
Elli Papadopoulou
Nikos Papakyriazis
Antonis – Dimitris Parissiadis
Despoina Siapkari
Theofilos Tomis
Aggeliki Valsamidou
Alexandros Vasiliou
Stella Vasilleiadou

Programme Committee

Abela, Charlie – University of Malta, Malta
Baptista, Ana Alice – University of Minho, Portugal
Basili, Carla – CNR (National Research Council), Italy
Borbinha, José – INESC-ID/IST – Lisbon Technical University, Portugal
Chan, Leslie – University of Toronto, Canada
Coleman, Ross – University of Sydney, Australia
Costa, Sely – University of Brasilia, Brazil
Engelen, Jan – Katholieke Universiteit Leuven, Belgium
Ginouvès, Véronique – Aix-Marseille Université – CNRS, USR3125
Hedlund, Turid – Hanken School of Economics, Finland
Hemmje, Matthias – FernUniversität in Hagen, Germany
Kapidakis, Sarantos – Ionian University, Greece
Lavesson, Niklas – Blekinge Institute of Technology, Sweden
Linde, Peter – Blekinge Institute of Technology, Sweden
Mac An Airchinnigh, Mícheál – University of Dublin, Ireland
Manola, Natalia – University of Athens, Greece
Martens, Bob – Vienna Univsersity of Technology, Austria
Méndez Rodríguez, Eva María – Universidad Carlos III de Madrid, Spain
Mornati, Suzanna – CINECA, Italy
Rauber, Andreas – Vienna University of Technology, Austria
Rosner, Mike – University of Malta, Malta
Ruiz-Perez, Sergio – DataCite, United Kingdom
Scharnhorst, Andrea – DANS-KNAW, The Netherlands
Schimmer, Ralf – Max Planck Digital Library, Germany

This page intentionally left blank

Contents

Research Articles

This page intentionally left blank

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-3

How Can Libraries and Other Academic Institutions Engage in Making Data Open?

Peter LINDE[1a], Bridgette A WESSELS[b], Thordis SVEINSDOTTIR[b], Merel NOORMAN[c]

[a] *Blekinge Institute of Technology, Sweden*
[b] *University of Sheffield, UK*
[c] *Royal Netherlands Academy of Arts and Sciences (KNAW)*

Abstract. In this paper we will address the questions of what and where the value of open access to research data might be and how libraries and related stakeholders can contribute to achieve the benefits of freely sharing data. In particular, the emphasis will be on how libraries need to acquire the competence for collaboration to train and encourage researchers and library staff to work with open data. The paper is based on the early results of the RECODE project, an EU FP7 project that addresses the drivers and barriers in developing open access to research data in Europe (http://www.recodeproject.eu).

Keywords. Open data, Libraries, Open Access

Introduction

During the last 30 years libraries have adopted to new demands while analogue media turned digital. Librarians have creatively adapted to passing fads, and/or long lived realities such as Archie, Gopher, NCSA Mosaic, FTP, SGML, XLM, Open Access, PDA etc. Today most university libraries have Institutional Repositories and a digital publishing department dedicated to supporting researchers' needs of dissemination, preservation and open access advice. Libraries do have long experience of advocacy, training and implementation of open access of publications and of dealing with digital information but now, when we are finally talking about a tipping point for scholarly Open Access documents[1], a new hot topic with a whole new set of demands on library skills, budgets and organization have arrived – Open data[2].

Open Access (OA) to research data is increasingly regarded as a positive development that should be encouraged and stimulated within the European research landscape. The European Commission is pushing for research data to be more open in its Framework Programme Horizon 2020[3], and the trend is also growing within the

[1] Peter Linde, Blekinge Institute of Technology, 37179 Karlskrona, Sweden. Peter.linde@bth.se

individual member states as well as the academic community. Several influential journals are now encouraging or requiring researchers to make the data that supports their publications freely accessible (for example all the BioMed Central journals, The Open Access Geoscience Data Journal Dataset Papers in Science, eLIFE, F1000Research etc) while national and private funding agencies list open access to research data as a condition for funding. However, achieving open access and realizing its benefits requires considerable work, as the growing literature on data sharing and open access shows.

There now seems to be a more general consensus about the value that open data can bring to science and society. According to its advocates, unrestricted and digitally facilitated access to data would enable faster progress in science through minimising duplication of effort and offering scientists a wider range of data to use for re-analysis, comparison, integration and testing. It would contribute to the quality and integrity of scientific practices, as it increases transparency and accountability. It would also improve the way science and scientific data can be used in relation to social goals, and thus enhance the value of the contribution that science makes to society. Moreover, there is a strong notion that open data will be beneficial to innovation and economic growth. The European Commission, for example, refers to open data as "an engine for innovation, growth and transparent governance[4].

But open access and the re-use of research data have proven to be a challenge in most disciplines. Many repositories, created to encourage data sharing, remain largely empty[5]). Despite the difficulties a few vanguard libraries have felt a need to support researchers in the management and dissemination of research data. We will take a closer look at some of these initiatives, which often started as 'new opportunities' projects aiming to expand library services in a time where classic university library activities like cataloguing, media acquisition, subscription services etc. are questioned or being replaced or automated. The barriers to open research data are many and it is not realistic to believe that one stakeholder can solve all the challenges single-hand. There is a strong need for cooperation inside as well as between organisations, sharing expertise and specialist knowledge.

The central question posed in this paper is: how can libraries handle this new service together with other open data stakeholders in the academic world?

The paper presents a review of policy documents, reports, scholarly literature and other relevant documents to provide an overview of current developments within the field. We provide an analysis of some of these approaches in order to identify good practices and potential barriers.[2]

In the current, very highly, competitive university climate, productivity and quality are buzz words, and increasingly funding for research is based on bibliometrics. In this environment it is becoming more important for university management to keep track of the productivity and quality of the research publications. At the same time more funders are mandating open access and universities are struggling to promote their brand in order to hire the best scientists and attract the brightest students.

[2] This paper is based on findings made in the ongoing work packages of the RECODE project[6].

In this landscape many librarians realize that their services, including repositories, is one of many that have to interconnect in order to support and make research more visible.

Today, university libraries are investigating possibilities of integrating institutional repositories with CRIS systems (Current Research Information System) usually run by university research offices or similar departments[7]. In Sweden this is being investigated on a national level where the national repository portal SwePub will possibly be integrated with the Swedish Research Councils CRIS system[8]. Universities like the University of Edinburgh have integrated all research service into one department (Information Services) which include classical library functions but also have divisions like IT-infrastructures, Digital Curation Center, the Jisc-designated national data centre (EDINA) and the Data Library[9].

In their Roadmap for Research Data the League of European Research Universities listed the library as a main source for data management and discovery[10]. It is evident that an important new role for the library going down the E-science road is to be a competent team player when it comes to build such support structures for researchers. This is best done together with other important players at the university - Research Office Services, Archive staff and Academic IT Services and of course data centre specialists.

The need for training & advocacy

Most researchers and university support staff are new to the task of open data management which implies massive amounts of advocacy and training. In the Opportunities for Data Exchange (ODE) project[11] it is spelled out: “Improving the skills and understanding of researchers in data management is essential. Training should begin in the institutions that train researchers, at the outset of postgraduate study and the latest, possibly even earlier”. It is pointed out repeatedly that discipline-focused education in data management best practice must be incorporated into student and researcher training at an early stage. So in order to play an active part in establishing open data libraries and to build competence for this, cooperation with other university stake holders is important as well as being pro-active in open data management advocacy and training.

One reason why data sharing and open access is still not the norm in most disciplines is due to the reason that researchers are reluctant to make their data public. Their concerns range from work being scooped or misused, to not having enough time or funding to make their data accessible, to maintaining the privacy and confidentiality of their research participants [5]. Researchers may also lack the expertise to share their data[12]. Scientists express a variety of concerns for the “amount of work and the time needed to make data meaningful and useful if made openly available. For instance, the time needed to annotate, create and apply metadata and document context. This extra work would take up time from other research activities such as data collection, analysis, publications and applications for funding, all of which bring clear and demonstrable rewards and benefits to scientists and their careers”[13]. Another key problem is that it requires considerable technical skills to translate data in to machine-readable formats and to use the software tools to access and analyze the data. Researchers that wish to make their data publicly and digitally available and re-usable have to become acquainted with software tools and data formats that might not easily fit their existing research practices. Re-using data, in turn, requires researchers to learn about how to

search and use data through web-based tools. It can also be difficult to find common standards and formats to share data, such that others can easily interpret and use the data. These practical barriers are also reflected in the European Commission's *Online survey on scientific information in the digital age[14]*. About 90% of the respondents in this survey disagreed with the statement: "Generally speaking, there is NO access problem to research data in Europe". Providing training to researchers and technical staff as well as creating awareness about the possibilities and limitations of data sharing will therefore be conductive to making more research data openly accessible in the various disciplines.

Academic institutions have an important role to play in training advocacy. The Commission's survey also included the question how the European Union could best contribute to access and preservation of scientific publications and data. Most respondents agreed strongly with the statements "supporting the development of a European network of repositories" and "encouraging universities/research institutes, libraries and funding bodies etc. to implement specific action"[15]. Since many funding bodies already place responsibility for data management policies and compliance with research institutions, this also increases the pressure on the academies to make data openly available.

Within the whole academic community there is a lack of professional preparation for data management and no one is really taking responsibility for the research data management function. In many ways libraries are in a good position to take on this responsibility but the standard curriculum of library schools do not prepare students for managing data. This has to change.

Different cultures and target groups

In the material reviewed it is a common observation that researchers are a very heterogeneous group. Not only discipline-wise but also between individuals within the same team. Therefore it is important to gain an understanding of the "culture" within any give set of researchers before considering how to influence their research data management behaviour[13].

Research data is different from publications. It is more diverse and often linked to project communities which calls for new ways of working, thinking and cooperating for librarians. Data diversity, tools and researcher needs should not be measured at the disciplinary level but at the research group level.

It is recommended that for advocacy and training purposes interviews, case studies and surveys are developed to understand researcher requirements and behaviour[16, 17, 20, 21, 23]. This must be the basis for developing advocacy/training materials that will motivate researchers, as well as making them understand the obligations to institutions, funders and the public. Preparing data management plans and training staff to accomplish them is new and mostly unchartered waters for universities and research institutions but there are some good examples of and reports on how to support these institutions in open data management.

Mark L. Brown and Wendy White tell the story of how University of Southampton through collaboration with UK Research Data Service and involvement in projects like the Institutional Data Management Blueprint Project (IDMB) started to improve and

formalize initiatives to support researchers at the university in managing their research data[18].

For training purposes, the use of automated- and web tools was set up. For example automated tools to support minting of DataCite DOIs and web based guidance to help interpret funders' requirements.

For data management planning service for researchers a training program was developed to engage with various groups from postgraduate researchers to senior scientists. Planning and realization of these courses, lectures, workshops and seminars were always done together with the researchers themselves.

In a consultancy report made for Jisc[19], the roles, rights, responsibilities and relationships of institutions, data centers and other stakeholders who work with data were explored. The conclusions regarding advocacy and training are very similar to the conclusions from Southampton: The importance to target and tailor measures to specific disciplines and sub-disciplines; Awareness of data curation and preservation good practice is generally low but it varies a lot between disciplines; Recommendations to data center and institutional repository staff to go out and promote their training programs with a mix of methods, seminars, workshops, lessons etc.

As reported in most of the literature an important target group for open data management advocacy and training are young scientists and students at master level and onwards. A first focus of advocacy should be on the postgraduate and the graduate student community since they are in the front line as data collectors and generators, and of course as future researchers[20].

Bottom up or top down?

The typical American data curation program is "devoid of top-level mandates and incentives, but rich with independent "bottom-up" action". A structure like this is based on enterprising individuals and makes for a slow speed of development[21]. In a recent American survey with the aim to identify current trends in research data management at research institutions only 9% of the respondents answered yes to the question "Does your institution have a DM policy"? Close to 90% agreed with the follow up statement "An institution-wide DM policy is important" which shows that university stakeholders like researchers, librarians, office of research staff, teachers etc. are keen to see such policies implemented[22].

The reason libraries have started data curation programmes at all is due to their vanguard position relating to open access publications repositories and the digital preservation initiatives early explored by university libraries. This is also said to give the library opportunity to leverage existing partnerships and engage in new ones to build skills and necessary alliances for data curation. Engaging with a few research communities as a start up pilot is a way to gain acceptance, formalization and getting program commitments from administrative levels. A successful project might well be a way of convincing university administrators of the benefits of a university wide curation policy and mandate[21].

In Southampton [19] the response to the insight that funders increasingly placed responsibility for data management policies and compliance with research institutions resulted in a bottom-up approach based on researchers needs and an incentive to design requirements for an institutional top-down approach policy and infrastructure. Their

experience with open access publishing repositories was that "researchers were open to new practice as long as it was researcher led, integrated into research workflow, reflective of discipline distinctions and supported by advice and training. Clarity over policy and responsive service support were essential".

It was very important that the institutions at the university felt that they were in command of the investments and service support regarding data management without feeling compelled by a set of requirements.

In this process the resulting data management policy was putting the responsibility for recording, maintenance, storage and security etc. and the compliance with relevant regulations on the researchers which is good news from a library viewpoint. It is sensible that the creators of the data also record it and that the library is there as a supporter of the process instead as an accountable enforcer.

Of the key components in the Southampton project, an institutional policy framework, a working institutional data registry, a one stop shop for data management advice and guidance and a sustainable business model it is the university policy on research data management that is considered the most important. In the end and because of the power balance there is a need for a formal mandate or policy from a higher university authority[18].

Librarians introduced and administer the institutional repository and the idea about open access with a great knowledge about scholarly communication issues but since they do not bring any funding into the university the library is mostly perceived as a service based unit without much influence. But in the meantime, and as a first step to a formal policy, when there is no clear guidance from government authorities and university administrations are withholding resources or initiatives on data management issues, the bottom up approach is a way to start where advocacy is the first step only.

New roles and partners

University of Southampton is one of many examples of how initiatives for data curation projects do not stop with collaboration inside the university departments. Many times necessary skills are only available through partnering with outside institutions or organizations[21, 22]

No matter how libraries approach the challenge of data curation an introduction of new skills in the library profession is sorely needed. Working in partnership with scientists' future job roles as "data librarians" must contain skills both on the technical side and the archival side of the data coin. Specialists like this will play a key role in the scholarly publication process and must be rewarded accordingly. Library schools need to introduce courses that fit these new job descriptions.

There is absolutely a need for convergence between library and archival skills in order to make university repositories a well functioning place for open data. This could also be a part of professional development and training[18]. This is also true for library professionals vis-à-vis research office professionals who are close to researchers supporting them with project applications, statistics etc. There might also be a chance for classic library roles such as liaison librarians to expand. Liaisons can help researchers depositing their data at the point of data creation. They can advice about standards applicable to the needs, create curation plans to the whole life cycle of the data in full compliance with funder mandates[23].

As stated earlier the skill levels of researchers regarding data management are variable and training is much needed. So parallel to advocacy there is a requirement for development of community skills. But since most of the expertise in data management is concentrated in data centers, there is a need to engage and formalize a flow of knowledge from data centers to institutions where staff now increasingly are being appointed to manage and develop repositories for data curation.

Since 1976, CESSDA (Consortium of European social science data archives) has served as an informal umbrella organisation for the European national data archives. The CESSDA data archives and other similar subject data archives are in a good position to work with universities libraries and negotiate with archives on training.

Sometimes there is a polarization of views regarding the role of institutional repositories for data. Data centers and data archives have a more long-term perspective than the institutional repositories, which are relatively new structures yet to prove their ability. But both data centers and libraries have a stewardship role in data curation activities. They both help and guide researchers depositing their data. Dividing the different roles on short-term, easily accessible storage taken care of by institutional repositories and long-term preservation by data centers could be one way to facilitate for better data management support and cooperation[18].

Conclusions and discussion

Underlying issue of the new roles for the libraries in open data management is of course the question about funding the new services. There is obviously a need for the university to make economic plans for the costs of storage, curation, training etc. for research data.

It can be a major problem to convince university administration to gather economic resources for developing data curation models. In fact most of the scarce funding for research data management is coming from libraries themselves[22]. Usually there is no extra seed money available inside the organization and libraries either have to reallocate internal resources or find external funding, e.g. cooperation with outside partners. Therefore the initiation of grants and funding for libraries on national or international levels will be an important factor for getting data curation to gain speed on a broader level at universities[21].

There will probably be no real increase in funding without institutional or national mandates implementing research data management plans. Bottom up practices are slow generators of change and general acceptance and will therefore have to be complemented with formal policies.

Among the major academic stakeholders in the open data eco system we have the funders of science – the councils and foundations; the creators of data – the researchers and we have the disseminators and curators of data – in this case the libraries, archives and the data centres. All these stakeholders with their organizations will need to cooperate, as the barriers are multiple and complex, that only joint forces can realize the idea of open data. Funders and policy makers need to clearly mandate data management and also earmark funds for training, infrastructure, data curation projects etc. Professional associations have to reflect on instigating new opportunities for

training of professionals. Librarians, IT-specialists and research office staff from the universities need to collaborate with archivists and curators from data centres and vice versa. Researchers need to find new priorities regarding the importance of data management, need to find ways to make data management pay career wise.

All this cooperation is already going on but it will have to spread and it has to be fuelled by governmental and academic authorities that issues policies that can facilitate cooperation and clear roadmaps for the way forward. Equally important are the non-governmental advocacy groups and other cross-professional organizations that have taken an interest in pushing the question of open data forward. Organizations like COAR, EUDAT, LIBER[24], RDA, SHERPA, SPARC, KE and many more are doing a fantastic job of advocating and informing about the importance of open data management and they are a giant resource for libraries that are about to start data curation schemes.

There is a current gap of technical knowledge and access to proper infrastructure but there is also among the libraries and librarians a lack of understanding of the complexity of the process of managing open data Using the experiences from the case studies performed in the RECODE project so far, we argue that the value of unrestricted access to research data depends significantly on the quality of the OA process. Our analysis of the values and motivations amongst researchers regarding OA showed that approaches to support and improve the development of open access to research data need to address at least the following issues:

- They should be sensitive to the different scientific practices to ensure that existing research rigour is maintained as well as facilitating OA.
- They should make the link between infrastructures, legal and ethical issues, and institutional frameworks, so that the OA ecosystem can support an appropriate approach to all types of data within their research areas.
- They need to provide safeguards for anonymity and privacy of research participants.
- They should provide ways to reference and attribute all open data correctly as part of ethical research practice.
- They need to pay attention to technological issues; such as the way technology drives the collection of vast datasets, the lack of technical infrastructure to store data and interoperability issues.
- Cultural barriers are significant, especially issues such as competition within science for reward and reputation, the lack of trust between scientists and the lack of career related rewards and prestige resulting from publishing and sharing data.

It is vital for libraries to realize that now is the time to be proactive regarding research data management – introducing professional preparation programs, starting up pilot programs, monitoring major data initiatives like DataCite, DataONE etc. and good examples of library initiatives like University of Edinburgh[9], University of York[25] University of Southampton or Purdue distributed data curation center[26] or else risk being bypassed by other players in the arena of establishing research data management programs. The role of libraries in data management training is not evident for everyone. Some researchers agree that libraries should have and increasingly important role as

data managers and experts based on their role in open access article publishing. Others argue that data centers could provide the support needed to handle the data correctly [11]. It is high time to start to reflect on these issues and to start studying experiences made so far in the urgent task of making research data openly available. If the library does not see the potential in the task of pioneering open research data, as it have in advocating open access to research publications, there is a major risk that other stakeholders quickly will fill that role and expand services visavi researchers and librarians will be left with the question, of how libraries can engage in making data open, unanswered.

References

[1] Archambault, Eric et al. Proportion of Open Access peer-Reviewed Papers at the European and World levels – 20014-2011. August 2013. Produced for the European Commission DG Research & Innovation by Science-Metrix Inc.

[2] By ”open data” we refer to research data defined as any material used as a foundation for research.

[3] Press releases database. Commission launches pilot to open up publcicly funded research data. 2013. http://europa.eu/rapid/press-release_IP-13-1257_en.htm. Visited 140131.

[4] 1 European Commission (2011). Open Data, an engine for innovation, growth and transparent governance, COM 882 final, Brussels, 12 December 2011. Retrieved from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF

[5] Nelson, Bryn. Data sharing: Empty archives. *Nature* 461, 160-163, 2009.

[6] Policy RECommendations for Open access to research Data in Europe. http://recodeproject.eu/

[7] Joint, Nicholas. Current research information systems, open access repositories and libraries. *Library Review* Vol. 57:8, 2008

[8] System för analys av svensk forskning. http://www.mynewsdesk.com/se/kungliga_biblioteket/pressreleases/system-foer-analys-av-svensk-forskning-947591. Visited 140125.

[9] Rice, Robin et al. Implementing the Research Data Management Policy: University of Edinburgh Roadmap. *International Journal of Digital Curation* Vol. 8:2, 2013.

[10] LERU roadmap for Research Data. League of European Research Universities, 2013. http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf

[11] Dallmeier-Tiessen S, et al. (2012). Compilation of Results on Drivers and Barriers and New Opportunities. Retrieved from [http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-CompilationResultsDriversBarriersNewOpportunities1.pdf].

[12] Borgman, Christine L. The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*, Vol 63:6, 2012.

[13] Sveinsdottir, Thordis et al. Deliverable D1: Stakeholder Values and Ecosystems. Policy RECommednations for Open access to research Data in Europe (RECODE), 30 september 2013. http://recodeproject.eu/wp-content/uploads/2013/10/RECODE_D1-Stakeholder-values-and-ecosystems_Sept2013.pdf

[14] European Commission (2012). Online survey on scientific information in the digital age, 2012. ISBN: 978-92-79-23170-4. DOI:10.2777/7549

[15] Ibid.

[16] Lyon, Liz et al. Final report – disciplinary Approaches to Sharing, Curation, Reuse and Preservation. Jisc 2009. http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf

[17] Schmidt, Lisa, Ghering, Cynthia; Nicholson, Shawn. Digital Curation Planning at Michigan State University. *Notes on Operations* 55(2), 2011. http://staff.lib.msu.edu/nicho147/Research/DigCur_LRTS_2011.pdf

[18] Pryor, Graham and Sarah Jones and Angus Whyte. *Delivering Research Data Managment Services: Fundamentals of good practice*. Facet Publishing, 2013.

[19] Lyon, Liz. Dealing with Data: Roles, Rights, Responsibilities and Relationships – Consultancy Report, 2007.

[20] Carlson, Jake R. and Bracke, Marianne S., "Data Management and Sharing from the Perspective of Graduate Students: An Examination of Culture and Practice at the Water Quality Field Station" (2013). *Libraries Faculty and Staff Scholarship and Research.* Paper 53.
[21] Walters, Tyler. Data curation program Development in U.S. Universitites: The Georgia Institute of Technology Example. *The International Journal of Digital Curation*, Vol. 4:3, 2009.
[22] *Research Data Management* – Principles, practices, and prospects. Council on Library and Information Resources. 2013. ISBN 978-1-932326-47-5. http://www.clir.org/pubs/reports/pub160
[23] Gabridge, T. The last mile: Liason roles in curating science and engineering research data. Research Library. Issues: A bimonthly report from ARL CNL and SPARC August 2009. http://old.arl.org/bm-doc/rli 265 gabridge.pdf
[24] Chrisensen-Dalsgaard et al. Ten recommendations for libraries to get started with research data management. Final report of the LIBER working group on E-Science/Research Data Management, 2012.
[25] Research data management at the university of York. http://www.york.ac.uk/about/departments/support-and-admin/information-directorate/strategy/projects/rdm/
[26] Distributed data curation center, D2C2. Purdue University Libraries. http://d2c2.lib.purdue.edu/ Purdue University

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-13

Attitudes towards open access: a meta-synthesis of the empirical literature

Aspasia TOGIA[a,1] and Stella KOROBILI[a]
[a] *Department of Library Science & Information Systems, ATEI of Thessaloniki, Greece*

Abstract. The aim of the present study is to report the results of a meta-synthesis of the empirical literature on scholars' attitudes towards Open Access (OA) journals. A total of 16 articles published in scholarly journals since 2002 (when the Budapest Open Access Initiative was released) were included in the study and five major themes emerged from their examination and analysis. The literature indicates that attitudes and perceptions of OA are varied across countries and across disciplines. Free access, which is perceived to facilitate wider dissemination of research outputs, is a strong incentive for publishing in OA. However, quality and reputation are the most important factors in selecting a journal and take priority over the availability of free access. Although OA is perceived to have many advantages over the traditional publication model, it raises some concerns too, especially in regard to the author-pays model, the quality of peer-review and the impact of the journals.

Keywords. Open access, attitudes, scholars, authors, meta-synthesis

Introduction

Since their appearance in the 17th century, scholarly journals have been the main vehicle for the dissemination of scientific knowledge. Researchers publish the results of their activities for a number of reasons: to expand the knowledge in their subject field, to contribute towards solving problems or to establish their reputation [1]. On the other hand, access of scientists and academics to scientific literature is essential for upgrading their knowledge; designing new research studies and writing research papers. Scholars have always been seeking efficient ways for communicating their thoughts to the larger possible audience and Open Access (OA) publishing model has the potential to meet this challenge, by making scientific information free for anyone to access. Open Access, as it is defined by the Budapest Open Access Initiative [2], is the free availability of scientific research publications, permitting users to read, download, copy, distribute, print, search, or link to the full texts of these publications without financial, legal or technical barriers. The concept of Open Access is not a new one, but it has really gained more attention in the last decade, mainly because it has begun to gain the support of governments, universities and funding agencies [3, 4].

The Open Access movement has many supporters within the scholarly community, partly due to the growing dissatisfaction with traditional publishing models. As Allen [3] points out the combination of three important factors encouraged experimentation with alternative publishing models: escalating costs of scientific journals; objection to a

[1] Corresponding Author.

model which restricts access to the results of publicly funded results; and Internet technology which allows widespread dissemination of information. On the other hand, opponents of OA raise serious concerns and point out several problems associated with the idea of making all scholarship available for free [5]. In recent years a number of efforts to publish OA journals on a larger scale have emerged [6] and quite a few studies on authors' opinions and perceptions of OA have been published. This study reports the results of a meta-synthesis of the empirical literature on the attitudes of scholars towards OA journals. In an effort to make a contribution to the international literature on OA publishing, it synthesizes and analyzes the recent articles as collective body of literature.

1. Methodology

A qualitative method, meta-synthesis, has been employed in order to systematically analyze and synthesize the findings of previous studies on scholars' attitudes towards OA. Meta-synthesis was advanced by Bair & Haworth [7] in an effort to develop a comprehensive understanding of doctoral student attrition and retention. As a method for the integration of multiple research studies on a specific topic, meta- synthesis is related to, though distinct from, meta-analysis and meta-ethnography. Whereas meta-analysis is applied to quantitative studies, and meta-ethnography is applied to qualitative studies, meta-synthesis "synthesize findings from a combination of both qualitative and quantitative studies" [7, p. 485]. Meta-synthesis is a qualitative research methodology, since it is not possible to synthesize data from both qualitative and quantitative studies, due to the absence of a common metric. Its aim is to integrate, compare and analyze in a constructivist way many previously unrelated studies, allowing interpretive themes to emerge from the synthesis. Through this method, the results from the literature were synthesized, in order to identify key themes on attitudes towards OA publishing and understand these emerging themes in relationship to each other.

The studies used in this meta-synthesis met certain selection criteria: (i) they addressed scholars' attitudes and perceptions about OA journals; (ii) they reported the results of empirical research; (iii) they were full length articles published in peer-review journals; (iv) they were published between 2002 (when the Budapest Open Access Initiative was released) and 2013; (v) they were written in English. In December 2013 literature searches were performed in LISA and LISTA databases as well as in Google Scholar. In addition, the authors conducted ancestral searches of the reference lists of the articles retrieved through the database searches. Titles and abstracts were screened for relevance by the one of the two authors. The full texts of potentially relevant articles were assessed independently by the two authors and disagreements were resolved by discussion. Studies that were identified but not included in the sample were removed for one or more of the following reasons: (i) the study could not be considered a research article (it did not report use of specific research methodology neither it presented specific findings); (ii) the study examined the attitudes towards OA publishing from the perspective of publishers or librarians; (iii) the research was concerned with open access venues other than journals, e.g., institutional repositories; (iv) the study dealt with one specific aspect of OA, e.g., copyright; and (v) the full text of the study could not be obtained. This procedure resulted in a total of 16 articles which met the selection criteria. After the studies were

gathered, they were read and summarized, and the following categories of information were noted: (i) author(s) and date of the study; (ii) national origin; (iii) research objectives; (iv) participants ; (v) research design; (vi) data collection instrument(s); (vii) major results. In the next step statements that explicitly described issues relevant to the attitudes towards OA journals were identified within each article. Using the QSR NVivo software package, interesting ideas were coded in a systematic way across the entire set of articles, and data relevant to each code were collated. The process of coding was part of the analysis, as data were organized into meaningful groups. After all data from the articles had been initially coded and collated, codes were analyzed and sorted into emergent themes. These themes represent the content of the entire set of articles examined. Theme identification followed an inductive approach, and codes were developed without adhering to a predetermined coding plan or the investigators' theoretical assumptions.

2. Results

The research objectives, participants, research design and the main findings of each study are presented in Table 1.

Table 1. The studies included in the meta-synthesis

Author(s)	Population	Research objectives	Participants	Research design	Data collection instrument	Main results
Coonin, 2011	US	To explore publishing practices and perceptions about OA publishing	1,293 Business faculty	Quantitative	Questionnaire	Limited awareness of OA journals. Low levels of self-archiving. Confusion regarding the issue of electronic journal versus print publishing.
Gul, Shah & Baghwan, 2010	Kashmir	To explore experience, attitudes and perceptions about the OA movement.	84 Science and Social Sciences faculty	Quantitative	Questionnaire	The concept of OA is still in the early stages. Differences in publishing practices between disciplines.
Hernadez-Borges et al., 2006	Spain	To evaluate familiarity with OA publishing and attitudes towards the author-pays model	100 authors of articles in PubMed	Quantitative	Questionnaire	Low awareness about OA publishing model. Respondents clearly rejected author fees due to lack of funding and knowledge about the prestige or reputation of OA journals.
Hoorn & van der Graaf, 2006	UK and Netherlands	To explore the authors' perspective on copyright issues in OA journals	355 authors of OA articles in biomedical journals	Quantitative	Questionnaire	A strong desire on the part of authors to change the present situation whereby authors transfer the entire copyright for their works to the journal publishers. The ideal license agreement is one that allows reuse for educational/scholarly purposes.
Ivwighreghweta & Onoriode, 2012	Nigeria	To examine the extent of researchers' appreciation of OA publishing	140 university lecturers	Quantitative	Questionnaire	High use of OA journals. The major constraints are unavailability of Internet facilities and lack of knowledge of the existence of OA journals.
Mammo & Ngulube, 2013	Ethiopia	To examine knowledge, use and attitudes towards OA journals	768 academics	Mixed methods	Questionnaire & interviews	High levels of knowledge and use. Positive attitudes towards OA journals, but some confusion about the issues regarding copyright and impact factor.

Author(s)	Population	Research objectives	Participants	Research design	Data collection instrument	Main results
Mischo & Schlembach, 2011	US	To examine faculty's OA practices and attitudes	54 Engineering faculty	Mixed methods	Questionnaire & interviews	A vast majority never published in author-pays journals and had limited plans to do so in the future. Concerns regarding the author-pays model and reluctance deposit in the institutional repository. An overwhelming consensus that commercial publishers should not pursue the Gold route
Nariani & Fernandez, 2012	Canada	To study the uptake of library support for author funding, the motivating factors and satisfaction with OA publishing	20 faculty who published in OA journals	Mixed methods	Questionnaire & interviews	Respondents were increasingly publishing in OA journals and were appreciative of library funding initiatives. Impact factor and readership were strong motivators for publishing.
Nicholas, Huntington & Rowlands, 2005	International	To explore the author views regarding OA publishing	3,787 authors from 97 countries	Quantitative	Questionnaire	General ignorance of OA publishing. Differences in opinion and practice between authors of different disciplines and countries.
Park & Qin, 2007	US	To explore motivating factors for publishing in an using OA journals	14 faculty members and doctoral students	Qualitative	Interviews	Perceived journal reputation, topical relevance, and availability are common incentives. Factors affecting publishing and use are interrelated.
Sanchez-Tarrago & Fernandez-Molina, 2009	Cuba	To assess knowledge about and attitudes towards the open access movement	160 researchers from 11 health institutions	Quantitative	Questionnaire	Low level of knowledge and unfamiliarity with OA initiatives and strategies. Low rates of publication in OA journals and self-archiving.
Schroter, Tite & Smith, 2005	International	To explore authors' attitudes towards OA publishing and author charges	28 international authors submitting to the *BMJ*	Qualitative	Interviews	Familiarity with OA. Low rates of publication in OA journals other than the BMJ. Positive attitudes towards OA publishing. Willingness to submit to OA journals. Dislike for author charges.
Schroter & Tite, 2006	International	To assess authors' knowledge and perceptions of OA publishing	468 international authors submitting to 3 biomedical journals	Quantitative	Questionnaire	Knowledge of OA publishing and author-pays models. Low rates of publication in author-pays journals. OA policies had little impact on authors' decision of where to submit papers.
Swan & Brown, 2004	UK	To compare opinions and experiences of OA authors and non-OA authors	311 authors	Quantitative	Questionnaire	High awareness of OA publishing opportunities. Less awareness of self-archiving. The main reason for publishing in OA journals was the principle of free access. The main concerns were grants and impact.
Utulu & Bolarinwa, 2009	Nigeria	To examine the extent of academics' awareness and use of OA initiatives as authors and readers	180 academic staff members	Quantitative	Questionnaire	High awareness of the pre-print and open access journal initiatives compared to the post-print initiative. Significant use of OA access initiatives. Differences in awareness and attitudes between disciplines.
Walrick & Vaughan, 2007	US	To identify motivating factors for publishing and attitudes towards OA	14 biomedical faculty	Quantitative	Interviews	Publication quality, free access and visibility are the most important incentives for selection of OA journals.

The majority of the studies employed quantitative methods to collect and analyze data: ten of them were questionnaire based surveys and one used interviews to collect

survey data. Two studies used qualitative approaches and other two employed mixed methods. Qualitative data were gathered through interviews, while a combination of interviews and questionnaires was used in the mixed methods studies. Five broad themes emerged from the analysis of the empirical literature.

2.1. Theme1: Awareness and experience of OA journals

Findings are mixed with respect to awareness of OA journals. Some studies [8, 9, 10, 11, 12] reported low levels of knowledge of the OA publishing model and the issues involved in it. In Gul, Shah & Baghwan's study [9] it was established that the majority of the respondents were aware of only two journals in their field, while in Coonin's study [8] the most popular answer to a question regarding how respondents became aware of OA was "first I've heard of it". On the contrary, in other studies [13, 14, 15, 16, 17] a significant percentage of the respondents were aware of OA publishing and the existence of OA journals. Even in these cases, however, there is evidence of unfamiliarity or confusion with the "author-pays" model [15] and with some features of OA, such as the open peer-review or the ability to attach supplementary data to the articles journals [18]. According to Nicholas, Huntington & Rowlands [11] researchers from the US, Australasia and Western Europe were more likely to report knowing nothing about OA, while researchers from Eastern Europe, South America and Asia were more likely to report awareness. The authors explained this difference by supposing that scholars based in countries with a strong publishing system do not need to know about alternative models. Hernadez-Borges et al. [19] and Sanchez-Tarrago & Fernandez- Molina [12] attributed the relatively low levels of knowledge among Spanish-speaking scholars to the fact that OA initiatives appeared initially in English-speaking settings. Colleagues seem to play an essential role in raising OA awareness [8, 9, 12, 13, 16, 18]. Self-knowledge is another common source of awareness [8, 12, 13, 16, 18]. Other sources of awareness are funding agencies [8, 9] and professional societies [8], while evidence about the role of the library is contradictory [8, 9, 12, 16].

A consistent finding of research in this area is the small number of authors actually publishing in OA journals. A number of researchers have found very low rates of publication in OA venues [9, 11, 12, 14, 15, 17, 20]. The most cited reasons for not publishing in OA journals were lack of familiarity with this type of publication and with methods for identifying OA journals to publish in [12, 15], economic constraints associated with the author-pays model [12], and perceptions that OA publications are of poor quality and not widely read [15]. Nicholas, Huntington & Rowlands [11] found that scholars' location was associated with the use of OA journals as a medium for research dissemination. The authors commented that low publication rates were observed in locations which had a poor commitment to OA publishing. The same study revealed a relationship between publishing in OA and previous experience of publishing on the web. Scholars making available their materials on the web or depositing them in institutional repositories were more likely to publish in OA journals.

2.2. Theme 2: Factors affecting decision to publish in OA journals

The principle of free access for all readers emerged as an important motivation for publishing in OA journals. Open access articles reach a much larger audience than any priced journal and increased usage means increased visibility for authors, raising their profile and the impact of their research, and creating opportunities for international

collaboration. As for the rest, scholars' decision to publish in OA venues is affected by the same factors determining journal choices in general. Of these, journal quality seems to be central to decision making and takes priority over the availability of open access.

> As reported by some of the participants, requests for electronic copies of their publications by individual researchers, often in foreign countries, remains quite frequent and in some cases has prompted an interest in the open access movement. [21, p.12]

> One faculty member (…) decided to publish in a BMC journal at the suggestion of her collaborators from a developing country. In another instance, a health science researcher, whose research has been focused on native communities in Ontario, wanted her paper to be read by aboriginal community researchers and hence decided to publish in an OA journal. The same author mentioned that she has started collaborating with researchers from Malaysia after reading their article in a BMC journal. [18, p. 187]

Theme 3: Perceptions of OA journals

Clearly in the mind of most scholars the strongest characteristic of OA journals is that they promote improved and more equitable access to knowledge by all kinds of readers and availability of research papers to the developing world [8, 11, 12, 14, 21]. In a number of studies open access is perceived to facilitate wider diffusion of research outputs and increase the impact of researchers work [9, 12, 16, 21].

Other benefits derived from open access are considered to be faster publication times [8, 9, 12, 14, 21] and reduced costs, especially in terms of subscriptions to traditional journals, but also in terms of time savings, photocopying and interlibrary loans [8, 15]. Swan & Brown observed differences in perceptions among OA journal authors and those who had not published in OA venues.

> Over 90% of OA authors published in this way because of the principle of free access. They also associate other values with publishing in open access journals: they perceive them to be faster than traditional journals, to have a larger readership and consequently to be cited more frequently, and to have high prestige and quality than traditional journals available to them. The perceptions of NOA authors tend to be opposed: they perceive open access journals as having a smaller readership and lower citation rates, and of generally being of lower quality and prestige than the traditional journals they publish in. [16, p. 223]

Copyright ownership, which is perceived to give more control over authors' publications, has been reported as an advantage from the faculty members interviewed by Nariani & Fernandez [18]. As one researcher put it "I like OA journals because anyone can download these papers and I can use them as examples for teaching purposes. Students don't need to pay for it" (p. 189).

