NEW AVENUES FOR ELECTRONIC PUBLISHING IN THE AGE OF INFINITE COLLECTIONS AND CITIZEN SCIENCE: SCALE, OPENNESS AND TRUST

This page intentionally left blank

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust

Proceedings of the 19th International Conference on Electronic Publishing

Edited by

Birgit Schmidt

University of Göttingen, State and University Library, Germany

and

Milena Dobreva

University of Malta, Malta

IOS Press

Amsterdam • Berlin • Tokyo • Washington, DC

ISBN 978-1-61499-561-6 (print)
ISBN 978-1-61499-562-3 (online)
Library of Congress Control Number: 2015948572

Cover photo courtesy of Elisa Von Brockdorff.

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

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

Preface

The International Conference on Electronic Publishing (Elpub) is just one year away from its 20th anniversary. Elpub 2015, the 19th edition of the conference, will continue its tradition of bringing together a wide range of stakeholders: academics, publishers, lecturers, librarians, students, developers, entrepreneurs, and users interested in issues regarding electronic publishing in diverse contexts. Three distinguishing features of this series of conferences are: a broad scope of topics creating a unique atmosphere of active exchange and learning on various aspects of electronic publishing; the combination of general and technical tracks. Lastly, a streamlined submission, revision and proceedings publication process guarantees the inclusion of current and cutting edge research in the programme.

Twenty years is a commendable lifespan for a conference in such a volatile area. Elpub has contributed to the ever-changing environment every year by focusing on a special theme. Elpub 2015 will explore the interplay of two dimensions of electronic publishing – the ever growing volume of digital collections, and the improved understanding of the widest user group: that of citizens. This exciting theme encompasses human, cultural, economic, social, technological, legal, policy-related, commercial, and other relevant aspects.

Are we currently facing the dawn of the age of citizen science? Not quite yet. However, open science is clearly on the horizon – not least through both the dedication of a growing group of researchers and the incentives set out by the European Commission and other research funders world-wide. The rationale for this approach is manifold: from promoting free access to all kinds of outputs resulting from publicly-funded research; a clear need to address data management, sharing and reuse; and open e-infrastructures for research to new forms of collaboration, publishing and user engagement. All this sets the foundations for a much wider interaction of researchers with their own community as well as professionals, industry representatives and citizens.

Citizen engagement will be by nature diverse. For example, communication via social media, or joint cultural heritage projects which explore the participation of citizens, e.g. patients and carers, in the effort to create better ways to discover, enrich and select information. The conference will present a range of results in this area, and explores opportunities for participation in old and new publishing paradigms.

The conference theme will be introduced by three inspiring keynotes: The main program on 1–2 September 2015 features Prof. Gowan Dawson (Leicester University, UK) who will deliver a keynote on "Constructing Scientific Communities: Citizen Science in the 19th and 21st Centuries". Prof. Gail Feigenbaum (Getty Research, USA) will explore "Electronic Publication: Intended and Unintended Consequences". Finally, a special intervention within the conference panel session will be offered by two local speakers from the publishing industry in Malta, Donald Tabone and Adrian Hillman.

Similar to previous conferences, Elpub offers a combination of themed sessions and a poster session. In addition, the discerning participants have a chance to follow satellite events focusing on topics which enjoy a great interest in the professional community.

Two pre-conference workshops will be held on 31 August 2015: “The evolving scholarly record: library stewardship roles in a fast changing multi-stakeholder ecosystem”, presented by Titia van der Werf, Senior Program Officer at OCLC Research, and “Upskilling for Research Data Management: How do you train the Data Librarian?” which will be delivered by Andrew Cox (University of Sheffield) and Anna Maria Tammaro (University of Parma). This is also co-organised with the Maltese Library and Information Association (MaLIA). Post-conference events feature a workshop organised by the EC-funded CRe-AM project entitled “Shaping the future for e-Publishing”, the launch of the DARIAH in Malta, and a workshop on “The role of knowledge maps for access to Digital Archives” organised by the KNOwESCAPE COST action.

The conference takes place as one of the academic facets of VIVA – the Valletta International Visual Arts Festival, which reinforces the creative edge of electronic publishing and strengthens the prominence of the electronic publishing domain in the host country, Malta. It is co-organised by the St James Cavalier Centre for Creativity which offers its unique premises to the 19th edition of Elpub, and by the Department of Library Information and Archive Sciences of the University of Malta.

We would like to express our gratitude to all members of the Elpub Executive Committee who, together with the Programme Committee, helped to put in motion an impressive range of ideas which allowed us to offer such a diverse and exciting programme. We also would like to thank our sponsors – Emerald, ProQuest, Copernicus, and Springer at the time of writing – for their support and interest in reinforcing the connection between academic discourse and the professional publishing community.

We wish you all an inspiring conference and look forward to the anniversary 20th occasion of Elpub in Göttingen!

Birgit Schmidt and Milena Dobreva

1 July 2015

Conference Organisation

Organising Chairs

General Chair: Milena Dobreva – University of Malta, Malta
Programme Chair: Birgit Schmidt – University of Göttingen, Germany

Executive Committee

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

Organising Committee: Volunteers

Aggeliki Varsamidou
Despoina Siapkari
Dimitris Iliadis
Elli Papadopoulou
Panagiota Karmpa
Sotiris Sismanis
Vicky Georgiadis

Programme Committee Members

Ana Alice Baptista – University of Minho, Portugal
Arūnas Gudinavičius – Vilnius University, Lithuania
Bob Martens – Vienna University of Technology, Austria
Caroline Sutton, – Co-Action Publishing, OASPA, Sweden
Charlie Abela – University of Malta, Malta
Eva María Méndez Rodríguez – Universidad Carlos III de Madrid, Spain
Fernando Loizides – Cyprus University of Technology, Cyprus
Herbert Gruttemeier – Inist – CNRS, France
Jan Engelen – Katholieke Universiteit Leuven, Belgium
Jenny Molloy – University of Oxford, Open Knowledge Foundation, UK
Laurent Romary – INRIA, Humboldt University, Germany
Leslie Chan – University of Toronto, Canada
Mikael Elbæk– Technical University of Denmark, Denmark
Natalia Manola – University of Athens, Greece

Sponsors


Emerald


Copernicus Publications
The Innovative Open Access Publisher


ProQuest


Springer

This page intentionally left blank

Contents

This page intentionally left blank

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-1

Lay Summaries for Research Articles: A Citizen Science Approach to Bridge the Gap in Access

Monica DUKE[1]
Digital Curation Centre, University of Bath

Abstract. The Patients Participate! project explored the feasibility of a citizen science approach to writing lay summaries for research articles. It involved a range stakeholders: funders of research (medical charities), service providers (the British Library), researchers and patients. Informed by practices within medical charities and the experiences of other citizen science projects, different methods were used to investigate trust, the skills required to produce a good lay summary, and the benefits of citizen science. A literature review into human factors was carried out and platforms for service delivery were analysed. The project was able to synthesise guidelines on participation in citizen science projects and the writing of lay summaries, and to identify challenges. This paper summarises the outcomes and lessons learned.

Keywords. Citizen science, patient participation, lay summaries, widening access.

1. Introduction

This paper describes a feasibility study that investigated some parallel trends in research in general, and in health in particular, to determine how to improve understandability of research articles for the general public. Patient and public involvement (PPI) is an approach that has arisen within delivery of healthcare and medical research. PPI describes processes in which non-professionals are included in medical decision-making that affects them. More broadly, the term citizen science is used when projects engage volunteers (the general public or enthusiasts) to tackle research tasks that might otherwise not be feasible due to scale. Benefits are claimed not only for the project, but also for the participants and society in general. With reference to research literature, including medical research, Open Access refers to a movement to make the results of research (as communicated chiefly through publications) more widely available, mainly by removing cost barriers. The Patients Participate project[2] asked the question: is it feasible to align these approaches for greater openness and involvement of the public in research by providing lay summaries alongside research articles, to improve accessibility, harnessing the effort of volunteers to power the effort?

[1] M. Duke@bath.ac.uk
[2] Patients Participate! Project website http://blogs.ukoln.ac.uk/patientsparticipate/

The rest of this paper describes the lessons that the project was able to learn from the practices of medical charities and the experiences of citizen science projects, using different methods such as workshops and literature review, to gather different perspectives on the feasibility of providing lay summaries to research articles using contributions provided by patients.

2. The Trend for Patient Participation

Whilst patients have been involved in medical research in different ways, such as participating in trials, donating tissue or analyzing their genes through services like 23andme[3], these can be considered passive roles. In contrast, patients[4] can take more influential roles such as setting research strategies and priorities for medical research charities, evaluating research by taking part in the peer review process alongside scientific experts and contributing to communicating the results of research. INVOLVE, a national advisory group, defines involvement as 'research being carried out '**with**' or '**by**' members of the public rather than '**to**', '**about**' or '**for**' them' [1].

Medical charities fund a significant amount of medical research (over £1 Billion in 2010/11) and see communication with their supporters about the research they fund as one of their key strategies [2]. They are keen to involve patients in the conversations about research so that it meets patient priorities, research is patient-focused and they consider the passionate and committed patient as being a powerful advocate who can act as an ambassador for the charity [3].

3. What Is a Lay Summary and How Are Lay Summaries Used?

Lay summaries are short accounts of research that are targeted at a general audience. Smith and Ashmore [4] recommended the INVOLVE definition as being the most succinct "A lay summary is a brief summary of a research project or a research proposal that has been written for members of the public, rather than researchers or professionals. It should be written in plain English, avoid the use of jargon and explain any technical terms that have to be included".

Besides members of the public, lay summaries can also be used by other researchers from nearby fields. They are often requested as part of grant application processes. Medical charities in particular are involving patients in decision-making: as members of funding panels, in parallel lay review processes, or simply by commenting on the value of research projects and their feasibility [5]. There are also reports that presentations that have been simplified for lay members are also easier to understand for other scientists in panels. Smith and Ashmore have suggested that lay summaries may be the only part of an application that a busy reviewer may ever read. Lay summaries can also be used for the recruitment of participants in clinical trials. CancerHelp UK is a website produced by Cancer Research UK using an experienced writing team to describe cancer trails and studies.

[3] 23andMe website https://www.23andme.com/

[4] The term patient and public describe a wide range of roles taken by people who may become involved, such as advocates, consumers, survivors or carers. 'Patient' is used to stand for an individual with an interest in a disease-condition from a personal perspective, and may not have had the condition themselves. [5]

4. Open Access and the Public Engagement Agenda – Researchers' Perspective

Due to changes in research communication, arising not only from the movement to open access to publications without charge, but also from funders' directives to consider the impact that research has on different beneficiaries, researchers need to widen their reach. Previously the main audience for research would have been considered to be other academics. With greater emphasis on engaging with different stakeholders, medical researchers have started to consider research patient groups as key stakeholders. Some of the aims of better communication include equipping patients to judge what the research means to them, helping them understand the investment in science, and keeping them better informed about advances.

Whilst academics accept that finding and accessing information can be challenging for the lay person, addressing the technical language and complexity inherent in their research can be a challenge. Researchers need to understand what patients want and what they value, and find the best routes for delivery and engagement [6].

However, the skills required for writing a lay summary are different from other writing tasks which may lead to difficulties when writing lay summaries [4]. Medical charities found that some researchers continued to write summaries that were not sufficiently clear or simplified [5]. Researchers require guidance on what should be provided in a lay summary and clarity about how summaries will be used (for example to make funding decisions) [3]. Other possible barriers that have been suggested are the variation in requirements across funders (e.g. word length) and directions that appear to be conflicting (e.g. brevity versus providing adequate explanation) [4].

5. Lessons from Citizen Science Projects

5.1. Human Factors in Citizen Science as Reported in the Literature

The project used a small selection of reports in the literature, with a focus on web-based citizen-science projects that conduct crowd-sourced data analysis or data collection (or reported experiences), to extract some factors that need to be considered in planning and delivering a citizen science project. Due to the limitations of space these can only be addressed briefly here, but the full report is available [7].

One of the findings was that involvement is affected by trust and credibility, and credibility is in turn influenced by ease of use and perceived risk. The site's look and feel can be used by users as an indicator of a site's credibility. The choice of factors that has been studied varies, but there is some consensus that perceived trustworthiness and credibility are a function of user judgements of various factors. These include user attributes: cultural factors (like nationality), attitude towards the activity being carried out, and the site usability (like ease of navigation and the level of guidance and support for the user).

Since several projects involve collection of data by participants (rather than writing), discussions on quality of contributions tend to focus on collection of data and techniques for data validity. Paulos [8] and Cooper [9] suggest complementary frameworks for carrying out citizen science projects. Besides planning around data collection, another important (and perhaps obvious) focus tends to be the participants: how to recruit them, train them and motivate them.

GalaxyZoo [10] had a dramatic increase in participation following a launch through the BBC Radio News item with the news spreading through print and online media. The participants were keen not only to contribute to the task, but were also active in helping each other through forums and collaborative research.

Although the subject of participant motivation has been somewhat understudied, Raddick et al were able to compile a set of motivation categories for GalaxyZoo volunteers. Nov et al [11] provide a number of pointers to literature that explores the motivation of contributors in citizen science communities and information-sharing communities (like wikipedia). They group motivation into intrinsic (improvement of skills, enhancement of status) and extrinsic (fun, intellectual stimulation).

5.2. Potential Benefits

Whilst projects that use a crowdsourcing model to increase the manpower available to the project derive benefit, it is suggested that the citizen participants will also enjoy possible gains such as:

- Empowerment: by becoming active participants and stakeholders
- Improved understanding
- Social contact: platforms for citizen science can act as a virtual meeting place, and can help to form communities and connect people who share interests. If researchers get involved in these communities, contact between scientists and citizen participants can also be facilitated.
- Inclusivity: by providing a level playfield where differences (physical or social) may be surpassed.
- Skills development: specific training, knowledge acquisition in a particular field, or confidence with technology or communication skills could be acquired.

5.3. Available Platforms

The features of platforms supporting projects which use a crowdsourcing model of engagement were analysed for suitability to the task of writing lay summaries [12], one class of projects are based around wiki-like platforms. Of these AcaWiki had a closely-matching aim of presenting summaries of academic papers. The WikiMedia medicine portal is written by volunteers and includes links to research as well as other medicine-related topics such as news and images. Of other sites linked to patients and medical information, PLoS Medicine offers lay summaries alongside research articles; the summaries are written by editors following a set of internal guidelines. PatientsLikeMe provided a model of a large site where patients are involved in providing information (mainly personal information on treatments, symptoms, progression and outcomes). Rather than use expert mediation, PatientsLikeMe puts the focus on patients interacting with each other.

Of some other platforms available, GalaxyZoo is an example of a successful initiative in the field of astronomy with many volunteers who help classify images of galaxies. The volunteers are supported through a blog and a forum. The software is available for setting up other projects, however the tasks undertaken by participants and

the input gathered are not similar enough to that of writing a lay summary. RunCoCo on the other hand had a focus on creating a community-donated collection of content, either by uploading content or adding information on existing resources. This suggested that the software might be better adapted for the input of a structured description that would be required in a lay summary.

6. Results: Synthesis of Lay Summary Guidelines and Practices by Medical Charities

Charities would like to find out about and access research publications that result from the research that they fund. Medical charities were surveyed about their practices, and the results are captured in some longer case studies [13], an overview of innovative ways being used by charities, for example, use of social media [14], and a table showing which charities are engaging in the production of lay summaries, and how: who writes the lay summaries, whether they provide guidance and the stages of research at which lay summaries are used [15].