Besides perceived advantages, OA usually raises some concerns too. A number of studies have demonstrated that open access publications are often considered to be of low quality and consequently less respected and prestigious than established, subscription-based journals [8, 10, 11, 15, 16, 21]. However, despite perceptions of poor quality, researchers believe that publishing in OA helps career development and should not be viewed as an obstacle to tenure and promotion [11, 18,

21]. Also, it appears that discussions about OA always bring up the issue of impact factor [16], and there is a belief that OA publications have lower impact factors than traditional journals or no impact factors at all, a problem discouraging many authors from publishing in such journals [15, 21]. In addition, there is evidence that OA is mixed up with peer review, with many researchers assuming that OA journals have inferior peer review, something that might lead to vanity publishing [11, 15].

2.3. Theme 4: Author charges

Views about author charges were found to vary, ranging from rejection to tolerance and even acceptance. Across several studies, respondents were mostly against author charges and would hesitate to submit to journals operating under the author-pays model [10, 14, 15]. The author-pays system is regarded as an additional barrier to researchers, and one that might reduce publishing opportunities for underfunded or young researchers and researchers from the developing world [16, 20]. The main concern of researchers seems to be how the fees will be paid, and support from grant agencies and institutions plays a significant role in shaping attitudes towards author charges. When funding agencies or universities cover the cost of publishing, author charges are acceptable [15, 18, 21]. In some instances respondents expressed concerns that author charges may deteriorate the review process [8, 20]. There is also evidence that quality of the journal might alleviate the unwillingness of many authors to publish under an author-pays model [14, 15].

2.4. Theme 5: The role of the discipline

Discipline appears to play a role in awareness of and attitudes towards OA journals. One of the most interesting findings of Nicholas, Huntington & Rowlands was that Science and Technology scholars were more likely to report knowing a lot about OA than their counterparts in Arts and Humanities who were more likely to report knowing nothing at all [11]. The authors attributed this difference to the fact that "scientists as a whole are more active in journal publishing and also in the frontline of OA developments" (p. 516). These findings are contradicted by a study conducted in Nigeria, which found that awareness of OA journals was higher among academics in the Humanities. In the same study, however, academics in Sciences showed increased willingness to adopt OA both as users and authors [17]. According to Gul, Shah & Baghwan Science scholars were more active in using and publishing in OA journals than their colleagues in Social Sciences. In the same study, Science scholars were found to be more familiar with OA content retrieval methods as compared to Social Science scholars.

The discipline also has an impact on how scholars view publication charges. Medical sciences authors seem to be less concerned with author fees, because many traditional, journals in these fields have long established pricing practices and charge authors without making their articles freely available. On the contrary, author publication fees are less common among social sciences publications.

3. Conclusion

OA has and continues to change the ways of scientific research, literature search, journal editing, publishing, and archiving [24]. The aim of this work was to synthesize the results of previous studies concerning scholars' attitudes toward OA and provide an overview of OA from the point of view of current and potential authors. This meta-synthesis reveals that scholars hold positive views towards OA journals. Ii also indicates that, although academic researchers are aware of the fact that OA journals can bring many advantages in research visibility and impact, OA publishing is not yet fully understood neither has it reached its full potential. Although academics and authors appreciate the benefits of free dissemination of information and advocate the moral argument of unrestricted access to scientific research, they have some concerns about the author-pays model and they question the quality, reputation and impact of OA journals. The issue of journal prestige is of great importance for authors because, among other things, is closely related to tenure and promotion. Journal reputation and perceived quality seem to be more important factors considered by scholars when selecting a journal to publish in than whether it is open-access. In the past few years there has been an ongoing debate over the quality of OA journals, and a frequent criticism of OA is that it will threaten the peer-review system, diminishing the overall quality of scientific journal publishing. Recent research indicates that negative perceptions of the quality of OA publishing are not well grounded. A study comparing the scientific impact of OA journals with subscription journals showed that OA journals indexed in Web of Science and/or Scopus are approaching the same scientific impact and quality as subscription journals [25]. Similar conclusions have been drawn from a SOAP study, according to which "OA publishing is a mature field with similar patterns and quality indicators as non-OA publishing" [26, p. 13].

The other unfavorable aspect of OA journals is author-payment. Many scholars appear to think that OA journals charge authors. They are generally unwilling to pay author charges and requiring them to cover publication costs is a serious disincentive to OA publication, especially in fields where the vast amount of research is self-funded, funded by the author's home institution or the funding is too small. It is interesting that in almost all studies open access and author charges were considered as identical, even though many OA journals waive publication fees for authors who do not have access to grants and funding, and many authors have claimed that most OA journals do not charge authors for publication [27, 28]. In fact, the majority of the journals listed in the Directory of OA Journals do not actually charge author-side fees but they rely on alternative sources of revenue [29]. The false assumption that all OA journals are fee-based adds to the misconceptions about open access and distorts the current OA publishing landscape.

The articles used in this meta-synthesis cover a time-span of twelve years. There is evidence that authors' understanding and practices concerning OA have changed over time. Rates of publication in OA journals and familiarity with OA publishing models seem to have increased. For example, there is a clear difference between the results reported nearly ten years ago by Nicholas, Huntington & Rowlands [11], and those reported in a very recent study by Nariani & Fernandez [18]. There is also evidence that the country differences observed by Nicholas, Huntington & Rowlands [11] continue to exist, with scholars in developing countries, like Ethiopia or Nigeria, demonstrating higher levels of OA journals knowledge and use.

Authors views should be taken into consideration by the key stakeholders of open access publishing. Attitudes and perceptions will determine the success and acceptance of this evolving model. Libraries can play an important role in connecting authors with OA movement by clarifying confusions, raising awareness of OA journals, and informing researchers of their publishing options. Librarians, who have long been calling for a change in the existing system, can communicate to both users and administrators the advantages of OA and its potential to address some of the problems surrounding the traditional publishing model.

One limitation of the present study is that it does not address all the issues associated with OA because the data are limited to the articles selected for inclusion. Another limitation has to do with the fact that quality assessment was not used for the exclusion of articles and so it is possible that some of the publications examined could be questioned for the quality of their methodology and the strengths of their findings.

References

[1] K.L. Palmer, E. Dill & C. Christie, Where there's a will there's a way? Survey of academic librarian attitudes about open access, *College & Research Libraries* **70** (2009), 315-335.

[2] Budapest Open Access Initiative. Available from http://www.soros.org/openaccess/read.html [Accessed 22 May 2008].

[3] J. Allen, Interdisciplinary differences in attitudes towards deposit in institutional repositories, *Education and Information Technologies* (2005). Available from http://eprints.rclis.org/6957/1/FULLTEXT.pdf [Accessed 12 December 2013].

[4] M. Kenneway, *Author attitudes towards open access publishing* (2011) Available from http://www.intechopen.com/public_files/Intech_OA_Apr11.pdf [Accessed 12 December 2013].

[5] R. Anderson, Author disincentives and open access, *Serials Review* **30** (2004), 288-291.

[6] B-C. Bjork, Open access to scientific publications: an analysis of the barriers to change? *Information research* **9** (2004). Available from http://InformationR.net/ir/9-2/paper170.html [Accessed 18 November 2013].

[7] C.R. Bair & J.G. Hawarth, Doctoral student attrition and persistence: a meta-synthesis of research, *Higher Education: Handbook of Theory and Research* **19** (2004), 481-534.

[8] B. Coonin, B., Open access publishing in business research: the authors' perspective, *Journal of Business & Finance Librarianship* **16** (2004), 193-212.

[9] S. Gul, T.A. Shah & T.A. Baghwan, Culture of open access in the University of Kashmir: a researcher's viewpoint, *Aslib Proceedings: New Information Perspectives* **62** (2010), 210 – 222.

[10] A.A. Hernandez-Borges, Awareness and attitude of Spanish medical authors to open access publishing and the ''author pays'' model, *Journal of Medical Library Association* **94** (2006), 449 – 456.

[11] D. Nicholas, P. Huntington, P. & I. Rowlands, I., Open access journal publishing: the views of some of the world's senior authors, *Journal of Documentation* **61** (2005), 497 – 519.

[12] N. Sanchez-Tarrago &J.C. Fernandez-Molina, The open access movement and Cuban health research work: an author survey, *Health Information and Libraries Journal* **27** (2009), 6 – 74.

[13] Y. Mammo & P. Ngulube, Academics' use and attitude towards open access in selected higher learning institutions of Ethiopia, *Information development* (2013), 1 – 14.

[14] S. Schroter & L. Tite, Open access publishing and author-pays business models: a survey of authors' knowledge and perceptions, *Journal of The Royal Society of Medicine* **99** (2006), 141-148.

[15] S. Schroter, L. Tite & R. Smith, Perceptions of open access publishing: interviews with journal authors, *BMJ* (2005). Available from http://www.bmj.com/content/330/7494/756 [Accessed 18 November 2013]. [16].

[16] A. Swan & S. Brown, S., Authors and open access publishing, *Learned Publishing* **17** (2004), 21– 224.

[17] S.C.A. Utulu & O. Bolarinwa, Open access initiatives adoption by Nigerian academics, *Library Review* **58** (2009), 660 – 669.

[18] R. Nariani & L. Fernandez, Open access publishing: what authors want, *College & Research Libraries* **73** (2012), pp. 182 – 195.

[19] A.A. Hernandez-Borges, Awareness and attitude of Spanish medical authors to open access publishing and the ''author pays'' model, *Journal of Medical Library Association,* **94** (2006), 449 – 456.

[20] W.H. Mischo &M.C. Schlembach, Open Access issues and engineering faculty attitudes and practices, *Journal of Library Administration* **51**(2011), 432-454.
[21] S.E. Warlick & K.T.L.Vaughan, Factors influencing publication choice: why faculty choose open access, *Biomedical Digital Libraries* **4** (2007). Available from http://www.bio-diglib.com/content/4/1/1 [Accessed 18 November 2013].
[22] O. Ivwighreghweta & O.K. Onoriode, Open access and scholarly publishing: ppportunities and challenges to Nigerian researchers, *Chinese Librarianship: an International Electronic Journal* **33** (2012) Available from http:// www.iclc.us/cliej/cl33IO.pdf [Accessed 18 November 2013].
[23] J.H. Park & J. Qin, Exploring the willingness of scholars to accept open access: a grounded theory approach, *Journal of Scholarly Publishing* **38** (2007), 55-84.
[24] A.Y. Gasparyan, L. Ayvazyan & G.D. Kitas, Biomedical journal editing: elements of success, *Croat Med J* **52** (2011), 423–428.
[25] B-C. Bjork & D. Solomon, Open access versus subscription journals: a comparison of scientific impact, *BMC Medicine* **10** (2012), 73, doi:10.1186/1741-7015-10-73.
[26] S. Dallmeier-Tiessen et al., *The landscape of Open Access Publishing today*, Talk presented at SOAP Symposium (2011). Available from http:// http://edoc.mpg.de/524966 [Accessed 2 March 2013].
[27] B. Hooker, If we won't sink in, maybe we can pound it in…, *Open Reading Frame* (2007, December). Available from http://sennoma.net/?p=555 [Accessed 20 March 2014].
[28] S. Shieber, What percentage of open-access journals charge publication fees?, *The Occasional Pamphlet* (2009, May). Available from http://blogs.law.harvard.edu/pamphlet/2009/05/29/what-percentage-of-open-access-journals-charge-publication-fees/ [Accessed 20 March 2014].
[29] M. Kozak & J. Hartley, Publication fees for open access journals: different disciplines-different methods, *Journal of the American Society for Information Science and Technology* **64** (2013), 2591 – 2594.

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-23

The implementation of the European Commission recommendation on open access to scientific information: comparison of national policies

Lisiane LOMAZZI[a] and Ghislaine CHARTRON[b]
[a] *CNAM-DICEN*
[b] *CNAM-DICEN*

Abstract. Two years after the publication of the European Commission recommendation on open access to scientific information, the critical threshold of accessibility to fifty percent of papers has been crossed. However, this figure is an average and the implementation of the EC recommendation varies from one country to another. The topical issue now is to observe the different steps of implementation and to wonder about the reaseons of such a disparity. In order to suggest many elements of the response, this research compares the different levels of implementation in the EU28.

Keywords. Academic publishing, open access, scientific information, policy, European Commission

Contrary to what the European Commission might expect further to its communication [1] and its recommendation [2] (concerning open access to and preservation of scientific information within the framework Horizon 2020) after being published dated 17th July 2012 its implementation by national governments and EU research funders have not led to a standardization of open access policies. This recommendation has undergone all manner of implementations concerning the level of incentive, the contents which are concerned, the embargo periods, etc.

First and foremost this paper propose doing a comparison between the national implementations of the CE recommendation in the EU28. The suggested analysis is a good example of its various interpretations and implementations. We compare the adopted action plans and their methods : mandatory deposit and national recommendation, delegation to each institution and research funder, national consultation of stakeholders'opinion, no policy at all.

1. Methodology

This study was conducted from bibliographical ressources on open access in the EU28 collected via the search engine called BASE [3] and other information from the OPENAIRE [4] portal and the UNESCO Global Access Portal [5].

2. The implementation of the recommendation at national level

Despite the EC recommendations we notice that there are four levels of implementation : no national open access mandate and policy, consultation in progress to implement a national policy, funders mandates and policy, coordinated national policy by a recommendation or an act.

2.1. No national open access mandate and policy

The european countries that have not implement a national open access policy are : Romania, Cyprus, Greece, Estonia, Bulgaria, Malta, Slovakia, Lithuania, Czech Republic, Luxembourg.

Those countries present some common characteristics that explain the status quo in the national implementation of the european open acces policy. First, there are all (except from Estonia, Luxembourg and Czech Republic) countries that have gross domestic expenditures on research and development as a percentage of gross domestic product less than 1 [6] while the lower percentage is 0 and the higher is 3.5. Second, they are countries that publish less than 1 000 scientific articles per year except from Greece and Czech Republic. In short, there are quite small stakeholders on the european research scene.

We can easily deduce that in spite of the later realizable budget savings thanks to an open access to scientific publications [7], those countries cannot afford to set up infrastructures and open access funds. In some cases, the needed infrastructures exist but the will to implement an open access policy comes up against the lack of researchers awareness or an insufficient demand caused by the number of published articles at national level.

2.2. Consultation in progress in order to implement a national policy

Four european countries have not implemented a coordinated national policy yet but are on the right track. Indeed they launched a national consultation with all the stakeholders that should lead to the proposition of a bill.

In Poland, a national consultation about open access to public ressources was set off by Minister of Administration and Digitalization in 2012 [8]. Its aim was to define open access policy guidelines that will be integrated in a bill including open access to educative, cultural and scientific resources which will be publicly funded : the “Act on Open Public Resources”. The fear not to afford open access gold in the long term leads to favour green open access.

In Slovenia, the Research and Development Act states that results from publicly funded research must be accessible. The aim of the first period from 2011 to 2014 of the *Resolution on the National Research and Development Programme 2011-2020* [9] was to launch a large national consultation with every stakeholder in order to establish some guidelines to a future bill that would include data too. The *Plan on the National*

Research and development Programme 2011-2020 [10] also mention the connexion of all national repositories in CRIS (SICRIS [11]).

In the Netherlands, since 2009, Universities Rectors clearly indicated their commitment in favour of open access by conversing about the means to encourage the open access implementation. The NWO, an independent research body which funds research and one of the biggest dutch funders, leads a strong policy in favour of open access notably the gold road by funding subsidies granting programmes to pay the author fees. For the time being there is no project of open access implementation policy but only a national consultation.

In France, even if the Geneviève Fioraso's speech, Minister of Higher Education and Research, delivered on the 24th January 2013, indicated that «the French government reaffirm[ed] its support to open access to scientific information principle », the implementation of a mandatory open access policy is not approved unanimously notably among publishers in SHS. A national consultation was launched recently by the ministry of Higher Education and Research in order to establish what is the optimal embargo period for SHS journals. Currently, there are five mandatory deposit policies (IRSTEA, IFREMER, CIRAD, INRA, INRIA) and two national funders incitative policies (CNRS, INSERM) [12].

2.3. Funders mandates and policy

Currently, in the UK, the gold road is more plebiscited than the green one even if the latter is not deserted. The Research Council UK, a consortium of seven independent research councils, set up a gold open access policy. This policy was examined and an intermediate report [13] and is going to be reconsidered in the autumn 2014. Sixteen others funders also have their own open access policy, the list is available on SHERPA/RoMEO [14].

In Denmark, on the 22nd June 2012, the five principal national funders (Danish Council for Independent Research (DFF), the Danish Council for Strategic Research, the Danish National Research Foundations, the Danish Advanced Technology Foundation, and the Danish Council for technology and innovation) decided a common open access policy. This policy requires the deposit of a digital version of research articles in open archives within the six or twelve months after the article acceptance. Seven universities out of eight have an open access policy. However, it is often more a declaration of intent than a real mandate.

In Finland, even if the open access principle has been encouraged for a long time, concrete actions came into being just recently. In 2011, Minister of Education and Culture launched a project named TTA with the aim to create an open access national scientific policy and to build the necessary infrastructure. Currently, a national bill has circulated among the different stakeholders so they can make comments on it. This bill recommends either gold road or green road but sets aside hybrid publications. An open access funding has been set off. The Science Academy that is the main funder recommands to researchers to publish in open access journals as often as possible.

In Sweden, two major funders, the Swedish Research Council and the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS), instituted open access mandate (green open access mandate concerning

peer-reviewed articles to deposit in open archives within six months after publication) and the Association of Swedish Higher Education (SUHF) which recommands open access to its 41 institutions members and encourages these to institute their own open access policy. Recently, the Swedish Research Council (SRC) has been appointed by ministery in order to establish the guidelines of a national policy in favour of open access. The first version of this report should be published by the end of 2014.

In Austria, the open access movement began in 2009. For two years, the rate has speeded up with the creation of some funders mandates notably the one of the Austrian Science Fund (FWF) that recommends to researchers to publish in open access journals, the fees being paid back by the dedicated fund and to deposit an electronic version in open archives within twelve months after publication. The Austrian Academy of Sciences (OAW) has a green open access policy but also has a publishing house that publishes gold open access journals and books. Others policies/institutional mandates should be set up soon but there is currently no expression of a need concerning the implementation of a national open access policy [15].

Hungary has a national research environment particularly active, fostered by the government and the Scientific Hungarian Research Funding (OKTA) which is the major funder. The OKTA policy encourages open access by requiring that the funded researchers publish in open access journals and deposit an electronic version into open archives. The only current open access government decree is about doctoral thesis (n°33, 7th March 2007).

2.4. National policy coordinated by a recommendation

In Belgium, it is really difficult to set up an open access national policy owing to the federalism that clearly complicates the coordination between different regional research environments, publishing stakeholders and linguistic issues. Nevertheless, the two major research funders FWO in the Flemish Community and FNRS [16] in the French Community both have a green open access mandate adopted in 2013 that needs a deposit of researchers'publications in open archives. A first step towards was the implementation of a national open access policy with the Brussels Declaration [17] on the 22nd October 2012 the signatories were the official ministers representatives of Walloon Region, Brussels Region and Flemish Region. This Declaration defines a belgian open access policy. The signatories committed themselves to encourage open access to the publicly funded research results by informing the researchers, by recommanding them to make their publications available at the latest six months (STM) and twelve months (SHS) after publication, by examining the possibilities for the public funds to pay the open access publication fees, by encouraging the creation and preservation of deposit infrastructures, by thinking about the risks and opportunities of each open access road with the stakeholders. This dialogue has turned into a national consultation and the publishers syndicate is going to sign an agreement with universities that could lead to embargo periods from six to twelve months and even more for the publications in Humanities and Social Sciences.

In Ireland, there are four national open access funders mandates (Higher Education Authority, Health Research Board, Irish Research Council for Science, Engineering and Technology) out of the seven national funders. The government announced on the

23rd October 2012 what were the open access national principles in *National Principles for Open Access Policy Statement* [18]. Among the major principles, we found a deposit obligation for scientific research publicly funded publications and an incentive to publish in open access journals. This recommendation favors the green road but does not definitely set aside the gold one. That fits with the creation of a dedicated fund in order to set up institutional deposits and a national portal whereas no specific fund has been launched to finance the gold road.

In Portugal, some open access initiatives have been set up since 2004. Although the Portuguese government, the public and private funders have officially not announced open access policies or mandates yet, the Conference of Portuguese University rectors (CRUP) recommends to the research bodies to implement a mandated repository policy for research publications and data. The CRUP trusts to the generalization of an only open access european mandate that could lead to a lack of a national mandate implementation.

In Croatia, there is a scientific open access community, particularly active through four institutional repositories and one national portal that makes accessible more than 250 scientific croatian journals (HRCAK). Currently, there is no croatian open access funder mandate. The document *Science and technology policy of the Republic of Croatia 2007-2010* issued by the ministry of Science , Education and Sports mentions that the publicly funded research results have to be accessible to the general public thanks to open access publications or databases. On the 24th October 2012, a national declaration was publicized [19].

2.5. National policy coordinated by a law

In Latvia, the adoption of the national reform programme for the implementation of european strategy « Horizon 2020 » by the Latvian Cabinet have not led to the adoption of open access policies or mandates by the funders or the government in the long term. However, this programme mentions an obligation to deposit publicly funded research publications into repositories (embargo period up to six months in STM and twelve in SHS) and the creation of subsidies for gold open access journals.

Spain was the first state to legislate on open access, from 2011, with the « Ley de la Ciencia, la Tecnología y la Innovación [20] ». The implementation of this law is not very much prejudicial to publishers insofar it maintains the editorial embargo as it is mentionned in article 37 paragraph 3.

In Germany, the law dated July 2013 about orphan and unavailable works includes a clause about open access. This clause gives to the authors a right of secondary publication. This allows to take similar but non-profit publication by the author twelve months after the article acceptance in STM and SHS. This right is applied if the research work is publicly funded and if the article is accepted in a journal that is published at least twice a year. This settlement affirmed its superiority on the contract.

In Italy, in March 2013, the major research bodies Presidents, associated with the Conference of Italian University rectors signed a declaration in favour of open access. In October 2013, the legislator intervened on open access regarding a decree-law about preservation and restoration of cultural goods. However, whereas the initial bill planned an open access to the articles six months after publication, the bill which was

adopted on the 8th October 2013 requires embargo periods of 18 months in STM and 24 months in SHS and books are not concerned. This modification of the first version of the bill is the consequence of an important work of lobbying that was done by private italian publishers who considers that a six months embargo period is insufficient to assure the economic viability of publications.

As a conclusion, it's important to be aware of the fact that this research is a snapshot of a situation at a given time. Indeed, the data evolve with time and need to be reactualized permanently.

However, at the end of this research, we notice that imbalances have emerged since the beginning of the EC recommendation implementation. That brings us to the question of who exactly is really benefiting from Open access, the countries that lead the world in scientific output or these that run behind ? In order to answer to this question, two specificities need to be considered : the specific language of papers production and the scientific discipline anchorage either in human sciences or in hard sciences. As a consequence, this issue will be the subject for further research on the future of non-English-speaking national publishing in the context of the EU recommendation.

References

[1] European Commission communication towards better access to scientific information. http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_fr.pdf

[2] European Commission recommendation on access and preservation scientific information. http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

[3] BASE - bielfield academic search engine. http://www.base-search.net/about/en/

[4] OpenAIRE - Open Access Infrastructure Research for Europe. http://www.openaire.eu/fr

[5] Global Open Access Portal. http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/access-by-region/

[6] OECD (2013),*Table 2. Dépenses intérieures brutes de R-D (DIRD) en pourcentage du PIB, inPrincipaux indicateurs de la science et de la technologie*, OECD Publishing. doi: 10.1787/msti-v2013-1-table2-fr

[7] Houghton, J., 'Open Access – *What are the Economic Benefits? A comparison of the United Kingdom, Netherlands and Denmark*', published by Knowledge Exchange, 2009. http://www. knowledge-exchange.info/default.aspx?id=316

[8] SPARC Europe news. http://sparceurope.org/polish-ministry-of-administration-digitization-great-debate-on-open-public-resources/

[9]Resolucija o raziskovalni in inovacijski strategiji Slovenije 2011–2020. http://zakonodaja.gov.si/rpsi/r08/predpis_RESO68.html

[10] Načrt razvoja raziskovalnih infrastruktur 2011–2020. http://www.arhiv.mvzt.gov.si/fileadmin/mvzt.gov.si/pageuploads/pdf/znanost/RISS/NRRI.pdf

[11] Slovenian Current Research Information System. http://www.sicris.si/default.aspx?lang=eng

[12] Jonchère, L., *Synthèse sur les politiques institutionnelles de libre accès à la recherche, 2013*. http://archivesic.ccsd.cnrs.fr/docs/00/80/11/88/PDF/Synthese-politiques-LA-Jonchere-fev-2013.pdf

[13] The implementation of open access Report, House of lords, Science and Technology Committee, 3rd Report of Session 2012–13 http://www.publications.parliament.uk/pa/ld201213/ldselect/ldsctech/122/122.pdf

[14] SHERPA-RoMEO. http://www.sherpa.ac.uk/juliet/index.php

[15] Austrian Government (2013) Chapter on "Open Access" in the Austrian Research and Technology Report 2013

[16] FRS-FNRS reglement on Open Access. http://www.frs-fnrs.be/uploaddocs/docs/SOUTENIR/FRS-FNRS_Reglement_OPEN_ACCESS.pdf

[17] Brussels declaration on Open Access. http://openaccessbelgium.files.wordpress.com/2012/10/brussels-declaration-on-open-access.pdf
[18] National Principles for Open Access Policy Statement, Committee of Irish research organisations. http://www.tcd.ie/Library/assets/pdf/National%20Principles%20on%20Open%20Access%20Policy%20Statement%20(FINAL%2023%20Oct%202012%20v1%203).pdf
[19] Croatian Open Access declaration. http://www.fer.unizg.hr/oa2012/declaration
[20] Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación, https://www.boe.es/boe/dias/2011/06/02/pdfs/BOE-A-2011-9617.pdf

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-30

Information needs of researchers in a bibliographic databases environment: a literature review

Morgana Carneiro de ANDRADE[a,1]
[a]Doctoral Program in Technology and Information Systems – University of Minho, Portugal
[b]Ana Alice BAPTISTA
[b]Algoritmi Research Center / Information Systems Department – University of Minho, Portugal

Abstract. This article presents a literature review whose aim was to identify the reported information needs of researchers when they consult bibliographic databases. Initially, 192 articles were retrieved using Scopus, Web of Science and Google Scholar databases. After applying the criteria for exclusion, the number of articles was reduced to 16, which is already an indicator of the small number of studies on this specific topic. The results show that it is hard to identify the information needs of researchers. They also show that the researchers have been requiring information with a higher degree of granularity. We conclude that although the available studies provide important information about the researchers' information needs and hints on how to address them, there is a need for more in-depth studies. The results of these deeper studies may be useful to serve as an indication for the creation of new procedures and tools, including those based on new metadata elements drawn to improve search results on Linked Open Data tools.

Keywords. Information needs. Bibliographic databases. Metadata.

Introduction

The theme of information needs is present since the first studies in the field of Library and Documentation Science, and subsequently in Information Science. With the advent of the Internet there was an increase of studies on this topic, especially with focus on information services, such as digital libraries and bibliographic databases.

According to Kuruppu & Gruber [1] "understanding information needs, information-seeking behavior and information use of researchers is challenging" and it gets more complicated as they play several roles (researcher, teacher, administrator,...), their needs and interests change over time and they are "continuously" affected by technological advances.

The huge amount of information present on the Internet and the diversity of services have paradoxically contributed to hinder the identification of the most relevant

[1] Information Systems Department, University of Minho, 4810-058 Guimarães, Portugal. Email: morganaandrade@hotmail.com

papers. In order to find relevant information in less time, it is required from the user that he knows "[...] what to get, from where to get, how to get it" [2]. These questions correspond to what, in the field of Library and Information Science (LIS), is called the informational need.[2]

One of the forerunners in this area was Taylor [3], who in the article "The Process of Asking Question" brought insights on information needs that are relevant till today. The author proposes four levels of information needs:

- First level - the conscious and unconscious needs, which when identified refer to the "perfect question";
- Second level - the conscious needs that are poorly defined which will be made understandable from interactions with other people;
- Third level –the conscious needs that are well defined, but that may not be properly input into the information system;
- Fourth level –The conscious needs that are well defined and may be "translated" into the language of the information system in a way they can be processed.

From these approaches, Taylor lists some aspects that affect the man-machine relation: a) system organization, which includes input and output characteristics; b) types, complexities and characteristics of the subject related to the question; and c) the researcher's competence.

The "internal organization", part of the system organization and its input characteristics, as the author understands it, corresponds to the access points, which in his view are related to the degree of sophistication in the use of terms, the depth of analysis and indexing and the level of specificity. These access points could assume a "multidimensional space", ranging from empirical data to theoretical concepts, by way of descriptive data, experimental evidence, historical material, results analysis, interpretation of descriptive categories of information. For Taylor, the way the information service can be exploited has implications on the way the researcher formulates his questions and on number of relevant answers he gets from the system.

In Taylor's second to forth level it is present the role of the librarian as an interpreter and a translator of user needs to the system. Due to technological advances and the familiarity of researchers with the technology, this role of the librarian is becoming less present. Because users do not resort so often to the presence of the librarian, information services need to offer functionalities that meet their needs by providing access points that allow them to retrieve relevant documents. Therefore, for this literature review, we are especially focused on one of the items identified by Taylor: the way the services are internally organized, classified, and indexed, and their access points. For this purpose we consider that the researcher is aware of what he wants, but he depends on the "internal organization" of information services for his information needs to be met.

In this sense, the identification of access points that meet the user's needs may be useful to serve as an indication for the creation of new procedures and tools, including those based on metadata elements drawn to improve search results on Linked Open Data tools. The tendency to increase granularity at the description level allows that initiatives such as the W3C's "The RDF data cube vocabulary", "Data Catalog Vocabulary", linked data can be more easily adopted.[4-5] Cyganiak et al. [4], by publishing guidelines at the W3C initiatives to the multidimensional publication of data, confirm the thought of Taylor [3], aired 52 years ago as Taylor also stressed the need for a multidimensional space to the empirical data and theoretical concepts. Also according to Ismail & Kareem [6], the Web does not provide support for the novice

researcher and thus one of the solutions might be that information services become semantically interoperable. One suggestion to resolve some of these issues was the use of semantic web technologies.

This article has the following structure: Section 1, Research Design, which contains the description of the strategy for conducting the search. Section 2, Results, which are presented and discussed in order to clarify what has been developed so far to meet the information needs of researchers. Section 3, Conclusion, which brings the final considerations.

1. Research Design

We started by identifying studies addressing the needs of researchers when consulting bibliographic databases and what has been developed to meet this need. The databases used for the search were: Scopus, Web of Science, Networked Digital Library of Theses and Dissertations (NDLTD), Library & Information Science, and Technology Abstracts (LISTA) and Google scholar. The keywords used were: information needs x bibliographic database; information seeking behavior x scientific communication; user profile x scientific information; scientific articles x user's needs; retrieval x bibliographic database x user's needs; user studies x bibliographic database; information needs x scientific communication.

The above terms were used in the fields of keywords/topics in databases, and in Google scholar the title field was used. When searching the databases, we have made the following choices: temporal limits: none; language: English, Portuguese and Spanish. The same procedure was used for Google scholar, but in this case the results' analysis was limited to the first 100 most relevant articles.

The papers resulting from this search were selected based on the titles and abstracts (step 1). The articles cited in these studies were also selected based on titles and abstracts (step2). We adopted the same procedure to select the articles that cite the ones retrieved in the first step (step3). This way, we sought an outcome with greater coverage but without losing relevance to the theme.

In order to maintain consistency of the proposed study we excluded articles that we considered to be out of scope, such as the ones on: software applications, information behavior, relevance analysis, technical analysis of information retrieval systems. In addition, we identified a few studies whose approach was more general, and which analyzed aspects like kind of used sources, language or subject.[7-8] In this case, we chose not to include these articles in the results, for not bringing relevant information as to achieve the purpose of this literature review. Inclusion was restricted to scientific articles. Based on these criteria, the number of articles that were actually aligned with the objective of the present study was reduced from 192 to16.

2. Results

The analysis of the 16 articles results in a panorama of what has been researched on the researchers' information needs when seeking in bibliographic databases as presented in Table 1.

Table 1. Subjects and aspects approached in articles

AUTHORS \ ASPECTS	Subject	Components	Domain	Indexing/Metadata	User Profile
Amato & Straccia, 1999	IN				X
Bates, Wilde, & Siegfried, 1993	ISB		X		
Bates, 1996	ISB		X		
Bishop, 1998	ISB	X	X	X	
Bishop, 1999	IN	X		X	
Borgman, 1986	IN			X	
Courtright, 2007	ISB				X
Crowston & Kwasnik, 2003	IN			X	
Dogan et al., 2009	IN		X		
Hjørland, Nielsen, & Williams, 2001	ISB	X		X	
Ismail & Kareem, 2011	IN			X	
Lee & Downie, 2004	IN			X	
Markey, 2007a	ISB		X		
Markey, 2007b	ISB		X		
Rowlands, 2007	ISB		X		
Sandusky & Tenopir, 2008	ISB	X		X	X

Note: IN – information needs; ISB – Information Seeking Behavior

The four main aspects that arise from this analysis are: components of articles, indexing and metadata, domain, and user profile.

- Use of components of scientific articles as a way to enhance search results

  In the context of this study, components of articles represent physical or logical structures of a document.[9-11] Tables and figures are examples of physical components, whereas the data resulting of some experiment represent logical structures or narratives.[10]

  Bishop [10] conducted a study that examined how the components of a scientific paper are identified, stored and utilized by users in digital libraries. In her study she used DeLIver which is aimed to allow researchers at the University of Illinois to search for components of documents. Bishop [11] reported that researchers appreciate the use of specific components in specific situations. The researchers also demonstrated the importance of using the components to decide which articles resulting from the search process should be read. Bishop's research, as she discusses, is aligned with the principles advocated by Paul Otlet.[10-11]

  According to Otlet, as chemistry researchers moved from the analysis of molecules to atoms, efforts should converge to think of ways to allow science to have access to specific parts of the content of the publications. In a visionary way, in the early decades of 1900, Otlet said: "methods will be found to index works quickly and completely in order to permit the retrieval, instantly and without trouble or difficulty, of the substance of what each publication contributes to knowledge" [12]. Rayward [13] adds that these statements referred to the atoms of information that could be reconfigured in order to meet the information interests and needs of users.

  Hjørland, Nielsen, & Williams [14] cite Bishop's results to highlight the "discussion of the need to replace the traditional linear structures in documents with a free combination of 'info-bricks'", which are "the extraction of

individual facts and ideas as separate units" [10]. The authors also mention the studies of Al-Hawamdeh et al., Al-Hawamdeh and Willett and Lalmas and Ruthven, whose studies support the usage of components of texts.

- Relationship with the domain

The domain, is considered by several authors to be a factor that interferes with the researchers' information needs and with the way the users proceed when searching. For example, it could be expected that faceted search conjugated with Boolean operators in online databases would return in better results independently of the domain of the researcher. However, according to Bates [15] research in the humanities, end users find it difficult to perform those kinds of searches, as well as finding optimal results.

This idea is reinforced by Markey [16-17], which considers that experts in a particular field seek high degree of accuracy, limiting the search field. Their strategies are based on identifying clues contained in any word or phrase in the title, the name of an author, a variable, a test or a particular research center, reducing the number of items retrieved, but getting high degree of relevance.

The domain is also related to the modification of the researchers' information needs through, for example, the arising of new more specific research fields.[18] An example is given by the study of Dogan et al. [19], who found that the majority of searches in PubMed in the subject field are performed by gene, protein and / or disease. This is particularly interesting if we take into account that just a few decades ago instead of gene or protein, the searches used terms related to some kind of treatment or diagnosis.

- Process of indexing and access points

The users' information needs are often articulated ambiguously not only in what concerns the terms, but also in what concerns the usage of the structure of the system being searched. The process of conducting a search involves issues such as syntax, semantics, structure and purpose of the search; how the access points are used to reduce and expand the results; search alternatives, and if the fault in the search derives from a personal or a system failure.[20]

Sandusky & Tenopir [9] claim that there is a great difficulty for researchers to identify relevant articles, as these are still indexed in a way that does not comprise detailed information about the document. Reference is made to ProQuest CSA, which developed a prototype system that provides detailed information by indexing individual components of the articles. This model allows the realization of Boolean searches using author, title, statistics, geographic and taxonomic terms. Searches can be refined by maps, figures, photographs, type of article or predictive models. This procedure is related to the increased level of granularity used in the design of databases. Bishop [10] complements Sandusky &Tenopir claims by stating that the indexing of articles is highly standardized, including the identification of the author, title, abstract and keywords. Bishop further argues that existing items in the articles that are not explored in the search fields may promote the retrieval of the most relevant documents.

Bates, Wilde, & Siegfried [21] present the results of the analysis of online bibliographic databases usage. These results contributed to improvements of (1) facets of the Styles and Periods Art And Architecture Thesaurus, with the inclusion of some terms that were previously neglected in the thesaurus; (2) database structure, with the inclusion of names of artists, based on the variation

of names and terms that identify the academic disciplines in the database. For these authors, only a good indexing enables the obtention of high precision and relevance search results.

Like Bishop [10-11], Crowston & Kwasnik [22] also mention the issue of indexing and underline its relationship with a critical factor: context. Crowston & Kwasnik identify the difficulty to meet the users' needs with relevant results due to issues such as inaccurate or incomplete information in the databases which are related to indexing.

Studies in which the indexing process is approached are still incipient in what concerns the possibilities of expanding the level of analysis and granularity. This does not go along with some initiatives in which the use of descriptive metadata ceases to be limited to the identification of title, author and subject. Items such as treatment outcome, risk factors and conclusion as topic begin to be explored either manually or automatically, especially with the use of Semantic Web tools.[23-24]

Hjørland et al. [14] add that the different structures that exist in texts have consequences in the search. The search strategy in scientific databases can vary according to specific access points such as methodological issues or findings, which are considered topics of greatest interest.

As regards to description and recovery, there are metadata or access points, which result from the activity of indexing documents, i.e., access points determine objective possibilities of document retrieval by users through algorithmic or automated procedures.[14] In this sense, studies have shown that software agents can process articles' information using semantic treatment as a new form of content exploitation.[25]

Lee & Downie [26] developed a study in relation to music information retrieval (MIR) in the field of music digital libraries (MDL). One of the quests was "What types of metadata or access points should be provided to users." Regarding this quest, they identified the need for new types of metadata as access points that include information about music or music objects and information that contextualize searches of users' the real-world.

Hjørland et al. [14] stress that the Subject Access Points (SAP) are critical to the retrieval of documents. Thus, if these access points are not supplied by the information services, users may not be able to retrieve the documents they need. Some studies also show that increasing the granularity of the information in the information services contributes to the improvement of the search, since it provides a wider range of access points.[9-10,14-15,20,26] By restraining these claims, we argue that a higher degree of granularity of access points is expected to promote high precision and relevant search results that are more aligned with users' needs. This is related to Otlet's claims about information atomization. As Menzel [27] points out, "the expressed and conscious wants of individuals in any area are constrained by their perception of what is feasible". As what is feasible currently in most information services does not include high granular searches and access points, researchers' needs may not be properly externalized, forwarding to Taylor's 3rd level of information needs. Perhaps this is what better explains why there are not enough studies on researcher's information needs regarding bibliographic searches. Hence the need to explore the perspective of the researcher, what he needs and what access points would meet his needs.

- User profile

  The concern with the user is externalized by Courtright [28], when she observed that there was a change in the direction of the studies in what concerns information needs. The studies used to focus on a system-centered model and were redirected to a user-centric model, where research focuses on the information of the participating actors.

  Amato &Straccia [2] and Courtright [28] consider that the information needs may vary depending on the type of user and on the context in which these needs are analyzed or required. Thus, one of the concerns in the development of information needs studies is to establish what user profile is to be analyzed.

  Additionally, researchers are users whose experience in developing searches presupposes information with the highest level of specificity and relevance. Sandusky &Tenopir [9] found that, for this type of user, the identification of relevant items in a short time reflects on certain types of behavior as, for instance, the time allocated for reading articles. These results are corroborated by the studies of Berghel et al. [29], Shotton [23],Tenopir et al. [30-31].

3. Conclusion

This literature review presented some interesting and promising aspects related to information needs of researchers and which are worth exploring: components of articles, indexing and metadata, domain, and user profile. It is clear that Otlet was indeed beyond his time in what concerns his suggestions regarding the need to access to specific parts of documents quickly and completely. However, most authors reported difficulties in identifying the information needs of researchers. Perhaps these difficulties are related to the difficulty that researchers themselves have in expressing their real needs when using the information services, forwarding them to Taylor's level 3 of information needs.

Another aspect that was identified with this literature review concerns Taylor's level 4 of information needs, i.e., the conscious needs that are well defined and may be "translated" into the language of the information system in a way they can be processed. There is a tendency to us a high degree of information granularity that can be optimized by the use of semantic Web and linked data services that are able to provide more relevant search results in less time and with less effort of the researchers.

Regarding the development of this study, it is worth noting the difficulties and limitations we found. Some of the difficulties we had are related with some problems identified by authors cited in this literature review. In particular, we had problems that are related to the indexing process of the databases (the inclusion of specific keywords), which caused a limitation in the number of retrieved articles. This implied the need to adjust the search strategy in order to identify other relevant articles for this literature review.

As future work we suggest that studies focus on the information needs that could be met by the use of semantic Web technologies. For example, How do researchers enjoy the potential offered by these tecnologies? How these technologies are being used by the information services? What new services could be created using these technologies to better meet the researchers' needs?

Acknowledgements

We thank for Espírito Santo Federal University, Brazil; CAPES Foundation, Ministry of Education of Brazil for financial support to our research activities.
Part of this work is funded by FEDER funds through the Competitiveness Factors Operational Programme - COMPETE and National Funds through FCT - Foundation for Science and Technology under the Project: FCOMP-01-0124-FEDER-022674

References

[1] Kuruppu, P. U., & Gruber, A. M. (2006). Understanding the Information Needs of Academic Scholars in Agricultural and Biological Sciences. *The Journal of Academic Librarianship*, *32*(6), 609–623. Retrieved from http://www.sciencedirect.com/science/article/pii/S0099133306001510 . doi:10.1016/j.acalib.2006.08.001

[2] Amato, G., & Straccia, U. (1999). User profile modeling and applications to digital libraries. In *Research and Advanced Technology for Digital Libraries* (p. 184–197). Berlim: Springer. Retrieved from http://link.springer.com/chapter/ 10.1007/3-540-48155-9_13. doi: 10.1007/3-540-48155-9_13

[3] Taylor, R. S. (1962). The process of asking questions. *American documentation*, 13(4), 391–396. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/asi.5090130405/abstract . doi: 10.1002/asi.5090130405

[4] Cyganiak, R., Reynolds, D., & Tennison, J. (2013). The rdf data cube vocabulary. *W3C candidate recommendation.* Retrieved from http://www. w3. org/TR/2013/CR-vocab-data-cube-20130625.

[5] Data Catalog Vocabulary (DCAT). (2014). [2014 Feb 20]. Retrieved from: http://www.w3.org/TR/vocab-dcat/

[6] Ismail, M. A., & Kareem, S. A. (2011). Identifying how novice researchers search, locate, choose and use web resources at the early stage of research. *Malaysian Journal of Library and Information Science*, 16(3). Retrieved from http://works.bepress.com/maizatulakmar_ismail/3.

[7] Calva-González, J. J. (2003). Las necesidades de información de los investigadores del área de Humanidades y Ciencias Sociales. *Revista general de información y documentación*, 13(2), 155–180. Retrieved from http://revistas.ucm.es/index.php/RGID/article/view/10799.

[8] Calva González, J. J. (2009). Las necesidades de información del usuario en la automatización de unidades de información. *Revista Biblioteca Universitaria*, 1(1), 6. Retrieved from http://www.dgbiblio.unam.mx/servicios/dgb/publicdgb/ bole/fulltext/volI1/necinf.html

[9] Sandusky, R. J., & Tenopir, C. (2008). Finding and using journal-article components: Impacts of disaggregation on teaching and research practice. *Journal of the American Society for Information Science and Technology*, 59(6), 970–982. doi:10.1002/asi.20804

[10] Bishop, A. P. (1998). Digital libraries and knowledge disaggregation: the use of journal article components. In *Proceedings of the third ACM conference on Digital libraries* (p. 29–39). Retrieved from http://dl.acm.org/citation.cfm?id=276679. doi>10.1145/276675.276679

[11] Bishop, A. P. (1999). Document structure and digital libraries: how researchers mobilize information in journal articles. *Information Processing & Management*, 35(3), 255–279. doi: http://dx.doi.org/10.1016/j.bbr.2011.03.031

[12] Otlet, P. (1990). The science of bibliography and documentation. In Rayward, W. B (Org.). *International organisation and dissemination of knowledge: selected essays of Paul Otlet*. Amsterdam: Elsevier for the International Federation of Documentation. Retrieved from https://www.ideals.illinois.edu/handle/2142/4004

[13] Rayward, W. B. (Org.). (1990). *International organisation and dissemination of knowledge: selected essays of Paul Otlet*. (pp. 7-10). Amsterdam: Elsevier for the International Federation of Documentation. Introduction. Retrieved from https://www.ideals.illinois.edu/handle/2142/4004.

[14] Hjørland, B., Nielsen, L. K., & Williams, M. E. (2001). Subject access points in electronic retrieval. *Annual Review of Information Science and Technology*, *35*, 249–298

[15] Bates, M. J. (1996). The Getty End-User Online Searching Project in the Humanities: Report No. 6: Overview and Conclusions. *College & Research Libraries*, 57(6), 514–23. doi :10.1108/eb024508

[16] Markey, K. (2007a). Twenty-five years of end-user searching, Part 1: Research findings. *Journal of the American Society for Information Science and Technology*, 58(8), 1071–1081. doi : 10.1002/asi.20462.

[17] Markey, K. (2007b). Twenty-five years of end-user searching, Part 2: Future research directions. *Journal of the American Society for Information Science and Technology*, 58(8), 1123–1130. doi:10.1002/asi.20601

[18] Rowlands, I. (2007). Electronic journals and user behavior: A review of recent research. *Library & Information Science Research*, 29(3), 369–396. doi : 10.1016/j.lisr.2007.03.005.

[19] Dogan, R. I., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed® user search behavior through log analysis. *Database: the Journal of Biological Databases and Curation*, 2009. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2797455/. doi : 10.1093/database/bap018.

[20] Borgman, C. L. (1986). Why are online catalogs hard to use? Lessons learned from information-retrieval studies. *Journal of the American society for information science*, 37(6), 387–400. doi : 10.1002/(SICI)1097-4571(198611)37:6<387::AID-ASI3>3.0.CO;2-8.

[21] Bates, M. J., Wilde, D. N., & Siegfried, S. (1993). An analysis of search terminology used by humanities scholars: the Getty Online Searching Project Report Number 1. *The Library Quarterly*, 1–39. Retrieved from http://www.jstor.org/stable/4308771.

[22] Crowston, K. & Kwasnik, B. H. (2003). Can document-genre metadata improve information access to large digital collections. *Library Trends*, 52(2), 345–361. Retrieved from http://works.bepress.com/cgi/viewcontent.cgi?article=1003&context=barbara_kwasnik.

[23] Shotton, D. (2009). Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing*, 22(2), 85–94. Retrieved from https://webvpn.uminho.pt/http/0/www.ingentaconnect.com/content/alpsp/lp/2009/00000022/00000002/art00002. doi:10.1087/2009202

[24] Scientific Data. (2014). Retrieved 2014 Feb 20 from http://www.nature.com/scientificdata/.

[25] Marcondes, C. H., Mendonça, M. A. R., Malheiros, L. R., Da Costa, L. C., & Santos, T. C. P. (2009). Ontological and conceptual bases for a scientific knowledge model in biomedical articles. *RECIIS*, 3(1). Retrieved from http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/240/251. doi:10.3395/reciis.v3i1.240en.

[26] Lee, J. H., & Downie, J. S. (2004). Survey Of Music Information Needs, Uses, And Seeking Behaviors: Preliminary Findings. In *ISMIR* (Vol. 2004, p. 5th). Retrieved from http://people.lis.illinois.edu/~jdownie/ismir2004_survey_downie_draft.pdf.

[27] Menzel, Herbert. "The Information Needs of Current Scientific Research." *The Library Quarterly* 34, no. 1 (1964): 4–19. Retrieved from http://www.jstor.org/stable/10.2307/4305417

[28] Courtright, C. (2007). Context in information behavior research. *Annual review of information science and technology*, 41(1), 273–306. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/aris.2007.1440410113/full. doi : 10.1002/aris.2007.1440410113.

[29] Berghel, H., Berleant, D., Foy, T., & McGuire, M. (1999). Cyberbrowsing: Information customization on the Web. *Journal of the American Society for Information Science*, 50(6), 505–513. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(1999)50:6<505::AID-ASI5>3.0.CO;2-R/abstract . doi:10.1002/(SICI)1097-4571(1999)50:6<505::AID-ASI5>3.0.CO;2-R.

[30] Tenopir, C., King, D. W., Edwards, S., & Wu, L. (2009). Electronic journals and changes in scholarly article seeking and reading patterns. In *Aslib proceedings* (Vol. 61, p. 5–32). Retrieved from: http://www.emeraldinsight.com/journals.htm?articleid=1766871&show=abstract

[31] Tenopir, C., King, D. W., Spencer, J., & Wu, L. (2009). Variations in article seeking and reading patterns of academics: What makes a difference? *Library & Information Science Research*, 31(3), 139–148. Retrieved from http://www.sciencedirect.com/science/article/pii/S0740818809000516.

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-39

Identifying User Behavior in domain-specific Repositories

Wilko VAN HOEK[a,1], Wei SHEN[a] and Philipp MAYR[a]
[a] *GESIS – Leibniz Institute for the Social Sciences, Germany*

Abstract. This paper presents an analysis of the user behavior of two different domain-specific repositories. The web analytic tool *etracker* was used to gain a first overall insight into the user behavior of these repositories. Moreover, we extended our work to describe an apache web log analysis approach which focuses on the identification of the user behavior. Therefore the user traffic within our systems is visualized using chord diagrams. We could find that recommendations are used frequently and users do rarely combine searching with faceting or filtering.

Keywords. User information behavior, web log analysis, information seeking, search process, online information services

Introduction

In recent years, many web analytics applications have been published to measure and analyze usage data in order to understand and optimize the information seeking in web systems. When designing a domain-specific repository, it is important to understand the ways in which users perform searches. A lot of studies were conducted to understand user behavior in the context of web search analysis. Bates proposed a dynamic search model, describing that searcher's information needs change over time [2]. She further extended her work by characterizing the common information seeking process that consists of sequences of search tactics [1]. To investigate the human information search process, Koch et al. [3] conducted a thorough log analysis, which grouped the session-based log entries into eleven different activities and used these activities to identify user behavior. Mayr [5] presented a quantitative, non-reactive measure for standard Apache log files focusing on typical navigation types which can easily be extracted from the referrer information in the log.

Domain-specific repositories always target at a certain user group, which has substantial domain expertise and aims to search for specialized, domain-oriented information. Typically specialized users develop individual search tactics [1] to operate in the repositories in an efficient way. The traffic of these users leave traces in the web server log which can be consulted for detailed analyses of navigational structures of a system. Russell-Rose et al. [6] categorized users into four types: double experts, domain expert/technical novices, domain novice/technical experts and double novices. In this sense, we assume that users of domain-specific repositories could be ranked as

[1] Corresponding author: Wilko van Hoek, GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany; E-mail: wilko.vanhoek@gesis.org

double experts or domain expert/technical novices. Double experts are individuals identified with high domain and technical expertise, which often use teleporting search strategy, formulating the queries precisely and jump quickly to the destination [7]. Domain expert/technical novices, on the other hand, are able to use their knowledge to formulate effective queries, but lack the technical confidence to explore unknown territory [4].

In this paper, we report preliminary findings of a user behavior analysis within three domain-specific collections. The three collections belong to two different repositories. Two of the three collections are part of the *Effektiv!*[2] portal and the third collections is the *Social Science Open Access Repositories (SSOAR)*[3].

1. Background

The *Effektiv!* portal is an academic online portal funded by the German Federal Ministry of Education and Research which offers descriptions of programs to support family friendliness at German education institutions and disclosed best practices to help the scholars and students to balance better between an academic career and their family life. The core parts of the *Effektiv!* portal are two collections. An online database with practice examples of family-friendly best practices in academic education institutions (herein after called *Effektiv! best practices*) and a bibliography of literature specialized in family-friendliness and gender topics (herein after called *Effektiv! literature*). Both collections are online since April 2013.

Founded in 2008 the *SSOAR* is a full-text server for open access publications in the field of social sciences. Furthermore, *SSOAR* offers the social scientists, scientific associations and publishers the opportunity to self-archive their publications, to enhance the visibility of their work on the web. There are currently about 27,600 digital papers archived in *SSOAR*.

The portals *SSOAR* and *Effektiv!* are both based on the same repository software DSpace. Different search user interfaces have been designed to help the user to apply specific search strategies. A guided search concept is applied to design the *Effektiv! literature* and *Effektiv! best practices* user search interface. The main design idea behind the *Effektiv! literature* search interface is: besides an overall search box, the user can enter further search terms for assumed popular attributes such as author and title in additional search boxes. Users are further invited to select values of two additional filters. A browsing of the "subject area" is provided at the right side of the site. In the *Effektiv! best practices* user search interface, filters are emphasized and presented at the top of the search form while the standard search box is at the bottom. In this case, users are encouraged to narrow down their searches quickly with the selection of these filters. On the contrary, the main search interface for *SSOAR*, called *browse and search*, is designed with a faceted search concept, in which attributes are displayed as links in a navigational menu. This approach facilitates the user to intuitively search by progressively refining their choices. In addition to the faceting,

[2] Effektiv! - For Greater Family Friendliness in German Higher Education Institutions (www.familienfreundliche-hochschule.org). Funded by the German Federal Ministry of Education and Research (BMBF) (grant no. 01FW11101). Any opinions expressed here are those of the author(s).

[3] http://www.ssoar.info

users can browse the system by disciplines. When a discipline is selected, the user can apply facets or search in the result list. In this way facetted search and browsing can be combined. Besides the *browse and search* interface, a traditional advanced search is also provided, supporting users to freely formulate their search queries.

In the following chapters we want to gain insights into the way the different search options are used in the three collections. We want to find out which concepts work well and which do not.

2. Web Analysis Using *etracker*

The *etracker*[4] web analysis software was used to identify the general user behavior of the two repositories. The investigated time period is from 1st April 2013 and 31st December 2013.

According to *etracker* there were 254,240 users visiting the SSOAR portal and 5,641 users visiting the *Effektiv!* portal during this time period. We examined the user number, page impression, visiting time of the user interface in each repository as shown in Table 1. We can see that the most SSOAR users (over 90% of all) went to the *browse and search* user interface to search for documents. The advanced search interface was rarely used. About 14% of the *Effektiv!* portal visitors used the best practices collection and only 5% of visitors viewed the literature collection. This effect may be due to the structure and multi-functionality of the *Effektiv!* portal. Besides the two collections, the *Effektiv!* portal also provides services like online advisory service, press information etc., which means that the main goal of many *Effektiv!* portal visitors may not be the best practices or literature search.

The parameters "page impressions per user" and the visiting time were calculated in both SSOAR search interfaces. However, the *Effektiv!* best practices user interface was apparently much more viewed than the literature user interface. And although the visiting time per page of best practices was adjacent, the users within the best practices user interface took more time visiting than the users within the literature. This may indicate that compared to *Effektiv! literature* collections, the *Effektiv!* users made more search queries in the *Effektiv! best practices* collection.

Table 1: Summary statistics of different user interfaces

Repository name/ User Interface name	Total users	Page impressions	Page impressions per user	Visiting time per user	Visiting time per page
Effektiv! best practices	788	4,100	5.20	00:02:27	00:00:28
Effektiv! literature	331	1,219	3.68	00:01:55	00:00:31
SSOAR browse and search	24,421	117,765	4.82	00:03:29	00:00:43
SSOAR advanced-search	3,616	15,574	4.31	00:03:00	00:00:42

The click path chart indicates the user's movement paths through the web pages. Figure 1 shows the click path chart of the *SSOAR browse and search* user interface (here *SSOAR/discover/*). The yellow node in the middle of the chart represents the discovery user interface. The grey *entry* node above represents the direct entry in the user interface from other pages (compare approach in [5]), while the grey *exit* node

[4] www.etracker.com

beneath represents requests where users left the page. The five top ranked referrer sites are listed at the left side of the graph and the five top following sites are listed at the right side. About 25% of the users went from *SSOAR* homepage (German and English) to the *browse and search* interface and 16.8% came directly from external pages. Nearly 30% of the users left the *SSOAR* portal after viewing the interface. Due to the fact that over 50% of the sites are not analyzed and grouped to the *others* node, this click path analysis in *etracker* is very limited and the sequential search process cannot be thoroughly displayed.

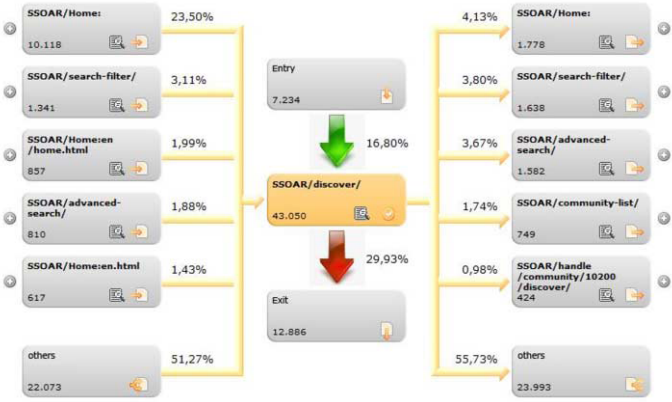


Figure 1: Click path chart

3. Apache Web Log Analysis

Although the analysis of the *etracker* data has given us some first insights of the users' behaviors, many details are missing. We don't know which types of interaction are most or least frequently used. The main problem seems to be that *etracker* cannot identify what type of action is performed on a page and what kind of information lies behind an URL. For instance, in the click tracker analysis most of the traffic has been grouped together as *others.* It is not possible to figure out in *etracker* how many document views, searches or browsing actions the group *others* are consisting of.

To overcome this lack of detail, we decided to analyze the raw log files of our web servers. On both systems we are using the Apache 2[5] web server with an identical logging configuration. During the time period between 1st April 2013 and 31st December 2013 we collected IP (anonymized), timestamp, requested URL and referrer of all visitors were collected. As the functionality of both systems relies mainly on http-requests, we can identify what page is viewed by analyzing the URLs given in the log.

To understand the users' behavior, we focused on analyzing the pairwise information of requested URL and referrer. So to say we looked at where users were coming from and where they were going to. Both systems are using the software Solr[6] as their search backend. This allows us to identify what pages where requested by analyzing the URL. For instance, we can see if a simple search, based on a single query, is conducted or whether a more complex search, using filters or facets has been

[5] Apache Server Project (http://httpd.apache.org/)
[6] Apache Solr (http://lucene.apache.org/solr/)

executed. We grouped the user traffic into different types of search interactions (see table 2 and table 3) and then calculated how many requests involved users to switch between these types.

When analyzing web server log files, it is important to clean out automated accesses e.g. by web spiders that usually generate the biggest amount of traffic. Spiders systematically request every part of a web page. Most of those spiders can be identified by their IP address. The software DSpace collects lists of spiders. We used these lists to clean out all know web spiders. In addition we truncated all requests regarding the hour at which the access was conducted and counted the number of request per IP address. Based on this data we could identify a small set of further IP addresses responsible for a large amount of traffic. We then filtered out the data generated by those IP addresses.

In the following subsections, we will describe our analysis of the web server log files. We will introduce a new visualization technique applied for log data. Thereafter we will present the results of our analysis for both *Effektiv!* collections and *SSOAR*.

3.1. Chord diagrams

In the following sections, we use chord diagrams[7] to visualize the traffic within the three collections (*SSOAR*, *Effektiv! literature* and *Effektiv! best practices*). We used the D3.js library[8] to create the diagrams. These diagrams have originally been used to visualize the movements of people between different neighborhoods[9]. We transferred this idea to our situation by interpreting different types of search interactions as neighborhoods and the transition from one to another as a movement between two neighborhoods. For example, using free *text search* is one type of interaction and assigning filters another. When a user changes the result list for a free text search by applying filters, we counted this as a transition between two neighborhoods. At first we defined two types of web pages that can be used in all interfaces (cf. table 2). Then we identified the different types of search interactions (cf. table 3 and table 4).

Table 2: General page types accessed by users

Page type	Effektiv! literature & best practices	SSOAR
G1	Initial search page	Initial search page
G2	Document URL	Document URL

Table 3: Search interaction types performed by users

Search interaction type	Effektiv! literature & best practices	SSOAR
S1	List-all	Repository overview
S2	Free text search without filter	Free text search without facet
S3	Filter search	Faceted search
S4	Free text search with filter	Free text search with facet
S5	Change query	Change facet
S6	-	Advanced-search

[7] Chord diagrams (http://bl.ocks.org/mbostock/4062006) have been inspired by circos http://circos.ca/.
[8] http://d3js.org/
[9] http://bost.ocks.org/mike/uberdata/

Table 4: Browsing interaction types performed by users

Browsing interaction type	Effektiv! literature & best practices	SSOAR
B1	Browsing	Browsing
B2	-	Browsing with free text search
B3	-	Browsing with facet
B4	-	Browsing with free text search and with facet

To understand the way chord diagrams work we will give an example. Figure 2 illustrates the traffic related to the *Effektiv! literature* collection. The chord diagram can be read as follows. The total amount of data is represented by a circle. The data is grouped around this circle. Each type of interaction or page is represented by an arc. The size of an arc represents the amount requests where the referrer URL was assigned to that type of interactions or pages. The area between two arcs illustrates the traffic between the corresponding types. For instance, in approx. one third of the requests where the referrer was assigned to browsing (type B1), the destination URL belongs to a document (type G2). In reverse, most of the traffic where the referrer is a document URL (type G2) has a destination URL that was assigned to browsing (type B1).

3.2. Log file analysis results for *Effektiv! literature*

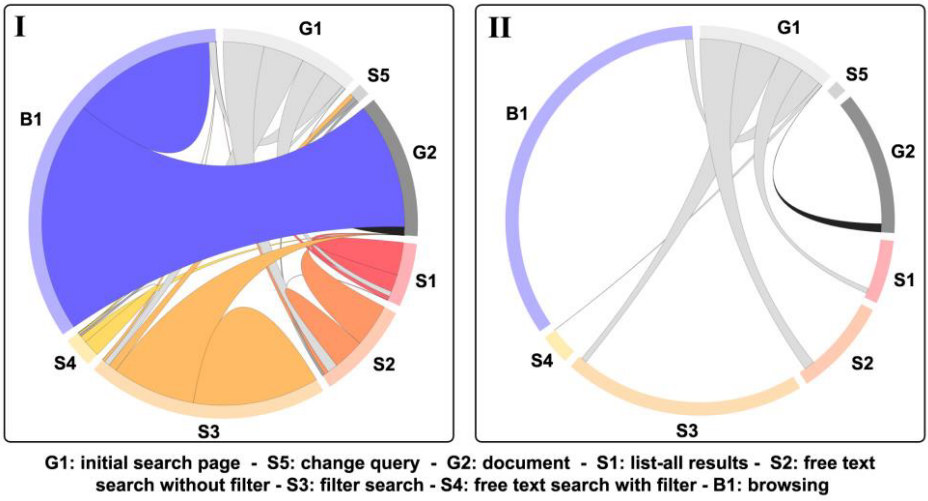


Figure 2: User traffic for different interaction types for the *Effektiv! literature* database – overview in I and II shows the traffic only for the initial search page (S1).

For the collection *Effektiv! literature* the search interaction browsing (type B1) is responsible for the highest amount of user traffic, as 35% of the referrers hold URLs belonging to browsing. This is followed by filter search (type S3) that takes up 21% of the traffic. Part II of Figure 2 shows which type of interactions users have used after the initial search page (type G1). Overall the users seem to continue quite equally with the different type from the initial search page. Simple search without search terms and searching without filtering are most often used in this situation. Combining search terms with filters is an exception as it is rarely the next step taken by the majority of users. After selecting a way of querying the users only rarely change their way of searching, there are only small amounts of requests in which users change the search

terms or switch from one type to another. Also only 10% of the traffic consists of requests where a search term has been entered.

Comparing a filter search and browsing, a basic difference in the users' behavior can be observed. Users who are filtering without query terms are viewing a document in 37% of the cases, but remain within the same search type in 55%. Staying in the same search type means that users are viewing a second result page or sorting their results. When browsing, the users access documents in 60% of the cases and stay in browsing in 37%. Users seem to find relevant documents by browsing more often than by just filtering. The second observation in this context is that many requests show movements from documents to browsing. This can be explained by links that are shown on a document page. For instance, users can proceed from document pages by browsing the system using the author name. This may also be an explanation for the dominance of the browsing related traffic in the log files.

3.3. Log file analysis results for *Effektiv! best practices*

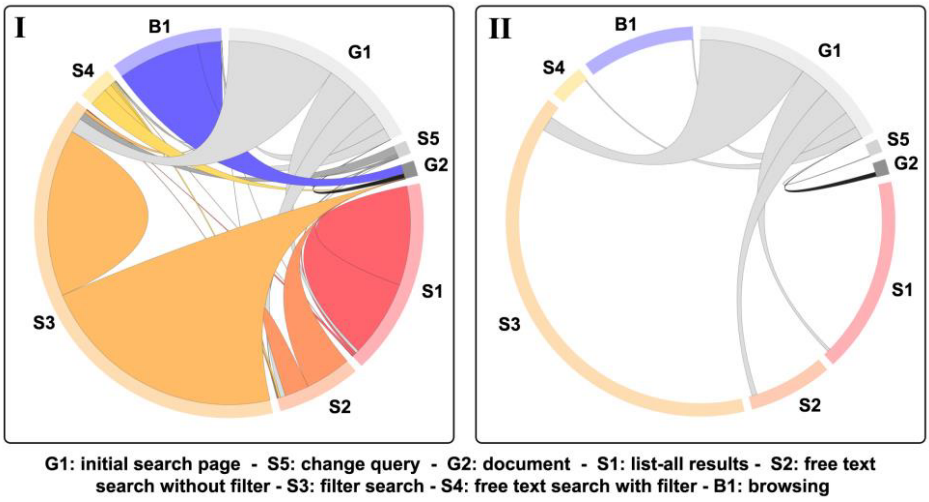


Figure 3: User traffic for different interaction types for the *Effektiv! best practices* collection – overview in I and II shows the traffic only for node A.

Figure 3 illustrates the user traffic related to the *Effektiv! best practices* collection. Here filtering without entering search terms (type S3) is the most often requested type with 45% of the total traffic. Listing all results (type S1) and browsing (type B1) are ranked second and third with 16% and 9% of the traffic respectively. The preference to filter without search terms can also be observed by looking at the traffic from the initial search page (type G1). 55% of the users are proceeding from the initial search page, by filtering the data without entering search terms. Overall users tend not to change or reformulate their search query as only a small amount of traffic is related to those cases and users rarely query the system using search terms.

When looking at the relation between users moving from browsing to documents or users moving from filtering without search terms and to documents, it can be observed that the users' behavior is slightly different. 48% of the users that are filtering without search terms access documents while 32% stay within the interaction type. In contrary 77% of the traffic that was generated by users browsing the data led to

documents, while only 21% of the users remain browsing. Users that are browsing the system seem to access documents more frequently than users that don't, although more users are filtering the system.

3.4. Log file analysis results for the SSOAR

In SSOAR there are two interfaces that allow the user to search in the collection. There is the advanced-search that allows querying by searching in specific metadata fields as well as searching over all fields and there is the browse and search interface in which search terms can be combined with facets. The browse and search also allows users to browse for documents and to search the result list of the browsing or apply facets to filter that result list. The browsing functionality has been discussed strongly during the development of the browse and search interface. We therefore decided to distinguish between faceted search (types S1-S5) and browsing (types B1-B4). Figure 4 shows the traffic between the interaction types advanced-search (type S6), faceted search, and browsing.

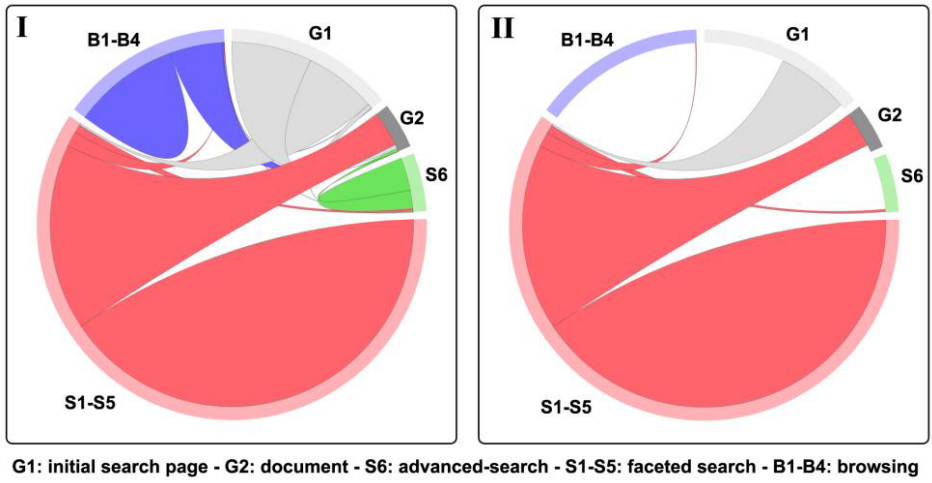


Figure 4: User traffic for different interaction types for SSOAR – overview in I and II shows the traffic only for node D.

62% of the users' traffic is concentrated on faceted search. And 60% of the movement from this interaction type is self-directed. This means that in those cases users conducted interactions like query reformulation or selecting facets. Nearly one third of the group's traffic is related to requests from faceted search to documents (type G2). The next largest amount of traffic for one type, with roughly 20 % of the total traffic, is browsing. Here similar to faceted search, 60% of the movement is self-directed and 40% represents movements to documents. The same observation holds for the advanced-search. When looking at the movement between the three search types, it becomes clear that only a small amount of users switch from faceted search to either browsing or advanced-search and an even smaller proportion of users request the opposite direct. The high amount of traffic from document to faceted search can be explained by the "more about" link. This link is shown on the document page and directly triggers a faceted search using metadata fields such as authors.

To better understand the users' behavior within the faceted search and the browsing; we generated two additional diagrams that show the traffic related to those two types. Interestingly the two do not interact strongly. Users that decide to first browse the system usually remain in that status. This is understandable as users are able to further query and facet in the filtered results and may not be interested in broaden their results again. Figure 5 shows the traffic for faceted search (part I) and browsing (part II).

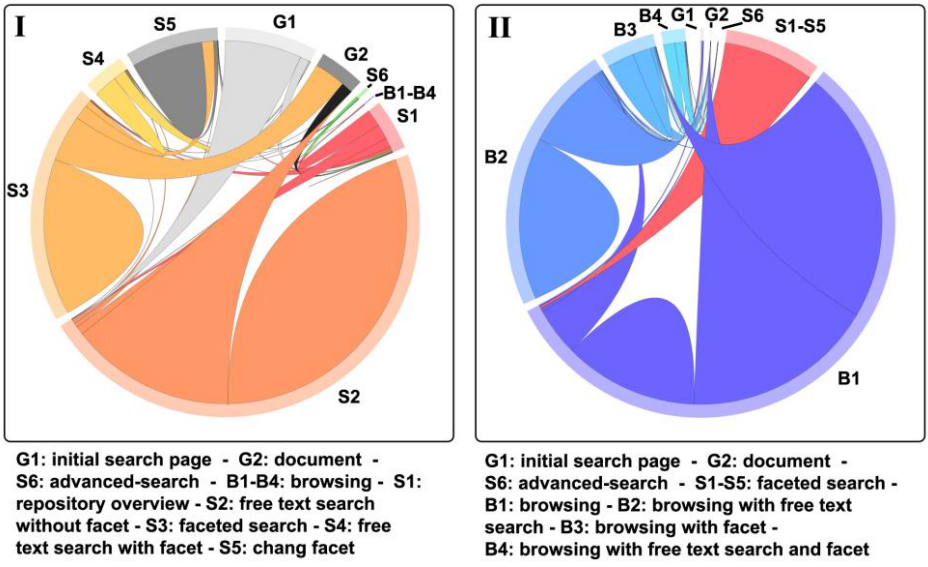


Figure 5: Different types of user traffic in SSOAR related to faceted search (I) and browsing (II)

Most of the traffic related to faceted search is generated by users that query the system by entering search terms without using facets (50%). In addition this is also the most often used starting point as 88% of the users proceed with this type after the starting page. The second highest amount of traffic belongs to search where only facets are applied (21%). Obviously users prefer to query the system by using search terms or use facets but rarely combine both.

A different situation can be observed when looking at the traffic related to browsing. Browsing without entering search terms not using facets is the most dominant type here with 60% of the traffic. It is followed by browsing with search terms without using facets (24%). An exception of the analyzed data lies in the movement from browsing without search terms to using facets. A high proportion of 40% of the traffic related to browsing without search terms and without facets is related to applying facets. But in total browsing without search terms with facets takes only 5% of the traffic. This way of querying the system seems to be a dead end that we cannot explain right now.

Conclusion

In this paper, we have presented the results of two analyses of user behavior in the repositories *Effektiv! literature*, *Effektiv! best practices* and SSOAR. In the first

analysis we tried to identify user tactics by looking at the information provided by the service *etracker*. In the second analysis we conducted an own evaluation of the raw log files generated by the web servers.

During the first analysis it became clear that the information provided by services like *etracker* do not suffice to identify the user's behavior. We therefore decided to evaluate the traffic generated by our users by ourselves, in form of a log file analysis. Based on our own log files analysis we could observe that users searching in the *Effektiv! literature* collection use the browsing and filtering opportunities intensely and rarely type in search terms. In addition we could see that the links provided on the document pages, where users can proceed within the system by browsing for authors or topics were used frequently.

For the *Effektiv! best practices* collection browsing is less frequently used but filtering is therefore used more often. The difference in the users' behavior in this collection to the *Effektiv! literature* collection can be explained by the fact that the provision of the browsing links presented on the document pages are less interesting to the users. To improve the search in this collection we will consider adding further browsing opportunities.

In the results for SSOAR we could see that advanced-search and browsing is less frequently used than faceted search. Looking into the detailed behavior for faceted search and browsing we could see that users do rarely combine facets and search terms. We consider to change our front end to improve this. The next observation is that faceting when browsing seems to be a dead end decision. We will need to examine this more closely to understand why this is the case. The third major observation is that, similar to *Effektiv! literature*, the opportunity to continue the search from a document by clicking on a link which allows searching for more documents with the same metadata entry is used frequently. We should improve our functionality regarding this feature in SSOAR.

Overall analyzing the log files is worthwhile and should be done more frequently. After setting it up it can be used regularly to analyze the effects frontend changes of a web site result in. Right now we are able to better understand how our systems are used. By improving our method and extending it to identify user-session, we are confident to be able to identify user search tactics and thus gain more information about our users. We will continue to analyze our data regularly in the future.

References

[1] Bates, M.J. Information Search Tactics. Journal of the American Society for Information Science, 30. July 1979, 205-214.

[2] Bates, M.J. *The design of browsing and berrypicking techniques for the online search interface*, Online Information Review, Vol. 13 Iss: 5, pp.407 – 424, 1989.

[3] Koch, T.; Ardo, A.; Golub, K., *Browsing and searching behavior in the Renardus Web service: a study based on log analysis*, Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference, pp.378, 7-11 June 2004.

[4] Jenkins, C. et al. Patterns of information seeking on the web: a qualitative study of domain expertise and web expertis. IT&SOCIETY,1(3), 64-89, 2003.

[5] Mayr, P. *Website entries from a web log file perspective - a new log file measure,* Proceedings of the AoIR-ASIST 2004 Workshop on Web Science Research Methods (2004)

[6] Russell-Rose, T. and Tyker T. *Designing the search experience: chapter 1: The User*. Morgan Kaufmann, 2013.

[7] Teevan, J., Alvarado, C., Ackerman, M., &Karger, D. *The Perfect search engine is not enough: a study of orienteering behavior in directed search*. ACM Press, 415-422, 2004.

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-49

Data Quality Issues and Content Analysis for Research Data Repositories: The Case of Dryad

Dimitris ROUSIDIS[a,1], Emmanouel GAROUFALLOU[b], Panos BALATSOUKAS[c], Miguel-Angel SICILIA[a]
[a] *University of Alcala, Madrid, Spain*
[b] *Alexander Technological Educational Institute of Thessaloniki, Greece*
[c] *University of Manchester, UK*

Abstract. Research Object (RO) repositories extend traditional forms of scholarly communication by providing scientists the means necessary to store, share and re-use datasets generated at various stages of the research process. Yet this shift to digital publication does not guarantee that outputs, results or methods are reusable. Data quality is absolutely vital for the dissemination, reuse and sharing of digital resources. Manual metadata quality control is practically impossible and as a result, many quality criteria, both semantically and structurally get overlooked and digital objects may become problematic. The aim of the research reported on this paper was to identify the data quality problems associated with the Dryad research data repository. In particular, three metadata elements (Creator, Date and Resource Type) were analysed and quality issues associated to these elements were identified. The paper concludes with some recommendations for improving the quality of metadata in research data repositories.