Furthermore, some of the charities were willing to share their guidelines for producing lay summaries. Generally speaking, guidance for writing lay summaries comprises a structure or sections to address the questions that patients would like answered about research (which can be in the form of a template) and directions on writing style. The content should include who funded the research and why, the impact expected, concrete everyday examples should be used and timescales given where relevant. Instructions for writing style consist of suggestions such as writing in the active voice, positive phrasing, using simpler everyday words, avoiding jargon, using correct grammar, punctuation and spelling and an appropriate tone.

7. Challenges

By investigating the current practice in lay summary writing and in running citizen science projects, the project was able to identify a number of questions that would need to be addressed:

Current practice suggests that patients are taking influential roles in directing research, however lay summaries are written in the main by researchers or highly trained writers. How could training to patient participants be delivered at scale? One possible model could involve collaboration between researchers and patients, with patients giving feedback to help refine a lay summary written by researchers so that it fulfills its purpose.

None of the available platforms were clearly geared at the task of writing a lay summary, although some offered features that could be adapted (to structure the writing). Moreover it is not known if some of the aspects (such as reputation and ratings to motivate contributions) would be suitable in this context. Further information is also needed about any special needs that need to be met in order to be inclusive.

Although between them charities have compiled guidelines for lay summary writing which were in enough agreement to allow a synthesis [3], it was not clear to what extent existing guidelines had been tested. How can quality assurance of summaries be implemented and what evaluation criteria can be applied? Evidence of impact would also need to be collected.

The infrastructure service delivery model still needs to be explored to find out how to associate summaries with the research article and make it available alongside.

8. Acknowledgements

The Patients Participate! Project was funded by the JISC eContent Programme during 2011-12. The project partners were the Association of Medical Research Charities (represented by Simon deNegri, Sarah Ellis and Kay Julier), the British Library (represented by Lee-Ann Coleman and Karen Walshe) and the DCC at the University of Bath (represented by Liz Lyon and Monica Duke with input from Emma Tonkin, Melanie Welham and Paul De Bank). The overview provided is the result of various activities and contributions by the different project partners during the running of the project and summarises outputs from the project produced by the whole of the project team The project is also indebted to the charities, patients, researchers and thought leaders who took part in the workshop and other activities; without their contributions the project would not have been possible.

References

[1] INVOLVE. What is public involvement in research? Available from: http://www.invo.org.uk/find-out-more/what-is-public-involvement-in-research-2/

[2] S. Ellis, Patient Perspectives on Research Findings [Powerpoint presentation]. Available from: http://www.slideshare.net/monicaduke/sara-ellis

[3] M. Duke, How to Write a Lay Summary. DCC How-to Guides, Edinburgh: Digital Curation Centre (2012). Available from: http://www.dcc.ac.uk/resources/how-guides

[4] M. Smith, Ashmore, The Lay Summary in medical research proposals – is it becoming more important? Poster presentation at Making an Impact – Annual Conference of the Association of Research Managers and Administrators, Manchester, June 2010.

[5] Association of Medical Research Charities. Natural Ground: Paths to patient and public involvement for medical research charities. AMRC (2009). Available from: http://www.amrc.org.uk/publications/natural-ground-paths-patient-and-public-involvement-medical-research-charities

[6] M. Welham, Communicating Research Beyond Academia A Researcher's Perspective [Powerpoint presentation] Available from: http://www.slideshare.net/monicaduke/melanie-welham

[7] M. Duke, E. Tonkin, Patients Participate! Literature review: Usability and human factors in citizen science projects, and trust and credibility on the Web, UKOLN: University of Bath (2012). Available from: http://blogs.ukoln.ac.uk/patientsparticipate/files/2012/09/PatientsParticipate-literature-review-v1.2.pdf

[8] E. Paulos, Designing for Doubt: Citizen Science and the Challenge of Change, Engaging Data (2009). Available from: http://senseable.mit.edu/engagingdata/downloads.html

[9] C. B. Cooper, J. Dickinson, T. Phillips, R. Bonney, Citizen Science as a Tool for Conservation in Residential Ecosystems, *Ecology and Society* **12**(2) (2007), 11. Available from: http://www.ecologyandsociety.org/vol12/iss2/art11

[10] M. J. Raddick, G. Bracey, P. L. Gay, C. J. Lintott, P. Murray, K. Schawinski, A. S. Szalay, J. Vandenberg, Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers, *Astronomy Education Review* **9**(1) (2009), 1. Available from: http://arxiv.org/abs/0909.2925

[11] O. Nov, O. Arazy, D. Anderson, Technology-Mediated Citizen Science Participation: A Motivational Model. *Proceedings of the AAAI International Conference on Weblogs and Social Media* (2011). Available from: http://faculty.poly.edu/~onov/Nov_Arazy_Anderson_Citizen_Science_ICWSM_2011.pdf

[12] Patients Participate! Citizen Science Platforms Available from: http://blogs.ukoln.ac.uk/patientsparticipate/

[13] Patients Participate! Case Study Report (2011). Available from: http://blogs.ukoln.ac.uk/patientsparticipate/files/2011/10/Case-study-report-Final.pdf

[14] Patients Participate! Perspectives from medical research charities (2012). Available from: http://blogs.ukoln.ac.uk/patientsparticipate/files/2012/10/PP-Charities-perspectives.pdf
[15] Patients Participate! Survey of members' activities on lay summaries (2012). Available from: http://blogs.ukoln.ac.uk/patientsparticipate/files/2012/10/PP-AMRC-members-writing-for-lay-audiences-summary-table.pdf

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-8

CIVIC EPISTEMOLOGIES – Development of a Roadmap for Citizen Researchers in the Age of Digital Culture

Antonella FRESA[a, 1], Börje JUSTREL[b], Valentina BACHI[a], Neil FORBES[c]
[a]*Promoter S.r.l., Italy*
[b]*Riksarkivet, Sweden*
[c]*Coventry University, UK*

Abstract. The CIVIC EPISTEMOLOGIES project investigates citizen science and crowdsourcing in the domain of the research in Digital Cultural Heritage and Humanities (DCHH). The ultimate aim is to produce a validated Roadmap indicating the suggested direction that the deployment of services and infrastructures should take, in order to support the participation of citizens in the research processes and the participation of creative industries in the exploitation of digital cultural content. The case of DCHH is particularly relevant because of the major cross-cutting role that the humanities play in European research and innovation, recently acknowledged in a clear way in the Horizon2020 Community Programme for Research and Innovation. Cultural heritage and humanities also represent a subject area in which citizens are particularly active, counting several – still spread - experiences of their involvement in recording, annotating and cataloguing activities on an individual or group basis, as volunteers and amateurs. The case of broadening e-Infrastructure deployment to support the participation of citizens to DCHH research, even if holding a strong impact potential for social cohesion and job development, is not yet fully explored. The paper discusses about the multidisciplinary approach to citizen science and how this method can contribute to the benefit of many scientific domains, research communities, and technology advancements as well as delivering novel social and economic impact.

Keywords. Civic Epistemology, research, DCHH, Citizen Science, Digital Cultural Heritage, Digital Humanities, e-Infrastructures, creative industries, roadmapping.

1. Introduction

The term Civic Epistemologies was introduced by Sheila Jasanoff in 'Designs on Nature' [5] in which she defines civic epistemologies as "the institutionalized practices by which members of a given society test knowledge claims used as a basis for making collective choices."

The CIVIC EPISTEMOLOGIES project (FP7 Coordination and support action, 2014-2015 [1]) is about the participation of citizens in research on cultural heritage and humanities. The participation of Europe's citizens in scientific research development has just started to be exploited, while it represents a big potential for improving European competitiveness. ICT are powerful drivers of creativity and can open

[1] Corresponding author. Promoter S.r.l. via Boccioni 2, 56038 Peccioli, Italy; E-mail: fresa@promoter.it.

interesting opportunities to support the participation of citizens, but specific technical know-how is still generally lacking in the creative industries sectors, and, in addition, humanities scholarship is not yet taking full advantage of ICT to engage with wider audiences. New skills are needed to enable the cultural sector to grasp employment and commercial opportunities; and on the other side e-Infrastructures need to broad their deployment with new services and access policies targeted to this scope.

The research on DCHH can play an important role in the development of the European Research Area, and can be leader in the discovery of new directions of cross-disciplinary research, and opening this research to the civil society can provide new perspectives, which are demonstrating very interesting results, but are not yet fully part of the current research practices.

The proposed solution endorsed by CIVIC EPISTEMOLOGIES is to empower the existing e-Infrastructures with new services, targeted to the needs of specific research domains, in order to broaden the communities of users, including citizen scientists as an integrated part of the communities. It should be possible to tailor the new services to the requirements of each research community; at the same time, it should be necessary to identify common layers and standards that can be shared among different domains. This scalable and modular approach to the e-Infrastructures deployment will allow to serve better the research and to reduce costs of development.

The major outcome of the project is represented by its Roadmap, where pilot experiences, case studies, focus group activities and workshop results will be integrated into a unique product, designed with a very multidisciplinary approach.

2. Co-producing and Co-creating Knowledge

In many areas, Cultural Heritage (CH) may exist in a fragile or degraded condition, may be rendered vulnerable by economic development such as urbanisation, or may simply be lost through neglect. But, there is considerable interest among the general public in exploring, recording and cataloguing their own CH or that of their community or locality. CH digital content is massively increasing. This data may be held within a dedicated online archive or it may be collected and form a contribution to an aggregated database or archive.

At the same time, an increasing number of citizens are engaged in and with online discussion fora, and social networking platforms. But the outcomes are not always easy to predict and are sometimes negative and undesirable. The danger is that, without the establishment of a civic epistemology, separate communities develop as exclusive and even elitist and, as a consequence, the range and scope of a common set of civic values and understandings related to CH is thereby diminished.

The project aims to investigate how the phenomenon of citizen science can be encouraged and facilitated in a way that a shared or common CH discourse develops, knowledge is advanced and the exchange of ideas remains open and participatory. Cultural institutions and academies should welcome and embrace the opportunities implied in citizen science, as it offers occasions to be closer to citizens who are actually their publics. Next to this, a participatory and co-creative approach is positive and benefitting for cultural institutions as it adds to the knowledge-base of their collections, and opens up new ways for their collections to be used. But it does also create challenges for institutions, raising issues about curatorial authority over interpretation and on skill development for making citizens satisfactorily participate in research. First,

the citizen who is a culture consumer has to realise that he or she can become a producer, taking a more active role. This calls for a broad awareness campaign, where cultural institutions as well as platforms such as Europeana and specialized research infrastructures can make users aware of their shared responsibility to become caretakers of the cultural practices they engage in. Being conscious that one is a stakeholder in what happens is a first requirement in order to feel the need to intervene, to contribute, to have a voice. This is of utmost importance since in many instances of Cultural Heritage data part of the knowledge is not with the institutions but with the general public, in the stakeholder communities that have a relation to the subject matter (e.g. in the context of Orphan Works).

Participation however also means having the skills to do so: this requires partly a rediscovery of dormant, forgotten skills (such as painting, drawing, creative writing) and activation of consumer-oriented skills (such as using a smartphone) into more active forms of creativity (such as making street photography). On a third level, there are needs on specific training on the technical and digital skills involved to participate in online culture: understanding the Web language (HTML, CSS, XML), having notions of metadata and the Resource Description Framework, learning about digital formats and documents, and learning how to code small apps.

Finally, citizen authorship skills need to get the right visibility and recognition. By stimulating knowledge and use of Creative Commons licensing models and a deeper understanding of IPR issues it is possible to tap the hidden economic power of citizen cultural activities.

3. New Services, New Deployment, New Skills

The design of new services for the research, tailored on the needs of each research area, should be planned with a concrete approach, based on practical case studies and pilots with real users who should provide experimental proofs of the concepts defined in the theoretical sphere. It is also necessary to consider a shift of mentality in the cultural heritage sector, in order to accept the participation of non-professional curators in the development of new knowledge and for this purpose, naturally, design appropriate procedures and guidelines to be applied by the different subjects.

The new deployment should be carefully planned by indicating the steps that each stakeholder must take: decision-makers, policy-makers and programme owners to make available the necessary financial resources, research communities to identify the protocols of interaction with citizen scientists, citizens to associate and organise themselves into representative bodies, e-Infrastructure providers to plan for the future deployments. These steps and the concerned communities should be described in a sound and shared Roadmap.

New skills are needed in our changing society. Underinvestment in skills renewal and knowledge / technology transfer and the loss of traditional skills leads to the risk of innovation deficit and of a general lack of diversity and choice across design, production and markets, resulting in missed employment and commercial opportunities. A Roadmap which offers new understandings and ways of grasping opportunity can also lead to economic as well as social benefits. CIVIC EPISTEMOLOGIES is a project which shares its commitment to the values of openness, collaboration and wide participation.

The project's over-riding strategic objective is to support the development of a policy on the role e-Infrastructures can play in encouraging and facilitating the mediation process of citizen science in the area of DCHH, in order to bring about a closer alignment between the private and public spheres. It seeks to identify and deploy new services and protocols enabled by e-Infrastructures, which will in turn support Europe's citizens, its creative enterprises and its wider cultural industries to enter into productive technology-enabled dialogue with cultural heritage institutions and Humanities research. CIVIC EPISTEMOLOGIES is engendering dialogue, which is still relatively infrequent, between the different actors in the Cultural Heritage (CH) and DCHH sectors – research bodies (creativity, digital humanities, digital libraries), e-Infrastructure providers and citizens' associations, all of which seldom share their specialist knowledge outside their immediate groupings, whether professional or interest-based. Larger industries in the cultural sector, including the owners of industry archives as well as national public heritage bodies, will be supported to open up their innovation potential through informal dialogue with interested volunteer users and experts.

4. Understanding Stakeholders Requirements

One important aspect is the establishment by research institutions of clear protocols for citizens' engagement, and shared research goals where these are achievable, thus not only enhancing their roles within communities of interest at local, national and potentially global levels, but also greatly increasing the reach and impact of their research. Similarly, the role of heritage institutions should enter into a phase of change, looking for becoming both content providers and service providers, and to explore new audiences and markets for DCHH.

In order to understand better the requirements of different stakeholders from both DCHH and citizen science communities, the CIVIC EPISTEMOLOGIES project applied a mixed methods approach to understand the different demands and expectations of citizens and stakeholders from the citizen science domain (cultural institutions, academic institutions, activist organisations, infrastructure providers). The project explored the existing body of knowledge featuring general examples of citizen science work as well as examples of citizen science integrated in the DCHH context; one of the areas of particular interest was the link to the concepts of impact and value of digital cultural resources, an area of considerable interest as discussed in Hughes et al. [4].

The project also undertook specifically designed user studies which were aimed to investigate the similarities and differences in requirements of various stakeholders. The methodology adopted was mixed methods combining a web survey with expert consultations within the project consortium with focus groups aiming to capture the opinions of different stakeholders/users (policy makers with a focus group held in Malta; citizen activist organisations with a focus group held in Sweden, and citizen scholars with a focus group held in Spain). The web survey gathered 85 responses mostly from European countries. The initial analysis of these user studies is presented in Dobreva [2] and Dobreva, Azzopardi [3].

The CIVIC EPISTEMOLOGIES project supports the Horizon2020 strategy, in which the research on cultural heritage and on social sciences and humanities is embedded in cross-cutting initiatives. Critically, the CIVIC EPISTEMOLOGIES

Roadmap will offer practical support for improved social cohesion arising from the sharing across all DCHH sectors of common and individual European cultures.