Keywords. Big Data, Data Quality, Descriptive Analysis, Open Access Repositories, Metadata, Research Objects, e-Research

1. Introduction

The parallel growth the availability of scientific data (big data) and the emergence of cloud computing has radically changed research activities. eScience and eResearch applications have extended traditional forms of scholarly e-infrastructure (such as institutional repositories and digital libraries) and enabled scientists to store, access, analyse, use and share datasets generated at various stages of the research process [1]. Given the big volume and diversity of scientific data, research repositories are becoming integral part of the communication and collaboration process between scientists and research groups. Although research on the technical and architectural characteristics of research data repositories has progressed (e.g.[2], [3], [4]), there is still a need to measure their growth and analyse their contents. This includes knowledge on the size, composition and growth dynamics of these repositories. Such knowledge would eventually result in insights on the behaviour of researchers and the usability of their research products for reuse, e.g. for experiment repetition.

It is well documented in the literature that measuring the growth and analysing the contents of digital repositories in general is important for the sustainability and

usability of this type of technology (e.g. [5]). Yet, data quality issues (e.g. in terms of metadata) may influence the effectiveness of the analysis of this type of repositories.

The aim of the research reported in this paper was to identify the data quality problems associated with the analysis of the contents of a research data repository, called Dryad. Being this a first attempt to measure research data repositories, the objectives were chiefly exploratory, concretely:

- To perform a descriptive analysis of the contents of the Dryad repository across different variables (metadata), such as the type and format of datasets, the size and submission date of data packages and files; and
- To identify data quality issues and challenges related to the analysis of these metadata elements;

This paper is structured as follows: First, a literature review on previous work is documented and the Dryad repository is described. Then the methodology and results of the analysis are presented. Finally, conclusions and suggestions for further research are reported on the last section of the paper.

2. Previous work

2.1. Quantitative analysis of Repositories

Since the concept of a Research Repository or Research Object is relatively new [6], our knowledge of analysing the contents of research repositories comes primarily from studies conducted in other types of data infrastructures, such as Learning Object Repositories (LORs). A series of seminal studies on the analysis of the contents and growth of digital repositories have been reported by Ochoa and Duval. The goal of these studies [7], [8], [9] was the application of techniques that measured and analysed the processes that create, publish, consume or adapt information in the context of learning object repositories. Several techniques and algorithms were employed in this respect. The purpose of these algorithms was: to apply a set of metrics that would facilitate the assessment of the quality of the learning object metadata within repositories and establish the potential relevance of the learning objects for a given user and situation; to assess the growth of the repositories by analysing the contents of metadata elements, such as the repository's size and growth over time, contributors' characteristics and the number of published material; to examine the relationship between the popularity of an object and its reuse [7], [8], [9].

The findings of the studies conducted by Ochoa and Duval [8], [9] showed some interesting patents regarding the growth, reuse and quality of metadata within repositories. For example, they observed abnormalities on the size distribution of repositories and surprisingly enough a linear growth over time regardless the size and popularity of the repositories. The number and the growth of contributors within repositories varied across repositories due to differences in the size and nature of each individual repository. Regarding the contributor's publication distribution (i.e. the amount of content deposited in the repository by a contributor), the conclusion was that it is relevant to the contributor's engagement with the repository. The issue of reusability of content within repositories was examined by Ochoa in a follow up study in the context of Learning Object repositories [7]. The results of the quantitative analysis were rather discouraging as on average a mere 20% of Learning Objects was reused. A very interesting and rather unexpected result was the lack of correlation

between the popularity of a learning object and its reuse. Finally, in terms of the quality metadata included in the repositories, Ochoa and Duval found the growth of repositories and changes in the nature of information deposited may have an effect on the actual quality of metadata used for the description of the contents of learning object repositories (e.g. inconsistencies in the use of metadata and vocabularies, different levels of completeness within and across repositories).

2.2. (Meta)data Quality

Data quality is defined as the state of completeness, validity, consistency, timeliness and accuracy that makes data suitable for a specific use [10]. Dekker [11] states that data is of high quality "if they are fit for their intended uses in operations, decision making and planning". There is no distinction between the data and metadata quality considerations [11]. Metadata quality is a vital factor for electronic interoperability [9], [12], [13], [14]. The growth, proliferation and evolution of digital objects are accompanied by an analogous transformation of their metadata which causes a consistency issue affecting at the same time their quality [9], [15]. In many cases, the larger the dataset, the greater the probability a problem will emerge [12]. Also, research has shown that there are effects of discipline of the quality of metadata, thus suggesting a cultural dimension on data quality (e.g. [16])

2.2.1. Quality Issues and Metadata Elements

Sokvitne[17] conducted a research about the effectiveness of the metadata elements of the Dublin Core for retrieval. Sokvtne was focused on the following metadata elements: title, creator, publisher, contributor and subject. The study showed problems with all the above elements. In particular, the DC.title and the DC.subject weren't adding any value for retrieval purposes, while the DC.creator, DC.publisher and DC.contributor presented inconsistent name formats. He concluded the study by questioning the suitability of the Dublin Core for information retrieval unless various problematic issues were resolved. The main issues were that the elements should be populated and used correctly, while precise instructions, descriptions and rules should be set.

Barton [12] outlined the areas where metadata element problems most commonly arise. These were:

- Spelling, abbreviations and other similar data entry errors and ambiguities.
- Author and other contributor fields. The most common issues are that the same name is entered differently (e.g. inconsistent entry of initial, first-last name ambiguity), a name can change (for instance if one gets married and adopts/adds the spouse's name) and finally synonyms especially in common names.
- Title. Many resources have more than one possible title, while others, particularly non-textual resources, may have no title at all.
- Subject – in the form of keywords and classifications. The main issue with the subject is who should enter the data; the author or the metadata specialist? The author can ensure the entry of the correct terminology but the metadata specialist can ensure the data consistency. The use of taxonomies and subject classification schemes is part of the solution.

- Date. Two main sets of problems are met in this element. Initially the format is the main issue as there are numerous formats that one could use. The format of the date entry should be strict, predefined and unique. The second issue is that date is often ambiguous as it may refer to publication date, submission date, date the record became available, etc.

2.3. The DRYAD Repository

Dryad is an open access repository that permits scientists – in pure sciences and medicine – to store, search, retrieve and re-use research data associated to their scholarly publications. Data are deposited as files with permanent identifiers (DOIs) and metadata. Collections of related files may be grouped into data packages with metadata describing a combined set of files. Currently the repository contains approximately 4500 data packages associated with scholarly articles published in almost 300 international journals [18].

Dryad's primary aim is to facilitate data discovery and reuse, thus guaranteeing the long-preservation of this [19]. Greenberg [3] established as the main two goals of the repository, "the one-stop deposition and shopping for data objects supporting published research" and "the support of the acquisition, preservation, resource discovery, and reuse of heterogeneous digital datasets". One of the strong and appealing characteristics of Dryad according to Peer [20] is that its curatorial team "works to enforce quality control on existing content".

Dryad's developers, by using the Singapore framework metadata architecture in a DSpace environment via an Extensible Markup Language (XML) schema [21], [22] and HIVE (Helping Interdisciplinary Vocabulary Engineering), implemented the infrastructure so that the automatically generated metadata inherit characteristics from their original sources by harvesting keywords assigned by authors and controlled vocabularies – ontologies[3]. The Singapore Framework for Dublin Core Application Profiles is a framework created in order to maximize interoperability and reusability (Dublin Core Metadata Initiative) by shifting from the "resource-driven legacy approach", representing an information package, to the granular component parts of a resource [22]. Dryad's metadata requirements are simplicity, interoperability and Semantic web compatibility [23].

Greenberg initially [24] and [25] performed quantitative studies which were focused on the reusability of the repository's metadata. The main findings of the studies, based on the study of two Dryad workflows, were that 8 out of 12 metadata elements (contributor, corresponding author, identifier citation, subject, publication name, description, relation is referenced by, title) had a reuse at 50% or greater. The researchers concluded that reuse was more common in the case of traditional bibliographic elements; and the generation of more accurate metadata earlier in the metadata workflow is necessary.As opposed to the studies conducted by Greenberg and colleagues on the re-usability of metadata, the research reported in this paper is focused on the identification of the main quality issues related to analysis of the metadata elements of the Dryad repository and how these may affect the measurement of the growth of its contents.

3. Methodology

A mechanism that involved the downloading of the metadata elements from the Dryad and their transformation to a proper format was employed. On January 2014, the metadata of the repository were harvested. At this point the Dryad was holding 4.557 packages, 13.638 data files, 287 journals, 16.595 authors and 751.658 times an instance of the repository was downloaded. The *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) Validator & data extraction tool* was used for the metadata harvesting[1]. A total of 516 xml files were downloaded (135MB). The xml files were merged into a single file using *Mergex*, a command line tool for merging xml files[2]. Finally, a method to use and analyse the data from the xml files had to be employed. Due to the descriptive nature of the statistical analysis performed it was decided to analyse the data using Microsoft Excel 2010. It was anticipated that the records per file would be more than 65.536. so using an earlier version of MS Excel would be rather problematic. Therefore the xml to Csv Conversion Tool[3] was used to transform the XML files into CSVs and import these to Excel. It is woth mentioning that importing directly the xml file to Excel provided very frustrating results. The converter provided 19 csv files: i) contributor, ii) coverage., iii) creator, iv) date, v) dc, vi) format, vii) header, viii) identifier, ix) listRecords, x) metadata, xi) record, xii) relation, xiii) request, xiv) responseDate, xv) resumptionToken, xvi) setSpec, xvii) subject, xviii) title and xix) type. A selected sample of metadata elements was analysed. These were: contributor, creator, date, subject, type, relation, coverage, dc, identifier and title. However, since the focus of this paper is on the presentation of the data quality issues, rather than a detailed description of the contents of the Dryad repository, a small subset of three metadata elements is presented: Creator, Type and Date. These elements represent typical cases where data quality issues can impede the quantitative and qualitative analysis of the Dryad repository.

4. Results

4.1. Creator

The number of contribution per author is depicted on table 1. In total 16.567 authors contributed 86.087 objects. As it is shown in Table 1, the majority of creators (i.e. authors of the research objects) contributed between one to five research objects in the repository.

Table 1. Amount of objects published by each contributor

Contributions	Amount	Contributions	Amount	Contributions	Amount
1	1422	11	248	21-30	286
2	6131	12	225	31-40	128
3	2282	13	137	41-50	129

1 http://validator.oaipmh.com/
2 https://code.google.com/p/mergex/
3 http://xmltocsv.codeplex.com/

4	1541	14	144	51-60	70
5	1060	15	84	61-70	46
6	773	16	92	71-80	35
7	601	17	100	81-90	25
8	396	18	82	91-100	15
9	362	19	55	>100	2
10	242	20	47	**Total**	**16567**

4.2. Date

This metadata element was assigned to various types of dates like date accessed, date available and date issued. For the purpose of this analysis we gathered the dates corresponding to the date issued (according to the cataloging guidelines of Dyad's[4] wiki, dc.date.issued is the official date of publication, inherited by dataset; the date of the formal issuance of the resource) of the 43.453 objects in the repository. The distribution per year is depicted in Table 2.

Table 2. Amount of Objects issued per year

Date	Amount	Date	Amount	Date	Amount
1995	1	2002	10	2009	416
1996	10	2003	11	2010	3172
1997	10	2004	13	2011	25411
1998	59	2005	12	2012	5035
1999	50	2006	13	2013	8005
2000	17	2007	27	1/1/2014-9/1/2014	176
2001	67	2008	97	Invalid input	841

It should be noted that there are two abnormalities in the flow of the records within the repository. On October 2010 2.572 publications where entered when the previous month the amount was a few dozens and on April 2011 the number was skyrocketed to around 23.000, more than half (52,67%) of the total publications of the repository. Since it is highly unlike that on a single month half of the input of the repository was published it seems that there is mix-up with date issued and the date input in Dryad.

4.3. Type

A total of 53.598 records were retrieved for the DC.Type element and their distribution is shown in Table 3. In the type field is shown the exact text that was found in the type field, except from blank were actually there was nothing inserted.

[4] http://wiki.datadryad.org/Cataloging_Guidelines_2009

Table 3. Type distribution of objects

Type	Amount	Percentage %	Type	Amount	Percentage %
Activity	4	0,007	Image	62	0,116
Article	4451	8,304	Map	1	0,002
Book	3	0,006	none	4086	7,623
Blank	4	0,007	oneyear	830	1,549
custom	109	0,203	protocol	11	0,021
Dataset	36708	70,167	untilArticleAppears	6429	11,995

As shown in Table 3, the Dataset type holds the vast majority of the dc.type element with 70,17%, followed by the Article with 8,30%. However, it is apparent that there are types in the table that should not appear in a first place like custom, blanks, none, oneyear. protocol and untilArticleAppears. According to the Dryad's Cataloging Guidelines dc.type is the "Code indicating the type of file. This is automatically detected by DSpace, but can be modified manually". Obviously there are issues with the automatic detection and irrelevant/unrelated with the dc.type entries are inserted. If we clean the data and leave only the suitable type files, then 42.129 records remain and the percentages change: Activity 0,009%; Article 10,565%; Book 0,007%; Dataset 89,269%; Image 0,147%; and Map 0,002%. Consequently, nearly 90% of the stored files are dataset and nearly 10% are articles.

4.4. Data quality problems

A significant number of major data problems were identified in the case of the Creator, Date and Type metadata elements. The methodology for the conversion and analysis of data was quite problematic. The noise accumulation and the incorrect assignment of the records to the proper fields were the main problems with the conversion. Data irrelevant to the fields and data misplaced made the initial files difficult to analyse and manipulate making a manual intervention essential. Furthermore, the quality of the data, an issue completely irrelevant with the conversion procedure, was not the anticipated one taking into account Dryad's development. The most common quality issues are summarised below.

4.4.1. Creator

The highest variety of issues was identified in this element. Out of 16568 records, a total of 1443 (8,71%) demonstrated the following issues:

- Additional names: Many authors were input with just their first name. The problem emerged in 614 (42,55%) cases when the authors' additional name were added as a different record and also by including additional ones (e.g. Aradhya, Mallikarjuna K. and Aradhya, Mallikarjuna).
- Using initials: Another serious issue was the use of initials instead of the whole name (11,64%). For instance Schim van der Loeff, M. F. and Schim van der Loeff, Maarten Franciscus.
- Differentiation of languages: A percentage of 12,06% occurred with this issue. There are numerous variations for writing a name in non-English language. Trying to convert a name by the English alphabet may be problematic as there are many symbols that do not exist. For instance, accent aigu or accent grave in French, umlaut in German, etc. make an error when writing a name very

possible. The most frequent mistakes were made in French, Spanish, Scandinavian, German, Chinese, Balkan and East Europe names. The use of short names and diminutives were also included in this category (e.g. Zach instead of Zachariah).

- Miswritten: With a percentage of 2,56% many errors due to typos were indemnified (e.g. Philipp instead of Phillip). In this category errors like when a first name was missing or when the name was inserted at the surname field were also counted.
- Dots and commas. The second most frequent mistakes (23,08%) were the absence of dots or the use of commas at the end of initials.
- Spacing: Different creator entries existed as in a few cases (2,36%) no or too many spaces were inserted during the name input.
- Miscellaneous: Issues like using irrelevant text (e.g. et al., PhD, status, code, etc.) were grouped in this category (0,83%).
- Ambiguous: There were around 71 cases (4,92%) were there was serious doubt whether different writings of a creator were belonging to the same person, mainly because they were very common (e.g. Gold, John and Gold, J. or Edwards, Mary and Edwards, M.).

It should be noted that in one occasion the names of a certain creator (that we will not write his surname) were input with six different ways (A Rus. – A. Rus – A. Russel – Alan R. – Alan Rus – Rus).The problems appeared in this element were also identified in the Contributor element; although an analysis was not performed, a rough review validated the same symptoms.

4.4.2. Date

Serious issues were also met at the DC.Date element. There was absolutely no consistency in the format and no control for the insertion of dates. As it is mentioned in section4.2 there were dates with invalid format: 4 dates from 1900-1904, 321 dates after the date that the metadata was harvested, 476 dates equal to 1/1/9999 and 40 dates that were blank or with text. Table 4 depicts the second main issue; the inconsistency in the date format. The length of the date varies from being blank up to 20 characters.

Table 4. Length of issued date

Length	Count	Example
4	156	2009
6 to 7	163	2009-03
8 to 10	42590	2009-09-07
20	503	2009-10-01T10:19:28Z
Various	41	Blanks, unacceptable format

4.4.3. Type

Almost twenty percent (21,4%) of the records in the DC.Type metadata element was jargon or blank or completely irrelevant to the element. The absence of data control and quality was more than obvious. As with the other elements a mechanism that will allow only correct data entry has to be employed.

5. Conclusions

The purpose of this study was to illustrate some of the main data quality issues associated with the use of metadata in the Dryad Repository. In addition to the reusability of research objects, addressing issues related to data quality of metadata in the Dryad repository is important for the accurate analysis and monitoring of the growth of the repository. In order to address this objective all the metadata from the Dryad repository were harvested and analysed. A plethora of data misuse issues were identified; issues that constitute the data inappropriate for text mining or data mining purposes. A mechanism that secures the metadata input from the issues that we identified needs to be employed. Data control would make repositories far more appealing and sustainable.

We propose a set of ideas that might enhance the quality of Dryad's metadata. For example, a solid format of the names should be specified. Each creator and contributor should be assigned with a unique ID that would hold their full name. When requesting an entry of the full name at the repository this unique ID should be inserted. To avoid any complications, the ID might be interlinked with an email. Possible synonymies can be resolved by the use of unique full names (e.g. different writing of first names, the use of initials, or the use of a father name should be implemented). If for any reason the creator wishes to change the name, then all of the records related with the name should be updated automatically, through the unique ID. In the case of dates, input should follow the same format (e.g. dd-mm-yyyy). Validation rules must be applied when each date is entered (e.g. it is more than obvious that a date cannot be posterior than the current date or prior than the creator's birthday). Finally, in the case of the type metadata element, inconsistencies can be fixed through the use of pre-defined list of values for authors to select from.

Based on the belief that "metadata solutions will become common-place for accomplishing various tasks" [26], our future work will be focused on Dryad Repository and the rest of its metadata elements. More elaborate statistical analysis by using R will be employed and data mining and text mining techniques will be applied in order to provide a better understanding of the repository's data, to identify any associations, clusters or hidden patterns and provide a visualization of these results.

References

[1] Garoufallou, E. and Papatheodorou, C. A critical introduction to Metadata for e-Science and e-Research, Special issue on Metadata for e-Science and e-Research. *International Journal of Metadata Semantics and Ontologies ((IJMSO)*, 9(1) (2014), 1 – 4.

[2]Bernard, J. et al. A visual digital library approach for time-oriented scientific primary data.*International journal on Digital Libraries*, **11(2),** (2010), 111-123.

[3] Greenberg, J. Theoretical considerations of lifecycle modeling: an analysis of the Dryad Repository demonstrating Automatic metadata propagation, Inheritance, and Value System Adoption. Cataloguing & Classification Quarterly, **47(3/4)** (2009), 380-402.

[4] Heery, R. Digital Repositories Roadmap review: towards a vision for research and learning in 2013. *JISC*. Available: http://kennison.name/files/zopestore/uploads/libraries/documents/reproadmapreviewfinal.pdf [29/12/2013]

[5] Kelly, B et al. Open metrics for open repositories. *In: OR2012: the 7th International Conference Conference on Open Repositories,* (2012).Available: http://opus.bath.ac.uk/30226/ [29/12/2013].

[6] Bechhofer, S., De Roure, D., Gamble, M., Goble, C., Buchan, I. Research Objects: towards exchange and reuse of digital knowledge. *FWCS2010*, (2010), Available: http://eprints.soton.ac.uk/268555/1/fwcs-ros-submitted-2010-02-15.pdf [10/03/2014]

[7] Ochoa X.Learnometrics: metrics for learning objects. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (LAK '11). ACM, New York, NY, USA, (2010, 1-8. http://doi.acm.org/10.1145/2090116.2090117

[8] Ochoa, Xavier and Duval, Erik. Quantitative Analysis of Learning Object Repositories. *IEEE Transactions on Learning Technologies*, **2(3)**, (2009), 226-238.

[9] Ochoa, X. and Duval, E. Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, **10(2)**, (2009), 67-91.doi:10.1007/s00799-009-0054-4

[10] Gordon, K. Principles of Data Management. Facilitating Information sharing, 2007, ISBN 978-1-902505-84-8. Available: http://www.bcs.org/upload/pdf/data-management-chapter1.pdf [13/03/2014]

[11]Dekkers, M., Loutas, N., De Keyzer M., and Goedertier, S. Open data and metadata quality. (2013). Available:https://joinup.ec.europa.eu/sites/default/files/D2.1.1%20Training%20Module%202.2%20Open%20Data%20Quality_v0.09_EN.pdf [25/01/2014]

[12] Barton, J., Currier, S., Hey, J.M.N. Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. *Proceeding of 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice – Metadata Research and Applications*, (2013), 39-48.

[13] Park, J. Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, **47(3)**, (2008), 213-228. doi:10.1080/01639370902737240

[14] Palavitsinis, N., Manouselis, N., Sanchex-Alonso, S. Metadata quality in digital repositories: empirical results from the cross-domain transfer of a quality assurance process. *Journal of the Association for Information Science and Technology*, (In press)

[15] Lee, D. Practical maintenance of evolving metadata for digital preservation: Algorithmic solution and system support. *International Journal on Digital Libraries*, **6(4)**, (2007), 313-326. doi:10.1007/s00799-007-0014-9

[16] Balatsoukas, P., O'Brien, A., and Morris, A. The effects of discipline on the application of learning object metadata in UK Higher Education: the case of the JORUM repository. *Information Research*,**16(3)**, (2011). Available: http://www.informationr.net/ir/16-3/paper481.html[1/2/2014]

[17] Sokvitne, L. An Evaluation of the Effectiveness of current Dublin Core Metadata for Retrieval.*Proceedings of VALA 2000*. Victorian Association for Library Automation: Melbourne, (2000).

[18] Dryad Digital Repository. Frequently Asked Questions.Available: http://datadryad.org/pages/faq[29/12/2013]

[19]Beagrie, N., Eakin-Richards, L., and Vision, T. Business Models and Cost Estimation: Dryad Repository Case Study*, iPRES2010,* (2010), Vienna.

[20]Peer, L. The Role of Data Repositories in Reproducible Research.Yale, (2013). Available:http://isps.yale.edu/news/blog/2013/07/the-role-of-data-repositories-in-reproducible-research#.UzINafmSxyM [12/0/2014]

[21] White, H. Carrier, S., Thompson, A., Greenberg, J., &Scherle, R. The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment. *DC 2008, the 2008 International Conference on Dublin Core and Metadata Applications*, (2008), Berlin.

[22] Greenberg, J., White, H., C, Carrier, S. and Scherle, R.A Metadata Best Practice for a Scientific Data Repository.*Journal of Library Metadata*, **9 (3)**, (2009) 194-212. Available: http://dx.doi.org/10.1080/19386380903405090 [15/2/2014]

[23] Greenberg, J. Linking and Hiving Data in the Dryad Repository. The Semantic Web: Fact or Myth. *CENDI, FLICC, and NFAIS Workshop.National Archives*, Washington, DC, (2009). Available: http://www.cendi.gov/presentations/11-17-09_cendi_nfais_Greenberg_UNC.pdf [9/1/2014]

[24] Greenberg, J. and Vision, T. The Dryad Repository: A New Path for Data Publication in Scholarly Communication. *OCLC*, Dublin, Ohio, (2011). Available http://www.oclc.org/research/news/2011-03-24.htm [22/1/2014]

[25] Greenberg J, Swauger S, Feinstein E.M. Metadata Capital in a Data Repository. *Proceedings of the International Conference on Dublin Core and Metadata Applications*(2013), 140-150

[26] Greenberg, J. and Garoufallou, E. (2013). Change and a Future for Metadata. In: Garoufallou, E. and Greenberg, J. (eds), *Metadata and Semantic Research: 7th Research Conference, MTSR 2013*, Thessaloniki, Greece, November 19-22, 2013. Proceedings. Communications in Computer and Information Science (CCIS), Vol. 390, pp. 1-5.

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-59

Developing the Greek Reference Index for the Social Sciences and Humanities

Victoria TSOUKALA,[a,1] Alexia PANAGOPOULOU [a], Giorgos STAVROU [a], Eleni ANGELIDI [a], Evi SACHINI [a], Alexandros NAFPLIOTIS [a]
[a] National Documentation Centre/National Hellenic Research Foundation, Athens Greece

Abstract. The Greek Reference Index for the Social Sciences and Humanities (GRISSH) is a service that collects, documents, stores and, where possible, provides access to peer-reviewed publications in the Social Sciences and the Humanities (SSH) by Greek publishers. It also provides long-term preservation for the digital and print files of the publications. The GRISSH was conceived by the National Documentation Centre (EKT) and the documentation and access platform is developed by the organization itself (OpenABEKT). The GRISSH is, in essence, a collaborative project that advances with the assistance and active participation of the publishing and scholarly community in Greece. It is intended as an essential reference service for the research and publishing community in Greece and abroad. The present contribution presents the goals, objectives and key benefits of the project; the evaluation criteria for the selection of content; the specifications for the development of the index; the methodology of documentation; the emerging collaborations with stakeholders. It also, present how the GRISSH project is aligned with the national and international agendas in view of a coordinated development and e-infrastructure that will support the sectors of academia, research and academic publishing.

Keywords. Greek Reference Index, Social Sciences and Humanities, peer review content, open access, documentation, services, e-infrastructures

Introduction

The National Documentation Centre (EKT)[1] is the national institution for the aggregation, documentation and dissemination of scientific information. Founded in 1980, the organization serves the country's research, education and business communities and the wider public. At EKT, scientific content e-infrastructures embrace technological innovation in fulfilling its main mission to aggregate, document, store and preserve digital content and disseminate it openly to the public in a way that promotes growth, research and innovation. Placing emphasis on multi-directional content reuse, EKT develops enabling factors for the creation, use and growth of digital content in its entire lifecycle.

Access to knowledge lies at the heart of EKT's activities. The organization is a strong supporter of open access as a means for social and economic development. It stands at the forefront of national and international open access initiatives at the policy

[1] Victoria Tsoukala, National Documentation Centre/ National Hellenic Research Foundation, Leoforos Vassileos Konstantinou 48, Athens 16635, Greece. Email: tsoukala@ekt.gr

and infrastructure level, such that allow the optimal circulation of scientific knowledge and digital culture for the benefit of society and economy.

In this context EKT develops e-infrastructures for scientific and cultural information with integrated value-added services of national and international impact. Such services of EKT's e-infrastructures are, for example, the National Archive of PhD Theses (www.phdtheses.ekt.gr), the ePublishing services (http://epublishing.ekt.gr), and repository services (Software as a Service). EKT's e-infrastructure development is co-financed by Greece and the European Union/European Regional Development Fund (Operational Program "Digital Convergence") through a large-scale project, the *National Information System for Research and Technology* (NISRT) (www.epset.gr).

This paper presents a new and innovative project in EKT's Humanities new agenda, the Greek Reference Index of Scholarly Publications in the Social Sciences and Humanities (henceforth GRISSH). The paper focuses on explaining the aims, scope and context behind this new project, the methodology used in implementing the project, the future planning and expected impact.

1. The GRISSH and its context: aims, scope and relevance

The GRISSH is a service that collects, documents, stores and, where possible, provides access to peer-reviewed publications in the Social Sciences and the Humanities (SSH) by Greek publishers. At the same time, it will provide long-term preservation for the digital and print files of the publications. The GRISSH was conceived by EKT and the documentation and access platform is developed by the organization itself. The GRISSH is a collaborative project that advances with the assistance and active participation of the publishing and scholarly community in Greece, whose needs it primarily addresses. The GRISSH is envisioned as part of a wider index that will gradually record the high quality publications of the country in all fields. At EKT we envision the GRISSH as becoming an essential reference service (hence the name Reference Index) for the research and publishing community, primarily in Greece but also abroad.

The project is currently close to launching the platform in a beta form that will be open for evaluation and comments by the stakeholder community in the early summer 2014. The platform will contain a sample of the material to start with.

With the GRISSH we aim to:

- Index and record in detail the Greek output in publications in the Social Sciences and the Humanities and provide access to it from a single point in the web
- Provide an essential research tool for the Greek and international research community in the SSH with enhanced search and browsing capabilities of the indexed content at the article level
- Promote the Greek scientific output as well as the research produced in the Greek language internationally and contribute to its increased international relevance
- Contribute to the long-term preservation of the aggregated content
- Afford the extraction of metrics and indicators regarding the Greek peer-reviewed output in the SSH, in such ways that they can serve evidence-based policies, and inform decisions of researchers, research performing and research funding institutions.

- Provide to Greek publishers indexing, repository and content management services for published work, as well as metrics regarding their use and impact
- Contribute to the development of a tight network of collaborating research and publishing institutions in the country in the SSH whom the GRISSH serves and through whom it is made possible
- Be part of a wider international network for similar indexes developed by European Member States and other countries in the SSH and other fields and interoperate with them.

EKT has, since many years, identified the fields of the Humanities and the Social Sciences as ones of great potential with regard to developing e-infrastructures and has, accordingly, directed numerous of its activities there,[2] which now also include the innovative pilot project GRISSH. This project addresses primarily the lack of systematic indexing of the Greek output in the SSH and the lack of international visibility, especially for the publications in the Greek language. Further, it essentially offers the structured information on which metrics and other indicator systems can be developed for the SSH publications in the future. The lack of indexing, and thus visibility, can be attributed to the particular research processes in the SSH, whereby the research and publication process is a much longer, and often a solitary, process, whose outputs in publications may take years to materialize. The SSH thus display a much slower dissemination process than, for example, the natural or medical sciences, and one that until recently has relied a lot less on technology than the first. Regional studies in the Social Sciences and the Humanities, often carried out in the local languages, in this case in Greek, further act as a factor of 'isolation', in the sense that this type of research is very difficult to include in international indices, such as the ISI or Scopus and disseminate widely. Finally, in contrast with other scientific output, Social Sciences and Humanities research is harder to evaluate using the standard methodology of referencing and impact factors[3]. For all of these reasons, EKT considers that recording the Greek publications in the SSH and making them widely available is of prime importance and urgency.

These issues have long since been identified and have prompted relevant efforts in Europe, turning the attention to the significance of national infrastructures that document and promote the dissemination of digital content in the Social Sciences and Humanities. Countries with strong presence in regional studies and languages other than English are engaged in producing similar indexes that record their output. In Norway, the Norwegian Association of Higher Education Institutions set up a national research database in 2004 containing the bibliographical details of all significant academic publications in all fields of science and scholarship, i.e. including the humanities. Also, in Flanders in 2009 work began to construct the Flemish Academic Bibliographic Database for Social Sciences and Humanities (VABB-SHW)[4]. Additionally, the European Science Foundation (ESF) started the initiative for the *European Reference Index for the Humanities* (ERIH)[5], aimed at presenting the Humanities's impressive track record and ongoing research achievements systematically to the rest of the world. It is within this context that the GRISSH is being developed; further, it has been considered of vital strategic importance that our effort align with such national and international agendas as described above in view of a coordinated development and e-infrastructure that will support the sectors of academia, research and academic publishing.

The GRISSH will offer some core benefits to the Greek research, academic and publishing community: a powerful single gateway to a bibliographic database of Greek

publications in the Social Sciences and Humanities, international exposure and greater visibility for the publications, and the ability to reuse open access content. It will facilitate statistical and evaluative information-gathering as there is provision for a systematic metrics service that will document the impact and use of this content. It will afford easier communication of this national database with those of other European countries and with the European database that has been envisioned by the ESF. Finally, a by-product of this endeavor will be a wider re-use framework of this content for business innovation, education and lifelong learning.

2. Implementing the Index: Methodology and status of work

2.1. Content selection

Recognizing that the selection of content is of critical significance for the GRISSH, EKT developed an acquisition policy in order to compile an initial list for its content. The first content target group, on account of its scholarly significance, has been the SSH journals published in Greece. Other types of materials will be included in the future, such as monographs. Aim of the acquisition policy was to initially identify the most important journals in the SSH produced in Greece. In doing so, it is understood that this initial list may change many times in the future, as GRISSH will depend on the characteristics of the publications themselves. The policy was broadly based on a number of quality criteria set by indexing services, such as the European Reference Index for the Humanities or Jstor, as well as criteria determined by the aim to reflect as accurately as possible the quality-assured journals of the country in the SSH in the Reference Index6. Those criteria are: that the journals implement specific evaluation/quality assurance processes for their publications; that they are 'active' journals and display timeliness of publication; that they may have very recently ceased to operate but have displayed timeliness and quality assurance for long periods in the past and are considered significant tools for research by their respective research communities; that their subject-matter is not strictly focused on a specific location in Greece (e.g. local city journals); that they are the official instruments of publication of the research of the important Greek universities and research centers. The starting point was a list of scientific periodicals published in Greece that was included in European Reference Index for Humanities in 2007. These periodicals were hailed as important in their respective area of expertise, by consortiums of science experts. Additionally, a thorough search in Ulrich's index, afforded an initial list of approximately 165 journals that had to be examined. With the help of experts in the various fields and using the aforementioned criteria approximately 85 journals were validated for inclusion in the GRISSH, an estimated total of about 40.000 articles to be documented in the bibliographic database.

Subsequently, a strategy was developed to involve the publishers in the project and spearhead the collection of publications. In communicating with the publishers due attention was paid to conveying to them the significance of the venture and the benefits for them, as well as being consistent with written and phone communication. As a result of a successful communication campaign, publishers have been motivated and have submitted their journals in print form to EKT. The organization now possesses most of the journals comprising the list in actual print copies. As part of our digitization and long-term preservation strategy, digitization will take place where necessary.

2.2. Technology

The specifications for the development of the Index are laid on two tiers: developing the bibliographic e-infrastructure that will be used for the documentation and developing the user interface for the end user.

The bibliographic system that supports the GRISSH is the OpenABEKT, an online tool developed by EKT. Developing the functionalities that will serve the purposes of the GRISSH was a very important stage in proceeding with the project. The OpenABEKT web platform serves for the documentation of periodicals, issues and articles using UNIMARC as the metadata schema of the system. Each record entry into the system is essentially part of the Index database. The core benefits of the web platform are: (a) it provides different documentation forms to cater for the documentation needs of both experienced (i.e. librarians) and non-experienced end users (i.e. publishers); (b) it is easy to access to, navigate and search in the content; Given that one of the key goals of the Index is to provide, in the future, to participating publishers the ability to insert metadata by themselves, care was giving in developing a user-friendly interface for non-specialized users that comprises forms with free text, drop down menus and controlled value lists facilities. The OpenABEKT has the following functionalities for the end users: view the journals, issues and articles lists; submit new items; edit past submissions; create authorities using simple, user-friendly forms; upload the digital copies of the articles as well as publication covers.

For designing the web interface and functionalities of the Index a Best Practices[7] research was implemented so that widely-used functionalities are adopted. The specifications for the development of the index platform were designed to cater for usability, content accessibility and satisfaction by delivering two-tiered user services: experienced and non-experienced user. The platform functionalities are: simple and advanced search, browsing by Author, Subject, Frascati subject category, periodical and publisher, search filters for defining results, refinement of the search results, keyword search through open-access digital files and statistical information on the use, content and Index users. Moreover, users will be able to view the full bibliographic metadata records and the article references, to view and save open access digital files of their choice, to export citations of an article in various bibliographic styles, such as Harvard and APA, and to print, send by email and share items with their social networks.

2.3. Documentation

For the documentation of the journals and articles it was necessary to collect all controlled vocabularies, thesauri and authority names and subjects in order to create a core database of authority files for the Social Sciences and Humanities.

Considering that thematic documentation of articles is of major importance, as it facilitates information retrieval from the Index, we designed and created a workflow for developing a controlled vocabulary of authority subject headings. Developing a thorough inventory of the fields within the Social Sciences and Humanities, we researched and compiled subject headings from the Library of Congress online catalogue. The chosen main subject headings were 31 records. These records include

coded information about wider and narrower relationships with other subjects. By using OCLC service tool we downloaded the 31 records and mapped them from Marc21 to UNIMARC. Finally we entered them into the database and linked them to the wider and narrower relationship terms, adding up to a final sum of 2100 subject authority records. The Frascati taxonomy, which also offers a structured thematic map of the Social Sciences and Humanities, was incorporated into our documentation system OpenABEKT in two languages, English and Greek. Thus, by achieving thematic structuring of our content we can draw useful metric information and offer additional search fields and browse functions on the Index platform.

Additionally, we utilized the existing EKT's ePublishing author Index (http://epublishing.ekt.gr) as a basis in compiling the authorities for personal names and publishers that work within the Social Sciences and Humanities. Following that, the authorities were enriched and curated with further elements drawn from the National Library of Greece, the National Book Centre and the Library of the Congress. The English translation of Greek names was performed using the ELOT 743 standard.

Finally, we documented in detail all the 85 journals which were initially selected to be part of the Index, all the issues that EKT received from publishers, along with some articles, in order to test the system and its functionalities in the beta stage.

3. Future plans

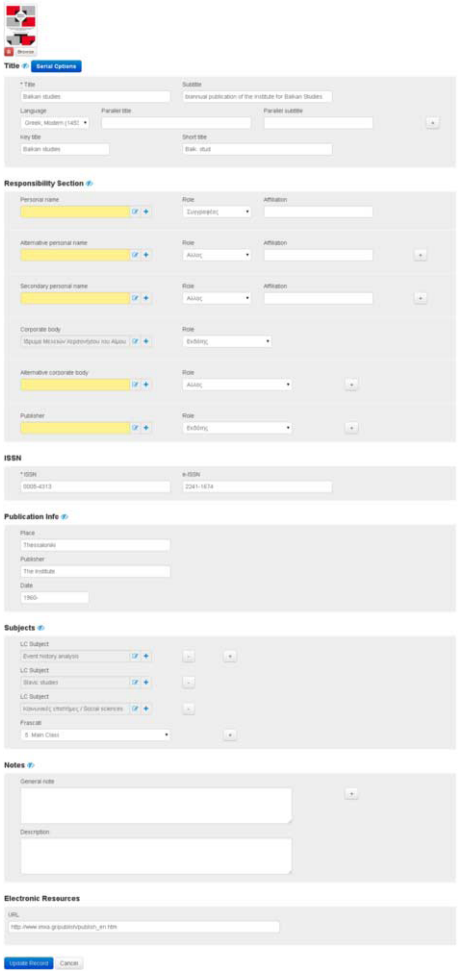
Having entered a mature phase in the project, whereby the planning, acquisition of content and development of the greatest part of the e-infrastructure is in place, EKT is soon to test the platform and expeditiously document the metadata of all articles in the bibliographic database. The metadata of all journals will be accessible through the platform in the beginning of 2015, along with openly accessible content, where possible.

As underlined above, this is a project whose value will be fully realized when it becomes useful to the scientific community, the publishers, as well as policymakers who are to benefit from it. EKT, therefore, intends to mobilize those stakeholder communities very soon. Their consistent involvement with the GRISSH will help shape its future direction according to the needs of the stakeholder communities it was designed to serve. To ensure that the Index addresses all the current sector thinking, EKT plans to call for consortiums of experts to help with the development and continuing evaluation of the service. In the following months, specialist committees comprising established researchers and academics will be called upon to act as an advisory board. Their role will be to evaluate the development of the Index, to suggest new services based on international best practices, to communicate the needs of the scientific community and to participate in the review process for new content.

The GRISSH in this initial phase is only the basis for the development of the bibliographic database with other types of content, such as monographs and proceedings, for the development of new services, such as statistical and indicator services and citation index services, among others. It is the intention of EKT to eventually incorporate the GRISSH into the central European database for the output in the SSH, once available, so that Greek research outputs are accurately and adequately represented internationally and so that key societal benefits can be harvested through this endeavor. Additionally, EKT intends to instigate close collaboration with other

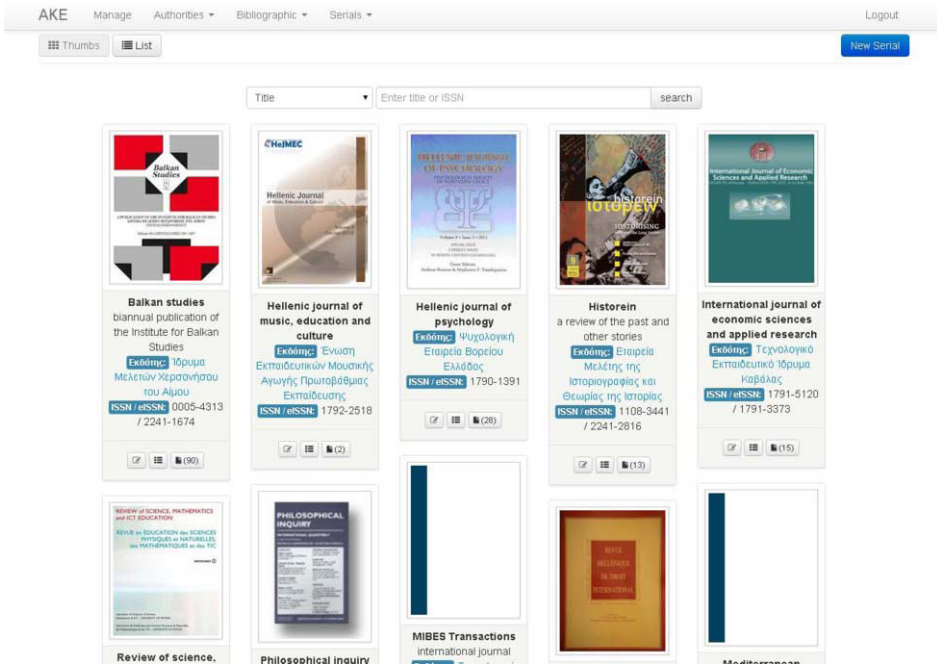
countries that have initiated similar efforts for mutual benefit, advancement and alignment of the initiatives. An integrated collaborative policy is being developed towards this aim.

Pictures

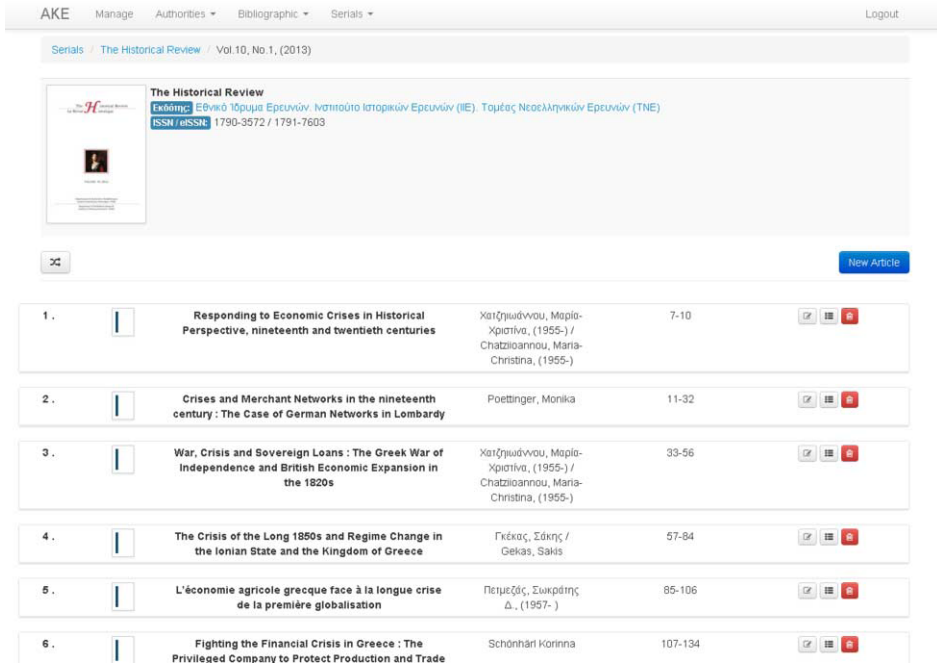


Picture1: documentation form for non-experienced users

Picture 2: documentation form for experienced users



Picture 3: main catalogue at documentation system Open ABEKT



Picture 4: articles of an issue at documentation system Open ABEKT

References

[1] National Documentation Center. Available at: www.ekt.gr

[2] Other important activities are: the repository of primary sources "Pandektis" (http://pandektis.ekt.gr), the repository of the Parthenon frieze (http://repository.parthenonfrieze.gr/frieze/), and the EKT's ePublishing service, which largely focuses on the Social Sciences and the Humanities (http://epublishing.ekt.gr)

[3] Royal Netherlands Academy of Arts and Sciences (2011) "Quality indicators for research in the humanities", Interim report by the Committee on Quality Indicators in the Humanities.

[4] Royal Netherlands Academy of Arts and Sciences (2011) "Quality indicators for research in the humanities", Interim report by the Committee on Quality Indicators in the Humanities.

[5] European Science Foundation, "European Reference Index for the Humanities (ERIH)". Available at: http://www.esf.org/index.php?id=4813

[6] Martin, B., Tang, P., Morgan, M., Glänzel, W., Hornbostel, S., Lauer, G., et al. (2010) "Towards a bibliometric database for the social sciences and humanities—A European scoping project" a report produced for DFG, ESRC, AHRC, NWO, ANR and ESF. Sussex: Science and Technology Policy Research Unit

[7] Ossenblok, T.L.B., Engels, T.C.E., Sivertsen, G. (2012) "The representation of the social sciences and humanities in the Web of Science - A comparison of publication patterns and incentive structures in Flanders and Norway (2005-9)", Research Evaluation, Vol. 21, Issue 4, pp. 280-290.

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-68

A Digital-First Authoring Environment for Enriched e-Books using EPUB 3

Ben DE MEESTER [a], Tom DE NIES [a], Hajar GHAEM SIGARCHIAN [a], Miel VANDER SANDE [a], Jelle VAN CAMPEN [a], Bram VAN IMPE [a], Wesley DE NEVE [a,b], Erik MANNENS [a] and Rik VAN DE WALLE [a]

[a] *Ghent University - iMinds - MMLab, Belgium*
[b] *KAIST - IVY Lab, Republic of Korea*

Abstract. The overall majority of books are currently being made with primarily a printed outcome in mind. To make a digital version of these books, most manuscripts need to be re-processed, which usually results in customary built e-books. This need for a customized authoring workflow for every electronic version of a book makes it impossible to build e-books in a cost-effective way. In this paper, we propose a novel workflow that incorporates both print and digital book authoring.

By charting the currently most widespread workflow Flemish publishers use to author print books and e-books, we are able to identify the most pressing problems. These are the print-first approach, the vendor lock-in situation of the e-reader market, and the high cost of updating and/or maintaining the content of an (e-)book. To overcome the aforementioned problems, we devise a new workflow that follows a digital-first approach using Open Web standards, separating content, structure, and layout. We evaluate the proposed workflow by building a proof-of-concept authoring environment.

Using this new workflow, both digital and print books can be built without significant additional costs. The proof of concept is evaluated using an experts group of Flemish publishers, and received general positive reception, with concerns on how to incorporate the proposed workflow into production environments. By not limiting the proof of concept to a fixed data model, it could handle content from more content providers, facilitating further research into the possibilities and future requirements of the EPUB 3 specification.

Keywords. Authoring environment, EPUB 3, Open Web standards

Introduction

In recent years, digitized and possibly enriched versions of print books (*e-books*) have evolved from a novelty to an integral part of the book reading market. More and more people are reading digitally, and there are many indications this trend will continue, resulting in a market where e-books will become at least as important as their printed counterparts [1].

Meanwhile, e-book authoring is usually seen as an afterthought, and the authoring of most books is still largely based on the same publishing and associated printing techniques of the last decades of the twentieth century (i.e., *print-first*) [2]. These workflows

follow a basic **linear progression**, from the author that provides the content to the publisher that prints that content, after which e-books are produced.

However, there are three main discrepancies between print book authoring and digital book authoring:

- Authoring for print books is primarily **page-centric** instead of content-centric, which results in a tight coupling between content and layout.
- Current publishing software enforces a **print mindset**, therefore lacking digital-only book elements such as interactive quizzes, animations, and videos.
- Print book authoring has only one outcome in mind, whereas digital book authoring requires the need for **multi-platform publishing**. Indeed, digital books are not only read on e-readers, but also on smartphones, phablets, tablets, and desktops [1].

Besides the great diversity among current reading platforms (e.g., smartphone versus desktop), there is also a great diversity in terms of e-book formats [1]. As *EPUB 2*, the previous standard for e-books, could not handle advanced interactivity and multimedia features, e-reader vendors developed **custom e-book formats** to provide advanced features to their readers. These custom formats are not portable between reading systems, and force customers to reside with one reading system [3], as customers would have to re-buy their e-books if they would switch between vendors. This situation is called *vendor lock-in*. Examples of vendors and their formats include (1) Amazon, with its proprietary format KF8 for Kindle, and (2) Apple, with its iBooks format. Note that Apple e-readers currently also support the open EPUB 3 format [4].

The above-mentioned three discrepancies, and the existence of multiple e-book formats contribute to the fact that authoring a digital book based on a print book format is cumbersome and labor intensive. First, the content has to be decoupled from the print layout and a new digital layout has to be devised. Second, books have to be enriched afterwards by experts on a book per book basis. And third, different e-book formats have to be ported to one another, and possible corrections in the manuscript have to be propagated to all used e-book formats.

This current situation leads to the so-called **e-book price paradox**. Publishers are motivated by the market growth to author e-books as well as print books, but the extra effort needed to author e-books increases their price [5]. In the meantime, the consumer market expects a digital book to be cheaper than a printed copy, assuming that there are practically no costs associated with the authoring of e-books. Publishers are thus forced to lower their prices, making it currently impossible to author e-books in a cost-effective way [6]. This holds especially true for small organizations, which typically have less technological and budgetary resources than market leaders. As such, a strong need exists among publishers for tools that allow overcoming the issues that contribute to the high cost of current e-book authoring.

The remainder of this paper is organized as follows. After a review of previous research efforts in Section 1, we analyze the current publishing workflow of Flemish publishers in Section 2, subsequently identifying and discussing the most prominent issues. In Section 3, we introduce a new workflow. Next, in Section 4, we discuss and evaluate the proof of concept authoring environment that we have built to validate the newly proposed workflow. Finally, we present conclusions and directions for future work in Section 5 and Section 6, respectively.

1. Related Work

Previous analysis of a print-first workflow has resulted in two general recommendations on accommodating this workflow to the new requirements of authoring e-books. In the first recommendation, the book is processed as usual, using current printing techniques, after which a conversion is performed to make the book ready for digital publication [7]. The second recommendation states to completely redesign the workflow to be adaptable to both the traditional print books and the newer e-books [8]. Because of the heterogeneous e-reader market [1], developing a general authoring tool to output all e-book formats is labor intensive, and many publishers have adopted the first alternative to lower development costs [9].

There are a number of digital book authoring tools that do exist (e.g., OERPUB, PressBooks, and Pubcoder, but none of them are built with general applicability in mind. Moreover, these tools typically have a limited support for interactivity and multimedia features, and have very limited customization options. There also exist digital-only distribution channels (Issuu/Open Edition) and hybrid sites that offer personal authoring and publishing facilities (Ourboox). However, all these alternatives[1] are targeting authors. As they lack the link with print book publishing, and are tailored to individuals publishing their own digital books, they cannot be used in current publishing environments.

Meanwhile, since the fall of 2011, the *IDPF*[2] finalized *EPUB 3*, the latest version of the open e-book standard [10]. The most important improvement of EPUB 3 in comparison with EPUB 2 is the **support for Open Web standards**, including *HTML5*, *CCS3*, and *JavaScript*. By fully supporting those standards, the possibilities of the e-book format have increased greatly, and adding interactivity and multimedia features to an e-book has become easier and more maintainable [11,12], foregoing the need for vendor-specific solutions to build enriched e-books.

With EPUB 3 supporting advanced features by default, custom formats are no longer necessary, making it possible for the e-reader market to become more homogeneous. Publishers can now choose to build valid EPUB 3 e-books, supporting interactivity and multimedia features, and the need no longer exists to port this EPUB 3 to custom e-book formats. This makes rethinking the current authoring workflow more advantageous than trying to extend the current workflow of print book authoring that uses e-book authoring as a last step [13,14].

2. Current Workflow for Digital Book Authoring

By querying an experts group of nineteen publishers, we identified the **most widespread workflow** in which books are currently being authored in Flanders, resembling the publishing chain used in Britain and America, as presented by J.B. Thompson [9] (Figure 1). The dashed arrows indicate where no automatic conversion between formats is possible.

Analyzing this workflow together with the experts group, we identified three major problem areas that need to be solved in order to be able to move towards cost-effective e-books: the current *print-first approach*, *vendor lock-in*, and the *costs of corrections and updates*.

[1] www.oerpub.org, www.pressbooks.com, www.pubcoder.com, www.issuu.com/openedition, and www.ourboox.com, respectively

[2] http://idpf.org/

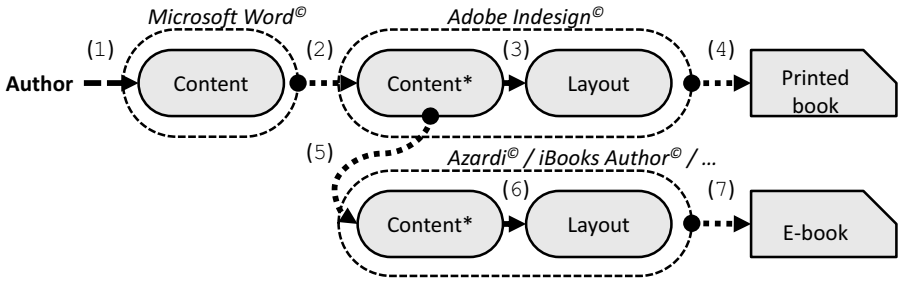


Figure 1. The current workflow books are usually produced in, identified by querying an experts group

Print first The making of a new book starts with the writing of a text by an author using text processing software (1). Then, once the author has finished his or her version of the text, this version is provided to the publisher (2). In a number of cases, a selection is made out of the provided content to appear in the book, which is denoted in Figure 1 by the asterisk at *Content*. Then, layout and structure are added to the content, using specialized software (3). After the layout has been completed, the book is ready for print (4).

However, when a proper electronic version of the book is requested, the **structure and layout usually need to be redone** (5, 6). This is because of the fact that the proprietary format of layout software is tailored to meet the strict requirements of a print book, whilst e-books generally provide a lot more possibilities than a print book. A digital book can have a dynamic layout, adjustable to the reading environment, with interactivity and multimedia features, and dynamic content. These are properties that are not present in a print book, and that have to be added afterwards, which is a cumbersome and expensive task.

Vendor lock-in Our inquiry made clear that a significant amount of authoring print books is done using proprietary software. *Microsoft Word* is the prevalent market leader among authors to write their content in, and *Adobe Indesign* is mainly used to edit the structure and layout of a print book. When authoring e-books, many software packages exist, but *Azardi* and *iBooks Author* are used primarily.

Also, most e-book authoring software is still tailored to output proprietary e-book formats. This vendor lock-in leads to very strict **limitations with regard to innovation and creativity**, since vendors often decide to only partially implement a standard, to guarantee their vision of what a good user experience constitutes [15]. For example, Apple poses significant limitations on their devices regarding interactive scripts and simultaneous video playback [16].

Corrections and updates Because of the uni-directional way of making a book (the author provides the content, after which the publisher puts the content in a certain layout, and where the result is finally published in different formats), it becomes very challenging to make adjustments once in the layout phase of the workflow. These adjustments can no longer be done by the author, but have to be done by the layout designer, as these layouts are made in a specialized application, and there is **no current way of updating the content without interfering with the layout**.

The situation worsens in a *multi-channel* publishing environment. As we see that usually no automatic conversion is possible between the content of a print book and

a digital book ((5) of Figure 1), the content does not only have to be adjusted in the layout software for the print book, but also in the specialized software for authoring the electronic versions of the book.

3. Proposed Workflow

As EPUB 3 fully supports the latest Web standards, it becomes clear that e-books are shifting from static content pages towards **packaged interactive web pages** that can be easily accessed offline. This opens the possibility to rethink the authoring of e-books as if it were designing Web pages. Many concepts of Web design can thus be reused to our advantage.

Our proposed workflow, as depicted in Figure 2, tackles the most prominent issues with the current way of authoring books, listed in Section 2. As also suggested by Silva and Borges [8], a dynamic information flow is devised, focusing on digital publishing with a clear division between content, structure, and layout.

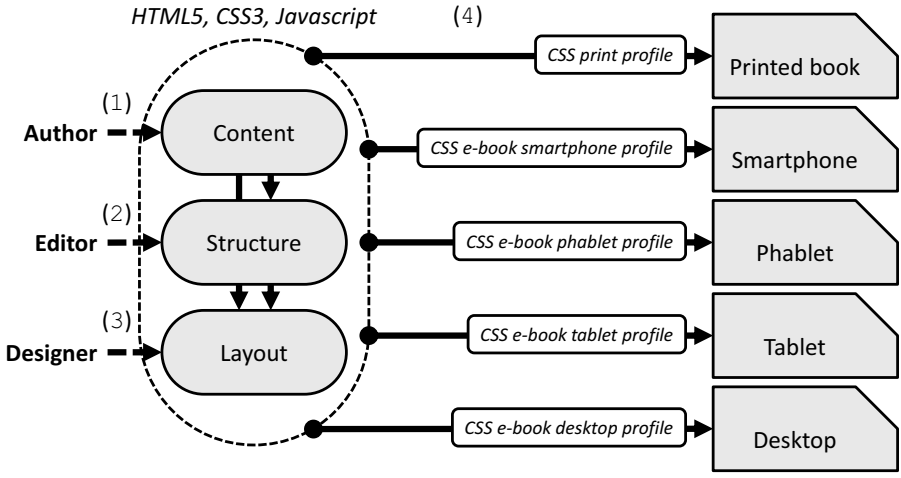


Figure 2. The proposed workflow, tackling the most pressing problems of the current most widespread workflow of (e-)book authoring

Focus on digital As e-books offer a lot more possibilities than traditional books, the proposed workflow focuses on the digital product. By using tools that focus on a digital outcome instead of a printed outcome, a book can be created that makes full use of the possibilities provided by the Open Web standards.

By using *graceful degradation*, a concept that originates from the area of Web design [17], a book can be authored once, and visualized differently in multiple reading systems. When using graceful degradation, the most advanced interactivity and multimedia features can be integrated as primary content, and *fall-backs* are introduced as a replacement when a reading system does not support those advanced features. For example, when a video is embedded in a book and a printed version is requested, a fall-back

can be created as a representative still from that video, a textual representation, or the video could be omitted in the printed version altogether.

Using this concept, only one version of the book is necessary, as long as good fallbacks are used for different reading platforms. The distributed versions of the book are then all alike for different reading platforms, except for the **different styling profiles** that are applied to the different versions ((4) of Figure 2).

Division between content, structure, and layout To address the problem of maintaining corrections and updates, we envision a strict division between content, structure, and layout, as it is the case in modern Web design. As shown in Figure 2, authors, editors, and designers can work collaboratively ((1), (2), and (3)), and their changes are **automatically passed on to each other** (as shown in Figure 2, using the filled arrows between *Content*, *Structure*, and *Layout*).

Open standards A logical consequence of focusing on a digital outcome is the embracing of the latest Web standards to produce an e-book. As is shown in Figure 2, no proprietary software is part of the publishing workflow. By using the same open standards that are used by EPUB 3, **no conversion is needed** between the content built by the authoring tool and the packaged e-book (hence the filled arrows between the application and the (e-)books). This idea results in an integrated design software which is purely Web based, where the author and publisher can both work in, without the need of proprietary software.

4. Proof of Concept

To validate the proposed workflow, we have built a proof of concept implementation, resulting in a **Web-based authoring environment** using a *Software as a Service* (SaaS) model, i.e., an application completely running in a web browser. This proof of concept is the result of multiple iterations of rapid prototyping [18]. These iterations have undergone multiple feedback sessions with the experts group, in the form of meetings and hands-on workshops.

A back-end provides the RESTful Web service [19] that can edit the different elements of the book, whereas a front-end is provided in the shape of a Web interface. Figure 3 depicts the system design, together with the distinctive views that are provided by the front-end. Three views are available in the application, accommodating the three most important roles in the book authoring process:

- a *content view* for the author, where individual blocks of content (e.g., a paragraph, a video, or a heading) are uniformly defined.
- a *structure view* for the editor, where (non-)linear links between chapters can be defined and content blocks can be assigned to the different chapters.
- a *layout view* for the layout designer, where visual adjustments can be made to the individual content blocks.

These separated views enable multiple people with different expertise to collaboratively work together on the same book.

Because both the application and the content are based on Open Web standards, searching content is automatically provided by any Web browser, and multi-platform

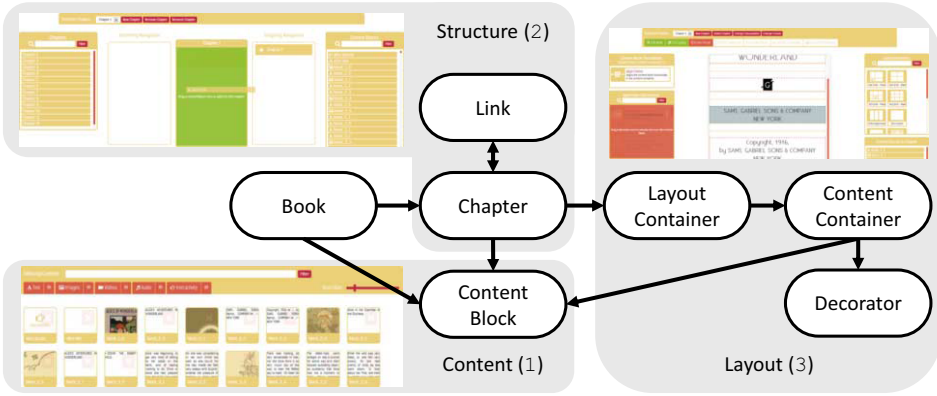


Figure 3. System design showing the division between content, structure, and layout, as well as their visualizations in the built proof of concept

compatibility is provided by default. A demo video of the proof of concept implementation is available[3], where an EPUB 2 file is loaded into the application, adjustments are made to the content, structure, and layout, and where the result is exported as a valid EPUB 3 file, retaining all interactivity and multimedia features.

Content (see (1) of Figure 3) In our authoring environment, every atomic part of a book, further referred to as a *content block*, is handled individually. By design, no distinction is made between different types of content blocks. In practice, this means no distinction is made between regular text, images, videos, and interactive widgets. Also, custom types can be defined using an XML-template, making the application extensible. The **homogeneous handling of heterogeneous content** encourages content providers to cut loose from the habits of the old book authoring process, where flat text is seen as the single most important aspect of a book, and multimedia and interactivity features are added afterwards. This also implies that adding an advanced feature requires the same effort or skill as adding regular text for a user of the authoring environment.

Structure (see (2) of Figure 3) By decoupling the structure from the content, the editor can decide how to divide the content blocks between different chapters. This makes it possible to prevent particular content blocks to appear in the eventual book. That way, unfinished content blocks can be present in the authoring environment during the editing process, whilst the publisher can remain sure the exported e-book will always be a *camera-ready* version. This enables *lean publishing* [20], in which books are published multiple times in intermediate versions, to receive feedback early-on from readers, and receive traction among the readers base of publishers.

Also, no linear direction of chapters is forced by the application. Instead, *Links* can be made between chapters in an arbitrary way. This makes it possible to make a hierarchical table of contents, and to **interlink chapters in multiple ways**. This is particularly interesting for cookbooks (where chapters can be clustered by type of dish), educational books (where chapters can be linked according to multiple lesson plans), or travel guides, where landmarks can be interconnected based on their relative distance.

[3] http://users.ugent.be/~bjdmeest/PotF_demo.mp4

Layout (see (3) of Figure 3) The layout is decoupled from the structure and content using two types of wrappers: *layout containers* and *decorators*. Layout containers can consist of multiple *content containers*, and each content container can contain exactly one content block. The layout containers decide how content blocks are positioned relative to each other using these content containers (e.g., a caption content container underneath a main content container). The decorators add special styling to a content container (e.g., changing the color of the background of a content container).

In the front-end, a *What You See Is What You Get* (WYSIWYG) editor is provided, where designers can do the markup of a book, and possibly extend the default layout containers or decorators with custom templates.

In the layout view, it is also possible to **preview the book in different e-reading environments** (i.e., switching between the resolutions of a smartphone, phablet, tablet, or desktop).

4.1. Qualitative Evaluation

The proof of concept of the proposed workflow was evaluated using a hands-on workshop with eight organizations that are stakeholders in the Flemish publishing landscape. This includes five publishers, two ICT organizations, and the metadata center for the Dutch book trade.

Using a 5-point *Likert scale* [21] and an open questionnaire, we surveyed the participants about their prospects for (interactive) e-books in general, the feasibility of the proposed workflow in a production environment, and the usability of the proof of concept. The potential of electronic publishing is rated highly among all participant, with little variance between e-book use cases (e.g., education versus fiction). However, the proposed workflow received mild acceptance among participants. Many questions have been received about how the workflow could be integrated into their current publishing processes, and significant concerns were raised at how well the proof of concept handles already published e-books. Positive feedback has been received regarding the ease of adding interactivity and multimedia features to an existing e-book. Concerning usability, reactions were largely positive, but no consensus was found on who would gain most benefit using the proposed proof of concept (e.g., authors versus publishers). Opinions differed on whether authors are willing to write and edit their content in the authoring environment.

5. Conclusions

Based on an inquiry of an experts group, we identified the most prominent issues with the current book authoring processes in Flanders as the **print-first approach, vendor lock-in, and cumbersome corrections and updates**. We addressed these issues by proposing a new workflow that uses a digital-first approach. As a consequence, open standards can be used to output truly interactive EPUB 3 files, with elegant fall-backs for printed media.

We developed a proof of concept as a RESTful web service using Open Web standards to facilitate an evaluation of the devised workflow by a group of Flemish publishers of different domains (mostly educational and children's books). Evaluating the proof of concept, we see that although the importance of electronic book publishing is understood

by the publishers, considerable concerns are raised about incorporating this new form of publishing in their current workflow and handling legacy content.

By **handling any type of content block uniformly**, advanced features that are supported by the EPUB 3 format can be integrated into a book with no extra effort. Using a **digital-first approach**, Open Web standards can be used fully without having to take into account the limitations of the print book format. By strictly **separating content, structure, and layout**, users are encouraged to collaboratively work in parallel on the same book, without the possibility of influencing the work of each other. Also, maintenance costs are lowered, as adjustments in one editing view (e.g., Content View) are automatically propagated to the other editing modes (Structure and Layout View).

However, by using a strict data model in our proof of concept (Chapter > Layout Container > Content Container > Content Block), we are limited in supporting legacy content without a conversion between formats. In addition, connections with other publishing services and workflows become more complex. Furthermore, limiting the workflow to a purely digital-first approach creates high concerns among publishers, as it is orthogonal to their current workflow.

6. Directions for Future Work

Using a more **flexible data model**, it should be possible to handle legacy content more robustly, as well as connections with content providers through the use of services. This could result in a hybrid approach, where print and digital are handled equally, allowing for an easier integration into current workflows, thus increasing production applicability.

Also, an important factor that is currently missing from the workflow is the addition of *metadata* to the content. A well annotated e-book opens the door for books being discovered more easily, dynamic on-topic extra content, semantic search engines, etc. [1, 22, 23].

As e-books are currently only being built as a last step in the publishing process, no effort is done to correctly annotate the content in any prior steps. Moreover, as manually annotating books is cumbersome and expensive, this step is usually omitted altogether. However, with the proposed workflow, (semi-)automatic **semantic annotation can be incorporated into e-books from the very first step** of the authoring process. Indeed, the proof-of-concept that separates the content from the other parts of the authoring process is an ideal base to analyze and annotate the content in a (semi-)automatic way.

Further research is encouraged to explore solutions into incorporating annotation techniques into the proposed workflow and proof-of-concept, as to investigate the possibilities these annotations bring with them.

Furthermore, the proposed generic authoring environment can be used to do further research on experimental features of future e-book formats, as proposed by the W3C Digital Publishing Activity[4].

Acknowledgements The research activities described in this paper were funded by Ghent University, iMinds, the IWT Flanders, the FWO-Flanders, and the European Union, in the context of the project "Uitgeverij van de Toekomst" (Publisher of the Future). Furthermore, we would like to thank *ADZ*, *Averbode*, *Crius*, *Die Keure*, *Intersentia*, *Lannoo*, *Meta4Books*, and *Van In* for their feedback.

[4] http://www.w3.org/dpub/

References

[1] Rüdiger Wischenbart. *Global eBook: A report on market trends and developments*. O'Reilly, 2013.

[2] Bill Martin, Hepu Deng, and Xuemei Tian. Expectation and reality in digital publishing: Some Australian perspectives. In *Openness in Digital Publishing: Awareness, Discovery and Access*, volume 11, pages 199–208, Vienna, Austria, June 2007. ELPUB.

[3] Christoph Bläsi and Franz Rothlauf. On the interoperability of eBook formats. Technical report, Johannes Gutenberg-Universitt Mainz Germany, April 2013. European and International Booksellers Federation.

[4] Nate Hoffelder. Its official: iBooks now supports Epub3. http://www.the-digital-reader.com/2012/05/23/official-ibooks-now-supports-epub3/#.Uq6-3vRDvoE, May 2012. Accessed December 16th, 2013.

[5] Sue Polanka. What librarians need to know about EPUB3. Technical report, Wright State University, January 2013.

[6] Jan Engelen. E-books: Finally there? In *Publishing in the networked world: Transforming the Nature of Communication*, volume 14, pages 444–448, Helsinki, Finland, June 2010. ELPUB.

[7] John B. Thompson. *Books in the digital age: The transformation of academic and higher education publishing in Britain and the United States*. Polity, 2005.

[8] Ana Catarina Silva and Maria Manuel Borges. Book design program: A transition to a hybrid publishing context. *Information Services and Use*, 31(3):189–197, 2011.

[9] John B. Thompson. *Merchants of culture*. Polity, 2010.

[10] Garth Conboy, Matt Garrish, Markus Gylling, William McCoy, Murata Makoto, and Daniel Weck. EPUB 3. http://www.idpf.org/epub/30/spec/epub30-overview.html, October 2011. Accessed December 16th, 2013.

[11] Kalin Georgiev, Nicholas Matelan, Ludmil Pandeff, and Holly Willis. Sophie 2.0 and HTML5: DIY publishing to mobile devices. In Yasar Tonta, Umut Al, Phyllis Lepon Erdo-an, and Ana Alice Baptista, editors, *Digital Publishing and Mobile Technologies*, volume 15, pages 20–27, Istanbul, Turkey, June 2011. ELPUB.

[12] Craig Weiss. HTML5: Game changer for e-learning? *Learning Circuits - American Society for Training & Development*, 10:5, September 2010.

[13] Gustavo Cardoso, Carla Ganito, and Cátia Ferreira. Digital reading: The transformation of reading practices. In *Social Shaping of Digital Publishing: Exploring the interplay between Culture and Technology*, volume 16, page 126, Guimares, Portugal, June 2012. ELPUB.

[14] Neelie Kroes. Books in the 21st century - opening address to representatives & members of Federation of European Publishers, Frankfurt Book Fair. http://europa.eu/rapid/press-release_SPEECH-11-660_en.htm, October 2011. Accessed December 17th, 2013.

[15] W. Brian Arthur. Competing technologies, increasing returns, and lock-in by historical events. *The economic journal*, 99(394):116–131, 1989.

[16] Safari HTML5 audio and video guide - iOS-specific considerations. https://developer.apple.com/library/safari/documentation/AudioVideo/Conceptual/Using_HTML5_Audio_Video/Device-SpecificConsiderations/Device-SpecificConsiderations.html, 2012. Accessed January 17th, 2014.

[17] Murielle Florins and Jean Vanderdonckt. Graceful degradation of user interfaces as a design method for multiplatform systems. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, IUI '04, pages 140–147, New York, NY, USA, 2004. ACM.

[18] Chee Kai Chua, Kah Fai Leong, and C Chu Sing Lim. *Rapid prototyping: principles and applications*. World Scientific, 2010.

[19] Leonard Richardson and Sam Ruby. *RESTful web services*. O'Reilly, 2008.

[20] Peter Armstrong. The lean publishing manifesto. https://leanpub.com/manifesto, February 2013. Accessed March 17th, 2014.

[21] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 22(140):55, 1932.

[22] Wei Shen and Ute Koch. e-books in the cloud: Desirable features and current challenges for a cloud-based academic e-book infrastructure. In *Digital Publishing and Mobile Technologies*, volume 15, pages 80–86, Istanbul, Turkey, June 2011. ELPUB.

[23] David Shotton. Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing*, 22(2):85–94, 2009.

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-78

EPISCIENCES – an overlay publication platform

Christine BERTHAUD[a], Laurent CAPELLI[a], Jens GUSTEDT[b,c], Claude KIRCHNER[b], Kevin LOISEAU[a], Agnès MAGRON[a], Maud MEDVES[b,d] , Alain MONTEIL[b], Gaëlle RIVERIEUX[b], Laurent ROMARY[b,d,1]

[a] *Centre pour la Communication Scientifique Directe (CCSD), CNRS : UPS2275*
[b] *Inria, France*
[c] *Université de Strasbourg, ICube, Illkirch, France*
[d] *Humboldt-Universität zu Berlin, Institut für Deutsche Sprache und Linguistik (IDSL)*

Abstract. This paper delineates the main characteristics of the Episciences platform, an environment for overlay peer-reviewing that complements existing publication repositories, designed by the Centre pour la Communication Scientifique directe (CCSD[2]) service unit. We describe the main characteristics of the platform and present the first experiment of launching two journals in the computer science domain onto it. Finally, we address a series of open questions related to the actual changes in editorial models (open submission, open peer-review, augmented publication) that such a platform is likely to raise, as well as some hints as to the underlying business model.

Keywords. Overlay journal – Editorial platform – Scholarly communication-Repositories – Open Access

1. Exploring new scholarly publication models

The recent debates on Open Access have mainly focused on opposing models, the so-called green model, where scientists deposit their (possibly published) research papers in open repositories and the gold model where publishers, usually following the payment of an author fee, freely release the publication online. This debate often misses two points. First, that what is at stake is to have a reliable and sustainable communication system for science where scientists themselves have the say and are provided with all means to quickly disseminate their results while receiving the appropriate feedback (usually embodied by peer-reviewing) from their communities. Second, that all data generated around the evaluation, the reviews and the associated discussions (forums, etc.) shall be monitored by the scientific community.

Still, we know that alternative models to the traditional publisher-owned journals are possible, and experiences carried out in the human sciences with the OpenEdition endeavour for instance have shown that research communities may react favourably

[1] Corresponding author
[2] CCSD is a joint service unit between the CNRS, Inria and the University of Lyon

when a real alternative is being offered. Such initiatives provide a systemic concept of publishing (from scholarly blogs to journal publications) comprising both new editorial frameworks and business models.

In this context, we present a new initiative to provide an overlay journal environment, i.e. a journal that is built as an additional peer-reviewing layer on top of a publication repository (see Smith, 1999). This environment offers a technical and editorial platform for existing or new journals operated within a multi-institutional and publicly controlled infrastructure based upon a large-scale publication archive. By sharing the technical settings with a publication repository and focusing on the core missions of a scientific journal we expect to both reduce costs dramatically and open possibilities of experimenting new certification mechanisms.

To quote 0: "The underlying vision is that of a research infrastructure where no fee is applied to its users (whether author or reader) and which offers a set of basic services facilitating an efficient dissemination and review of scholarly papers. Like traditional journals, scientific quality is ensured by the recognition of the editorial committee that carries out the peer-reviewing process." Part of the uniqueness of the Episciences endeavour is the strong commitment of national institutions in ensuring both the quality of the service and its anchoring within a sustainable infrastructure.

In the remaining sections of this paper we will first show how an overlay journal is homothetic to the traditional journal publication principles. We will then describe the role of the publication archive in providing a set of core services for the deployment of a peer-reviewing environment and see what additional functionalities have been designed for the Episciences platform. We will identify which core mechanisms are required to provide a reliable certification service and which may be more peripheral. Finally we will present the first experiment carried out while launching two journals, namely DMTCS (Discrete Mathematics & Theoretical Computer Science) and JDMDH (Journal of Data Mining and Digital Humanities)[3], onto the platform and discuss various topics related to the potentialities offered by overlay journals.

2. Overlay journals seen as a specific case of scholarly journals

2.1. The main functions of a scholarly publishing platform

In his 2009 and 2010 papers, M. Mabe outlines the role of scholarly publishing along the following dimensions:

- *Registration*: the process of submitting a paper, which establishes the author's precedence and ownership of an idea
- *Certification*: where quality control is ensured through peer-review, and consequently scholarly reward is provided to the author
- *Dissemination*: the communication of the findings to its intended audience

[3] Jointly launched by the two French research organizations INRA (Agronomics) and Inria (computer science)

- *Archival record*: preserving a fixed version of an article for future reference and citation.