5. The CIVIC EPISTEMOLOGIES Roadmap

The Roadmap will permit the implementation of an e-Infrastructure to enable creation, access, use and re-use of DCHH content, to provide learning resources, to provide communication services to multidisciplinary research teams located in different geographic places, and finally to enable citizens to participate in a range of research goals established at European level together with cultural institutions and universities.

The ultimate aim is to address the scientific processes in DCHH and to bring citizens, through their associations, into the process of planning research.

The Roadmap is organized in building blocks addressing the following questions:

- WHY – overall objectives for making a roadmap
- WHO – who to address: target groups and/or user groups, stakeholders in general, members of society
- WHERE – where to go (specific objectives and goals for the roadmap to be a basis for requirements like improving access, enhance quality of holdings/collections, social enclosure or…?)
- WHEN – when shall these specific objectives and goals be reached (time line for implementing the roadmap)
- WHAT – what to produce: a roadmap, but what are the basic drivers and the added values of citizen research and crowd sourcing as a method and what are benefits of using distributed e-infrastructure
- HOW – how shall the roadmap be structured (address each targets groups and/or user group or be structured on general level)

The CIVIC EPISTEMOLOGIES Roadmap should make it possible for each institution in the DCHH domain to define its own practical action plan with a realistic timeframe for the implementation of its stages. The purpose of proposing a short-term action plan (2014-2015) is to initiate the development of e-infrastructure services on a level that will be self-sustainable and continue to progress on its own. This further progress is defined in terms of two further proposed time spans:

- Medium-term (2017-2018), i.e. two years after the end of the Civic Epistemologies project), and
- Long-term (2019 and beyond) for the logical continuation of the work.

The main components of the Roadmap can be summed up as follows:

- Empowering existing e-Infrastructures with new services: targeted to the needs of specific research domains;
- Tailoring new services to the requirements of each research community;

- Improved interoperability: includes better integration of internal and external digital resources within the overall workflows for handling research data; in a way this is a set of measures to avoid building 'digital silos' within the organisation;
- Establishment of conditions for cross-sector integration: a key condition for maximising the efficiency of successful solutions, transferring knowledge and know-how; a scalable and modular approach to the e-Infrastructures deployment is needed that will allow serving research better and reduce costs of development.
- Governance models for infrastructure integration: a necessary condition for successful institutional participation in larger e-Infrastructure initiatives, and aggregation and re-use of digital resources.

For each area a set of prioritised actions will be suggested (this is currently under development) and the current version of the Roadmap cannot be considered as a final step. It has on the contrary to be considered as a living document that needs to be continuously maintained, updated and improved as time passes, technology changes, new requirements have to be taken into account, and so on.

Therefore, the CIVIC EPISTEMOLOGIES project has created a dedicated webspace where it is possible to download the last version of the Roadmap, but also where it is possible for everyone to provide feedback and comments, a kind of Forum dedicated to the use of e-infrastructure services and facilities for citizen science and crowd sourcing targeting the DCHH domain. The webspace is already online[2] and feedback is extremely welcome.

6. Conclusions

A ground breaking part of the concept that the CIVIC EPISTEMOLOGIES project is aiming to introduce, is the possibilities to customise the citizen science focus services provided by e-Infrastructure, i.e. tailoring the service portfolio and characteristics to the actual tasks and requirements. However, even if the e-Infrastructure resources, on a general level, seems to be allocated in ways that could support citizen science activities quite well, the general conclusion must be that the market for that kind of distributed services is still in its infancy.

E-infrastructure services for citizen science and crowd sourcing are normally structured around development of tools, but need also to involve policy instruments necessary to achieve efficient intervention in the DCHH sector. Another important issue is the level of maturity in the DCHH domain to handle distributed services for citizen science and/or crowd sourcing, increasing technical know-how.

E-Infrastructures can reach their maximum potential in serving the DCHH domain in practice only if the domain is prepared to exploit the opportunities offered by using e-Infrastructures. From contacts with different stakeholders it is seems that parts of the DCHH domain is not yet taking full advantage of technologies to engage with wider

[2] http://www.civic-epistemologies.eu/outcomes/roadmap/

audiences. Services for supporting citizen research and crowd sourcing have not only to be flexible, but also easy to adapt and utilise, and address several areas. This is a clear message from most stakeholder groups.

The adapted infrastructure as outlined in the CIVIC EPISTEMOLOGIES project, including active user participation into professionally maintained cultural heritage knowledge systems and skills learning materials and community knowledge exchange platforms, should transform the current, web 1.0 information-dissemination platforms in the cultural heritage sector into 21st knowledge hubs that foster creative reuse, interaction and innovation from consumers that become more and more pro-sumers and active determinators of the cultural practices they participate in, and its collective memory.

References

[1] CIVIC EPISTEMOLOGIES website: http://www.civic-epistemologies.eu/

[2] M. Dobreva, Collective Knowledge and Creativity: The Future of Citizen Science in the Humanities, In: KICSS 2014 (Post-)Proceedings , Springer AISS, 2015. ISSN 2194-5357 (in print).

[3] M. Dobreva, D. Azzopardi, Citizen Science in the Humanities: A Promise for Creativity, In: G.. Papadopoulos (ed.) Proceedings of the 9th International Conference on Knowledge, Information and Creativity Support Systems, Limassol, Cyprus, November 6-8, 2014, ISBN: 978-9963-700-84-4, pp. 446-451, 2014. To appear as well in Springer series.

[4] L. Hughes, P. Ell, M. Dobreva, G. Knight, Assessing and Measuring Impact of a Digital Collection in the Humanities, 2013. LLC: The Journal of Digital Scholarship in the Humanities.

[5] S. Jasanoff, Designs on Nature: Science and Democracy in Europe and the United States, 2007. Princeton University Press.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-15

Collaborating on Open Science: The Journey of the Biodiversity Heritage Library

Jane E. SMITH[a,1] and Constance A. RINALDO[b]
[a]*Natural History Museum, London, UK*
[b]*Ernst Mayr Library Museum of Comparative Zoology, Harvard University, Cambridge, MA, USA.*

Abstract. The Biodiversity Heritage Library, BHL[2], is an established and successful digital library, formed by a global consortium of natural history libraries, with engaged and enthusiastic users. The extensive partnerships, curated content, innovative tools and services, the ease of mining the data all combine to establish an open science resource that advances scientific progress through linking, use and reuse. The aim of BHL as stated on the web page is: "Inspiring discovery through free access to biodiversity knowledge. The Biodiversity Heritage Library works collaboratively to make biodiversity literature openly available to the world as part of a global biodiversity community. BHL also serves as the foundational literature component of the Encyclopedia of Life (EOL)". BHL and EOL are linked via taxonomic names and bibliographies. BHL is linked in a similar way to the Global Biodiversity Information Facility (GBIF) and thus has broad exposure to scientists across the globe as well as a global public.

Keywords. Data reuse, digital library, open access, outreach, global partnership, Taxonomic Intelligence

1. Introduction

The purpose of this paper is to describe how the Biodiversity Heritage Library (BHL) has become an established and successful digital library with engaged and enthusiastic users [1, 2]. The extensive partnerships, curated content, innovative tools and services, the ease of mining the data all combine to establish an open science resource that advances scientific progress through linking, use and reuse. The aim of BHL as stated on the web page is: "Inspiring discovery through free access to biodiversity knowledge. The Biodiversity Heritage Library works collaboratively to make biodiversity literature openly available to the world as part of a global biodiversity community. BHL also serves as the foundational literature component of the Encyclopedia of Life (EOL)". BHL and EOL are linked via taxonomic names and bibliographies. BHL is linked in a similar way to the Global Biodiversity Information Facility (GBIF) and thus has broad exposure to scientists across the globe as well as a global public. Both EOL and GBIF present bibliographies that lead back to the BHL portal.

[1] Corresponding Author. Natural History Museum, Cromwell Road, London, SW7 5BD, UK, E-mail: jane.smith@nhm.ac.uk
[2] http://www.biodiversitylibrary.org/

2. BHL Content

The Biodiversity heritage Library (BHL) currently provides scientists, scholars, citizen scientists and the public free and open access to a critical mass of over 46.2 million pages of digitised text and grey literature on biodiversity.

BHL partners have worked collaboratively since 2005 to make biodiversity literature openly available to the world as part of a global biodiversity community. The partners and contributors include libraries, natural history, botanical and research institutions that collectively hold a substantial part of the world's published literature and original material including published scientific papers and books, grey literature such as field notebooks and extraordinary illustrations all related to biodiversity. The core audience for BHL is scientists, particularly taxonomists who need access to literature spanning all publication years from pre 1700 to currently published material that is often related to active specimen collections. Until recently, this content was available to scientists only by travelling to libraries and museums with rare and unique collections and this slowed scientific study and collaboration. BHL partner libraries addressed the challenge of this "taxonomic impediment" by building a digital library of biodiversity literature, designed for taxonomists and systematists and utilising their language of taxonomic names for accessibility. Content additions can be suggested through the BHL portal's feedback tool.

3. Copyright

While the bulk of the BHL repository includes public domain literature, by negotiating with publishers and other rights holders BHL also provides relevant literature that is current and in-Copyright. Many Natural History and related Learned Societies have contributed content directly or provided permission for their titles to be digitised by BHL partners for inclusion in the repository. Doing so has exposed the content of these often much specialised publications to a much wider audience. To date, nearly 400 publishers have given permission for the inclusion of their titles.

Copyright impacts our ability to contribute and ingest content and is complicated by the variation in legislative frameworks within countries and regions across the globe. Therefore, the BHL partners have been careful to establish a Copyright and licence framework that reflects those regional differences while providing a common approach to which all contributors can agree. It is a framework that can be updated as the Copyright landscape changes. The framework supports the open access philosophy underpinning BHL, but also aims to respect the rights of contributors including an explicit take down policy. For example, an early piece of work undertaken by the BHL-Europe regional members, funded by the EU, was the development of a Copyright framework and guidance for EU partners intending to contribute content. BHL partners follow due diligence practices before digitising titles and written permissions are actively sought from rights holders for material still in Copyright. Out of Copyright material within BHL, in the Public Domain can be used, with appropriate attribution, or reused for multiple purposes. Reuse of the material in-Copyright is subject to Creative Commons Attribution Non-Commercial Share Alike Licence.

4. Supporting Research

By continuing to add relevant content, curating that content and enhancing access through the application of innovative tools, BHL supports scientists and researchers in other disciplines in their day-to-day work and collaboration with others. The availability and reusability of the scientific data accessible via BHL ranges from taxonomic study and training, biodiversity research, biodiversity conservation and maintenance of diverse ecosystems, animal and plant disease control through to audiences beyond core science constituencies, including historical and cultural research, exploration, and global commerce.

In addition to the core Science audiences, it became rapidly apparent that other users were drawn to the biodiversity treasure trove in the BHL. The heritage natural history volumes are rich in extraordinary illustrations that are coveted for study and re-use by art historians, citizen scientists, commercial artists and designers and the general public. Innovative projects and services result from interest in illustrations, field notebooks and journals formerly found only in museum and library archives. For example, the Art of Life Project has developed an algorithm to extract illustrations within volumes and provide enough metadata to make them discoverable [3]. These illustrations have been used to make greeting cards and wedding invitations, as well as for scientific study. Another example of how BHL data can be enhanced is the development of a game to aid in the crowdsourcing of transcription for handwritten field notes and complex seed and nursery catalogues.

The BHL technical team is working with collaborators to bring new perspectives to the interpretation of the material held in BHL. A current example, using crowdsourcing, is the joint project Science Gossip, a collaboration among BHL partner Missouri Botanical Gardens, *Zooniverse* and the UK's Arts and Humanities Research Council (AHRC) funded Constructing Scientific Communities: Citizen Science in the 19th and 21st Centuries. The project is an investigation into the making and communication of science in the Victorian period and today, and in the process provides metadata to enable discovery of illustrations from BHL.[3]

In addition, tools have been developed so that citizen scientists and the general public can consult virtual exhibitions of curated content on broad interest topics such as exploration, spices and women in science.[4] Other projects support data mining and the improvement of Optical Character Recognition (OCR).[5]

5. User Engagement and Feedback

User engagement and feedback is critical to the advancement and sustainability of BHL. As well as supporting communication and collaboration in a virtual organisation, BHL has a highly developed social media strategy which supports user access and engagement. BHL is well connected and respected within the blogosphere, Twitter, Pinterest, Flickr and Facebook communities. Campaigns such as "Monsters are Real" [6]attracted attention from the general public and the news media. Social media is also

[3] http://www.sciencegossip.org/#/classify
[4] http://latinonaturalhistory.biodiversityexhibition.com/
[5] http://miningbiodiversity.org/
[6] http://blog.biodiversitylibrary.org/2014/10/monsters-are-real.html

how BHL shares the outcomes from the feedback received through issue-tracking software on the website with a direct avenue for comments and content requests. All feedback receives an answer. User surveys and conference panels provide opportunities for targeted feedback and guidance in the development of services and content [4].

6. Sustainability

The sustainability of BHL depends on a mixed funding model, including direct support by BHL partners, and single and jointly awarded grants, and more recently a dues structure for members and fees for services to non-members to support basic administrative costs. BHL operates as a virtual organization and its strength is the participation and long-term commitment of the members through alignment of strategic goals of the constituent institutions and the contribution of collection and technical expertise from across the globe.

7. Conclusion

Working in partnership has enabled the participating organizations to bring together and link their collections in ways that provide a more complete research resource and negotiate with publishers and other rights holders to include material still in copyright. Collaboration on standards, best practice and infrastructure solutions has enabled higher quality images, metadata and support tools to be produced, long term digital storage solutions to be achieved and the sharing and cost reduction of scanning operations and best practices.

References

[1] N.E. Gwinn, C.A. Rinaldo, The Biodiversity Library: Sharing biodiversity with the world, *IFLA Journal* **35** (2009), 25-34.

[2] C.A Rinaldo, J.E. Smith, Moving through time and culture with the Biodiversity Heritage Library. In Innocenti P (Ed), Migrating Heritage: Experiences of Cultural Networks and Cultural Dialogue in Europe. Ashgate Publishing Group, Surrey, 2014.

[3] T. Rose-Sandler, N.E. Gwinn, C. A. Rinaldo, The Art of Life: Merging the Worlds of Art and Science. Libraries, Citizens, Societies: Confluence for Knowledge. In Session 149 - Art Libraries with Science and Technology Libraries. In IFLA WLIC 16-22 August 2014 Lyon France. 2014.

[4] G. Costantino, B. Crowley, R. Morin, E.Thomas, Heeding the Cal. *User* **40**(4) (2011), 146–157. ISSN (Online) 2190-541X, ISSN (Print) 2190-0752, DOI: 10.1515/mdr.2011.019.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-19

An Open Access E-journal: How to Find Out Readers' Preferences? The Case of the "Sciences Eaux & Territoires" Journal

Caroline MARTIN [a,1], Valérie PAGNEUX[b], Alain HENAUT[c]
[a] *Chief editor, Direction de la prospective, de la veille et de la valorisation de l'information scientifique et technique, Irstea, France*
[b] *Editor, DP2VIST, Irstea, France*
[c] *Member of Editorial Board "Sciences Eaux Territoires" Journal, Irstea, France*

Abstract. How may we best evaluate an open access e-journal that is not intended to be cited in rank "A" scientific journals? In this study, we took the example of a journal that connects research and professionals workers in the environmental sciences. We compared information from downloads with readership surveys. The main finding was that readers remember the best articles from a given issue and classify the issues based on this memory. A clear dichotomy can be observed: some readers are particularly interested in the management of biodiversity and pollution and others reject all that links to it.

Keywords. Open access, readers' preferences, transfer of knowledge, environmental science, public policies support

1. Introduction

The editor of a journal should understand the journal's readers and their expectations. It is particularly important to evaluate the success of the articles the journal publishes, if only to justify its budget [1].

A whole set of methods exists to do this [2]. The most widely used information [1, 3, 4, 5, 6, 7] is:

- the number of times the article is cited in the scientific literature (citation frequency);
- the number of times the item is downloaded (download frequency);
- the number of links on the Web page of the article (PageRank);
- readership surveys;
- expert opinions.

The number of citations on social networks has recently been added to this set of tools [8, 9]. Problematically, experience shows that it is still difficult to obtain a correct evaluation for interdisciplinary journals [10].

We are involved in the editorial team and board of interdisciplinary journal which is entitled "Sciences Eaux & Territoires"[2] (published by Irstea (National research

[1] Corresponding Author. Irstea, 1, rue Pierre Gilles de Gennes CS10030, 92167 Antony Cedex, France E-mail: caroline.martin@irstea.fr

institute of science and technology for environment and agriculture) [11] and we need to better understand the expectations of readers so as to produce the most appropriate content. This poses a big challenge because the mission of the journal is to transfer scientific and technical knowledge to professional stakeholders and practitioners at the local level involved in engineering projects concerning rural development and the environment. Articles published in the journal are not intended to be included in scientific journal ranking. So how may we reach concretely the target audience and how may we best understand readers' preferences in order to determine the content for each issue?

Given the nature of the journal, we can't base our analysis on the frequency of citations to assess the quality of articles and the importance of readership. However, we can use the number of downloads for the journal because it is freely accessible on the Web, and we have compared the elements of information from downloads to those obtained by two surveys of readers conducted at a one year interval.

The purpose of this paper is to report the first our investigation to find out the preferences of the readers and consequently to link these with the thematic issues published by the journal.

After the presentation of the materials and methods of our analysis, we introduce the first results of our case study and we explore some themes of discussion, before our concluding observations.

2. Materials and Methods

2.1. Data Available

The analysis is based on different sources of data.

2.1.1. The Scope of the Data Used in the Case Study

The website of the "Sciences Eaux & Territoires" was created in May 2010, with the first issue was published that month. The data gathered came from all publications on the web site up till 30 January 2014. This represents 17 thematic issues and 14 occasional papers ("articles Hors série") comprising to the 249 articles published on line [12]. The data analysed in this article are:

1. the uploading date of texts on the website (www.set-revue.fr);
2. the number of PDF of the articles downloaded;
3. the number of clicks on the icon "PRINT" of the web page of the article.

We analyse the number of downloads as the sum of items 2 and 3 above. These data came from the Google Analytics account of the website www.set-revue.fr.

2.1.2. The Data from the Qualitative Survey Concerning Reading Habits and Preferred Topics

We conducted two surveys of subscribers to the e-mail alert system[3] of the journal. In the first survey, in spring 2014, subscribers were asked about their reading habits and expectations. There were 145 responses, or around 10% of subscribers at that time. In

[2] www.set-revue.fr
[3] http://www.set-revue.fr/creer-une-alerte-mail

the second survey, in spring 2015, we asked subscribers for their preferences by marking from 1 to 3 (1 was the best rating) the issues published between May 2010 and January 2014. The special edition articles and the English version of "*Public policy and biodiversity / Scientific topics, political issues and local action*" were not included in this survey. The responses are still pending but we had received 35 at the time of writing. Both surveys included questions to identify the professional activity of subscribers.

2.2. Statistical Analysis

The statistical analysis was developed using the XLSTAT software [13].

2.2.1. Calculation of the Number of Downloads

The data came from the Google Analytics' account of the journal's website [14]. The statistics provided by Google Analytics are very rich. For the present analysis we used only the number of downloads. The number of downloads of an item is equal to the number of times the PDF of the article has been downloaded and the number of clicks on the icon "PRINT" of the web page of the article[4] .

2.2.2. Calculation of the Readers' Preferences for the Issues of the "Sciences Eaux Territoires" Journal

The website doesn't allow download of a complete issue in one go. The interested reader must download article by article.

There are several ways to define the success of a thematic issue. We chose the average downloads of articles comprising each issue. We obtained the average downloads of the three articles most downloaded articles, and the average downloads of the three least downloaded articles.

2.2.3. Evaluation of the Readers' Interest for a Special Type of Article

Analysing the number of downloads on its own doesn't reveal reader's preferences. In fact, this is because it is impossible to distinguish the article's intrinsic quality from the attractiveness of the thematic issue of the journal. This can be an important source of error because the attractiveness of a thematic issue has an effect on both the most downloaded articles and those who are the least downloaded.

One solution is to classify articles in their thematic issue based on the number of downloads and calculate the average of the ranks of similar articles in all issues. Then a Wilcoxon signed-rank test allows us to know if the gap between the average and the median rank is statistically significant.

The results are easier to analyse if the typology is accurate, in other words it covers only ten or twenty articles.

2.2.4. Analysis of the Survey Concerning the Readers

The survey carried out during Spring 2014, included the question "What do you get from reading of the 'Sciences Eaux & Territoires' journal?". The respondent could choose among five answers, with the possibility to retain more.

[4] see the website www.set-revue.fr

For this analysis, we gathered the respondents per categories and then calculated for each question the percentage of those who answered "yes" in each category.

3. Results

3.1. Information Provided by the Downloads

3.1.1. Analysis of Potential Statistical Biases Due to the Date of Online Posting

To detect any bias due to the time from which the numbers are posting online, we compared the statistics of visits to the website over six monthly intervals. Two-thirds of downloads are made during six weeks after publication. Some time after this initial interest has died down, the number of downloads of an article increases by an average of 1% per month.

Another element in favor of the representativeness of the data is that the main themes occur regularly since the creation of the "Sciences Eaux & Territoires" journal (e.g. *"Biodiversity management and pollution, state of the art technical, techno-economic approach a societal problem*").

3.1.2. Preferences Concerning the Thematic Issues

The readers' interest for a particular topic can be estimated from the number of downloads (Table 1).

Table 1. Classification of issues based on the number of downloads. The table is ranked according to the column "The top 3 best articles". *Nota Bene*: we brought together the fourteen articles in a special issue entitled "ARTICLES HORS SERIE" (special edition articles published one to one).

The top 3 downloaded articles	Average articles	The bottom 3 downloaded articles	Title of the thematic issues of the "Sciences Eaux & Territoires" journal
788,0	467,0	320,3	L'évaluation du risque toxique dans les milieux aquatiques (*Evaluation of the toxic risk of aquatic environments)*
708,0	410,2	238,3	ARTICLES HORS SERIE (*articles special edition / one to one*)
687,3	394,4	279,7	Restauration écologique /Nécessité de construire des indicateurs pour un suivi efficace (*Ecological restoration / Necessity of constructing indicators for effective monitoring*)
653,0	289,3	127,0	Politiques publiques et biodiversité / Problématiques scientifiques, enjeux politiques et actions locales (*Public policy and biodiversity / Scientific topics, political issues and local action*)
579,3	302,3	202,7	Les invasions biologiques en milieux aquatiques - Stratégies d'action et perspectives (*Biological invasions in aquatic environments - Strategies and outlook*)
442,7	244,4	133,0	Les éco-indicateurs au service de l'agriculture durable - Méthodes d'évaluation et mises en œuvre (*Eco-indicators at the service of sustainable agriculture - Evaluation methods and implementations*)
393,0	270,3	177,0	Des recherches pour une gestion durable des forêts (*Research for sustainable forest management*)

The top 3 downloaded articles	Average articles	The bottom 3 downloaded articles	Title of the thematic issues of the "Sciences Eaux & Territoires" journal
386,7	150,2	158,3	Méthanisation agricole - Éléments de réflexion pour une intégration territoriale réussie (*Agricultural methanisation - Thinking Elements for successful regional integration*)
371,0	192,3	131,3	120 m3 - Le consommateur d'eau en question (120 m3 - The water consumption in question)
337,0	161,6	119,7	L'irrigation en France / État des lieux, enjeux et perspectives (*Irrigation in France / overview, issues and prospects)*
306,0	133,7	84,0	Recherche et Ingénierie au service des acteurs de l'assainissement / Avancées et perspectives (*Research and Engineering of urban water treatment serving local stakeholders/ Progress and Prospects*)
298,0	185,7	123,0	Estimer et réduire la consommation d'énergie à l'échelle de l'exploitation agricole (*Estimate and reduction of energy consumption across the farm*)
240,0	183,5	144,5	OPTIBAN, des solutions innovantes pour le traitement des bananiers (*OPTIBAN, innovative solutions for the treatment of banana trees*)
219,0	130,8	68,7	Risques naturels en montagne - Nouveaux outils, nouvelles connaissances et nouveaux savoir-faire (*Risks in the mountains - New tools, new knowledge and new skills*)
182,0	142,6	124,0	Le bassin de l'Orgeval : 50 ans de recherche au service des acteurs de terrain (*The basin of Orgeval river: 50 years of research for local stakeholders)*
146,7	77,1	53,7	Géosynthétiques - Un monde durable (*Geosynthetics - A sustainable world*)
52,0	26,2	12,7	Public policy and biodiversity / Scientific topics, political issues and local action

The most immediate statistic is the average number of downloads of articles components an issue (Table 1 second column from the left). However, this figure is not without bias. The number of time that an article is downloaded depends "a priori" on two factors. First, the success of the thematic issue (give the reader wants to browse the issue), second the quality of the article (give the reader wants to save). Can we get a ranking of thematic issues that is not biased by the quality of any particular article? One solution is to compare the rankings obtained by taking, i) three articles that have had the most success, in terms of downloads (Table 1 first column from the left), ii) three articles that have had the least success (Table 1 third column from the left). We consider that the rankings are reliable and trust if they are the same in both cases.

The three classifications are very consistent. Spearman's rank correlation coefficient gives the following results:

- "Average" vs "The top 3 downloaded articles" = 0.90;
- "Average" vs "The bottom 3 downloaded articles" = 0.84;
- "The top 3 downloaded articles" vs "The bottom 3 downloaded articles" = 0.79.

We find on the top of the ranking, those thematic issues which dealt with biodiversity management and pollution.

1. *"Evaluation of the toxic risk of aquatic environments";*
2. *Special edition articles in preference order: 1) "Mitigation measures for biodiversity: Improving environmental issues and governance?" 2) "Agricultural practices and soil fertility in France"; 3) "Assess the environmental performance of a highway system for biodiversity";*
3. *"Ecological restoration / Necessity of constructing indicators for effective monitoring";*
4. *"Public policy and biodiversity / Scientific topics, political issues and local action";*
5. *"Biological invasions in aquatic environments - outlook and strategies".*

The lowest ranking were issues deal with the technical and economic approach to a social topic or the technical overview:

6. *"Estimate and reduction of energy consumption across the farm";*
7. *"OPTIBAN, innovative solutions for the treatment of banana";*
8. *"Risks in the mountains - New tools, new knowledge and new skills";*
9. *"The basin of Orgeval river: 50 years of research for local stakeholders";*
10. *"Geosynthetics - A sustainable world".*

The "*Public policy and biodiversity / Scientific topics, political issues and local actions*" issue may be considered an outlier. This is the English version issue of "*Politiques publiques et biodiversité / Problématiques scientifiques, enjeux politiques et actions locales"* and was posted online five months after the French version. We presume the most interested readers have already downloaded the French version and that they did not see the interest to do it again for the English version.

3.1.3. Preferences Concerning Articles

Synthesis papers with real pedagogical qualities were favoured by readers. These are the most downloaded articles in each issue (31 articles, p-value <0.01%). Papers dealing with the examples in another European country, the presentation of tools, and specific case studies that are downloaded as often as the others (13 articles, p-value = 10%).

3.2. Information Provided by the Survey from the Subscribers of the "Sciences Eaux & Territoires" Journal

An e-mail survey was conducted in spring 2015 among subscribers to the e-mail alert system[5] of the journal. The survey is still open and relatively few answers have been received to date. However, the preliminary analysis has already provided some interesting information.

3.2.1. Concordance of Opinions between the Answers to the Survey and the Downloads

The subscribers to the journal were invited to indicate their preferences by marking from 1 to 3 the issues (1 = the most appreciated content; 2 = moderately appreciated content; 3 = the less appreciated content). Special edition articles ("Articles Hors-série") and the English version of "*Public policy and biodiversity / Scientific topics,*

[5] http://www.set-revue.fr/creer-une-alerte-mail

political issues and local actions" weren't included in the questionnaire. Therefore, the analysis below concerns only the fifteen thematic issues.

The answers are representative because their preference is consistent with Table 1. Spearman's rank correlation coefficient gives the following results:

- "Survey" vs "The top 3 downloaded articles" = 0.68;
- "Survey" vs "Average" =0.63;
- "Survey" vs "The top 3 downloaded articles" = 0.43.

The classification issues based on the answers to the survey is very similar to that which was obtained by taking the top three articles of each issue. Figure 1 below shows that the consistency is almost perfect for the first seven. The only exception is the issue "*Evaluation of the toxic risk of aquatic environments*" which was seventh in the survey results but was first for downloads. We note that this is the first issue published online in May 2010 when the "Sciences Eaux & Territoires" e-journal was launched, almost five years before sending the survey. However, there is no match even for the top of ranking tail issues.

Figure 1 in the right frame deals with the ranking from the worst three articles of each issue (data from Table 1). We see that in this case the match isn't strong, even for the top ranking issues.

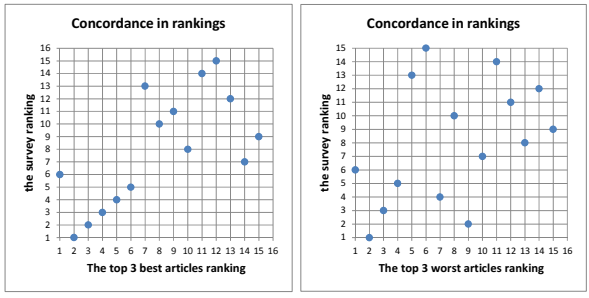


Figure 1. In the left frame, concordance of opinions on the ranking of the best three articles in each issue and ranking of the answers to the survey. In the right: concordance of opinions on the ranking by the three worst articles of each issue and the ranking of answers to the survey.

3.2.2. Different Categories of the Issues

Agglomerative Hierarchical Clustering (AHC) of all ratings (see above section 3.2.1) shows that we can distinguish three types statistically different thematic issues (see below Figure 2 left):

1. Management of biodiversity and pollution themes: this group includes the issues of *Evaluation of the toxic risk of aquatic environments; Ecological restoration / Necessity of constructing indicators for effective monitoring; "Public policy and biodiversity / Scientific topics, political issues and local actions"; Biological invasions in aquatic environments - Strategies and outlook; Eco-indicators at the service of sustainable agriculture - Evaluation methods and implementations; Research for sustainable forest management.*
2. Technical and economic approach of a societal topic, with the issues: *Agricultural methanisation; "120 m3, the water consumption in question"; "Irrigation in France"; "Research and Engineering of urban water treatment"; "Estimate and reduce energy consumption across the farm".*

3. A technical state-of-the-art review, with the issues like *"Optiban innovative solutions for the treatment of banana trees"; "Risks in the mountains"; "The basin of Orgeval river: 50 years of research for local stakeholders"; "Geosynthetics, a sustainable world"*.