Whereas this description nicely and conservatively describes the current publisher based setting of scholarly publishing, it may be subject to discussion when considering which new models should be experimented or further deployed.

A first element of discussion is whether all four functions should be situated within the same platform to be fulfilling the researchers' expectations. For instance, managing trustful affiliations is typically part of the competence of a research institution rather than that of a publisher. In the same way, archiving and managing a reference corpus of scholarly papers may be part of the core missions of a community, as exemplified by the initiatives carried out by scholarly associations such as the Association for Computational Linguistics (ACL) or the Association for Computing Machinery (ACM). Finally, it is easy to imagine that certification and dissemination can be completely disconnected from one another in a mediated world where social networks are more and more used to convey daily scientific news.

More importantly, we can see how this frozen scenario may be counter-productive to the very essence of scholarly publishing, namely to ensure the appropriate convey of knowledge between scholars, but also to the wider public. First, it subordinates the dissemination of scholarly papers to the peer-reviewing process, whereas we know how much the two can live independently from one another (see Gentil-Beccot et alii, 2009), but also how much danger there is when a selective review process prevents the dissemination of useful results[4]. This situation leads scholars to submit their papers iteratively to multiple settings and reviewers to get drowned under a deluge of useless refereeing work.

The whole idea of the Episciences initiative is to decompose the process to ensure maximal efficiency at the service of scholarly communication. In particular, we now see how publication repositories can play a core role for an open publication process.

2.2. Publication repositories as an infrastructure for scholarly publishing

Open archives are now widely available and can be used by any researcher to store, index and make any of their research documents freely available, whether or not these have been published in peer-reviewed channels (journals or conferences). Even more, these documents can range from research papers to experiments, data, computer programs or videos. Such archives as the e-print archive arXiv or Hyper Articles en Ligne (HAL) are widely accessible and provide a free and sustainable service. In the case of the HAL platform for instance, papers are associated with precise affiliation information for each author, and are supported by long-term archiving facilities. Additional services like the creation of personal or institutional web pages are also offered.

Seen from the point of view of scholarly publishing we can see how most existing publication archives provide an adequate environment for supporting several of the core functions related to traditional journals (see Romary & Armbruster, 2010):

[4] See for instance:
http://www.nature.com/news/half-of-us-clinical-trials-go-unpublished-1.14286

- They provide a reliable registration environment whereby both attribution (authors and their affiliations) as well as time-stamping[5] are attached to the registered documents;
- Dissemination is naturally ensured not only through the built-in open access nature of the archive but also because large scale publication repositories such as HAL[6] or arXiv are highly visible within search engines and their content are followed (alert mechanisms) by the research communities;
- Finally, archival record is also a natural component of publication repositories, with an additional advantage here, namely that papers from a given author or institution can be gathered within a coherent setting rather than being spread across various publishers' portals, whose long-term existence or accessibility is far from being ensured.

Beyond these standard functionalities, institutional publication archives often come with various additional features that make them even more powerful than usual publishers' environments. First, being hosted by sustainable institutions, they offer some guaranties that the technical environment and thus the corresponding content will be made available for a long period of time. This is even more the case for central repositories such as HAL, where a consortium of institutions, or even a national policy[7], is backing up the service. Research libraries also often curate the content, thus ensuring coherent metadata descriptions associated with authority lists of institutions or funded projects.

From a technical point of view, it is also important to apprehend how much versioning is an essential feature from the point of view of the academic process since it allows researchers to trace the processes when writing a document and, possibly, integrating the comments received from their colleagues, anonymously or not.

As a whole, we see that only a core set of mechanisms have to be implemented to fulfil the role of a scholarly journal environment, namely a) the management of the review process and b) the provision of more or less fine-grained copy-editing support. The following sections will describe how the Episciences project fulfils these.

3. Main functions of the Episciences publication platform

The Episciences platform is conceived in the spirit of traditional peer-reviewed journals, with additional facilities resulting from it leaning against a publication repository. The editorial team and the reviewing and publication workflow are standard, with the difference that the paper is managed by the author and not by the editors in charge, the labelling of the paper as accepted being of course fully handled under the control of the editorial board. This impacts on copy-editing because the

[5] This is for instance essential if these documents are to be used to assess the anteriority of a discovery as is the case for prior art search in patent organizations.

[6] HAL was ranked 5th in the Webometrics portal ranking as of Jan. 2014 (see http://repositories.webometrics.info/en/top_portals)

[7] see http://www.enseignementsup-recherche.gouv.fr/cid71277/partenariat-en-faveur-des-archives-ouvertes-plateforme-mutualisee-hal.html (in French)

author is responsible for the layout (unless he gives over some rights) and versioning with all versions of the paper (at least the submitted and accepted ones) being available on the repository.

3.1. Editorial services

In order to support the editors-in-chief and editorial boards in their day-to-day business, a support in terms of editorial management is provided. This comprises:

- Management of the peer-review process, comprising the channelling of community based feedback and the plagiarism detection;
- Handling the management of the journal volumes and issues;
- Contribution to some basic quality checking tasks (bibliography, metadata, cross-references, automatic detection of the state of the art);
- Communication and community management: advertising journals and papers through various channels and social networks (twitter, blog, academic social website), moderation of online discussions (made possible by the commenting functions and display of tweets related to an article)[8];
- General visibility: interaction with major indexing services and databases (Digital Bibliography and Library Project, Thomson Reuters, Scopus...), as well as adequate mirroring on relevant thematic repositories (ArXiv, PubMed Central, Research Papers in Economics, etc.).

3.2. Technical services

Through the hosting on the French national repository infrastructure HAL, all journals benefit from a high quality technical environment comprising 24/7 services, long term archiving of all versions and proper authentication and authorisation infrastructure. Other platforms such as arXiv offer similar facilities.

The platform offers web design tools so that each journal can customise its own website while their generic graphical identity retains features of the Episciences design.

Long term archiving of the reviewing information is also assured: the ratings as well as the exchanges between authors and reviewers are securely stored on the platform and are accessible to the editorial team at any time. According to the journal policy, reviews may be published as well as the reviewer's names (see discussion in section 6).

3.3. Intellectual property management

The Episciences model impacts at several levels on intellectual property issues. First, the Episciences platform leaves all rights to the journals concerning the ownership of the title. The basic idea here is that the platform will not be the publisher. In cases

[8] Such services have already been experimented for HAL: cf. http://fronthal.imag.fr/noModule

where there may be difficulties to manage such an ownership[9], the consortium of institutions in charge of Episciences will upon request temporarily host the ownership of the title.

From an author's point of view, a simple non-exclusive licence will be requested. As a matter of fact, given that the papers are available through a publication archive, they actually bear the associated open licence (in the case of HAL-Inria for instance a strong recommendation is made to have papers issued under the Creative Commons Attribution (CC-BY) licence).

3.4. Copy editing

Copy-editing is left to the editorial board of each journal, which will also decide of the submission format and style. Typically, submission in TeX or LaTeX may ensure that the formatting instructions will be slightly better met in most cases without any need for further copy-editing related to the actual formatting of papers. Still, we are aware that copy-editing is a question. The quality that is provided by author sources is very much varying, and there is not only a quality control job involved, but many authors definitively need help and guidance, and for some much of the work may have to be provided. Part of the developments we will have to consider (see section 6 below on the budget break-out) is to be able to support journals with such needs.

4. Managing the Episciences journal portfolio

The journals hosted on the Episciences platform are organised as thematic portfolios. The objective is to ensure quality and coherence on a discipline based rationale. In order to achieve this, each scientific domain that will have journals on Episciences will form a pool coordinated by a so-called meta-committee, a group of internationally recognised experts whose duty will be to select new incoming journals, check out their overall operation and quality, but also be the contact to attract new journals within their respective communities. Part of the duties of a meta-committee will also be to control the thematic coherence of the various journals, so that clear guidance can be given to authors as to where their papers should be optimally submitted.

Two such meta-committees are currently being set-up in Mathematics and Computer Science, which correspond to the communities that have started to show interest for Episciences.

5. Two initial experiments

We started the platform with two journals from different sub-domains in computer science. One of the journals, JDMDH[10], is a new creation, corresponding to an emerging domain with a scientific committee that has collectively decided to go for an open journal and to join efforts with Inria on the new platform. The other one,

[9] when not properly hosted by an academic institution or a scientific society

[10] Note that no official launch has been made yet at the time of submission of the paper and that the site is still in test phase

DMTCS, is an established open journal for which we designed a transition scheme to Episciences.

JDMDH covers all aspects of data mining methods for the humanities. The first launch issue is in preparation with all submitted papers already deposited in the Episciences framework (namely deposited on HAL and arXiv prior to submission to the journal). There is already a strong support within the editorial committee for the post-publishing peer-review process (see also discussion in section 6).

DMTCS is a well-established scientific journal. Placed at the cross-section between computer sciences and mathematics, it covers both, but emphasizes on work that profits for or from both. In the late nineties, DMTCS was one of the first open access journals that came to life, in a then rapidly growing context of the still new and chilling Internet. At first managed by a commercial editing house, the DMTCS title was quickly transferred to the scientific editors. DMTCS is structured in volumes and issues, though they are only formal remainders of ancient publishing traditions. De facto the journal is published continuously.

The online system[11] evolved from a collection of simple web pages and an editorial process managed through mail, over a home-brew server software, to the Open Journal System (OJS). Without dedicated specialised staff, the journal is clearly vulnerable and lacks reactivity and quality of service.

One of the main challenges when migrating DMTCS from OJS to Episciences was to manage legacy papers. First, it was necessary to keep two platforms alive in parallel for a while, namely until the peer-review process of the articles submitted in OJS is over (while new articles are submitted in Episciences). Second, it proved challenging to import all legacy papers into HAL with the expected level of metadata precision.

6. Issues raised by an overlay journal platform

The Episciences model is not a simple replacement of the traditional scholarly publishing environment. Its integration within the services of publication repositories in particular makes it bear specific characteristics, which we would like to analyse in this section, being aware that many consequences of the model are likely to appear when processing a larger portfolio of journals.

6.1. A low-cost platform

The economic study[12] of the EU-funded Publishing and the Ecology of European Research (PEER) project evaluated (p.48) the cost in a repository to range between 2 and 50 € per reference and between 2,5 and 53,2 € per full text[13]. It also showed that a baseline for managing the peer-review process alone lies around 200 € per article for most commercial journals. Such costs usually correspond to the manpower related to editorial secretariat and is planned to be one of the possible duties of future librarians within research institutions, as anticipated in (Guédon, 2001).

---

[11] With the support of Inria and Loria laboratory

[12] http://www.peerproject.eu/fileadmin/media/reports/PEER_Economics_Report.pdf

[13] Note that for HAL the average cost per paper has been evaluated to 14.73€

For such cost we need to be open as to the possible business models that may allow our initiative to break even in the long run. We basically see three main possible components for a balanced funding scheme:

- Following the model adopted for HAL, we have started to pool some core resources within a consortium of partners. The stability of such national institutions will ensure sustainability for the platform;
- We also need to unite forces with initiatives such as OpenEdition which sell additional services (cataloguing, smart formats (ePub)) to university libraries, whose benefits directly finance the journals themselves;
- We should not reject author processing charges when there is a request for additional copy-editing services, such as suggested by the Copernicus publisher for its open access journals.

6.2. Leaving away the post peer-review publishing paradigm

One important consequence of the overlay journal model is that papers are made public right at the time of their deposit on the publication repository, which means that the peer-review process actually takes place *after* the actual publication[14]. There are several consequences that derive from this principle:

- Having the paper online before peer-review obviously prevents author anonymity. Whereas this is not necessarily part of the cultural background of some scholarly communities, there are strong arguments to see this as a benefit for the scholarly process (see 0 and next section on open peer-review)
- Whatever the time and the duration of the review process, the paper benefits from a high visibility right from the onset. This may allow colleagues to comment at an early stage and even for the document to be cited if already relevant as background for another research. This aspect has become normal practice for many communities like in physics or astronomy with arXiv as a pre-print server;
- The paper remains available whatever the success of the peer-review, which guaranties the continuous availability of the corresponding results independently of the outcomes and possibly incidents of the certification process. This is important to circumvent the dramatic loss on non-published information that science currently faces (see Jones et alii, 2013);
- The experience gained from other open reviewing environment (see Pöschl, 2004) has shown that open manuscripts reduce the number of poorly written submissions, thus leading to a more efficient peer-review process;
- The paper may evolve further if new elements validating or invalidating the paper are discovered. An overlay publication system thus facilitates the management of versions (or errata in the mathematical domain).

[14] See the position blog entry by J. Velterop: http://theparachute.blogspot.co.uk/2013/11/essence-of-academic-publishing.html

The issue at stake is how much such a model will be accepted by a variety of scholarly communities or if we may have to allow "invisible" papers in publication archives to cover more publication scenarios.

6.3. Towards new peer-review models

Once the psychological barrier of post peer-review publication has been overcome, a platform such as Episciences is the ideal place to convince scientific communities that peer-review can take other forms than those known in traditional journal settings. There are indeed two complementary directions that we would now like to pursue:

- Open peer-review, whereby reviews become openly accessible with, possibly, the identification of the reviewers. By doing so, we encourage reviews to become publication objects of their own and be part of a publication bundle together with the paper itself;
- Community feedback: by linking papers to scholarly blog entries or pushing submissions to external reviewing platforms (e.g. PeerEvaluation) to offer further commenting environments.

6.4. Towards new documentary services

Linking a journal platform to a national publication repository opens up a wide range of potential services that would not be affordable for such a dedicated peer-review platform. In the context of our current developments on the HAL platform, such services include automatic PDF to metadata recogniser[15] (title, author, affiliation, keywords and abstract information) to simplify the submission process for an author, or the automatic detection of bibliographical references for linking the paper to other relevant publications.

An important disruptive step will be to systematically create a reference XML version of all papers[16], which in turn can be used to produce different publication formats (HTML, ePub, PDF with a specific layout, etc.).

6.5. Episciences for putting together data journals

Finally, we can see that the Episciences workflow is designed independently of the nature of the initial document. It may indeed not be a textual object but a compound of notes, programs (possibly active) and data that could benefit from the same kind of certification process. The way towards data journals, which only a handful of communities have tackled so far, can be part of the realm of overlay certification processes, when anchored on data or program repositories[17].

---

[15] Based on https://github.com/kermitt2/grobid

[16] Compliant with the TEI guidelines (cf. http://www.tei-c.org)

[17] In the computer science domain, the IPOL journal (http://www.ipol.im) for instance deals with the assessment of executable computer programs.

7. Overview

We think that putting together such a platform for overlay journals, and making it widely available to research communities, will offer a whole wealth of features for scholars by providing fast and efficient dissemination of scholarly results. Beyond the maths and informatics communities that are now involved in this endeavour, we expect a wider range of domains to benefit from this service.

The experiment carried out with our two initial journals has allowed us to secure most of the features on the platform and validate that a quick, and cheap, deployment of an overlay journal is possible. We can now identify our roadmap for the future in two complementary directions: bring in more journals in the informatics and applied mathematics domain, where we have already felt a strong demand, and attract a wide range of interested institutions to join efforts in securing the long-term sustainability of the endeavour.

References

Gentil-Beccot A., S. Mele and T. C. Brooks (2009) "Citing and Reading Behaviours in High-Energy Physics. How a Community Stopped Worrying about Journals and Learned to Love Repositories." CoRR abs/0906.5418

Guédon J.-C. (2001), "In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing", *Association of Research Libraries*, ISBN 0-918006-81-3

Jones C. W., L. Handler, K. E. Crowell, L. G. Keil, M. A. Weaver, T. F. Platts-Mills (2013) "Non-publication of large randomized clinical trials: cross sectional analysis". *BMJ* 2013;347:f6104

Mabe M. A. (2009). "Scholarly Publishing". *European Review*, 17, pp 3-22. doi:10.1017/S1062798709000532

Mabe M. A. (2010) Scholarly Communication: A Long View, New Review of Academic Librarianship, 16:S1, 132-144, doi:10.1080/13614533.2010.512242

Pöschl, U. (2004) "Interactive journal concept for improved scientific publishing and quality assurance", *Learned Publishing*, 17(2), pp. 105-113.

Romary L., Armsbruster, C (2010). Beyond institutional repositories, in International Journal of Digital Library Systems **1**, 1, pp.44-61 http://hal.archives-ouvertes.fr/hal-00399881

Romary L.,, P. Guitton & C. Kirchner (2013) "A French perspective on Open Access and the Episciences Initiative", in T. Hey (Ed.) *Tony Hey on eScience* (tonyhey.net), June 2013, http://tonyhey.net/2013/06/03/a-global-view-of-open-access-part-1/

Smith, John W T "The deconstructed journal : a new model for academic publishing". *Learned Publishing*, 1999, vol. 12, n. 2.

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-88

Publish your Data and Model Code: Research Output is More Than "Just" a Research Paper

Martin Rasmusen[a,1]
[a] *Copernicus Publications (Copernicus GmbH), Bahnhofsallee 1e, 38081 Göttingen*

Abstract. Research output is not only research articles. To provide outlets for publishing other outputs than articles, Copernicus Publications, an innovative open access publisher based in Göttingen, Germany, currently publishes the journals Earth System Science Data and Geoscientific Model Development. The first journal is dedicated to the peer-reviewed publication of articles on original research data sets in the Earth System Sciences. The second journal is dedicated to publish the description, development and evaluation of numerical models of the Earth System and its components. Both journals apply an innovative interactive open access peer-review with public referee reports, public comments from the community prior to editor's decision, and public author's responses. The motivation is to make the whole research output from data, to models, to the scientific findings and novel interpretations freely accessible, to foster scientific discussion, to increase transparency in scientific quality assurance, and to give credit to all involved contributors.

Keywords. Data publication, model code publication, open access, public peer-review

Introduction

In the Earth System Sciences, as well as in many other disciplines, the final interpretation of new scientific findings is the result of a long process of data collection, data interpretation, model calibrations, model runs, interpretation of these results, and conclusions regarding novel aspects. And it is teamwork of many people contributing to these results, not only scientists, but also engineers, data specialists, and many other groups of learned contributors.

When open access publishing started, the idea has been raised quickly to expand this principle to many other scientific sources than "just" to the final revised, peer-reviewed article. For decades, readers of scientific articles had to settle for graphs resulting from data interpretation or model runs without knowing much more about the data provenance and structure, without getting access to this data, without a broader understanding of the used models, and without a deeper insight into the model code. Neither reviewers nor readers could have ever reproduced the work of the author of a scientific manuscript.

[1] Corresponding Author.

Fortunately, the open access principle became a politically widely accepted strategy and liberal copyright and license agreements like Creative Commons' CC-BY license fundamentally reinvented the idea of access to scientific work and options for the reuse of, in most cases, research outputs financed through taxpayers money.

In 2008, two groups of scientists raised, independently, the ideas of a data publication journal on the one hand and a model development journal on the other hand. Copernicus Publications started these two titles applying the interactive open access publishing approach with public peer-review and interactive public discussion, established in 2001. Public referee reports, public comments from the community prior to editor's decision, and public author's responses are published alongside the discussion paper, an access-reviewed version of the author's manuscript. The motivation was to make the whole research output from data, to models, to the scientific findings and novel interpretations freely accessible, to foster scientific discussion, to increase transparency in scientific quality assurance, and to give credit to all involved contributors.

The following sections describe the journals Earth System Science Data (ESSD) and Geoscientific Model Development (GMD) in more detail, and explain the concept of interactive open access publishing.

1. Earth System Science Data (ESSD)

1.1. Aims, Scope, and Motivation

Earth System Science Data (ESSD) is an international, interdisciplinary journal for the publication of articles on original research data(sets), furthering the reuse of high (reference) quality data of benefit to Earth System Sciences. The editors encourage submissions on original data or data collections which are of sufficient quality and potential impact to contribute to these aims. The journal maintains sections for regular length articles, brief communications (e.g., on additions to datasets) and commentary, as well as review articles and "Special Issues".

Articles in the data section may pertain to the planning, instrumentation and execution of experiments or collection of data. Any interpretation of data is outside the scope of regular articles. Articles on methods describe nontrivial statistical and other methods employed, e.g. to filter, normalize or convert raw data to primary, published data, as well as nontrivial instrumentation or operational methods. Any comparison to other methods is out of scope of regular articles. Review articles may compare methods or relative merits of datasets, the fitness of individual methods or datasets for specific purposes or how combinations might be used as more complex methods or reference data collections.

This journal aims to establish a new subject of publication: to publish data according to the conventional fashion of publishing articles, applying the established principles of quality assessment through peer-review to datasets. The goals are to make datasets a reliable resource to build upon and to reward the authors by establishing priority and recognition through the impact of their articles.

The peer-review secures that the data sets are at least plausible and contain no detectable problems, that they are of sufficiently high quality and their limitations are clearly stated, that they are open accessible (toll free), well annotated by standard

metadata (e.g., ISO 19115) and available from a certified data center/repository, and that they are customary with regard to their format(s) and/or access protocol, however not proprietary ones (e.g., Open Geospatial Consortium standards), expected to be useable for the foreseeable future.

The articles in this journal should enable the reviewer and the reader to review and use the data, respectively, with the least amount of effort. To this end, all necessary information should be presented through the article text and references in a concise manner and each article should publish as much data as possible. The aim is to minimize the overall workload of reviewers, e.g., by reviewing one instead of many articles, and to maximize the impact of each article. [1]

The initiators of ESSD were David Carlson, director of the programme office of the International Polar Year (IPY) in 2007-2008, and Hans Pfeiffenberger, head of IT infrastructure at the Alfred Wegener Institute for Polar and Marine Research (AWI) in Bremerhaven, Germany.

1.2. Manuscript Submission

The precondition to submit a manuscript for publication in Earth System Science Data is that the data sets referenced in the manuscript are submitted to a long-term repository. Such a repository has to fulfill the following basic criteria under all circumstances [1]:

- Persistent Identifier: The data sets have to have a digital object identifier.
- Open Access: The data sets have to be available free of charge and without any barriers except a usual registration to get a login free-of-charge.
- Liberal Copyright: Anyone must be free to copy, distribute, transmit and adapt the data sets as long as he/she is giving credit to the original authors (equivalent to the Creative Commons Attribution License).
- Long-term Availability: The repository has to meet the highest standards to guarantee a long-term availability of the data sets and a permanent access.

1.3. Review Criteria

For future reuse and reinterpretation it is mandatory for the user to be assured about research data quality. It is the aim of ESSD to provide the quality assessment for datasets which already reside in permanent repositories. Is the article itself appropriate to support the publication of a dataset? Is the dataset significant – unique, useful and complete? Is the dataset publication, as submitted, of high quality? Reviewers are asked to decide how well the respective datasets presented by an article and the article itself meet the criteria significance, data quality, and presentation quality. [1]

1.4. ESSD Facts & Figures

By the end of March 2014, ESSD had 127 manuscripts submitted from which 99 have been published in the discussion forum of ESSD and 85 in ESSD as final revised journal articles. The final articles have an average length of 12 pages (median) and the

review takes on average (median) 29 days from submission to publication of the discussion paper, and 33 days from revised submission after public discussion to publication of the final revised and fully peer-reviewed paper. In the discussion forum, 433 comments have been posted, 207 of which are referee comments, 194 are author comments, eight comments are published by the journal editors, and 24 comments by members of the scientific community prior to the final acceptance of the manuscripts [2]. ESSD is indexed by Scopus.

2. Geoscientific Model Development

2.1. Aims, Scope, and Motivation

Geoscientific Model Development (GMD) is an international scientific journal dedicated to the publication and public discussion of the description, development and evaluation of numerical models of the Earth System and its components. Manuscript types considered for peer-reviewed publication are [3]:

- Geoscientific model descriptions, from box models to GCMs;
- Development and Technical papers, describing development such as new parameterisations or technical aspects of running models such as the reproducibility of results;
- Papers describing new standard experiments for assessing model performance, or novel ways of comparing model results with observational data;
- Model intercomparison descriptions, including experimental details and project protocols.

GMD is owned by the European Geosciences Union (EGU, http://www.egu.eu) and started in 2008. The main drivers and Executive Editors have been in alphabetical order James Annan and Julia Hargreaves, both from the JAMSTEC Research Institute for Global Change in Yokohama, Japan; Dan Lunt, University of Bristol, UK; Robert Marsh, University of Southampton, UK; Andy Ridgwell, University of Bristol, UK; Ian Rutt, Swansea University, UK; and Rolf Sander, Max Planck Institute for Chemistry in Mainz, Germany.

Since the scale and complexity of computer modelling tools increased, it was no longer practicable to describe models in papers. Furthermore, the normal journal peer-review focusses on the scientific results and the model description, and the technicalities are less well presented. However, the GMD initiators saw the needs to fully describe models and model developments in peer-reviewed publications. They aimed to guarantee reproducibility, traceability, transparency, and access [4]. Two nice quotes given on the GMD website are:

"I believe that the time is ripe for significantly better documentation of programs, and that we can best achieve this by considering programs to be works of literature."
(Donald E. Knuth, Literate Programming, 1984)

"Essentially, all models are wrong, but some are useful."
(George E.P. Box, Robustness in the strategy of scientific model building, 1979)

2.2. Review Criteria

Reviewers are asked to rate on the scientific significance, the scientific quality, the scientific reproducibility, as well as on the presentation quality. The reviewers decide whether substantial new concepts, ideas, or methods are described, whether the approaches and applied methods are valid, whether the models have the potential to perform calculations leading to significant scientific results, and to what extend the modelling science is reproducible. Fellow scientists have to be able to reproduce the sciences. Therefore, a focus is the completeness and the preciseness of the descriptions [3].

2.3. GMD Facts & Figures

By the end of March 2014, GMD had 596 manuscripts submitted from which 495 have been published in the discussion forum of GMD and 351 in GMD as final revised journal articles. The final articles have an average length of 16 pages (median) and the review takes on average (median) 33 days from submission to publication of the discussion paper, and 46 days from revised submission after public discussion to publication of the final revised and fully peer-reviewed paper. In the discussion forum, 2,129 comments have been posted, 1,018 of which are referee comments, 963 are author comments, 67 comments are published by the journal editors, and 81 comments by members of the scientific community prior to the final acceptance of the manuscripts [5]. GMD is indexed by Scopus and the Web of Science, and received the Thomson Reuters Impact Factor of 5.030 in 2012 [3].

3. Interactive Open Access Publishing

The Interactive Open Access Publishing aims to bring more transparency into scientific quality assurance by publishing the reviewer reports and the author's response freely accessible. In the first stage, the submitted manuscript is access-reviewed by one of the topical editors of the journal. It is a rapid review and involves only technical corrections. Then, the manuscript is typeset and published as so-called Discussion Paper. It is fully citable, receives a classical citation and pagination, as well as a DOI. The publishing platform is called the discussion forum.

The Discussion Paper is then subject to Interactive Public Discussion, during which the referees' comments (anonymous or attributed), additional short comments by other members of the scientific community (attributed) and the authors' replies are also published in the discussion forum alongside the Discussion Paper. Different from other initiatives experimenting with Public Peer-Review, the comments in this concept are also fully citable, paginated, typeset automatically by an online application, and remain online permanently.

In the second stage, the peer-review process is completed and, if accepted, the final revised papers are published in the journal itself. The latter is then the fully peer-

reviewed publication platform which is subject to indexing in the Web of Science, Scopus, and other databases.

The concept of Interactive Open Access Publishing started in 2001 and traces back to Ulrich Pöschl and Nobel laureate Paul Crutzen, both at the Max Planck Institute for Chemistry in Mainz, Germany. It was first applied to the journal Atmospheric Chemistry and Physics (ACP) [6], a very successful title owned by the European Geosciences Union (EGU) and published by Copernicus Publications.

Ulrich Pöschl described his concept in many publications [7], [8], [9].

References

[1] ESSD journal website, available at: http://www.earth-system-science-data.net. Access on 27 March 2014.
[2] ESSD paper statistics, taken from Copernicus Publications' manuscript review system Copernicus Office Editor. Access on 27 March 2014.
[3] GMD journal website, available at: http://www.geoscientific-model-development.net/. Access on 27 March 2014.
[4] GMD Executive Editors: Editorial: The publication of geoscientific model developments v1.0, Geosci. Model Dev., 6, 1233-1242, doi:10.5194/gmd-6-1233-2013, 2013.
[5] GMD paper statistics, taken from Copernicus Publications' manuscript review system Copernicus Office Editor. Access on 27 March 2014.
[6] ACP journal website, available at: http://www.atmospheric-chemistry-and-physics.net/. Access on 27 March 2014.
[7] Pöschl, U., Interactive journal concept for improved scientific publishing and quality assurance, Learned Publishing, 17, 105-113, 2004
[8] 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, 19, 293-314, 2010
[9] Pöschl, U., Multi-stage open peer review: scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation, Frontiers of Computational Neuroscience, 6, 33, doi:10.3389/fncom.2012.00033, 2012

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-94

Shared service components infrastructure for enriching electronic publications with online reading and full-text search[1]

Nikos HOUSSOS, Panagiotis STATHOPOULOS, Ioanna-Ourania STATHOPOULOU, Andreas KALAITZIS, Alexandros SOUMPLIS

National Documentation Centre / National Hellenic Research Foundation

Abstract. A major requirement for electronic publishing systems is the availability of rich and intuitive mechanisms that enhance the user experience of viewing and searching online electronic documents such as books, monographs and journal papers. This work concerns a set of infrastructural components and their utilization for the creation of related coherent services and features for end users. We present a set of sophisticated platforms, tools and mechanisms that have been employed in real-life cases for implementing document viewing and full-text search features, shared among application instances of various types. Challenges encountered and the provided solutions are discussed.

Keywords. electronic publishing, user experience, document viewers, online readers, image servers, jpeg2000, full-text search, optical character recognition.

1. Introduction

Providing users with a rich, intuitive and meaningful interface for accessing online electronic resources, e.g. books, monographs, journal papers is a significant factor for improving the user experience in electronic publishing systems and consequently for their increased adoption and usage. Traditionally, electronic journals, and scientific publications, have been published and viewed online, using the widespread PDF format, due to the rigid but portable formatting it provides, and HTML for versatility.

Recently, initiatives such as the "Google Books" and "Google Art" project, the Internet Archive "Open Library" and advanced repository systems, have paved the way for novel online reading capabilities and experience, with features such as "page by page" viewing of electronic resources and tile-based image viewing systems, exploiting advanced codecs such as JP2000 and corresponding online viewers.

[1] The work presented in this article has been partly supported by the project "Platform for provision of services for deposit, management and dissemination of Open Public Data and Digital Content" (Ref No 327378) which is co-funded by Greece and the European Union-European Regional Development Fund through the Operational Programme "Digital Convergence" (NSFR)

These advances are becoming gradually available in electronic publication systems [1] and, if incorporated in Open Access systems (e.g. journals, repositories) can contribute to even wider adoption by users and publishers. Such incorporation will enable, apart from intuitive reading capabilities, the efficient viewing of large data sets visualisations, maps and/or images of cultural artifacts.

In this contribution, we present the technologies, mechanisms and open source components employed to achieve this functionality. The corresponding infrastructure and tools includes, among others, an interactive online reader with dynamic zooming, thumbnail view, full-text search and hit highlighting capabilities, a JPEG2000 image server, a highly scalable back-end common infrastructure for storage, batch image processing, OCR, indexing / search and access to digital files by multiple applications.

The rest of the article has the following structure: Section 2 describes the environment of real-life services and applications where the solutions proposed in this article have been applied. Section 3 elaborates on the development of advanced document viewing user experience based on a range of back-end and front-end services. Section 4 concerns the support of full-text search in various environments. The paper ends with a summary section.

2. Services and applications context

This section briefly presents the environment of real world services and applications where the solutions proposed in this article have applied. Essentially, we have implemented them in the context of scholarly communications infrastructure and in particular multiple e-publishing systems and repositories that are operated by the National Documentation Centre of Greece (EKT).

2.1. The EKT electronic publishing platform

The National Documentation Centre of Greece has launched its own electronic publishing platform (EKT ePublishing at http://epublishing.ekt.gr), aiming to provide a single open access entry point to the content of scholarly eJournals, eBooks and eProceedings which have been produced through e-publishing services offered by EKT. Currently, the EKT ePublishing platform hosts fourteen academic eJournals with more than 3.000 open access articles, 14 eBooks and 43 conference proceedings.

EKT acts as the electronic publisher of Greek academic eJournals in various thematic areas, using the Open Journal Systems (OJS at http://pkp.sfu.ca/ojs/) platform which is an open source journal management and publishing system. E-Journals are peer reviewed, and each one is served by a separate, OJS installation, tailored made to the requirements and user needs of each Journal.

Along with the automated process of harvesting and publishing content from external web platforms, EKT ePublishing platform enables also specified authorized users to manually import data for eJournals, eBooks and eProceedings. Figure 1 illustrates the architecture of the EKT ePublishing platform.

The content of eJournals is periodically harvested and published in the central EKT ePublishing platform, through REST-style interfaces exposed by OJS installations. The service is triggered once per day and propagates metadata updates that have been recorded in the OJS journals to the central EKT ePublishing platform.

EKT ePublishing is a web platform built with Drupal CMS in conjunction with PostgreSQL database, nginx web server and the local file system. The central platform harvests metadata from the OJS instances via REST interfaces and, in addition to that,

receives data directly from authorized users to manually import content directly into the central system. A central Solr instance is used as a search engine indexing, while Apache Tika extracts the textual content from PDF files in the individual journals. As for the authentication and authorization processes, they are addressed by an LDAP Server populated with data about users organized in groups.

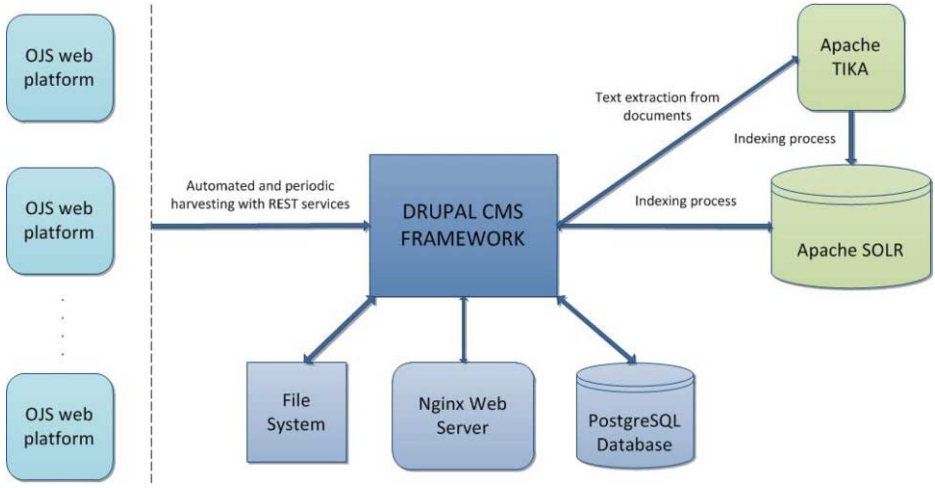


Figure 1: EKT ePublishing platform architecture.

2.2. EKT repositories

EKT operates more than 10 open access repositories with more than 80.000 metadata records, including the Hellenic National Archive of PhD theses, the institutional repository of the National Hellenic Research Foundation, a funder repository storing output of projects supported by the Greek Ministry of Education and a range of cultural repositories. The operating repositories use the shared infrastructure presented in this article for online document viewing and full text search.

3. Online reader: infrastructures, services and tools

3.1. Introduction

Currently, journal articles, books, monographs and other scientific electronic resources can be published and viewed online in various formats, such as PDF [2], HTML, EPUB, ODT. Traditionally, the most common and used format in the majority of scientific publications is PDF, because of the preserved printed format and the property protection capability, although there is an effort towards adopting the new version of EPUB format (EPUB version 3 at http://idpf.org/epub/30). Furthermore, electronic readers can be implemented in order to provide better user experience with page-by-page document reading and other significant advantages such as quicker browser loading times, abstaining from downloading large pdf files. National Documentation Centre (EKT) has integrated in all electronic publishing systems an eReader, offering greater reading capabilities. The EKT eReader is developed with the open source software Internet Archive BookReader (https://openlibrary.org/dev/docs/bookreader)

and can provide a more attractive reading experience with high quality content presentation in various views (single page view, two page view, thumbnail view), allowing zoom, full text-search with hit highlighting and integration with a variety of image servers.

Certain important advantages of page-by-page online readers compared with offline reading of PDF files and embedded PDF readers (e.g. in browsers) are the following:

- Huge files (e.g. files well over 50MB are commonplace in the systems described in Section 2) can be opened and navigated very quickly (low response time) and without overloading the end user's computer (low computer memory consumption), since the document is streamed on demand (one page at a time), not downloaded in its entirety and therefore opens instantly after a user click.
- Document opens for reading with one click (no need to download). Notably, embedded PDF readers of modern browsers have this feature as well, however this is commonly not available for files of big size (e.g. over 50MB), which instead get automatically downloaded by browsers or cause crashes.
- Each page has its own URL, which enables bookmarking and sharing at the page level, a feature particularly important in large texts such as books (although useful also for shorter documents like journal articles).
- Easy and fast visual navigation within the document using a grid-shaped thumbnail view screen.

On the other hand, embedded PDF readers have the advantage of copy-and-paste functionality (even for image PDFs with OCR'ed text embedded in the document), they allow easy bulk printing at the level of the entire document and of course they do not require the conversion from PDF to JPEG2000 images which is a process that needs to be executed by sophisticated infrastructure (see Section 3.3.1 and 3.3.2) and takes time and resources.

3.2. System architecture

EKT eReader presents the content of a PDF document page-by-page as a set of high resolution images, which are served via Djatoka image server. The process of converting a PDF file into JPEG2000 images is a fully automated process with specific workflow and requirements and can be triggered at will, from remote electronic publishing platforms as a service. The fact that the content presented in eReader is in an image format, makes full-text search process more difficult. However, EKT eReader, in its current release, supports full-text search with hit highlighting of search results. In order for the eReader to support full-text search functionality, an OCR (Optical Character Recognition) software for text extraction is required as well as a search platform for content indexing. Currently, ABBYY FineReader (http://finereader.abbyy.com/) is used for the OCR process and Apache Solr (http://lucene.apache.org/solr/) as a search server, enabling among others full-text indexing and search, hit highlighting, faceted search. Figure 2 illustrates the architecture of the EKT eReader infrastructure.