The correspondence between theme and popularity is clear. In the first category stated above, we find the top six issues based on downloads of the three best articles from each issue. The second category contains the next five issues and the last four issues belong to the third category.

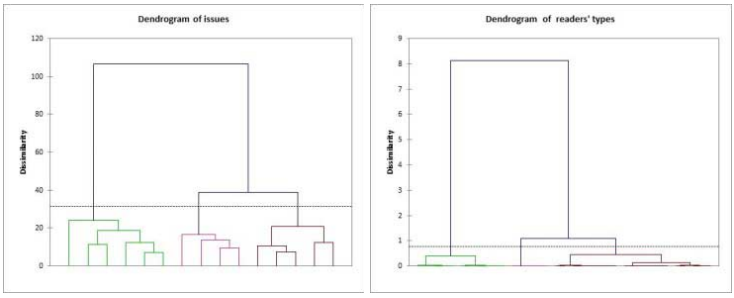


Figure 2. In the left dendrogram: automatic ranking issues based on the answers to the survey. In the right one: automatic ranking of readers' interests according to the answers to the surveys. The dotted line indicates the statistical significance level. The sections above the dotted line are significantly different at the 5% threshold.

3.2.3. Types of Readers

A second Agglomerative Hierarchical Clustering (AHC) was used to gather the answers to the surveys based on the scores assigned to each type of issue. The analysis shows that we obtain three classes with statistical significance (see Figure 2 right). The following typology can be observed:

The left group on the dendrogram is composed of readers who, in general, do not give good marks. Also they give systematically a bad rating to the issues in the category "Management of biodiversity and pollution" as defined above. On the other hand, some of them appreciate the issues belonging to the category "technical and economic approach to a societal problem" or "technical state-of-the-art review".

The middle group comprises readers who are interested by all issues but with a preference for those in the category "Management of biodiversity and pollution" which they consistently give top marks.

The right group on the dendrogram covers those readers rated highly the issues in the category "Management of biodiversity and pollution". Half of them are also interested in the issues of the type "Technical and economic approach to a societal topic". However, few of these readers focus on the issues of the category "technical state-of-the-art review".

3.2.4. Focus of Interest of the Readers with the Spring 2014 Survey

The survey sent in spring 2014, asked readers what they expected to read in the "Sciences Eaux &Territoires" journal (Table 2).

Table 2 Questions about the interests of readers and shortcut used in the others tables below.

What did you expect from reading "Sciences Eaux & Territoires" journal ?	Focus of interest	% Of those who checked box
Concrete elements to improve your professional practice	Professional practice	55%
Information to fuel a decision making process	Decision making	27%
General information on a topic	General information	55%
Information on technical and technological advances in your field	Technical advances	53%
Synthesis and analysis of the elements to better understand the issues of a topic that interests you	Understanding the issues	59%

On average, each reader has answered "yes" to two or three questions (average = 2.5 answers). All interests have nearly the same success, except the "Decision making". It is possible to refine the analysis by gathering the answers according to the occupational organisations to which belong to the readers. We calculate for each question the percentage of those who answered "yes" within each occupational structure. Table 3 shows the difference between this percentage and the percentage calculated on all answers described on Table 2.

Table 3 Differences between the percentage of "Yes" for a given occupational organisation and the percentage in all responses to the survey (see Table 2). The category "Not determined" corresponds to people who have not indicated their professional organisation.

Occupational organisation	Professional practice	Decision making	General information	Technical advances	Understanding the issues
Private companies	-0	-7	-5	-8	-14
Local authorities	1	5	-3	-5	-27
Water agency	45	-27	-5	-3	-59
Education	17	-10	-11	19	24
consultancy	15	-4	11	3	19
Regional services of central administration	-5	15	28	-3	24
research	-15	-7	-15	-3	11
Central administration	-35	53	5	7	1
Not determined	-17	27	-17	-14	-13
Irstea	-15	-27	25	-13	-19
Abroad	-11	-27	0	14	-3

In conclusion, the readers of the journal are very heterogeneous. Three types of occupational organisations of readers appear with a high proportion of "yes" to more than half of the answers: "Education", "Regional services of central administration", and "Consultancy". They have in common an interest in environmental challenges and issues. According to these results, "Private companies" and "local authorities" have the same concerns. They are in the average for all except for the "understanding of the issues", a topic that does interest them not very much.

On the contrary, some organisations give strong priority to a specific type of information: these categories include "Water Agency", "Not determined", "Irstea", and "abroad".

Some others organisations attach great importance to the improvement of professional practices (“Water Agencies”, “Education”, “consultancy”).

The contribution of the journal for the decision making process is a high expectation for three organisations: “Central Administration”, “Regional services of central administration”, “Not determined”. It may be noted here that the group “Not determined” is very homogeneous and has a single interest: the decision making process. We suppose probably that the people who compose this category did not indicate their professional organisation.

4. Discussion

4.1. How to Know the Preferences of the Readers: Interesting Results and Their Limit

We are involved in editing an open access e-journal “Science Eaux & Territories”. We want to know better our readers better and their expectations, and in particular we want to evaluate the relevance of the editorial line of the journal. We used two sources of information: the number of downloads and two surveys of subscribers to the e-mail alert of the journal. Several studies have shown that the number of downloads is reliable information that overlaps largely the information obtained by the frequency of citations [1, 3, 4, 5]. Information is almost exhaustive; in return the question is superficial: how many readers are interested in an article to the point of downloading?

The surveys help us to find out more details about a small number of readers: those who answer to the questionnaires. But, if the sample is representative, it gives information that we could not have obtained otherwise, for example about the influence of professional structures to which readers belong on their reading preferences (Table 3).

There is a bias more difficult to eliminate. The opinions of the respondents depend largely on their concerns at the time t [1]. Our work highlights another source of bias: the selective memory of those who answer to surveys. As shown in Figure 1, it is as if readers don’t remember the best articles of an issue and they classify the issues based on their single memory. Consequently readers are unable to classify the issues when no article catches their attention. The rankings given by the readers don’t replace the statistics on the number of downloads. However, they still remain useful because they help define the profiles of readers, and we can answer to this question: how many distinct and statistically significant categories of readers exist if we gather them according to the concordance in ranking? This work is possible even with a small sample if a very small number of types of readers is noticed.

Furthermore, we have shown a clear dichotomy between those readers who are particularly interested in the management of biodiversity and pollution and others who reject everything that relates to the subject (Figure 2). It is virtually impossible to obtain this information by analysing only the number of downloads, because it assumes that people who download the articles, have stable e-mail address and we are able to track their visits during several years on the journal’s website.

This work uses a very small part of the information provided by Google Analytics. We plan to refine and enrich the analysis by using for example the number of clicks on the pages of the journal's website.

4.2. How to Measure the Appropriation of the Scientific Knowledge for Practitioners in the Specific Context of the Open Access "Sciences Eaux & Territoires" Journal?

Irstea, the National Research Institute of Science and Technology for Environment and Agriculture, is a public institute under joint supervision of the Ministry of Research and the Ministry of Agriculture [11]. Irstea has built a multidisciplinary and systemic approach to three domains water, environmental technologies and land, which today form the basis of its strength and originality. The Institute is establishing itself as the spearhead of the environment in support of public policies.

To answer to this challenge, in May 2010, Irstea launched a project called "*Sciences Eaux & Territoires*", an e-journal to complete the mission of supporting public policies. The journal is completely subsidized by the institute with a full-time equivalent (1 FTE) dedicated. The choice of open access was determined by the mission of facilitating access to scientific and technical advances and knowledge from research to fuel action and decision process for practitioners at the local level concerning the following topics: water resources management (uses and risks), freshwater systems, quality and pollution, land management, ecotechnologies, rural management. The appropriation of scientific results by readers of the journal has become the core of its mission. "Sciences Eaux & Territoires" journal wants to be a link between practitioners and scientists. It represents a collaborative space dedicated to the co-construction of knowledge. The journal is a tool to promote wider the dialogue between science and society which is not often well-to-do and natural. Furthermore, in a constrained budget period, the social responsibility of science is under scrutiny and the efficiency of the tools dedicated to this demand, are assessed constantly.

The first answers of the analysis introduced to this paper, have determined a work plan to improve the correspondence between the inputs as suggested interests (from the surveys) of the readers and the outputs as thematic issues of the journal published on the website. Three approaches are drawn and are currently the object of a work in progress:

- the choice of the themes of issues to publish, this supposes refining the editorial line of the e-journal concerning the both topical and format levels (e.g. authors' guidelines);
- the role of the editorial board (for example, we have worked on modification of the assessment grid of papers);
- the ergonomics of the website www.set-revue.fr by improving the navigation on the website.

5. Conclusion

The general context of this study is the social accountability of science, especially, the assessment of public policies support leaded by Irstea.

The preferences of the readers of the "Sciences Eaux & Territoires" journal are very significant because they determine in large part the subsidies of the institute to keep going the journal. Some information was obtained with the analysis of the number of downloads from the website and the two surveys concerning the readership of the journal.

At this moment, we have not exploited the whole of available data from the Google analytics account of the website and the entire survey. Subsequent, fuller results and analysis are expected to help us understand better the preferences of our readership and consequently build close interactions between researchers, policy makers, decision makers and different stakeholders and practitioners at national and local levels through our open access journal.

References

[1] A. Serenko, M. Dohan, Comparing the expert survey and citation impact journal ranking methods: Example from the field of Artificial Intelligence, *Journal of Informetrics* **5** (2011), 629–648.
[2] J. Bar-Ilan, Informetrics at the beginning of the 21st century - A review, *Journal of Informetrics* **2** (2008), 1–52.
[3] J. Bollen, H. Van de Sompel, JA. Smith, R. Luce, Toward alternative metrics of journal impact: A comparison of download and citation data, *Information Processing & Management* **41** (2005), 1419–1440.
[4] D. O'Leary, On the relationship between citations and appearances on "top 25" download lists in the International Journal of Accounting Information Systems, *International Journal of Accounting Information Systems* **9** (2008), 61–75.
[5] DE. O'Leary, The relationship between citations and number of downloads in Decision Support Systems, *Decision Support Systems* **45** (2008), 972–980.
[6] L. Xue-li, F. Hong-ling, W. Mei-ying, Correlation between Download and Citation and Download-citation Deviation Phenomenon for Some Papers in Chinese Medical Journals, *Serials Review* **37** (2011) 157–161.
[7] D. Fiala, L. Šubelj, Žitnik S, M. Bajec, Do PageRank-based author rankings outperform simple citation counts?, *Journal of Informetrics* **9** (2015), 334–348.
[8] T. Kortelainen, M. Katvala, "Everything is plentiful - Except attention". Attention data of scientific journals on social web tools, *Journal of Informetrics* **6** (2012), 661–668.
[9] S. Haustein, I. Peters, CR. Sugimoto, M. Thelwall, V. Larivière, Tweeting biomedicine: an analysis of tweets and citations in the biomedical literature, *Journal of the Association for Information Science and Technology* **65** (2014), 656–669.
[10] I. Rafols, L. Leydesdorff, A. O'Hare, P. Nightingale, A. Stirling, How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management, *Research Policy* **41** (2012), 1262–1282.
[11] Irstea, Institut national de recherché en sciences et technologies pour l'environnement et l'agriculture, available at http://www.irstea.fr/ English version http://www.irstea.fr/en/accueil.
[12] Sciences Eaux & Territoires, la revue d'Irstea, available at http://www.set-revue.fr.
[13] XLSTAT, available at http://www.xlstat.com/en/.
[14] Google Analytics, available at http://www.google.com/analytics/.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-31

Sustainable Software as a Building Block for Open Science

Timo BORST[1]
ZBW – Leibniz Information Centre for Economics

Abstract. In the context of Open Science, almost every 'traditional' research activity and output has been affected and transformed by means of web based technology. New forms of research output have emerged, among them software as an important means and method for data driven science. But how can software be treated as scholarly work, and how can it be integrated into a digital research infrastructure? The paper depicts software development related to Open Science and points out some future directions for software to become part of a sustainable research infrastructure.

Keywords. Research software, research infrastructures, Open Science

1. Software as a Factor for Open Science

In recent years, the digital transformation of the system of scholarly communication has often been recognized as one of the basic outcomes and challenges in modern information society. Almost any stage of the scientific process has been affected, be it workflows like the generation, distribution and sharing of scientific results, or their reviewing and communicating. Central, if not any of the scientific processes happen by means of digital environments including tools and applications, while core processes like dissemination and communication of scientific results preferably take place in web based environments.

At the same time, the process of software development including producing, distributing, sharing and modifying program code has undergone a similar change, which at first sight does not look too surprising. Where else than in computer science and industry would you expect first a change or shift in the use of digital environments? Have the first emails not been sent by computer experts, the first user groups communicating via mailing lists not been established by those experts, and the sharing, modifying and enhancing of e.g. LINUX been conducted by them? Where in general the answer may look obvious, in the scientific realm it becomes a bit more tricky: The systematic distribution and sharing of software as both a basis for and outcome of a digital system of scholarly communication has neither been recognized so far, nor explicitly been put on the agenda like comparable workflows in the scientific publishing process. Rather, it may be still conceived as in an early stage of incubation.

However, there are indicators that software development for research purposes has already adapted itself to requirements or symptoms of Open Science. One striking

[1] ZBW – National Library of Economics, Düsternbrooker Weg 120, 24105 Kiel, Germany; E-mail: t.borst@zbw.eu.

observation may be that in the context of research data it has been demanded to publish not only the data itself, but also the program code to generate or manipulate it [1]. Publishing under the conditions of Open Access, and publishing Open Source Software converge from a both conceptual and operational point of view, especially when it comes to the reproducibility and replicability of scientific results by means of program code. Moreover, the following trends may indicate the importance of sustainable software as both an enabler and a result of Open Science [2]:

Under the general label of 'openness', we can observe a convergence of the three movements Open Science, Open Data and Open Source. Apart from the more recent endeavors in managing and provisioning research data and algorithms to calculate them, namely the third topic has its roots in a quite early practice independent from the other two, but with a certain impact on them. A significant difference still can be seen in the openness of Open Source Software towards commercial purposes, so these purposes have become an important driver.

With the emergence of social networks, graph based approaches towards the modelling of relations and collaborations between (social) entities have become very prominent and successful. But long before that, concepts and tools for graph-based version management have been introduced into distributed software development with an explicit history of software releases. Version management tools like Apache Subversion (SVN) or Concurrent Version System (CVS) arouse before or parallel to the World Wide Web (WWW), while web-based platforms like Sourceforge [3] and particularly GitHub [4] have fostered the distribution, tracking, monitoring and reuse of software in terms of awareness and collaboration. Regarding the variety of uses of GitHub as a popular distribution platform for code, data and text in sciences [5], one may seriously take these adaptations into account as steps towards an operational Open Science. For instance, it has been suggested to adapt traditional bibliographical metrics and to measure the impact of software distributors by means of some kind of 'page rank algorithm' to calculate the most cited (=forked) git repositories [6].

In analogy to 'Citizen Science' as a synonym for an open and collaborative science activity, one may regard the character of a 'Citizen Developer' contributing code in an open environment for like-minded persons. The basic idea is that in a role as a 'citizen developer', one may still contribute to Open Software projects without an institutional or professional background. In contrast and in a perhaps more sophisticated enhancement of the role as a 'citizen scientist' – where participants are mostly committed to mass data contribution –, the 'citizen developer' acts in a both collaborative and individualistic environment, leaving his or her digital footprints in software repositories being part of a global software environment for science and research.

Now, what to infer from these observations? If software is becoming more and more constitutive and public as crucial contribution to Open Science, it will be essential to provide an environment for managing this work similar to 'traditional' research output.

2. Upcoming Challenges: Research Software vs. Infrastructure Software from a Stakeholder's Perspective

From the point of view of infrastructure providers, software developed by scientists may be called 'immature' in the sense, that its origin is primarily individual, local and

temporal. Being absorbed by a specific research question, a researcher normally will not care very much about the engineering aspects of his or her software, nor will he or she have the time to 'harden' the code for potential reuse. Even for professional, full-time committed software developers this requirement is still bothersome and definitely not first-order activity. On the other hand, research software built by scientists is a constitutive part of a future environment for Open Science, in the sense that it will be needed for later reuse, validation and reproduction of scientific results in connection with other scientific outcomes, e.g. research data. Hence, the crucial question can be put as how to integrate (individual) research software with something like 'sustainable software infrastructure', so it can finally become part of the latter?

3. Future Directions and Recommendations for Sustainable Research Software

In the following, we recommend some principles and steps to be taken, which are obviously adopted from existing workflows in order to align the development of research software with other research and publishing activities. The principles are formulated as requirements towards the authors resp. the publishers of research software from the point of view of research infrastructure providers. They may be in line with requirements of other stakeholders like research funders, but this has to be negotiated yet.

3.1. Software Code as Open Source

As already stated, Open Science relates to Open Source Software in an intuitive, but not yet operative way. For the purpose of the transition from idiosyncratic code to reusable and adaptable, quality assured packages as part of a common software infrastructure, it is essential to fully provide transparent source code plus the software environment (libraries, runtime environment, virtual machines, container etc.) to run that code. Consider, e.g. the calculation of a regression analysis on the basis of a self-developed algorithm despite the fact that there are already reliable packages e.g. at CRAN [7]: in case of inconsistencies, a reviewer or adopter of the results would not be able to distinguish between wrong data or wrong code resp. algorithms. To be 'open' is not just an attitude or style adopted from another context, but a *conditio sine qua non* for reproducing, reviewing and revising research results based on software.

Publishing software as Open Source implies some legal, organizational and technical aspects which must be cleared in advance, best on an institutional level. In contrast to other research output like publications or data, a reuse of software for commercial purposes might be more likely – but there are a couple of proven Open Source licenses suitable for regulating these concerns.

3.2. Publishing and Sharing of Code in Conjunction with Data and Additional Material

It is good practice to publish all material related to a scientific work if not simultaneously, but in one location – at least with references to their physical location –, so the user gets a quick overview on and access to the total scholarly work and its components. A Digital Object Identifier (DOI) may be useful to resolve to the overall work ('jump page'), but can also be attached to its constituents (publication, data, code, blog posts, slides, tweets …).

For publishing and sharing of code, GitHub has become a trendy and widely used platform for 'social coding' [8]. Adopting common vocabulary and comprehensive workflow steps partly derived from other code versioning systems ('commit', 'fork', 'branch'), plus the typical 'social media' characteristics of global visibility, tracking and awareness have lifted GitHub to the most popular and widely used platform for code management and sharing, forcing e.g. Google to shut down their corresponding service and users to migrate their code repositories to GitHub [9]. However, several objections have been made towards GitHub as a service for global code management: (a) As an internet service, the platform is exposed to hacking, (b) the provider – a start-up company from San Francisco – is heavily dependent from external venture capital and potential market interests, and (c) the cloning of a GitHub repository is no real substitute for a decentralized, 'officially' redundant backup infrastructure [10]. Hence, GitHub may be rather regarded as a model for publishing and sharing research software, not as the only or primary host for publicly financed research output. It serves as a platform for disseminating code, but should be backed up by local code git repositories.

3.3. Making Code Citable

Although this aspect may be conceived as an integral part of publishing code, it deserves special attention and handling: So far, the citing of software code similar to traditional research output has not been put explicitly on the agenda of stakeholders like publishers or research funders, nor does it yet belong to the 'impact story' of researchers. On the other hand, citing is already supported by platforms like Figshare [11] or Zenodo [12], both of them integrating GitHub repositories as the original platform for code publishing. In Zenodo, each version of software can be referenced by its own DOI (cf. Figure 1), hence associated with the research supplement generated with that version (data, publication). This is especially important in the case of software packages which include non-standard algorithms, and where the results from the calculation are dependent on the distributor or even the version of software.

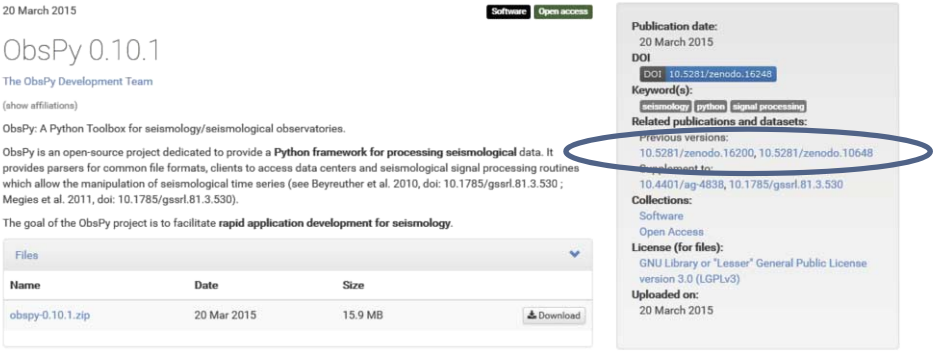


Figure 1. Screenshot from a Zenodo record.

Solutions like the one from Zenodo may be a good way to publish software by means of more 'official' platforms, while at the same time pointing to continuous development and deployment on platforms like GitHub.

3.4. Research Representation, Analysis and Evaluation

To become more visible as genuine research output, references to software code should become part of the scholarly record implemented by researcher identifier systems like ORCID [13], VIAF [14] or – on a more national level – DAI [15] and GND [16]. For instance, in ORCID the creation of an XML instance for describing a piece of published code should follow the XML scheme for describing a publication by using e.g. the APA style [17]:

```
<orcid-work>
...
<work-citation>
       <work-citation-type>formatted-apa</work-citation-type>
       <citation>
       Gregory Jefferis, James Manton, & Ben Sutcliffe. (2015).
       nat: nat 1.6.5. Zenodo.
       http://doi.org/10.5281/zenodo.17558
       </citation>
...
</orcid-work>
```

4. Conclusions

Considering software as explicit research output is still in its very beginning, and so is the integration into existing research infrastructures. Software development is not a primary research activity, but becoming more crucial especially in data driven sciences. From the point of view of a research infrastructure, the formal basics for identifying, citing and integrating software as research output are already there – what we now do need is more investigation on the curation, maintenance and preservation of this software, so it can become integral part of future research.

References

[1] http://openeconomics.net/principles/ (accessed 2 June 2015).
[2] M. D. Hanwell et al., Sustainable Software Ecosystems for Open Science: 15 Years of Practice and Experience at Kitware, *arXiv:1309.2966v1* (2013), DOI: 10.6084/m9.figshare.790756
[3] http://sourceforge.net/ (accessed 2 June 2015).
[4] https://github.com/ (accessed 2 June 2015).
[5] P. Krill, GitHub rolls out the red carpet for scientists (2014), http://www.javaworld.com/article/2157321/open-source-tools/github-rolls-out-the-red-carpet-for-scientists.html (accessed 2 June 2015).
[6] F. Thung et al., Network Structure of Social Coding in GitHub, 17th *European Conference on Software Maintenance and Reengineering (CSMR) 2013* (2013), 323-326. DOI: 10.1109/CSMR.2013.41
[7] http://cran.r-project.org/ (accessed 2 June 2015).
[8] L. Dabbish et al., Social coding in GitHub: transparency and collaboration in an open software repository, *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (2012), 1277-1286. DOI: 10.1145/2145204.2145396

[9] http://google-opensource.blogspot.jp/2015/03/farewell-to-google-code.html (accessed 2 June 2015).
[10] http://blog.printf.net/articles/2015/05/29/announcing-gittorrent-a-decentralized-github/ (accessed 2 June 2015).
[11] http://figshare.com/ (accessed 2 June 2015).
[12] https://zenodo.org/ (accessed 2 June 2015).
[13] http://orcid.org/ (accessed 2 June 2015).
[14] http://viaf.org/ (accessed 2 June 2015).
[15] https://www.surf.nl/en/themes/research/research-information/digital-author-identifier-dai/digital-author-identifier-dai.html (accessed 2 June 2015).
[16] http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html (accessed 2 June 2015).
[17] http://blog.apastyle.org/apastyle/2015/01/how-to-cite-software-in-apa-style.html (accessed 2 June 2015).

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-37

Using EPUB 3 and the Open Web Platform for Enhanced Presentation and Machine-Understandable Metadata for Digital Comics

Pieter HEYVAERT [a,1], Tom DE NIES [a], Joachim VAN HERWEGEN [a], Miel VANDER SANDE [a], Ruben VERBORGH [a], Wesley DE NEVE [a,b], Erik MANNENS [a] and Rik VAN DE WALLE [a]
[a] *Multimedia Lab, Ghent University – iMinds*
[b] *IVY Lab, Korea Advanced Institute of Science and Technology (KAIST)*

Abstract. Various methods are needed to extract information from current (digital) comics. Furthermore, the use of different (proprietary) formats by comic distribution platforms causes an overhead for authors. To overcome these issues, we propose a solution that makes use of the EPUB 3 specification, additionally leveraging the Open Web Platform to support animations, reading assistance, audio and multiple languages in a single format, by using our JavaScript library comicreader.js. We also provide administrative and descriptive metadata in the same format by introducing a new ontology: Dicera. Our solution is complementary to the current extraction methods, on the one hand because they can help with metadata creation, and on the other hand because the machine-understandable metadata alleviates their use. While the reading system support for our solution is currently limited, it can offer all features needed by current comic distribution platforms. When comparing comics generated by our solution to EPUB 3 textbooks, we observed an increase in file size, mainly due to the use of images. In future work, our solution can be further improved by extending the presentation features, investigating different types of comics, studying the use of new EPUB 3 extensions, and by incorporating it in digital book authoring environments.

Keywords. Dicera, Digital comics, EPUB 3, Enhanced presentation, Linked machine-understandable metadata, Open Web Platform

1. Introduction

The market for digital comics is growing, consisting of both people familiar with print comics and newcomers. The revenue numbers of the past years, as obtained for the North American market, testify to this statement [1]: (1) a growth between $640 million in 2011 and $680 million in 2012 for print comics and (2) a growth between $25 million in 2011 and $70 million in 2012 for digital comics. After a decline in 2010 for print comics, the revenue for these comics is increasing again, together with the revenue for digital comics, as shown in Figure 1. This figure also proves that the growth of the digital market does not make the print market shrink.

[1]Corresponding author. E-mail: pheyvaer.heyvaert@ugent.be

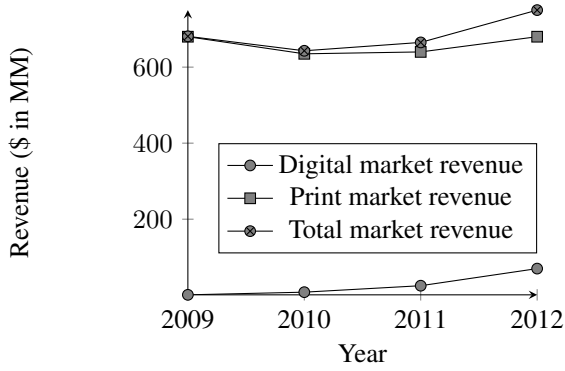


Figure 1. After a decline in 2010 for print comics, the revenue for print comics is increasing again. The revenue for digital comics is increasing as well (data gathered from [1]). The growth of the market for digital comics does not make the market for print comics shrink.

Authors have a number of distribution platforms at their disposal, including comiXology [2] and the iBooks Store [3]. However, these platforms use different (proprietary) formats to distribute comics, thus storing the same information in different packages. In addition, several platforms assume that all publications are textbooks, which can result in bad on-screen formatting. Furthermore, several platforms also require authors to add additional information (e.g., panel annotations), so that comics enable every feature of the dedicated reader application (e.g., for tablets and smartphones). All these factors contribute to an increased *production overhead*, while authors primarily want to focus on telling stories.

Tools [4, 5] exist to convert a comic to different formats used by the different distribution platforms. However, the use of these tools might still require intervention by the authors, because the *feature sets of the different formats differ*. This is *cumbersome for authors*. Other problems are related to the *extraction of information* from comics (e.g., separate panels, characters, text, and so on), both in the domain of presentation and metadata. While this information is present in a number of comics, it is not available in a machine-understandable way, which can improve discoverability.

We argue that the aforementioned problems are caused by the digital formats in which the comic books are stored. As such, we propose an Open Web Platform-based solution that has the following merits:

- it **circumvents the conversion** of a comic to the different formats required by different distribution platforms;
- it **alleviates the use of extraction methods**, which are currently powered by image processing, by providing the descriptive information, i.e. information about characters, pages, panels, and so on, through machine-understandable metadata, facilitated through our newly proposed ontology Dicera;
- it **works towards truly digital comics**, i.e. comics that allow for the use of animations, audio, reading assistance, multiple languages and machine-understandable metadata, facilitated through our newly proposed JavaScript library *comicreader.js*.

The remainder of this paper is organized as follows. In Section 2, we review related work. Next, in Section 3, we provide an outline of the technical requirements identified and the solution proposed. In Sections 4 and 5, we discuss the presentation and metadata elements of the proposed solution in more detail, respectively. We subsequently discuss and pay attention to an evaluation of the proposed solution in Section 6. In Section 7, we give an overview of future work. Finally, we provide concluding remarks in Section 8.

2. Related Work

Arai and Tolle [6] propose a method that allows for the automatic extraction of frames and their content, which includes text balloons and their text, resulting in better accuracy and processing time, compared to other methods.

Hoashi et al. [7] suggest to create thumbnails for each comic to more easily identify the content of a single episode of a comic series. Thumbnails are created by extracting the different frames and selecting the ones with the highest scores (using the frame's features and a linear regression model). Their results state that the proposed method enables users to search comic episodes faster.

The main problem Yamada et al. [8] handle is how to display high-resolution comic pages on low-resolution cellular phones. The authors developed a system that consists of frame detection, text extraction and layout analysis.

Print versions restrict the access for a number of people: the visually impaired, the motor-impaired and the users of mobile devices. To enhance the navigation through a comic, Ponsard and Fries [9] propose a solution that performs the following three steps: sequential ordering of the page-wide files, segmentation of each page into panels and sequential ordering of the panels in the right reading order.

Arai and Tolle [10] propose a new online method for automatically extracting text from text balloons in digital comics. This method has an accuracy of 100% for frame extraction from flat comics and an accuracy of 100% for balloon detection. Text extraction achieves an accuracy of 93.75%. The authors suggest that the new method is useful for the automatic translation of Japanese comics into international comics.

Morozumi et al. [11] summarize the basic requirements for a metadata framework for manga (the Japanese version of comics): different levels of description for manga elements, a clear difference between the intellectual entity and the publication of a manga, and the identification and description of the elements that make up a manga. This metadata framework is implemented by Mihara et al. [12].