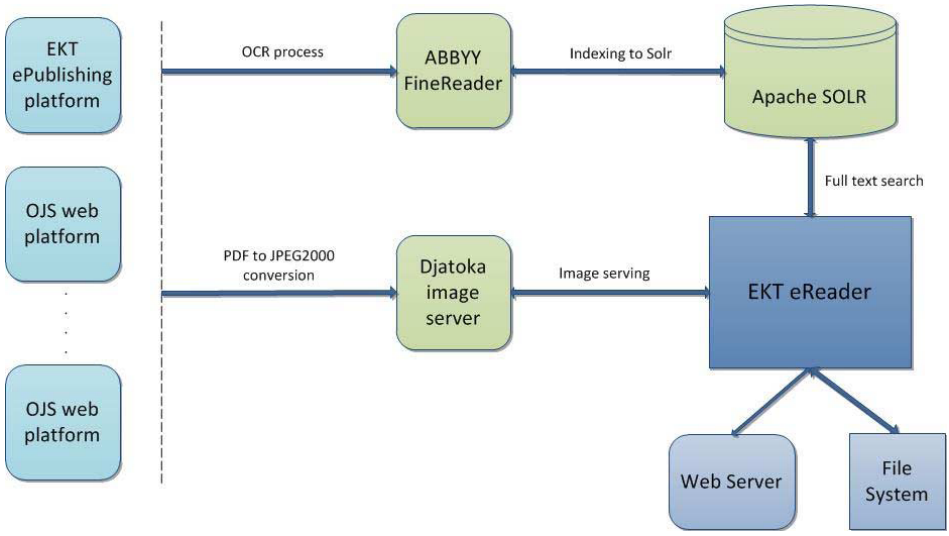


Figure 2: EKT eReader architecture.

3.3. Infrastructure and services for automatic preparation of material for the online reader

3.3.1. Back-end image server infrastructure architecture

While advances in PDF viewers have been made, with some of them having an increased level of interactivity and the capability to support on line reading (e.g. Multivio, https://www.multivio.org/) [3], JP2000 based viewer and JP2000 image servers still offered unrivaled browsing responsiveness, very fast browsing times and dynamic thumbnails generation among others. Since a significant amount of content is produced as "born digital", a conversion from text-PDF to text search enabled series is required in order to enable fast and responsive online reading. JP2000 online reading with search highlighting can be supported by a mature open software stack, comprising the IA Internet Archive Book reader and the DJatoka image server. However, a critical element is missing, namely the facility to convert born digital PDFs to a series of corresponding JP2000s. Furthermore, in order for such conversion to be suitable for a large environment, some basic requirements must be met, from the software implementing them, such as:

- Capability to convert thousands of documents, with no human involvement required
- Workflow support and seamless integration with IA book reader
- Compatibility with the DJatoka Image Server
- Support of batch mode operation and online on demand conversion
- Parallel processing (in one server) and desirably distributed (over a cluster of servers_) conversion

In order to provide such page by page viewing experience the following backend PDF conversion and image delivery components were integrated with the epublishing platform:

- A conversion platform comprising a multithreaded distributed conversion system, in order to interface with the epublishing platform and manage the

batch conversion process from PDFs to JP2000 files. The conversion platform is required to be decoupled for the delivery layer. EKT has developed the JP2K-Distiller, which is Python based set of scripts that manage the deployment, scheduling and distribution of conversion jobs, and the interfacing with the external components, i.e. the epublishing platform, OJS eJournals and DSpace repositories. JP2K-Distiller is highly distributable, in its second version, and can scale to an arbitrary number of processing nodes. File conversion is broken into several batch processes according to the active processing nodes and the conversion process is automatically delegated to the processing nodes. Furthermore functionality for the monitoring of the conversion process and log messages is provided.

- A highly scalable image delivery layer. The DJatoka image server is utilised, over a clustered installation forming a shared clustered deployment, which exploit multiple virtual servers over a shared storage, with advanced load balancing and failover capabilities.

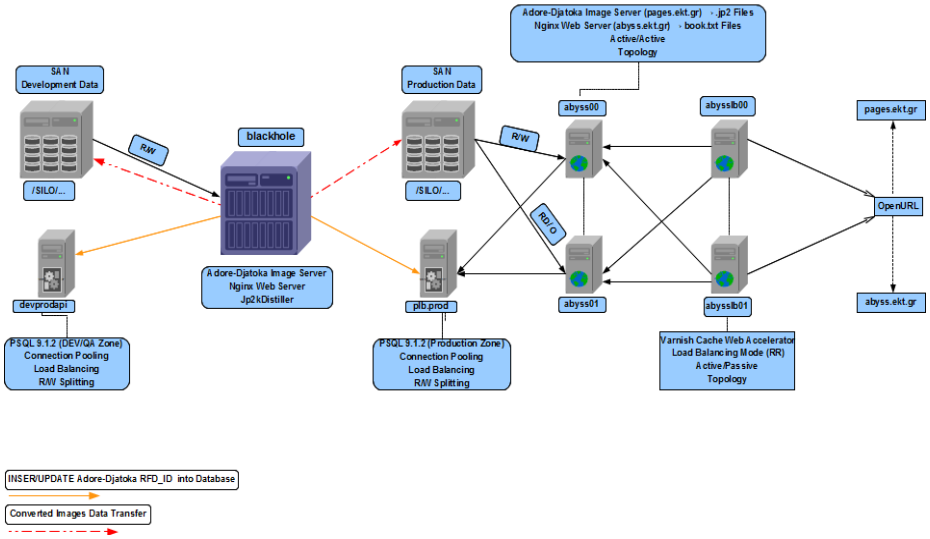


Figure 3: EKT eReader architecture.

This two-tier platform (delivery and conversion tiers), depicted in Figure 3, comprises a large scale image conversion and delivery platform, which is capable of driving large scale digital content systems and is exploited as a common service element among the publishing platform and other digital content systems (repositories and eJournals). The platform is implemented in the system level of a Nginx/Varnish failover load balancer/caching to 2 node tomcat/djatoka cluster serving JPEG2000 over a shared SAN storage and utilizes a 3 node Postgres 9 cluster.

3.3.2. Integration with external systems (OJS, Drupal, DSpace)

EKT has integrated online reading capability in several platforms, such as the central ePublishing platform (http://epublishing.ekt.gr/) which is developed with Drupal, eJournals management systems developed with OJS (e.g. http://www.medit-mar-sc.net/) and digital repositories (e.g. http://repository.edulll.gr/) developed with DSpace.

The EKT eReader is provided as a common service shared across platforms and interoperates with them using a loose coupling approach.
An important part of this interoperation is the workflow which is executed when new material (e.g. PDF files) is uploaded in any of the aforementioned platforms. The result of the workflow is that the generation and appropriate placement in the system of the entire set of artifacts (i.e. JPEG 2000 images, OCR results, indexing results) needed to publish the new material through the online reader. Furthermore, the links to the documents in the reader are made available to end users within the respective platform (Drupal, OJS or DSpace).

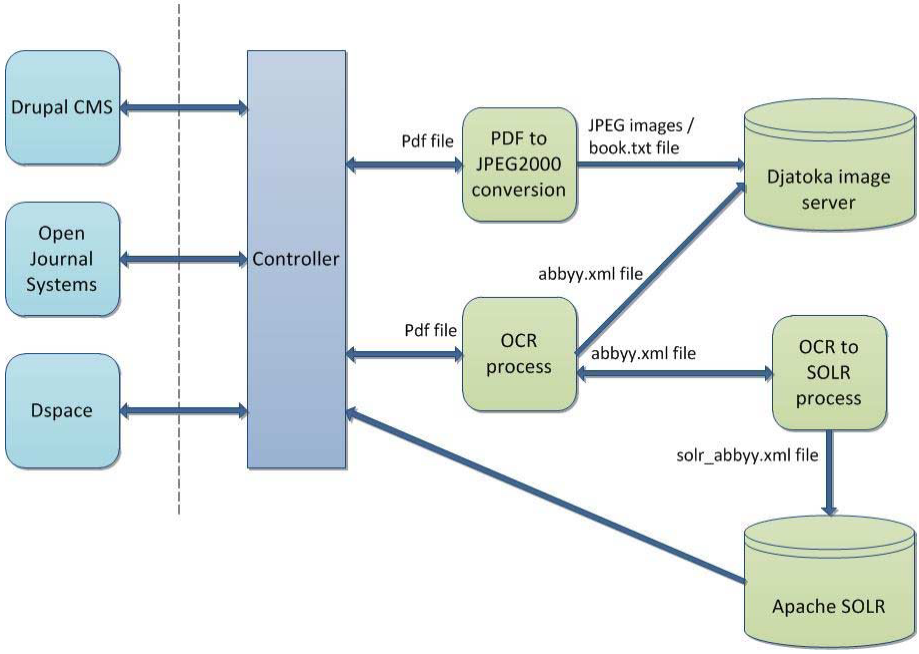


Figure 4: Integration with external systems.

This workflow is depicted in Figure 4 and can be described as follows: Each platform triggers the process via a REST request to a central "controller" which receives specific data for the new material, such as document title and authors, PDF file URL and the corresponding publishing platform (Drupal, OJS or DSpace). The controller triggers the process of converting PDF files to JPEG2000 images using the solution described in Section 3.3.1. Images are transferred to the Djatoka image server with a specified structure (e.g. file paths, identifiers) for each document and publishing platform, accompanied with a relevant text file describing this structure (book.txt file). The information in the text file includes data such as the number of images/pages per document, list of file names (reflecting order of images/pages in the document). Once the conversion process is successfully finished, the controller is notified by the conversion systems and informs the corresponding publishing platform through a callback invocation which includes as a parameter the URL of the document in the online reader. After successful conversion, the process to provide full-text search for the new document(s) is triggered, with Optical Character Recognition (OCR) as the first step; a commercial software server (ABBYY) is used for that purpose. In case the

output file (abbyy.xml file) from the OCR process meets minimum quality requirements (i.e. the text is recognizable by OCR), the file is further processed by a python script (solr_abbyy.xml file) so that it can be indexed in Apache Solr. When the indexing process is finished, the controller informs the corresponding platform in order to enable full-text search in the online reader. The whole process is triggered every time a new pdf file is uploaded in the publishing platforms for specified content types. The described EKT eReader extension has been implemented in Drupal, Open Journal Systems and DSpace and can be extended in order to meet the requirements of further systems, since the controller hides much of the complexity of the workflow implementation.

3.4. EKT eReader – additional features

3.4.1. Addressing diversity of images sizes and aspect ratios

A page by page reader presenting digital content in an image format, has to maintain uniformity of image aspect ratio and it is particularly important that image dimensions are determined by the current screen size, enabling better reading experience in mobile devices too. Moreover, the images in a document are not always in the same dimensions, considering that they might come from scanning process and be further processed. EKT eReader adopts an image processing algorithm in order to present all images in the same dimensions, by using each image width and height information provided by Djatoka image server. Image dimensions are determined according to the computation of the average aspect ratio for each document and change dynamically when screen size is altered, showing always the image with the appropriate resolution.

Furthermore, in many cases, there are documents which consist of pages with both portrait and landscape orientation, illustrating mostly large tables and wide images. Usually, a page-by-page reader implements the appropriate algorithms in order to publish and present images with standard dimensions in portrait orientation, resulting in the distortion of landscape pages. Other readers avoid using an image processing algorithm, thus, enable portrait and landscape pages to be published in their real dimensions. However, EKT eReader uses an algorithm that recognizes if there is a page in landscape orientation within a document. If a document contains at least one landscape page, then all the landscape pages are rotated 90 degrees (in two-page view) so that they are presented as portrait, avoiding images distortion and retaining the same dimensions for all pages. Furthermore, a relevant button appears next to the rotated landscape page, allowing to view the page in its real dimensions by clicking on it. Implementing this algorithm, consistency is preserved as regards to the size of pages, without distortion and quality reduction. Figure 5a illustrates a document with both portrait and landscape pages embedded in EKT eReader.

3.4.2. Bookmarking at the page level

EKT eReader is also customized in order to provide the ability to create a distinct URL not only for the entire document, but also for each page, allowing bookmarking and sharing on popular social media, at page level. A unique OpenURL-compatible URL is available for every page, both at the JP2000 image level and at the reader page level. This capability is provided by the features of the Djatoka server where a unique

identifier exists for each image (page). The format of the address is configurable through the Book Reader.

3.4.3. Printing the page level

EKT eReader enables the user to print each page in different image resolution, based on the current page zoom level. This is accomplished with the ability of Djatoka image server to determine multiple resolutions for each image, thus, printing a page in EKT eReader depends on the selected page zoom level, which corresponds to a specific image resolution in Djatoka.

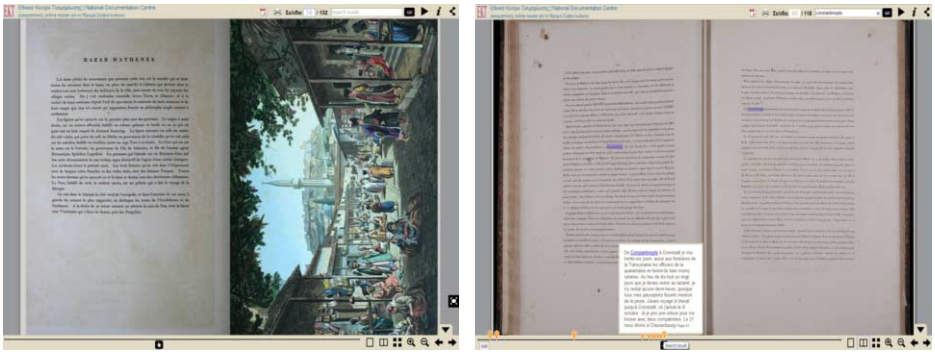


Figure 5: (a) EKT eReader with both portrait and landscape pages. **(b)** EKT eReader search results with hit highlighting

4. Infrastructure and services for full-text search

4.1. Full-text search integrated into the online reader

Full-text search service requires initially the initiation of the OCR process for a specific PDF file and the creation of the output file in XML format, which contains the coordinates of every single character, word, row or paragraph in the whole page. This process is critical for the success of full-text search service, thus, the quality of the output file must meet the appropriate requirements so that the process can proceed.

Thereafter, the output XML file is properly processed in order to be indexed in Solr search platform. Once the processed XML file is created and transferred into the Solr instance, Apache Solr is triggered to start the indexing process for all the new files that exist in Solr instance. EKT eReader implements full-text search functionality if the above actions are successfully completed, enabling the user to search inside the document. EKT eReader search process executes a query in Solr search platform and the returning results are matched with the initial XML file in order to provide targeted highlighting of search results in a quite interactive way. Figure 5b illustrates the search results for a specified keyword with hit highlighting in EKT eReader.

4.2. Distributed full-text search in the EKT ePublishing platform

The content of EKT ePublishing platform is constantly growing in a way that the integration with a search platform for distributed indexing is essential, in order to

quickly retrieve specific information. Furthermore, EKT ePublishing platform provides access to the full text of open access electronic publications which is in pdf format, in addition to the rest available metadata. Currently, one of the most popular and robust search platforms is Solr (http://lucene.apache.org/solr/) from the Apache Lucene project. Solr is an open source standalone search server which can be used as a replacement for traditional content search, providing features such as faceted search, hit highlighting, full-text search, near real-time indexing and dynamic clustering, and boosting the performance of a web application. Integrating Solr with EKT ePublishing is a common process, which leads to Drupal' s core content search replacement and the implementation of all the extra features of Solr. Indexing process of EKT ePublishing platform with Solr is triggered once each day, keeping the indexes up to date when new content is added in the platform or existing content is modified or removed.

Moreover, Apache Tika (http://tika.apache.org/) has been integrated in Solr search platform, for document text extraction, in order to provide full text search functionality. Tika is a project of the Apache Software Foundation (http://www.apache.org/) for metadata and text content extraction from various documents and can be easily implemented with Solr and Drupal CMS. EKT ePublishing platform allows electronic publications' pdf files to be uploaded in the corresponding installation file system. As for the electronic publications that are imported in the platform from external electronic journals systems, the source files of the publications are not transferred to the EKT ePublishing platform' s file system, for better performance and statistics management. As a result, not only the local files that are manually uploaded to the platform, but also remote pdf files from various external web platforms, are indexed in the same search platform, enabling distributed full text search in the EKT ePublishing platform with hit highlighting and faceted search, among other features. Implementing Solr provides also more advanced search capabilities and makes it possible to create relevant content blocks for each page showing similar content based on specific attributes. EKT ePublishing platform 2 supports search functionality with hit highlighting, facet filters and autocomplete search box.

5. Summary

This work concerns a set of infrastructural components and their utilization for the creation of related coherent services and features for end users. We present a set of sophisticated platforms, tools and mechanisms that have been employed in real-life cases for implementing document viewing and full-text search features, shared among application instances. Challenges encountered and the provided solutions are discussed.

References

[1] Tzoc, E. Document Viewers for Non-Born-Digital Files in DSpace. *Journal of Digital Information*, *13*(1), 2012.

[2] Zhou, Y. Are Your Digital Documents Web Friendly?: Making Scanned Documents Web Accessible. *Information technology and Libraries*, 29(3), 2010, pp. 151-160.

[3] Moreira, M. Multivio, a flexible solution for in-browser access to digital content. *7th International Conference on Open Repositories*, Edinburgh, 2012.

[2] http://epublishing.ekt.gr/

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-104

ELPUB Digital Library v2.0
Application of semantic web technologies

Anand BHATT[a] and Bob MARTENS [b]
[a]*ABA-NET/Architexturez Imprints, New Delhi, India*
[b]*Vienna University of Technology, Vienna, Austria*

Abstract. This paper presents the ongoing efforts to further develop the ELPUB digital library, which highly supported the dissemination of published materials within the community in the past years. elpub.scix.net has been serving the ELPUB-community since a decade, predominantly aiming to archive the output of the annual conferences. Doubtless, there is still a need for a platform to maintain the "collective memory" of the association and so far over 700 entries from 17 conferences were recorded. The repository, which utilized the SciX-technology, delivered the user access to the individual papers and made the data available via an OAI-interface as well. However, in the course of time digital library technology evolved and an association dedicated to Electronic Publishing ought to be at the forefront of novel developments. For this reason a shift was performed towards the Architexturez platform (library.elpub.net) aiming to implement advanced semantic web features. Especially the display of evolving topics and their gradual development is appealing and moreover the aggregation of individual bibliographies. Many of the features were designed in consultation with research communities in, among others, architectural computing and real estate. While deploying features and capabilities are well established in the digital library domain, the system is designed to support further research by the ELPUB community and this paper will elaborate on the transition and deliver an overview on the current prospects along with the technical capabilities.

Keywords. Digital repository, Open Access, Semantic Web, Mining, Social interaction

1. Introduction

Shortly before the turn of the millennium, comprehensive web developments took place and solutions to create topical repositories popped up. The rationale targeted to work out digital libraries - on a shoe–string budget -, which would not disappear shortly after their launch. Furthermore, there was no business model in the background aiming to keep the digital libraries running. Contrariwise, a closed pocket model based on volunteering capacities in academic environ–ments defined the starting point (likewise in ELPUB, which has no formal legal entity), which would then again donate an open access contribution to the community.

The initial work to setup a digital library can be characterized in these days as the extraction of metadata. Fortunately hardly any back digitization had to be handled for ELPUB, but even after a couple of elapsed years it took some efforts to gather the digital data. Indeed by its original launch a critical mass was already available and the ELPUB Digital Library has been extended after every annual crop [1].

Given these framework conditions the *SciX-technology* was a convenient working environment, however, trends were not covered, as further development did not take place. Nonetheless, the maintenance as such has been secured with insignificant disruptions in the course of time [2]. For this reason a search towards alternatives was started. In terms of quantities, we're moreover not talking about millions of records. However, even a couple of hundred recorded entries require a solution, which would go far beyond the basic idea of sustainable archiving.

2. Shift from SciX to the Architexturez Platform

The origin of the system goes back to 1999 when the creators of the system were required to develop experimental technology demonstrators, while writing technical standards for digital information management. The system has been developed over the years as the creators assisted in writing new standards, such as Unstructured Information Management Architecture (UIMA) and Extensible Resource Indicators (XRi) for the internet and provided the first statements of use.

The requirements for Architexturez were defined by way of a grant for Vienna University of Technology to initiate collaborative projects to initiate collaborative projects and proposed to the CumInCAD research community for review [3]. At the outset, it was determined that the novel platform should have added features such as named entity disambiguation, required to build lists of papers by entities such as authors and keywords, and semantic web features to better facilitate discovery and relationship mining activities in future.

2.1. New System Architecture

Figure 1 illustrates the reference architecture for the Architexturez platform. It is built with a *Drupal* web applications framework for the end-user facing tasks (items 5-7 in figure 1), from a heterogeneous set of modules, primarily based on *Apache Solr* and *Mahout libraries* for the applications part (layers 2-4), and a storage layer for data required by system. Proprietary data may be retrieved at runtime from external sources, such as *social network containers*, *citations databases* and *institutional repositories*. Unstructured Information Management Architecture (UIMA) and Mahout libraries provide, respectively, for search and advanced text analysis and document clustering capabilities.

Figure 1 illustrates a *reference architecture* for the Architexturez platform, reading from bottom-to-top, ① the data storage layer, including data connectors to external services and repositories, ② a database abstraction layer, ③ system core modules such

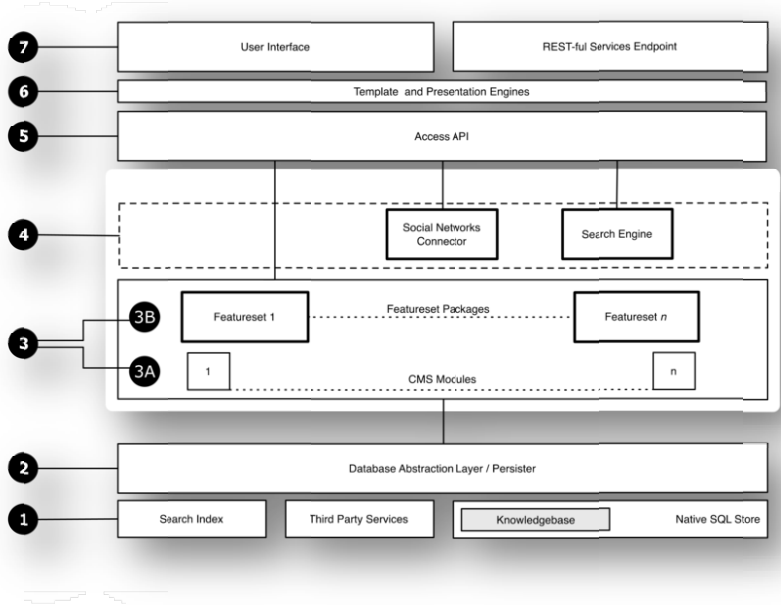


Figure 1. The Architexturez repository system architecture.

as (3a) rate calculation logic and user finder and (3b) other modules; ④ system services engines, such as faceted search engine, recommendation engine and people relevator engine, ⑤ an access API; ⑥ template engine for HTML 5 user interface and RDF/JSON, OAI-PMH and URN-NBN serializations for external customers, and ⑦ user interface and a REST services endpoint for data extraction.

If we take one instance, author names and keywords, as over time, there will occur instances where an author has used different names or abbre–viations, and synonyms for the same keyword are provided. At other times there may occur multiple authors of the same name. SciX relied on human judgment in disambiguating named entities. Comparable systems, such as *CiteSeer* [4], provided machine-learning frameworks of named entities [5] and others use rank-based algorithms [6].

Architexturez preferred to combine automated methods and individual human judgment in disambiguating named entities. It unites automated and human-edited methods and clustering algorithms that build list of authors that may be related and then provides human editors with tools to make the final decision, shown in figure 2.

3. Novel Functionalities

The Architexturez-interface differs very much from the previous SciX-solution. When entering the upgraded ELPUB Digital Library environment, the home page displays the

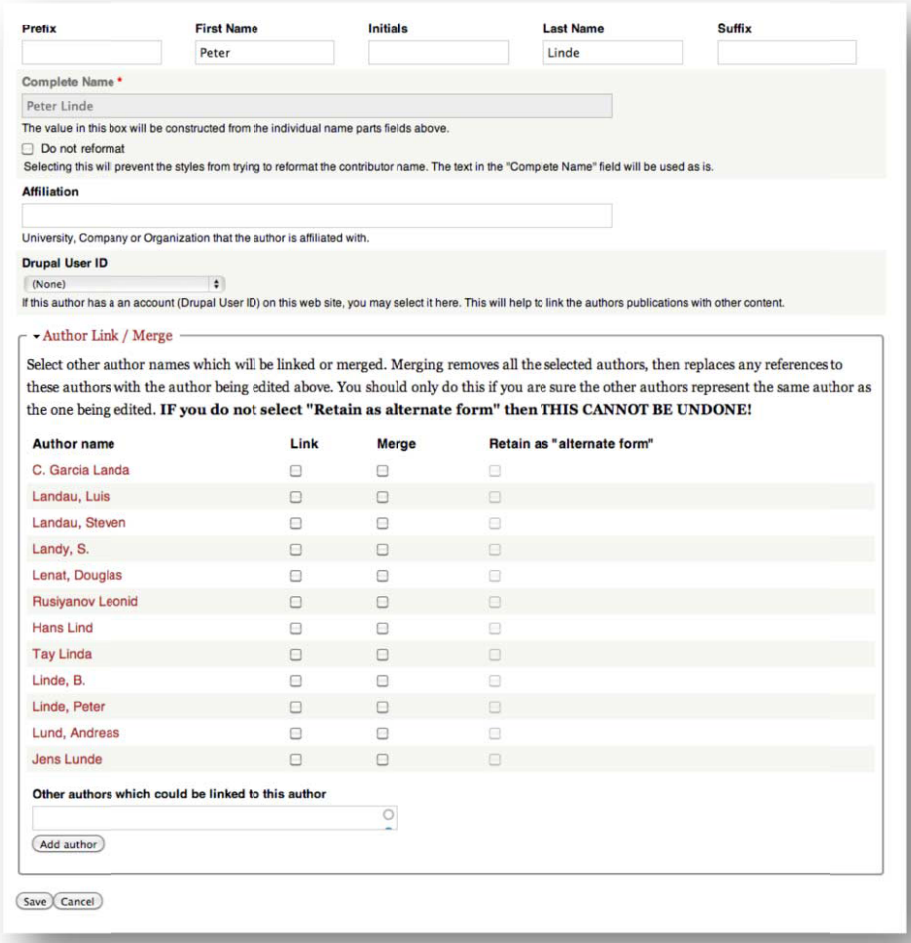


Figure 2. Author disambiguation screen – the system proposes clusters of authors with similar names, and human editors can (1) merge authors (2) retain author names as alternate forms or (3) do nothing.

previous set of conferences by way of the published proceedings. However, a user can also enter a search string at the top right. The bottom region of the page allows to embed aggregated information packages, such as news, calls for papers, or link collections, etc.

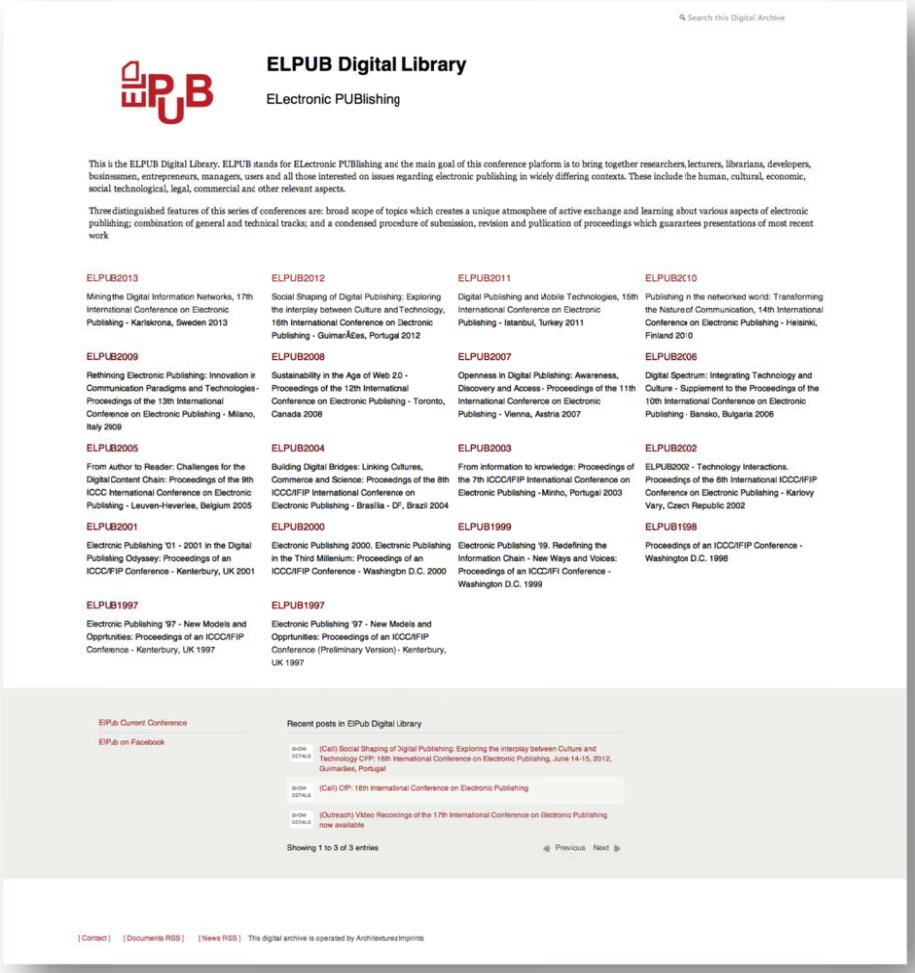


Figure 3. Architexturez: Screenshot of the front page.

Once a user has selected certain proceedings, a new screen will open, where in the central part the recorded entries are displayed. On the left side, there is a collation of keywords, and on the right a list of contributing authors in these proceedings appears. Paper abstracts and keywords are searchable on this screen as well (see figure 4).

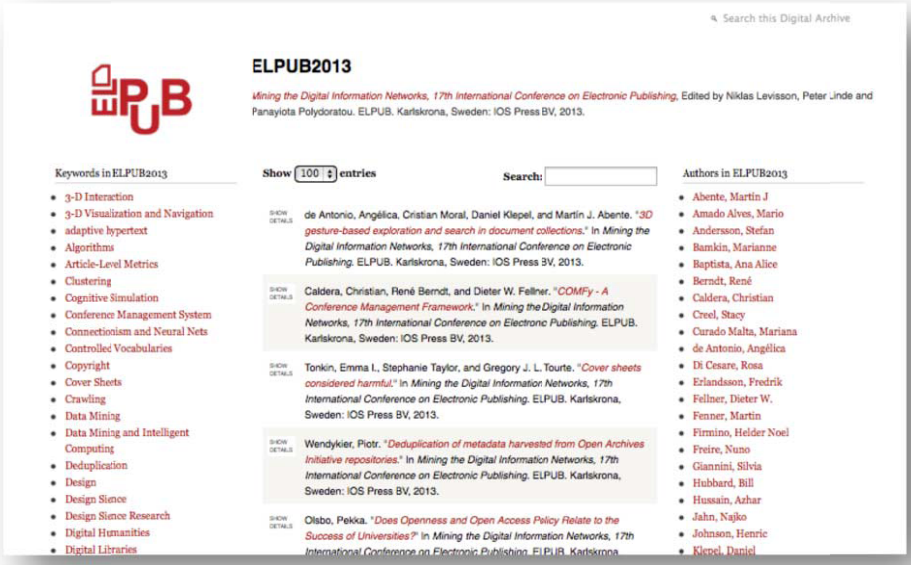


Figure 4. Display of proceedings with collated keywords and authors.

At this point the next step allows to follow a certain concept on the one hand to follow a keyword or, on the other hand, to trace back the publication output of a selected author. Especially novel members in the ELPUB community may take real advantage, in order to retrieve relationships between the evolving ELPUB topics and the involved authorship. At present predominantly PDF-files are serving as "multimedia" documents. However, other media types could be attached as well.

In term of intuitive searching/browsing, two features are being offered: (1) off-device navigation displaying citations and networks, (2) system-generated vector-space model to assist users in discovering new relationships. For the further development of the Digital Library as it stands, feedback from the side of the ELPUB-community is more than welcome.

The system embraces tools for ontology extraction for offline analysis of the documents and citations. It also provides a vector space model, for filtering and ranking relevant documents.

An important check such proposed and system-generated knowledge models must pass is the test of intuition, when authors and subject-experts review the related content. They should affirm that the related content contains documents that they would consider the most similar to the record being displayed.

Figure 5. Document record, with (1) a system generated preview of the full content and downoad link, (2) document title and abstract, (3) links to the author and keyword page, (4) related content provided by the vector space model, and (5) citation links.

4. Conclusion and Outlook

This paper described the switch of the ELPUB Digital Library towards a novel environment. The design parameters have been pointed out and a depiction of the system's interface was presented. It is to be regarded as an open invitation to the ELPUB community to use the novel platform, i.e. to experiment and research beyond the annual conferences.

The contribution intends to think about possible options for further developmental tracks. Above all the ELPUB community should be on the edge of novelties as far as the own publication output is concerned. For example data aggregation can be regarded as an important issue, where “internal” and “external” data sets are combined in a structured way.

A system to harvest citations and references to ELPUB papers has been created, and it is currently processing references from external sources. It has been stated already many times, that the amount of information is exponentially growing and even so, the need for expe–dient navigation (sorting) is becoming more and more important.

References

[1] Martens, B., Linde, B. and Turk, Z. (2003) A Digital Library for ELPUB Proceedings: The Use of a Web-Based Prototype ELPUB2003. From information to knowledge: Proceedings of the 7th ICCC/IFIP International Conference on Electronic Publishing held at the Universidade do Minho, Portugal 25-28 June, 2003.

[2] Martens, Bob; Linde, Peter; Klinc, Robert; Holmberg, Per (2008) Enhancing the sustainability of electronic access to ELPUB proceedings: means for long-term dissemination ELPUB2008. Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing held in Toronto, Canada 25-27 June 2008.

[3] Bhatt Anand; Martens, Bob (2006) ON-TO-CAAD: Investigating the Knowledge Contained within the Corpus of CAAD Research CAADRIA 2006 [Proceedings of the 11th International Conference on Computer Aided Architectural Design Research in Asia] Kumamoto (Japan) March 30th - April 2nd 2006.

[4] http://web.mit.edu/seyda/www/Papers/IST-TR_DisambiguationCiteseer.pdf

[5] http://www.kdd.org/kddcup2013/sites/default/files/papers/track2place2.pdf

[6] http://research.microsoft.com/pubs/154452/CIKM_CameraReady.pdf

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-112

EKT ePublishing: Developing an open access publishing service for the Greek research community

Alexandros NAFPLIOTIS, Victoria TSOUKALA, Nikos HOUSSOS, Andreas KALAITZIS and Evi SACHINI [1]
National Documentation Centre/National Hellenic Research Foundation, Athens Greece

Abstract. The present contribution concerns a case study of open access scholarly publishing in Greece, its history and effect in helping the local researcher community transition from a print-only mode of work to online working environments and in rendering Greek publications and scholarship more relevant to the international scholarly community. The paper elaborates on the goals of the project and the challenges that were encountered and addressed during its implementation. The project, which started in 2007 with the transition of three print journals in the humanities to an online and print format and online working environment, culminated in the development of an online platform that provides access to content and services from a single point in the web, ePublishing.ekt.gr. As part of the National Documentation Centre (EKT)'s services, we systematize and upgrade the journals' policies according to international standards, provide an online working platform and training, digitize and release in open access academic articles (more than 3,000 articles in established journals, published by small, non-profit, academic/scholarly society publishers, so far), provide DOIs, as well as concentrate on electronic books and conference proceedings – also to include purely online books in the future, starting with a born-digital monograph in a Humanities subject (onlineBook). In a nutshell, we have focused on providing publishers of scientific journals a range of comprehensive services which are constantly updated and improved in the light of the developments in scholarly communication, and which foster the internationalization, visibility, and preservation of research in these fields.

Keywords. electronic journals; open access; Greece; Open Journal Systems (OJS); epublishing

Introduction

The present paper presents a case study of open access scholarly publishing in Greece, its history and effect in helping the local researcher community transition from a print-only mode of work to online working environments and in rendering Greek publications and scholarship more visible and more relevant to the international scholarly community. The paper elaborates on the goals of the project and the challenges that were encountered and addressed during its implementation. One of the main reported successes of the project is the increased awareness among Greek

[1] National Documentation Centre/National Hellenic Research Foundation, Leoforos Vassileos Konstantinou 48, 11635, Athens, Greece, E-mail: nafpliotis@ekt.gr; tsoukala@ekt.gr; nhoussos@ekt.gr; akala@ekt.gr; esachin@ekt.gr

researchers of the capabilities and potentials of modern scholarly communication systems and the creation of a demand originating from the corresponding research communities themselves for the continuation and expansion of similar activities in the future. It should be noted here that most current content belongs to various fields of the Humanities and the Social Sciences; the significance of this cannot be overstated, especially for regional studies, often in the Greek language. In this respect, the contribution to this particular community in Greece is very important. Notwithstanding that, it should be stressed that our services and content are not limited to these specific scientific fields, but are also addressed to a wide variety of academic and scholarly publications from across the disciplines.

1. History of the service

The scholarly ePublishing service has been developed for and in collaboration with the Greek research community by the National Documentation Centre (EKT at www.ekt.gr), the backbone organization of the Greek national infrastructure for scientific documentation, online information and support services for research, science and technology. EKT is the national institution for documentation, information and support on science research and technology issues. Founded in 1980, EKT is part of the National Hellenic Research Foundation (NHRF). The latter also comprises three research institutes, one in Humanities (Institute for Historical Research) and two in the Natural Sciences (Biology/Biotechnology, Theoretical and Physical Chemistry). The platform was created as part of a wider project for the implementation of repositories and electronic publishing and was co-funded by the EU and the Greek state. More specifically, EKT's e-infrastructure development was co-financed by Greece and the European Union/European Regional Development Fund (Operational Program "Digital Convergence") project, called the National Information System for Research and Technology (NISRT). The central scope of the project was to increase support for OA in the Greek research and academic community by establishing infrastructures such as repositories and electronic journals that afford the digitization, permanent storage and free world-wide dissemination of the scientific output produced at NHRF [EKT also developed the IR of NHRF, Helios, Pandektis, A Thesaurus of Primary Sources for Greek History and Culture and the National Archive for Ph.D. Theses.

The project, which started in 2007 with the transition of three print journals in the humanities to an online and print format and online working environment, culminated (in early 2013) in the development of an online platform that provides access to content and services from a single point on the web, http://epublishing.ekt.gr. It is important to stress that EKT's ePublishing platform constitutes unique service in Greece, providing open access scholarly content and professional services to the academia, publishers and the wider public. As part of EKT's services, we systematize and upgrade the journals' policies according to international standards, provide an online working platform and training, digitize and release in open access academic articles (more than 3,000 articles in 14 established journals, published by small, non-profit, academic/scholarly society publishers, so far), gradually provide DOIs, and concentrate (apart from articles) on books and conference proceedings – also to include purely online books in the future, starting with a born-digital monograph in a Humanities subject (*onlineBook*).