With eDBtheque, Guérin et al. [13] want to create the first comics database with a ground truth for descriptive metadata. This database is built by using a visual segmentation protocol with guidelines regarding text lines, balloons and panels, and by annotating these text lines, balloons and panels.

Given the above research efforts, we can identify a number of trends. First, we can conclude that the need for information about the content of a comic steers current research efforts towards the development of extraction methods that make use of image processing techniques. The use of these computationally heavy techniques can be circumvented. Second, we can conclude that the research efforts regarding comics metadata are limited to basic administrative metadata (e.g., title, author and genre) and descriptive metadata (e.g., the characters appearing in a comic and the number of panels on a page). These descriptive metadata are stored outside the comic. However, the inclusion of de-

scriptive metadata in a comic can aid in the presentation of the comic content to different types of users (e.g., impaired users and users of mobile devices).

3. Requirements and Proposed Solution

The proposed solution consists of two major parts: presentation and metadata. The requirements for the first part, as identified through an analysis of related work (see Section 2) and the features of the (comic) applications of Marvel [14] and comiXology, are (1) support for *animations*, (2) support for different types of *devices*, (3) support for *reading assistance*, (4) support for *audio* and (5) support for *multiple languages*.

The requirements for the second part, as identified through on an analysis of related work, are (1) support for *extended administrative metadata* and (2) support for *extended descriptive metadata*.

To meet the aforementioned requirements, the EPUB 3 format [15] was chosen as the foundation of the proposed solution. Our motivation is as follows: (1) EPUB 3 is widely used [16], (2) EPUB 3 targets digital publishing, hence also digital comics, (3) the Open Web Platform (that is, HTML5, CSS3 and JavaScript) allows for a large range of possibilities, and (4) Ghaem Sigarchian et al. [17] concluded that EPUB 3 supports the most desirable attributes of an enhanced publication (that is, a publication enriched with multimedia and interactivity features). Note that the requirement to support audio is immediately fulfilled by EPUB 3, given that EPUB 3 makes use of HTML5.

In what follows, we discuss the presentation elements (Section 4) and the metadata elements (Section 5) of the proposed solution in more detail[2].

4. Presentation

In Section 4.1, we detail the use of layering. In Sections 4.2 to 4.4, we discuss the support for animations, reading assistance, different types of devices and multiple languages. Next, we explain the use of scripting languages in Section 4.5. In Sections 4.6 and 4.7, we discuss the consequences of scripting and layer unavailability, respectively. Finally, in Section 4.8, we look at the use of additional EPUB 3 specifications.

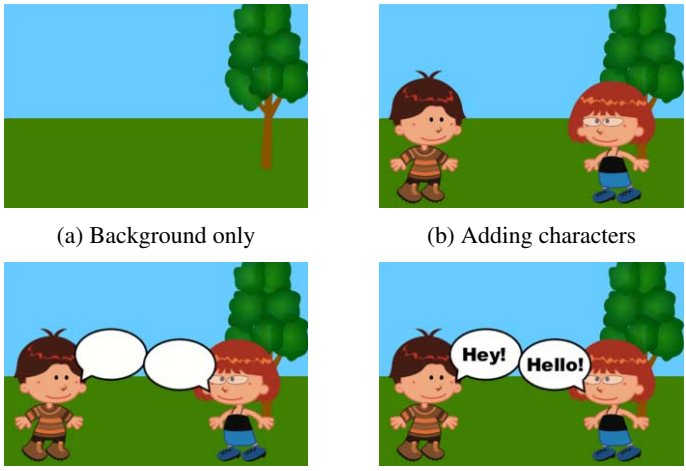
4.1. Layering

Before tackling the requirements, we define the way the graphical information is stored, which will enable the solution to fulfill the requirements. A page is not stored as one image. Instead, each panel is stored separately in a `<div>` element. This is taken even further: by using layering, every panel is stored as a group of images, and thus not as a single image. Every image is called a layer, hence the name layering. A layer can be created for the background, for each character and for each text element (e.g., a speech balloon, a caption or an effect), by representing each layer by a `<div>` element. This makes it possible to do manipulations on separate parts of the panel itself.

4.2. Support for Animations

To support animations in our solution, we rely on the Open Web Platform. JavaScript and the CSS property `display` are used to show the different layers. Figure 2 illustrates the concept of animations through an example. First, the background of a panel is shown, followed by the characters, speech balloons and the text.

[2] A screen cast of the presentation part of the proposed solution can be found at http://users.ugent.be/~pheyvaer/digital_comic.mp4.



(a) Background only (b) Adding characters

(c) Adding speech balloons (d) Adding text

Figure 2. A panel is built by adding layers ((a)-(d)) incrementally.

4.3. Support for Reading Assistance & Different Device Types

Reading assistance refers to zooming in on the panel that the user is currently reading and navigating to a new panel once the whole panel has been read (determined by user input). To accomplish this, a jQuery [18] plugin called Zoomooz [19] is used. This JavaScript library makes it possible to zoom in onto elements of Web pages, while taking into consideration the screen size of the device. In our case, these elements are the panels of a comic. An important remark is that our approach towards facilitating reading assistance and support for different types of devices is *independent of the reader application*, because it is achieved through scripting in the EPUB 3 file itself, and can, hence, be *different for every comic*.

4.4. Support for Multiple Languages

To allow for multiple languages in the same digital comic and to allow for switching between the languages available, the different translations of a piece of text are stored in different 'sublayers' (`<span>` elements instead of `<div>` elements are used, because span elements are used for inline elements and the different translations can be viewed as the inline elements of the text `<div>`). This allows showing the (text) sublayers of the currently selected language and hiding the other languages. Hiding and unhiding the different sublayers is again accomplished by making use of JavaScript and CSS.

In Listing 1, an example can be found of the use of two languages, Dutch and English. One layer represents all the text, which means both languages. Inside this layer, another `<div>` element represents the text object, needed for the addition of metadata. This `<div>` element has two `<span>` elements (the sublayers previously mentioned), one for each language. The `xml:lang` attribute is used to denote the Dutch and English translation. For example, if the user wants to read the comic in English, the CSS property `display` of all the `<span>` elements, with the value of the `xml:lang` attribute set to `nl`, are set to `none`, which makes those elements hidden, leaving only the English translation visible. In addition, using CSS, the font of the text, together with the size and the position on the panel, can be set.

Listing 1: Example of multi-language support

```
<div id='panel_text'>
    <div id='panel_text_obj'>
        <span id='panel_text_nl' xml:lang='nl'>Hoe het begon ...</span>
        <span id='panel_text_en' xml:lang='en'>How it began ...</span>
    </div>
</div>
```

4.5. Use of Scripting Languages

Most of the JavaScript code is not inherent to a single comic. Therefore, it is useful to create a JavaScript library to bundle all the code that is reusable. First, we list the requirements of such a library. Second, we present our implementation of such a library.

First, the library should contain functions that can be used to provide animations. Besides the standard 'appear' animation, it should be possible to add other (basic) animations, such as a slide, a dissolve, and so on. Furthermore, the functions need to be designed in such a way that they can be reused if a developer wishes to create his or her own animations. Second, the library should have support for reading assistance. To that end, we rely on the Zoomooz library (and jQuery). Third, the library should be able to handle different languages. In particular, the library should have support for switching between languages by manipulating the different layers. Given our proof-of-concept EPUB 3 file, we grouped the common functionality of its XHTML pages in an extensible JavaScript library called *comicreader.js*[3].

4.6. Unavailability of Scripting

According to the EPUB 3 [15] specification, it is not allowed to rely on scripting to deliver content to users. Scripting can only be used to enhance the user experience. Our fallback method, in case scripting is not possible or (temporarily) disabled on a device (or in an application), is to display the whole page at once: layering is not used anymore, no support for animations, no support for reading assistance, no support for audio and no support for multiple languages. For every panel, one image is included as a fallback, and this fallback image will be displayed when it is not possible to execute scripts.

4.7. Unavailability of Layering

It is possible that layers are not available, for instance when print comics are converted to our solution. This has the following consequences. Adding animations to a single panel will not be possible, given that the different layers to work with are not available. Support for multiple languages is not possible in the way it is defined in this paper, because the text is hard coded on the image. However, it is possible to define a `<div>` element that covers the original text with the translation needed. Audio can still be added, together with reading assistance and support for different types of devices.

4.8. EPUB Region-based Navigation and Multiple-Rendition Publications

The EPUB Region-based Navigation specification [20] allows adding region-based navigation through a visual rendition of a publication based on Regions of Interest. The major problem with this specification is the lack of layering, hence the inability to deal with animations. Although the specification incorporates metadata about regions of interests, these metadata are limited to the name (in string representation) of one character.

[3] comicreader.js and our proof-of-concept EPUB 3 file are available at https://github.com/mmlab/comicreader.js.

The EPUB Multiple-Rendition Publications specification [21] defines the creation and rendering of publications consisting of multiple renditions. In what follows, we discuss the effect of this specification on the solution proposed. Considering the use of multiple languages, the difference between our approach and the approach proposed in the specification is that every language should point to a different rendition. Our solution uses one file for each page that includes all the languages. Splitting those languages in different files would force a developer to duplicate all the other code inside the files, with the exception of the actual text of the comic. Hence, redundancy is created and the file size of the EPUB publication will increase (unnecessarily). Studying different approaches is proposed, such as using renditions with multiple languages using one file that includes all the languages, together with the use of JavaScript. Each rendition uses a different script that displays the correct languages and hides the other languages (cf. the approach previously discussed in Section 4.4). The downside is that problems may arise when scripting is disabled (see Section 4.6). This is not an issue when using the approach described in the specification.

5. Metadata

In Section 5.1, we discuss the different types of metadata. In Section 5.2, we discuss how the metadata is modeled, paying attention to the support for extended administrative and descriptive metadata in Sections 5.2.1 and 5.2.2, respectively. Finally, we look at the consequences of layer unavailability in Section 5.3.

5.1. Different Types of Metadata

All metadata can be divided into two groups based on the location where they are stored: *local metadata* or *remote metadata*. With remote metadata, we denote all the metadata that are not necessarily stored inside the EPUB file itself. A publisher can store all the metadata of a certain comic character on a server (e.g., biography, latest comics, and so on). We leave the decision about where to store what metadata to the user.

Local Metadata The metadata stored locally enables the user to gain a number of advantages of linked data, such as discoverability, shareability and re-usability.

Remote Metadata The motivation for also storing metadata remotely is twofold. First, storing all metadata locally would increase the size of the EPUB file. Second, storing all metadata locally would create redundancy, making it more difficult to keep the metadata up to date.

5.2. Data Modeling

To structure the metadata, we developed the ontology Dicera[4] (DIgital Comic book ERA vocabulary). Classes and properties of Dublin Core Metadata Initiative Metadata Terms [22] and the DBpedia Ontology [23] that could be reused are not redefined in Dicera (e.g., characters and locations). The use of the schemas provided by schema.org [24] and Friend of a Friend [25] have been considered, however, they are lacking the necessary concepts.

To add metadata to a digital comic, we use RDFa [26]. This tool, which was recently added to the EPUB 3 specification, allows adding metadata to Web pages.

[4]Dicera can be found at `http://semweb.mmlab.be/ns/dicera`.

5.2.1. Support for Extended Administrative Metadata

As part of the local metadata, we extended the current metadata (e.g., title, author, genre, and so on) available in EPUB 3 with (more detailed) information regarding story arcs, issues, genres and content ratings, through Dicera. Indeed, as an example, every comic can be associated with a certain story arc. This happens through an issue entity, which also includes the issue number. For each genre (a single comic can have more than one genre), the percentage it matches to the story can be denoted, given that genres can be quantified, i.e. a story can be only 25% drama (not necessarily 100%).

5.2.2. Support for Extended Descriptive Metadata

We added the following descriptive metadata, through Dicera: covers, pages, panels and text elements. The pages are connected to the contained panels. These panels are connected to their characters, locations, objects and text elements. Every character, location and object can also be connected to the cover if they appear on it.

5.3. Layer Unavailability

Layer unavailability puts limitations on the presentation part of the solution. The limitations on the metadata part are less severe. Each panel can still be annotated with corresponding characters, however, all characters will be included in the same layer. This is also valid for the locations, objects and text elements. Even if the text is hard coded on the image, text objects can still be used. These objects will solely be used for metadata purposes and not for both presentation and metadata as is normally the case. The most important limitation is the following: if the metadata are used to determine all the panels where, e.g., a character is present, it is only possible to get the complete panel/image. When layers are present, it is possible to retrieve images containing only the requested character, without the background, the other characters, the objects and text elements.

6. Discussion & Preliminary Evaluation

In this section, we discuss the need for extraction methods, the change in file size, the support for different publication platforms and the support for different reading systems.

6.1. Extraction Methods

The need for extraction methods, as mentioned in Section 1, can become superfluous if all the required information is already available in the solution. However, extraction methods and metadata information can be used together. Initial metadata can be added by using the information acquired from the extraction methods, and can subsequently be enhanced by manually correcting or adding information.

6.2. EPUB File Size

Comparing the file size of our comic to the file sizes of textbooks in EBUP format shows that our file size is more than 60 times larger than the average file size of a textbook. The sole reason for this is the presence of pixel-based image files in a comic. A consequence is that applications can no longer assume that an EPUB file comes with a rather small file size.

6.3. Distribution Platforms

From the work conducted here, we conclude that the features offered by different distribution platforms are replicable using the proposed solution.

The use of the proposed solution by distribution platforms eases the publication of content for authors to those platforms, because a single, standardized and actively de-

veloped format is used. Note that publishers might desire their own dedicated application to offer a special and unique experience to their readers. However, this is not valid. The presentation functionality does not depend on the reading system. It is embedded in the EPUB file itself, allowing it to be read on every system supporting the EPUB 3 specification. Based on these benefits, the companies behind these platforms might consider the solution, and eventually replace their current format by an open format such as ours. It also allows them to discontinue the development of their custom application for reading comics or to work towards a reading system supporting the (complete) EPUB 3 specification.

6.4. Reading System Support

Readium [27] (with the use of scripting), developed by IDPF [28], and the Kobo Glo [29], a dedicated e-reader (using the fallback panels) allow reading comics that make use of the proposed format. Unfortunately, iBooks [3], an iOS application, and Calibre [30], a desktop application, are not able to display these comics in a correct way, both with and without scripting. By testing the solution on these four different types of reading systems (that is, a web app, a dedicated e-reader, an iOS application and a desktop application), we have a basic global view of the reading system support for our solution.

7. Future Work

Future work is possible, both in the direction of the enhancement of the proposed solution and the creation of applications, reading systems and frameworks that work with the solution, so to enhance its adoption. Enhancing the solution can be done by investigating the support for different types of comics, by extending and optimizing the presentation features, by studying the use of the new (draft) specifications of EPUB 3, and by studying how the solution can be used to recommend new comics to readers. Furthermore, work needs to be conducted to create a remote server to work with the solution and a framework that links comics to other multimedia content such as movies and games, and to develop a reading system that fully supports the solution. While the advantages of enhanced digital comics are clear, authoring them in an efficient manner remains a major challenge, which prohibits their wide adoption. However, as digital book authoring solutions evolve, opportunities arise to adapt them towards digital comics. For example, part of our future work is to adapt the authoring environment described in [31] to create and manage enhanced digital comics using our proposed approach.

8. Conclusion

In this paper, we tackled problems regarding the extraction of information from (digital) comics by making use of EPUB 3 and adding metadata by making use of RDFa and the ontology Dicera. By offering presentation features using the Open Web Platform the proposed solution, with comicreader.js, circumvents the use of different (proprietary) formats by different publication platforms (causing overhead in the production process), the cumbersomeness of the use of format conversion tools for the different distribution platforms, and the differences in the feature sets of these formats. The increased file size compared to textbooks, and the currently limited EPUB 3 support by reading systems, prevent large-scale deployment of the proposed solution for the time being. However, since a standardized format is used that is being actively developed, reading system support will only increase in the future.

9. Acknowledgements

The research activities were funded by iMinds, Ghent University, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research Flanders (FWO Flanders), and the European Union.

References

[1] Digital comics nearly tripled in 2012. http://www.icv2.com/articles/news/26202.html, Retrieved 2014/05/08.
[2] Comics by comiXology. https://www.comixology.com/, Retrieved 2014/11/12.
[3] iBooks. http://www.apple.com/ibooks/, Retrieved 2014/11/12.
[4] EPUB to MOBI. http://www.epub2mobi.com/, Retrieved 2014/12/05.
[5] EPUB Converter. http://www.epubconverter.com/, Retrieved 2014/12/05.
[6] Kohei Arai and Herman Tolle. Automatic e-comic content adaptation. *International Journal of Ubiquitous Computing*, 1(1):1–11, 2010.
[7] Keiichiro Hoashi, Chihiro Ono, Daisuke Ishii, and Hiroshi Watanabe. Automatic preview generation of comic episodes for digitized comic search. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1489–1492. ACM, 2011.
[8] Masashi Yamada, Rahmat Budiarto, and Shinya Miyazaki. Comic image decomposition for reading comics on cellular phones. *IEICE transactions on information and systems*, 87(6):1370–1376, 2004.
[9] Christophe Ponsard and Vincent Fries. An accessible viewer for digital comic books. In *Computers Helping People with Special Needs*, pages 569–577. Springer, 2008.
[10] Kohei Arai and Herman Tolle. Method for real time text extraction of digital manga comic. *International Journal of Image Processing (IJIP)*, 4(6):669–676, 2011.
[11] Ayako Morozumi, Satomi Nomura, Mitsuharu Nagamori, and Shigeo Sugimoto. Metadata framework for manga: a multi-paradigm metadata description framework for digital comics. In *International Conference on Dublin Core and Metadata Applications*, page 61, 2009.
[12] Tetsuya Mihara, Mitsuharu Nagamori, and Shigeo Sugimoto. A metadata-centric approach to a production and browsing platform of manga. In *The Outreach of Digital Libraries: A Globalized Resource Network*, pages 87–96. Springer, 2012.
[13] Clément Guérin, Christophe Rigaud, Antoine Mercier, Farid Ammar-Boudjelal, Karell Bertet, Alain Bouju, Jean-Christophe Burie, Georges Louis, Jean-Marc Ogier, and Arnaud Revel. ebdtheque: a representative database of comics. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013.
[14] Marvel. Downloads and Extras. http://marvel.com/mobile, Retrieved 2014/05/08.
[15] IDPF. EPUB 3 Overview, February 2014. http://www.idpf.org/epub/301/spec/epub-overview.html, Retrieved 2014/05/06.
[16] Matt Garrish. *What is EPUB 3*. O'Reilly Media, September 2011.
[17] Hajar Ghaem Sigarchian, Ben De Meester, Tom De Nies, Ruben Verborgh, Wesley De Neve, Erik Mannens, and Rik Van de Walle. EPUB 3 for integrated and customizable representation of a scientific publication and its associated resources. In *Proceedings of the 4th Workshop on Linked Science*, October 2014. URL http://ceur-ws.org/Vol-1282/lisc2014_submission_3.pdf.
[18] jQuery. http://jquery.com/, Retrieved 2014/11/12.
[19] Zoomooz.js. http://jaukia.github.io/zoomooz/, Retrieved 2014/11/12.
[20] IDPF. EPUB Region-Based Navigation, . http://www.idpf.org/epub/renditions/region-nav/epub-region-nav.html, Retrieved 2014/05/08.
[21] IDPF. EPUB Multiple-Rendition Publications, . http://www.idpf.org/epub/renditions/multiple/epub-multiple-renditions.html, Retrieved 2014/05/08.
[22] DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/, Retrieved 2014/11/28.
[23] The DBpedia Ontology. http://wiki.dbpedia.org/Ontology, Retrieved 2014/11/28.
[24] Schema.org. omicvine.com/api/, Retrieved 2014/12/05.
[25] The FOAF Project. http://www.foaf-project.org/, Retrieved 2014/12/05.
[26] W3C. RDFa Core 1.1 - second edition. http://www.w3.org/TR/2013/REC-rdfa-core-20130822/, Retrieved 2014/05/08.
[27] Readium. http://readium.org/, Retrieved 2014/11/12.
[28] IDPF. http://www.idpf.org/, Retrieved 2014/11/12.
[29] Kobo Glo. http://www.kobo.com/koboglo, Retrieved 2014/11/12.
[30] Calibre. http://calibre-ebook.com/, Retrieved 2014/11/12.
[31] Ben De Meester, Tom De Nies, Hajar Ghaem Sigarchian, Miel Vander Sande, Jelle Van Campen, Bram Van Impe, Wesley De Neve, Erik Mannens, and Rik Van de Walle. A digital-first authoring environment for enriched e-books using epub 3. *Information Services and Use*, 34(3):259–268, 2014.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-47

From Print to Ebooks: A Hybrid Publishing Toolkit for the Arts

Digital Publishing Toolkit Collective [a], M. RIPHAGEN [b 1], M. RASCH [b], F. CRAMER [b]

[a] *Amsterdam University of Applied Sciences, Amsterdam, the Netherlands*
[b] *Institute of Network Cultures and PublishingLab, Amsterdam, the Netherlands*

Abstract. This article is an excerpt of the outcome of a two-year research and development project on hybrid publishing. The DPT Collective [1] developed a *Toolkit* which consists of the publication *From Print to Ebooks: A Hybrid Publishing Toolkit for the Arts* [2] and an online software kit [3] – which is meant for publishers who publish visually oriented books in mostly smaller print runs. This Toolkit focuses particularly (but not exclusively) on EPUB3 as an electronic publication format, and on Markdown [4] as a word processing format. The recommendations stem from our practical experience in collaborating on electronic publication projects with four Dutch art, design and research publishers: BISPublishers, Valiz, nai010 uitgevers and the Institute of Network Cultures .

Keywords. EPUB, hybrid publishing, markdown, pandoc

1. Introduction

'You must change your life' – borrowing from the philosopher Peter Sloterdijk [5], this could be the summary of our message to art-oriented and design-oriented publishers, writers, editors and designers who are currently transitioning from traditional book making to electronic publishing or – more typically – hybrid publishing of print and electronic formats. Hybrid publishing will sooner or later confront them with the need to re-think traditional publication formats, editorial and production workflows, as well as distribution opportunities.

Having said that, there are exceptions. Workflow changes can be also minor for publishers who already do all their editorial work in highly structured digital document formats such as XML (acronym for Extensible Markup Language, a version of HTML based on XML) or databases (an organized structured collection of information: common examples are address books, library catalogues and retail inventories); but this is typically only the case in scientific and technology-oriented publishing. Changes may also be minor for larger publishers who can afford outsourcing. Generating an electronic (digital) publication in parallel to a printed publication is then simply a matter of paying an external service provider, such as a document engineering company or a media design agency, to turn a Microsoft Word or Adobe InDesign file into an ebook. This process

[1] Corresponding author. E-mail: margreet@networkcultures.org.

can be quick if the book is visually simple – such as a novel or a textbook with few illustrations – and economically worthwhile if many ebooks will be sold.

We propose an alternative to the process mentioned above. Neither a complex internal IT infrastructure, nor costly outsourcing will be viable solutions for art and design oriented publishers. Unfortunately, there is no 'magic' software button that will turn a print book design into an electronic publication just like that. Since the two media are so different, each with its own specific editorial and visual design needs, such a button is unlikely to materialize in the future either. Hybrid publishing will ultimately require changes in the way the editorial work is done. The good news is that such change is possible. This research includes instructions on how to deal with the many issues that arise when making the transition from traditional to hybrid or electronic publishing.

We are not claiming that all ebooks will follow, or indeed should follow this path. We are simply laying out one of the many directions ebook creators can now take with their publication, by using simple and inexpensive tools, and without having to buy into the industry's gloss promises of multimedia and interactivity.

2. State of the Art

In this section we consider a more in-depth reading of industry promises vs. reality, the basics of a text, and three levels of electronic publishing.

2.1. Industry Promises vs. Reality

For art and design publishers, the challenge of 'going electronic' with their publications is greater than that faced by other fields of publishing, for a number of reasons:

- Visually oriented publications are still more difficult to realize technically in the electronic medium, particularly when designing for a multitude of different reading devices and ebook platforms.
- Small publishers are under a great deal of pressure to keep project costs low, often due to smaller budgets.
- In order to make the investment in an electronic publication durable, electronic publications must be sustainable: they should not require constant investment in technical maintenance and version updates.

There is a stark contrast between the fanciful promises of the computer industry and the often harsh reality of the new digital medium. On one hand, publishers, editors, designers and artists tend to overestimate the interactivity and multimedia possibilities of electronic publishing. These extra possibilities do exist, but in most cases bring with them higher development costs and remain specific to one particular technical platform.

On the other hand, publishers tend to underestimate how even technically simple and seemingly trivial types of electronic publications can in fact lead to a re-thinking of established publishing practices and formats. When traditional publishing formats are replaced by electronic formats, there is a real possibility for transformation. Once the book becomes electronic or hybrid, the permanence, immutability and stability typical of physical books is likely to mutate into dynamic, modular, and participative forms. Such publications can greatly benefit from the networked environment in which ebooks exist.

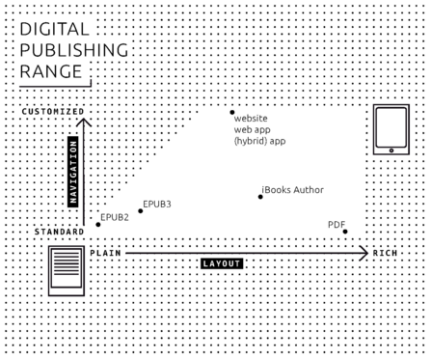


Figure 1. Digital publishing range.

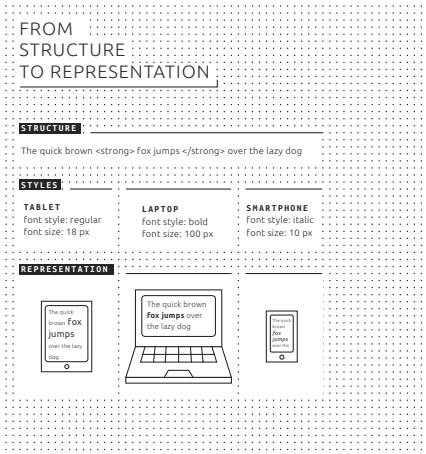


Figure 2. From structure to presentation.

Various types of electronic publications may be subject to different kinds of change. Still, the change will always be radical. For instance, an exhibition catalog can be split up into interrelated micro-monographs, which readers can download and read as individual ebooks. An ebook can be assembled from a variety of sources selected by individual readers, as is currently the case with Wikipedia, where visitors can compile their own collection of Wikipedia articles and export this compilation to an EPUB or PDF.

Users can choose from a multitude of hardware e-reader devices and software e-reader applications. The possible combinations of software and hardware are complex and virtually unlimited (see Fig. 1). This requires a certain adaptation level of not only the publisher, but also the designer. The possibilities for change can go beyond the rethinking of existing publishing formats, eventually even redefining what a book actually is.

2.2. The Basics of a Text

Throughout the historical development of writing, characters beyond the basic alphabet have played an increasingly important role, starting with blank spaces between words, then punctuation marks, and later markup [5] for formatting. In the electronic processing of texts, hierarchical ordering of words into sentences, sentences into paragraphs and so on, as well as additional reading aids such as bold or italic text, is made possible by using specific formatting codes. This process is called *markup* and the codes are called markup elements. All these markup elements require stable definitions and clear relationships if they are to be of any use. In order to establish which markup is allowed and how it should be used, markup languages were defined.

This is especially important and relevant in the context of hybrid publishing because it makes it possible, at the later stage of visual design, not only to define how each of these markup elements will be displayed, but also to provide different definitions for each specific output. For example, we can decide that for output A (say the printed book), text marked as 'chapter heading' will be centered on the page, in a different font and larger font size than the running text, while text marked as 'quotation' will be rendered in the same font and size as the running text, but in italics; while for output B we could instead decide to render chapter headings as bold and quotations as underlined text. By

combining the structured text with a different style sheet for each output format, a variety of end products can thus be generated using one single structured text. However, in order to make this possible, the source text must be as systematically structured as possible (see Fig. 2).

In electronic texts, markup has developed into two general types: 'What You See Is What You Get' [7] or visual-typographic (as in the markup tag 'bold type') vs. logical-semantic (as in the markup tag 'emphasis'). The logical-semantic markup is the foundation of hybrid publishing, since it can be translated into whatever visual formatting is most suitable for each particular medium. One of the main advantages of electronic books is that the same content can be published in a wide variety of formats.

All these new possibilities will require thorough, and potentially more labor-intensive, editorial and production strategies. Not only because of the possibility of representing the same content in a variety of forms, but more importantly because once they are properly edited and stored electronically, the content and its constituent parts can be endlessly used and re-used in different ways, now and in the future. This also means that electronic publishing will, in most cases, not bring any significant decrease in production costs. Though if one produces a series, you can benefit from for example, earlier made paragraph styles and predefined structures.

2.3. Genres of Publication

In the present research project we deal with a variety of publication products. In art and design publishing, the most common genres are: research publications, art/design catalogues, artists'/designers' books, and art/design periodicals [12]. The opportunities and challenges of electronic publishing are different for each genre. Common opportunities include searchability, ease of access and distribution, and modularization of content; common difficulties include layout consistency, page numbering and referencing, and potentially large file sizes. Independent of the different genres listed above, we can distinguish three levels of electronic publishing:

1. One-to-one, where the book is one single product published in different media.
2. One-to-many, where the book has different appearances in different media.
3. One-to-database, where the book is based on the content of a database which can be used in a number of ways.

The scope of electronic publishing ranges from the simple conversion of a paper book, to an electronic publication (for example, a PDF of the print edition as an ebook), to full-scale electronic publications which incorporate advanced digital formats such as video, or are published as 'native apps' (applications developed for a particular platform or device).

3. Towards a Hybrid Workflow Based on Markdown

Creating a workflow that is both structured and flexible enough to cater a variety of demands is a key step towards establishing an efficient electronic or hybrid publishing strategy. What we propose here is a hybrid workflow based on the need for publishing across different media, while keeping the main part of the work process in-house rather than outsourcing it.

3.1. Electronic Publishing Workflows: Desktop Publishing and Markdown

Instead of developing a digital publication based on the printed book at the end of a production process, as is common practice by publishers, the main workflow should be adapted at an earlier stage, and made efficient and practical for hybrid publishing. So rather than working separately on the PDF for the print book, the EPUB version, and a Kindle edition, the workflow is instead focused on a single source file (in the Markdown format) which can easily be converted into these different output formats using a relatively small number of digital tools.

3.1.1. Desktop Publishing (Traditional) Workflow (From Word to InDesign to Ebook)

A brief description of the desktop publishing (DTP) workflow currently used