2. EKT ePublishing: aims and services

We believe that the primary purpose of scholarly communication is the promotion and distribution of knowledge and we are committed to the principles of open access in providing publishing services (software, tools, knowledge, technical expertise, consultation services, web hosting and documentation) to the country's scientific communities, organizations and institutions. More than a dynamic application, EKT ePublishing embodies a business process which reinforces EKT's vision to contribute to the process of transition to new models of e-science. Adhering to the principles of Open Access to scientific information, EKT ePublishing services enable the transition of prestigious scientific publications into an online mode of operation. At EKT, we work together with publishers, editors and authors to deliver the electronic edition of accredited journals, books and conference proceedings so that they are openly available to the research and academic community in Greece.

At all stages of the publication process, EKT ePublishing supports publishers, institutions and research bodies with a range of comprehensive services which are constantly updated and improved in the light of the developments in scholarly communication. Services include, most significantly, the organization, documentation and organized dissemination of metadata and content of scholarly journals, training and consulting services on issues such as the standardization of editorial processes according to internationally accepted standards, intellectual property, the inclusion of content and metadata in international content indexers and harvesters via interoperable systems, and retroactive digitization and ingestion of legacy digital content into the platform as well as production of metadata for past issues.

We also offer open source interoperable technology and continuous IT support to the publishers we're working with. ePublishing services are addressed to public institutions and publishers of accredited scholarly journals as well as to the wider public.

3. Technical platform and services

In terms of creating a dedicated e-publishing tool for our eJournals, Open Journal Systems (OJS, by far the most popular open source platform at the international level) was identified as the most suitable system. We decided on customizing the OJS platform in ways that meet the requirements of each journal implementation. Some of the modifications developed on OJS were of the following types: additions and enhancements to the article and user metadata, including full support for hierarchically controlled vocabularies; functionality and journal workflow customisations; online reading through a sophisticated image server infrastructure; batch importing of previous journal issues; layout, appearance and usability enhancements; handling the case of articles that are complemented with supplementary files (e.g., images); multilingual support.

EKT's ePublishing web platform itself was developed in late 2012 using a distributed architecture based on the Drupal framework, the Solr indexing engine, the Apache Tika content analysis platform and the multiple OJS e-journal installations. The majority of content of EKT the ePublishing platform is harvested and published from external e-journal instances, with authorized users also having the right to

manually import content directly into the system. The connection between EKT ePublishing platform and the various OJS web platforms is accomplished through appropriate REST style web services.

The platform was customized in order to showcase the different categories of publications (eJournals, eBooks, eProceedings) in the most effective and presentable way. Our eJournals are presented in the form of a directory, with every journal being accompanied by basic information (focus and scope, scientific fields covered, editorial policies and author guidelines, announcements etc.) and, of course all the articles, grouped in issues and presented in pdf format within the platform itself. There are also separate directories for eBooks and eProceedings, with each publication both available for online reading through a page-by-page viewer (using a special infrastructure and set of services developed by EKT's Information Systems department [14]) and to download in PDF format. An important functionality of the ePublishing platform is the index of authors, which helps the reader to easily find all the works penned by the same author from across all electronic publications on the platform.

The EKT ePublishing platform allows electronic publications' PDF files to be uploaded in the corresponding installation file system. As for the electronic publications that are imported in the platform from external electronic journals systems, the source files of the publications are not transferred to the platform's file system, for better performance and statistics management. Instead, the full-text content is retrieved on demand for download or online viewing.

As a result, not only the local files that are manually uploaded to the platform, but also remote PDF files from various external web platforms, are indexed in the same search platform, enabling distributed full text search in EKT ePublishing with hit highlighting and faceted search, among other features.

An important feature of the EKT ePublishing platform is full-text search to all the material that is presented through the platform, including not only the books and proceedings that are hosted within the system but also the text of journal publications which reside in the individual e-journals (separate instances of OJS). Locally and remotely stored files are processed in the same central index, implemented using the Solr platform. A system of automatic extraction of textual content from PDF files and periodic synchronization with the central index ensures that the search function is updated when full-text material is inserted or modified within journals. The Apache Tika tool has been integrated in the Solr platform to achieve text extraction. Furthermore, hit highlighting in search results has been implemented utilizing and integrating the relevant facilities of Solr and Drupal.

4. Content and impact

Currently ePublishing hosts quality-assured scientific content from 12 Greek scientific publishers (the Institute of Historical Research of the National Hellenic Research Foundation, the National Centre for Marine Research, the Christian Archaeological Society and other respectable academic or scholarly publishers). The platform includes 12 peer-reviewed Open Access journals (following scientific evaluation processes and indexed in international databases) –each with its own, dedicated OJS installation-, as well as 2 other scientific publications, comprising more than 3,000 scientific articles in various languages, and more than 50 e-books and conference proceedings. Readers can browse an index of about 2,550 authors or search

for material that interests them by choosing from an extensive list of articles and books in 48 scientific fields. The integrated ePublishing environment has been developed with open source interoperable technology and, therefore, feeds other platforms and portals, such as www.openarchives.gr, the largest portal providing a single point of access to Greek scientific and cultural digital content of high quality (also developed by EKT). The platform is constantly updated with new issues, journals and other publications, as well as new services, and EKT is expanding its collaborations with scientific publishers who are active not only in the Humanities and Social Sciences but in other scientific fields, as well. As far as the promotion of the publications is concerned, we have developed a dissemination strategy for wide variety of outlets and recipients, for example creating leaflets and emailing them to lists and individuals with specialized interests; networking; presenting in conferences; and using social networking capabilities (on blogs, Facebook, Twitter, Flipboard etc.).

The statistics on the use of the journals are encouraging. They show a steady interest in them, for the moment centered in, but not exclusively focused on Greece. Greek and foreign researchers use the journals in their new form to conduct research. A noteworthy fact is that a more or less important part of the traffic for all journals originates from outside Greece, ranging from one-third to two-thirds of the visits. A large number of countries are represented, demonstrating the power of electronic publication in the worldwide dissemination of content. More focus will be directed in the promotion of the journals abroad, aiming at an increase of their significance as research tools for the global scholarly community. Our efforts towards achieving that goal (including targeting specific scholarly groups and communities) have already borne some fruit, with statistics for January 2014 showing a large increase in access from countries other than Greece, in comparison to figures for the first year of the platform's operation (approx. 33% vs 13%). As far as the impact of specific journal pages is concerned, there has been a steady growth in registered users for our eJournals, as well as in numbers for visits per year for all ePublishing journals, with 3,700 unique users per month on average for each eJournal.

5. New services and future plans

Since 2008, when EKT's ePublishing begun with the launch of one journal, this activity has now developed into a full-blown service that continuously grows. With a view to improving our services, increasing and diversifying the publishers served through ePublishing and the content that becomes available, new services are planned and the future of EKT ePublishing is seen in the context of relevant European services and infrastructures. Imminently, a dedicated HelpDesk will be developed and available in the spring of 2014. There, registered users will be able to submit electronically their requests and track their status, receive the response and even access the history of the communication with EKT. All users visiting the HelpDesk will be able to find answers to the most common questions that arise from new and old EKT ePublishing users by visiting the "FAQs". Users will be able to leave a comment or some input, which does not require a reply or support from EKT's personnel, by filling-in the specifically created form.

Further, EKT is seeking to enrich the types of publications offered through ePublishing. A major endeavor that will begin in the fall of 2014 is the *onlineBook* service, a specialized electronic publishing service for open access monographs. While

this, as all ePublishing services, is offered to the entire research and publishing community of the country, the *onlineBook* is being planned specifically with the needs and scholarly communication trends of the Social Sciences and Humanities research communities in mind.[1] This service is expected to be of particular interest for these scientific fields, since the monograph is a significant means for communicating their in-depth research. The aim of the *onlineBook* service is to enable publishers and Greek scientists to publish digital-born cutting-edge peer-reviewed research monographs in open access (also providing the ability to print on demand). It will comprise a suite of services for the publishers, namely, consulting that will help them develop systematic policies and specifications and improve their processes, as well as technical for the publication and dissemination of their work. As a first step, pilot-publications will be initiated with publishers already collaborating with EKT ePublishing. In 2015, it is expected that the service will become more widely available to accredited publishers. We are, additionally, exploring the possibilities of experimenting with the implementation of an open access data journal, in order to provide help the research and academic community transition to a culture of sharing their research data.

Conclusively, EKT ePublishing has gradually and steadily grown since its inception in 2007 to become a unique service for the Greek research and education community. The recent launch of the single access point platform, http://epublishing.ekt.gr, in the beginning of 2013, further contributed to the wide dissemination of the service among researchers and publishers in Greece. Apart from specific plans that aim at expanding collaborations with Greek publishers, diversifying types of publications and extending the technology, EKT is aware of the need to be part of international networks and of e-infrastructures for conducting and communicating research and is thus swiftly moving towards this direction. Finally, particular emphasis is placed on exploring appropriate income models that will enable a sustainable growth of the service in the future as one directed to accredited Greek not-for profit scientific publishers. This should be based on diversifying its incoming resources and increasing inclusion of the full spectrum of the Greek academic and research community, which it has been developed to serve (universities, research centers, and scientific societies).

References

[1] National Hellenic Research Foundation (NHRF) at www.eie.gr.
[2] National Information System for Research and Technology (NISRT) at www.epset.gr.
[3] Institutional repository of NHRF, *Helios*, at http://helios-eie.ekt.gr/.
[4] *Pandektis*, A Thesaurus of Primary Sources for Greek History and Culture at http://pandektis.ekt.gr/.
[5] *The National Archive for Ph.D. Theses* at http://phdtheses.ekt.gr.
[6] *Open Journal Systems* (OJS), more information available at http://pkp.sfu.ca/ojs/.
[7] *Drupal*, more info at https://drupal.org/about.
[8] *Apache Solr*, more info at http://lucene.apache.org/solr/.
[9] *Apache Tika*, more info at http://tika.apache.org/.
[10] Tsoukala, V. and Sachini, E-journal and Open Access Journal Publishing in the Humanities: Preliminary Results from a Survey among Byzantine Studies Scholars', in *Proceedings International Conference on Integrated Information*, Kos, 29 September-3 October 2011. http://helios-eie.ekt.gr/EIE/handle/10442/8755 (preprint and presentation)- 2012 (peer-reviewed).
[11] Sachini, E., Tsoukala, V., Houssos, N., Stathopoulou, I.-O., Paschou, Ch.-E., Paraskevopoulou, A., «Open Access in the Humanities: a case study of developing three open-access electronic journals in Greece», 13th International Conference on Electronic Publishing: Innovation in Communication Paradigms and Technologies, 10-12 June 2009, Milan, Italy, pp.543-556. http://helios-

eie.ekt.gr/EIE/bitstream/10442/8157/1/EKT_ELPUB_2009_paper_as_published.pdf (self-archived file of published paper) -2009 (peer-reviewed).

[12] Laakso M, Welling P, Bukvova H, Nyman L, Björk B-C, et al. (2011) The Development of Open Access Journal Publishing from 1993 to 2009. PLoS ONE 6(6): e20961. doi:10.1371/journal.pone.0020961

[13] Houghton J, Rasmussen B, Sheehan P, Oppenheim C, Morris A, et al. (2009) Economic implications of alternative scholarly publishing models: Exploring the costs and benefits. Loughborough University, 256 pp., available online at: http://ie-repository.jisc.ac.uk/278/. Accessed 2014 January 26.

[14] Stathopoulos, P., Houssos, N., Stathopoulou, O., Stavrou, G., & Soumplis, A. Enhancing OJS journals with advanced online reading and viewing capabilities. In PKP Scholarly Publishing Conference (2011)

[1] On this, see the recommendations of 'Open Access Monographs in the Humanities and Social Sciences Conference Report' by Jisc Collections and OAPEN, and specifically the one stating that '[W]e need incentives to encourage researchers to embrace OA enough to digitize and make available their primary sources and to embrace the opportunities that technology provides to make the future of the book a truly interactive and collaborative venture', available online at: jisc-collections.ac.uk/reports/oabooksreport.

Posters

This page intentionally left blank

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-121

A new paradigm for the scientific article

Anne-Katharina WEILENMANN[a]
[a] *BIBLINK.ch*

Abstract. Information overload is a great problem for the scientific community. To deal with the abundance of new scientific articles there are methods of sophisticated information retrieval tools and text mining tools. Little is known about the relationship between document structure and the structure of thought. This paper describes a project in this matter.

Keywords. scientific article, deconstruction, reading behavior, new writing model

Introduction and Purpose

Considering information overload and the abundance of new scientific articles, which is becoming a greater problem for the scientific community, there had to be a new discourse for finding innovative ways to cope with the growing amount of information. There are different solutions for optimizing the research process and literature review including alerting services, personalization and smart technologies, sophisticated information retrieval and text mining tools.

But is this enough to support the scholars and scientists of the future? There should be a way to go a step further and to find a new approach. It is no longer the search and retrieval process that should be the main focus of interest, but the granularity of the information, the scientific article itself.

A new pattern is emerging... New forms and ideas are being developed to improve and to reorganize the scientific article.

The purpose of this study[1] is to investigate two main questions:

1. Which cognitive processes can be observed while reading hypertexts and hyperlinks?
2. Is there another way to write a scientific article than the IMRD model (Introduction – Method – Results – Discussion)?

1. Literature review

In the digital age the process of reading and writing has changed; there are just information snippets, fragments, interlinked to each other, thus new forms of texts are created (blogs, microtexts...). New formats are emerging: In 2009, Elsevier launched

[1] This study is at the same time my doctoral thesis.

the "Article of the future"[2] as a model for embedding source data. On the one hand you see the integration of supplementary materials to enhance and enrich articles [1], on the other hand you will get new methods of construct and deconstruct scientific articles (Nowakowski et al. [2] and Amado Alves [3]). Groth et al. [4] define the future of documents as nanopublications. These scenarios require different distribution models. Priem/Hemminger [5] propose the idea of the „decoupled journal (DcJ)": „The DcJ brings publishing out of its current seventeenth-century paradigm, and creates a Web-like environment of loosely joined pieces – a marketplace of tools". Kircz [6] postulates the end of the traditional journal article.

There is very little research on the interaction between document structure and the structure of thought [7]. What effects can hyperlinks have on reading and understanding? Are the cognitive processes changing? According to Nielsen [8]: „Hypertext presents several different options to the readers, and the individual reader determines which of them to follow at the time of reading the text...".

2. Method

A suitable method for observing different cognitive processes is eye tracking. Cole et al. [9] show that you can recognize existing domain knowledge of a person with this method. Further they find: „Of particular importance is the fact that eyes fixate until the meaning of the word(s) is acquired." [10]. Eye tracking can visualize the different passages of a text already known. In this way it should be possible to delete these passages, to deconstruct and re-construct a text to generate new and shorter versions of it, to fit the right knowledge level of the researcher with the very essence of the content. How small is the smallest unit for understanding?

2.1. Study design

The study design contains determination of the test persons, which means the number of scientists and their reading patterns, and the papers/articles they have read in a defined time scale. Tenopir/Volentine [11] show that there are different reading patterns for each academic discipline and that medical/health scientists have to deal with the largest amount of papers. Therefore this study concentrates on life sciences. The number of scientists and papers is still subject to definition.

3. Conclusion and further research

The results of the eye tracking sample show the scientific article in a different and innovative way: to create an adaptive tool that can recognize a scientist's level of knowledge and to find the smallest unit for writing a scientific article.

Further research is needed concerning the granularity of a text: how far can deconstruction go to write an understandable text, can this go to the word level?

[2] Elsevier Introduces Article of the Future Project. URL: http://newsbreaks.infotoday.com/Digest/Elsevier- Introduces-Article-of-the-Future-Project-55322.asp (29.01.2014).

Combined with smart technologies there are many possibilities to develop a new and innovative eco system for scientific articles; perhaps to reach the status of Kelly´s vision: „In the new world of books, every bit informs another; every page reads all the other pages.“ [12].

References

[1] Shotton, D. The Five Stars of Online Journal Articles – a Framework for Article Evaluation. *DLib Magazine* **18** (2012), 1/2. URL: http://www.dlib.org/dlib/january12/shotton/01shotton.html (29.01.2014).

[2] Nowakowski, P., Ciepiela, E., Harężlak, D., Kocot, J., Kasztelnik, M., Bartyński, T., Meizner, J., Dyk, G., Malawski, M. The Collage Authoring Environment. *Procedia Computer Science* **4** (2011), 608–617. doi: http://dx.doi.org/10.1016/j.procs.2011.04.064.

[3] Amado Alves, M. The Shattered Document Approach to Adaptive Hypertext: Design and Evaluation. *Mining the Digital Information Networks: Proceedings of the 17th International Conference on Electronic Publishing*. N. Lavesson et al. (Eds.). Amsterdam : IOS Press, 2013.

[4] Groth, P., Gibson, A., Velterop, J. The anatomy of a nanopublication. *Information Services & Use* **30** (2010), 51–56. doi: 10.3233/ISU-2010-0613.

[5] Priem, J., Hemminger, B. M. Decoupling the scholarly journal. *Frontiers in Computational Neuroscience* **6** (2012), 1–13. doi:10.3389/fncom.2012.00019.

[6] Kircz, J. G. New practices for electronic publishing 2: New forms of the scientific paper. *Learned Publishing* **15** (2002), 1, 27–32. doi: http://dx.doi.org/10.1087/095315102753303652.

[7] Bishop, A. P. Document structure and digital libraries: How researchers mobilize information in journal articles. *Information Processing & Management* **35** (1999), 3, 255 – 279.

[8] Nielsen, J. Multimedia and hypertext: the internet and beyond. Boston [u.a.]: AP Professional, 1995.

[9] Cole, M. J., Gwizdka, J., Liu, C., Belkin, N. J., Zhang, X. Inferring user knowledge level from eye movement patterns. *Information Processing & Management* **49** (2013), 5, 1075 – 1091. doi: http://dx.doi.org/10.1016/j.ipm.2012.08.004.

[10] Cole, M. J., Gwizdka, J., Liu, C., Belkin, N. J., Zhang, X. Inferring user knowledge level from eye movement patterns. *Information Processing & Management* **49** (2013), 5, 1075 – 1091. doi: http://dx.doi.org/10.1016/j.ipm.2012.08.004.

[11] Tenopir, C., Volentine, R. UK Scholarly Reading and the Value of Library Resources: Summary Results of the Study Conducted Spring 2011. Center for Information and Communication Studies, University of Tennessee, USA, 2012. URL: http://tinyurl.com/73pr6eq (29.01.2014).

[12] Kelly, K. Scan This Book! The New York Times Magazine, May 14, 2006. URL: http://www.nytimes.com/2006/05/14/magazine/14publishing.html (29.01.2014).

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-124

Main actors in provision of fiction e-books in a small language market: a Swedish case

Birgitta Wallin and Elena Maceviciute
Swedish School of Library and Information Science
University of Borås, Sweden

Abstract. One of the consequences of the "small language" phenomenon is that the Swedish book industry is prey to the negative effects of globalization, since books have an international market and a Swedish multilingual citizen can buy e-books from international online booksellers. Publications in the local language are potentially in competition with books in English, and a local publisher or bookseller is competing with international publishers and Amazon.com

Keywords. E-books, small language market

Introduction

Today, a major technological revolution is taking place that affects every element in the total publishing, distribution and use system.

The distribution chain has changed markedly since the arrival of the e-book. Other actors than the traditional booksellers have become distributors of e-books. Large chains of bookshops, such as Borders and Barnes & Noble, are forced to shut down many of their shops. Libraries face the rivals on the Internet that started loan of e-books for small subscription fees.

An outline of the present system is shown in Figure 1. An author submits a manuscript to a publisher; the publisher assesses the market potential of the book and a contract is signed between author and publisher. The publisher produces the book, subcontracting the physical production to a printing company, or using in-house printing capacity, and retails it through booksellers. Libraries buy books through a combination of specialist library supply firms and local booksellers and individual readers either borrow from libraries or buy directly from bookshops. The alternative lines of interaction are the dotted red lines and some of the factors that affect each player.

This basic model has variants, of course: some publishers own bookshop chains, and some sell directly to the public and to libraries. The invention of the e-book, however, has the potential fundamentally to change not only the technology of book production but also how authors decide to publish their work and how readers decide to read. It also has the potential to remove the small bookseller completely from the system, although it is possible that large chains will survive by becoming more diverse. It has a potential to obliterate all other links in the chain leaving a bare carcass of direct communication between the author and the reader, though it is more likely that other mediators will enter the changing book sector as is already happening [5].

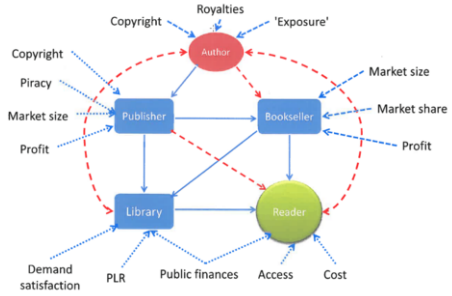


Figure 1. Interactions in the e-book field (Wilson 2013).

1. Small language and Swedish e-book fiction

The situation in Sweden with regard to the successful introduction of e-books into the national book culture is complicated by a number of factors. Sweden is a so-called small language market, but with a population that are technologically literate and, more or less, proficient in English and other languages. Therefore, many Swedes buy e-books in foreign languages provided by international booksellers. Swedish publishers are affected by piracy and the availability of free download sites as the rest in the world.

2. Swedish fiction e-book producers

A recent study [2] notes that publishing is particularly vulnerable to the impact of the e-book because of the 'long tail' nature [1] of the market for books. The arrival of the e-book means that it is now easier to satisfy low demand because an e-book can be produced on demand, rather than necessitating a print run that might find few buyers. Dyck and Sturgess [2] point to the disintermediation potential of the e-book, with authors selling directly to readers.

Swedish publishers of e-books are the largest in the country, but a number of small publishers sell their e-products on publishers' websites or through the major Internet booksellers. A variety of projects and organisations, such as libraries, museums, and archives digitize older books and disseminate them free of charge. E.g., the catalogue of the Project Runeberg lists 2762 titles of old books digitised with the help of volunteers. The website runeberg.org redirects authors who wish to self-publish to Project Gutenberg (self.gutenberg.org).

3. Swedish fiction e-book sellers

E-books are big news for libraries in the West reporting significantly increased demand. Their impact elsewhere appears to be somewhat less, but can be expected to grow as the number of devices that can be used for access increases. But as the music industry attempted to limit the distribution of music tracks, so book publishers seek to maximise the return on investment in authors and minimise the possible exposure to piracy. There are other social and cultural factors that influence the slower or quicker rate of e-book acceptance in different countries and regions. Swedish e-books do not enjoy the same

status of cultural objects as printed books. They are treated as software and/or service, therefore the VAT on them is as high as 25% (compare with 6% for print books). This is supposed to be one of the main factors slowing down their introduction to the market.

In Sweden, the sale of e-books is under control of four major publishers running the Elib network bookseller. Elib is the main channel of distribution for their e-books but also sells e-books produced by other publishers. There are other Internet sellers of e-books, such as Adlibris and Bokus. Adlibris has developed an e-reader Letto for Swedish books and Bokus provides Dito software for e-book reading. Swedish fiction can be bought through AppStore, Amazon and other international Internet sellers.

4. Swedish fiction e-books in public libraries

The provision of e-books to libraries in Sweden is dominated by one provider, Elib. The model that Elib offers to libraries includes a firm sum for each loan of an e-book (20 SEK= approx. €2.4). The books are always loaned for 28 days and only for offline reading. The access through local libraries is advertised not only through the websites of public libraries, but also on the website of Elib. There it is displayed prominently. The loans of e-books through libraries are free of charge to library users and this is also advertised on the Elib website with explanation that a reader just needs to acquire a library card for easy download of e-books.

This is an attractive model for readers and in comparison to more restricted models of e-books provision to libraries, as it does not limit multiple use of an e-book or provide other severe restrictions of use. Part of the offered collection is available with watermarks instead of restrictive DRM. Since April 2013, the company also offers a streaming service ElibU for school libraries and schools.

Table 1. Sales vs library loans of e-books

Year	Sales of e-book titles (Elib)	E-book loans from public libraries
2010	1 969	466 000
2011	3 018	647 000
2012	4 125	-

However, the Elib model is expensive for library loans of popular titles and prevents libraries from managing their collections effectively. Many Swedish public libraries were forced to stop loans of e-books as the part of the budget allocated for the service was used up. The negotiations between librarians and publishers ended without results and at present the Association of Regional and Local Authorities started this process again on behalf of public libraries. In addition, the Library Law that is in force from October 2014 demands that public libraries provide access to all literature regardless of format free of charge.

In 2013, several new actors, such as, Publit (distributor) and Axiell (digital service developer for libraries – with their platform Atingo), Adlibris, and Bokus (internet book shop) with Dito for e-books have entered e-book market and started providing books to libraries. Atingo offers a differentiated price for older and new books, but so far only few libraries have adopted their service (with Stockholm City Library taking the initiative) [3].

5. Conclusion

The diversification of the actors present in the market of Swedish fiction e-books is increasing especially in commercial and public distribution system. Though at present the power lies mainly with the publishers who dictate the conditions of e-book dissemination, the dynamics of the situation make it rather unpredictable. Public libraries are major customers for e-books and book loans through libraries are increasing more rapidly than e-book sales. Thus libraries acquire more bargaining power with the producers who cannot abandon e-books altogether because of the competition from the international book distributors.

References

[1] C. Anderson, *The long tail: why the future of business is selling less of more*, Hyperion, New York, 2008.
[2] J. Dyck and T. Sturgess, *The emerging importance of the e-book and its impact on publishing,* 2012. http://teresasturgess.files.wordpress.com/2013/01/the_emerging_importance_of_the_e-book_and_its_impact_on_publishing_dyck_and_sturgess1.pdf
[3] E. Maceviciute and M. Borg, *The current situation of e-books in academic and public libraries in Sweden*, In International Conference on Publishing: Trends and Contexts, Pula, Croatia, 6-7 December, 2013.
[4] T.D. Wilson, *The e-book phenomenon: a disruptive technology*, In International Conference on Publishing: Trends and Contexts, Pula, Croatia, 6-7 December, 2013.
[5] T.D. Wilson et al, *E-book in a small language culture: project description,* University of Borås, Borås, 2012.

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-128

Similarity between text and RDF

Marcelo Schiessl[a,1], Rita Berardi[b], and Marisa Bräscher[c]
[a] *Universidade de Brasilia, DF - Brazil*
[b] *Pontifícia Universidade Católica do Rio de Janeiro, RJ – Brazil*
[c] *Universidade Federal de Santa Catarina, SC – Brazil*

Abstract. Recently, sources of structured and unstructured data have been made available on the web, and gained attention among researchers from several areas. They are become more interested in using this global dataset due to its size and variety of information. In the Semantic Web field, many studies have translated structured data into unstructured data, and vice-versa, to make them comprehensible to machines and humans. However, we argue that we can take advantage of the existing information, in both text and RDF format. In this paper we focus on finding a way to compare them, and discovering which available text can represent an existing RDF. Hence, we propose a strategy to check whether a text represents the same knowledge that is shown in RDF format.

Keywords. Semantic Web, Natural Language Processing, Similarity measurement, RDF, Knowledge representation.

Introduction

Information is everywhere in a variety of formats. The Linked Open Data cloud (LOD) provides great sources of structured data in RDF (Resource Description Framework) format that can be consumed by humans and machines. Though, texts are still the more expressive and natural way of consuming information by people. Ideally, if we had sources of information in RDF and textual formats, we could take advantage of the benefits of each format.

Unfortunately, the correspondence between these two formats is not always straightforward. Thus, the task of translating the complex natural language (text) into a simple structure machine-readable (RDF) must be improved.

We argue that we can take advantage of the existing information, in both text and RDF formats, finding a way to compare them, and discovering which available text can represent an existing RDF. Hence, we propose a strategy to check whether a text shows the same information that is encoded in RDF format. This comparison accelerates the process of having the same information in text and RDF since it finds in advance a text to represent the RDF and both may be enriched or updated with new information. In addition, this comparison helps to decipher how patterns in natural language are represented in RDF format and vice-versa, contributing to an improvement in conversion techniques.

[1] E-mail: schiessl@unb.br, *rberardi@inf.puc-rio.br, and Marisa.Brascher@ufsc.br*

1. The similarity measure problem and the approach for estimating the similarity

We focus on one direction of the similarity, summarizing the measurement problem in the following question "Can you define which text better represents the information contained in an RDF?" Consider the motivating example in the Fig 1. Based on it, how can we objectively decide which text represents the RDF data more accurately?

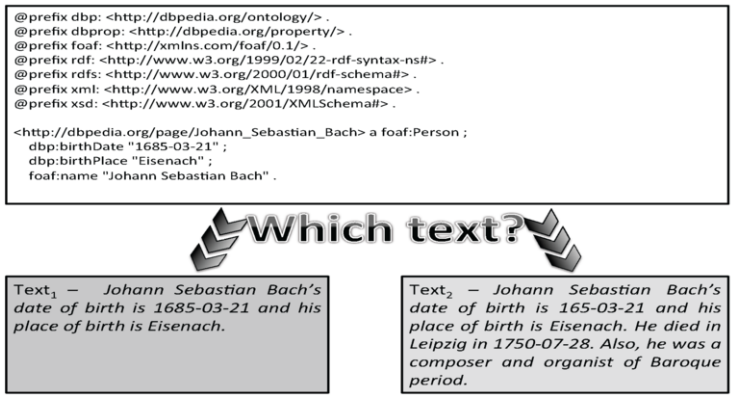


Figure 1 - RDF and texts

We consider only the triples where the object is a *literal type* to be compared to a text. Then, we take a bag-of-words from both data, and apply Dice Coefficient or Cosine Similarity, to quantify how similar they are.

1.1. Experiments

For the experiment, we focused on music composers' biographies domain. The experiment consists of setting up a corpus with 3 different textual documents that correspond to an RDF file and detecting which text is more related to this RDF. We collected 7 arbitrary names of music composers, which are Heitor Villa-Lobos (HV), John Cage (JC), Johann Sebastian Bach (JSB), Claude Debussy (CD), Ludwig van Beethoven (LvB), Richard Wagner (RW), and Wolfgang Amadeus Mozart (WAM). Figure 2 illustrates the steps to accomplish the goal.

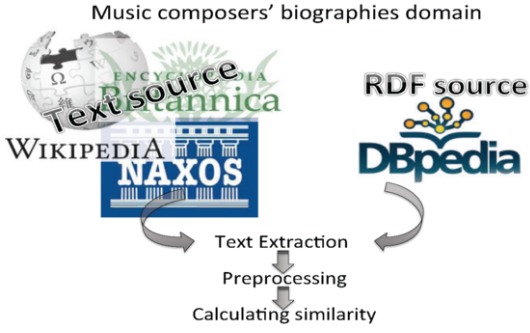


Figure 2 – Steps to accomplish the goal

The total of comparisons was 147 times, which corresponds to 1 RDF compared to the 3 versions of each txt file, i.e., Wikipedia, Naxos and Britannica. After the

preprocessing, the texts collection is composed of 28 files, 191,704 characters, 31,267 words, 1690 sentences and 564 distinct words of which the vocabulary is comprised.

1.2. Results

Figure 3 shows the result for the experiments. We present only the four highest similarity measures.

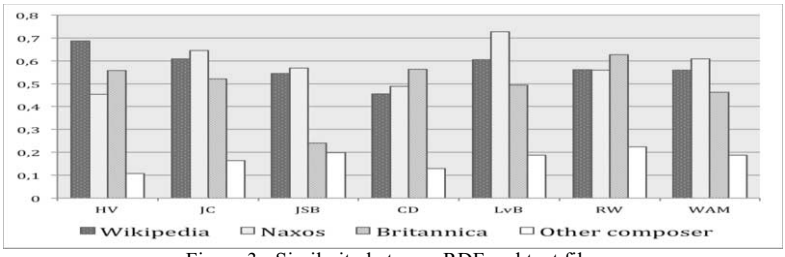


Figure 3 - Similarity between RDF and text files

Each group represents the RDF and the four bars, the texts corresponding to the highest similarity. As an illustration of how to interpret the graph, the first group is related to HV (Heitor Villa-Lobos), and the bars indicate degree of similarity.

For every group, the highest similarity measurements corresponded to the same composer in the RDF. Most of the top measures are above 0.6, which indicates an acceptable similarity. The similarity related to a different composer is in all cases the smallest one. This shows that we are able to detect a text or texts in a collection, which represents the same subject of an RDF file.

2. Conclusion and Future works

We have shown an approach to detect a text correspondent to an RDF data calculating the similarity between RDF and text files. The approach is based on a strategy of extracting the literals from an RDF and comparing them to a collection of texts by using the cosine similarity measurement. Our results seem to be a promising avenue to find out how well one format is represented in the other. As future research, we also intend to extract properties from the RDF, which can provide extra information like “birthPlace”, “hasAge” etc. and transform them to natural language format to aggregate more contents to the corpus to be compared. Besides, we will deal with negation in natural language, and semantic relatedness to improve our results.

Finally, as our work looks encouraging, in the next steps, it is essential to submit it to a more extensive proof of concept to demonstrate that our work is consistent with different domains and larger datasets.

Acknowledgments

This work was supported by CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil.

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-131

Policy Recommendations for Open Access to Research Data in Europe – Stakeholder Values and Ecosystems

Thordis SVEINSDOTTIR[a], Bridgette A WESSELS[a], Rod SMALLWOOD[a], Peter LINDE[1b], Vasso KALA[c], Victoria TSOUKALA[c], Jeroen SONDERVAN[d]

[a]*University of Sheffield, UK*
[b]*Blekinge Institute of Technology, Sweden*
[c]*National Documentation Center, Greece*
[d]*Amsterdam University Press*

Abstract. RECODE will leverage existing networks, communities and projects to address challenges within the open access and data dissemination and preservation sector, and produce policy recommendations foropen access to research data based on existing good practice. The open access to research data sector includes several different networks, initiatives, projects and communities that are fragmented by discipline, geography, stakeholder category (publishers, academics, repositories, etc.) as well as other boundaries. Many of these organisations are already addressing key barriers to open access to research data, such as stakeholder fragmentation, technical and infrastructural issues, ethical and legal issues, and state and institutional policy fragmentation. However, these organisations are often working in isolation or with limited contact with one another. RECODE will provide a space for European stakeholders interested in open access to research data to work together to provide common solutions for these issues. RECODE will culminate in a series of over-arching policy recommendations for a policy framework tosupport open access to European research data targeted at different stakeholders and policy-makers. (http://www.recodeproject.eu).

Keywords. Open data, Open Access, Open Research Data, Policies

Introduction

The objectives of the first RECODE work package are to:

- Identify and map the diverse range of stakeholder values in Open Access data and data dissemination and preservation.

[1] Peter Linde, Blekinge Institute of Technology, 37179 Karlskrona, Sweden. Peter.linde@bth.se

- Map stakeholder values on to research ecosystems using case studies from different disciplinary perspectives.
- Conduct a workshop to evaluate and identify good practice in addressing conflicting value chains and stakeholder fragmentation.

Three related actions were used to address the objectives:

- An analysis of policy and related documents and protocols, in order to map the formal expression of values and motivations.
- Five case studies in particle physics, health sciences, bioengineering, environmental research and archaeology. These explored issues of data size; quality control, ethics and data security; replication of large datasets; interoperability; and the preservation of diverse types of data.
- A validation and dissemination workshop that sought to better understand how to match policies with stakeholder drivers and motivations to increase their effectiveness in promoting Open Access to research data.

The Definitions of and Vision for Open Access

- The European Commission definition of "Open Access" is "free ... access to and use of publicly-funded scientific publications and data".
- The Berlin Declaration states that Open Access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.
- The Berlin Declaration's vision is that Open Access to data has the potential to create "a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community".

The Stakeholder Taxonomy

We have identified five basic functions in the Open Access ecosystem: Funders & Initiators; Creators; Disseminators; Curators; and Users. These functions are represented by different performers (stakeholders). Each performer undertakes activities and produces records in relation to Open Access Data. The functional taxonomy of Open Access stakeholders was constructed in parallel to the review work and mapping of stakeholder values in WP1. The construction work of the taxonomy and the wide search for data management protocols, Open Access policies, and ethical protocols have fertilised and re-fertilised both tasks to ensure that we cover all aspects of the Open Access ecosystem.

Conclusions

There is a clear overall drive toward Open Data Access within the policy documents, which is part of a wider drive for open science in general. This is underpinned by the view of science as an open enterprise. In societal terms, data is valued as a public good in the sense that its production is funded by public money and thus should be accessible to the general public.

Discussions of Open Data tend to refer to science as a single sector, leading to differences between disciplines being ignored in policy making. Each discipline has different methods for gathering and analysing data, some disciplines deal with sensitive data, others deal with data that may have IPR or legal issues. It is important that these differences are recognised, as they will inform the debate about subject specific requirements and common infrastructures for Open Data Access.

Future work

Work package 2 – Infrastructure and technology will focus on issues surrounding open access and data dissemination and preservation infrastructure and technologies, including issues such as standards, interoperability, metadata, etc.

D2.1 – Infrastructural and technological challenges and potential solutions (*Submitted March 2014*)

Work package 3 – Legal and ethical issues in open access and data dissemination and preservation will identify legal and ethical obstacles, barriers and solutions in relation to open access and data dissemination and preservation across Member States and third countries from the perspectives of a range of different stakeholders.

D3 – Legal and ethical barriers and good practice solutions (*Submitted April 2014*)

Work package 4 – Institutional evaluation and support for open access data will focus on institutional practices and barriers in open access and data dissemination and preservation, including examining measures for evaluating data quality, integrity, impact and trustworthiness of data as well as barriers such as training, funding and infrastructure.

D4 – Institutional barriers and good practice solutions (*Expected June 2014*)

Work package 5 – Policy guidelines for open access and data dissemination and preservation will consolidate the information from WPs 2-4 and review relevant open access and data dissemination and preservation policies at the European and Member State level and in third countries. It will identify policy gaps where the grand challenges discussed in WPs 2-4 are not being addressed and consolidate a series of policy recommendations.

D5 – Draft guidelines for different stakeholder groups on supporting open access to and preservation of research data (*Expected Sept 2014*)

Work package 6 – Stakeholder engagement and mobilisation will create a taxonomy of open access stakeholders and consider how open access and data dissemination and preservation stakeholders might be best mobilised to implement the RECODE policy recommendations and maintain collaboration activities between different types of stakeholders.

D6 – Using existing open access networks to support policy harmonisation across Europe (*Expected Jan 2015*)

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

Subject Index

This page intentionally left blank

Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

Author Index

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank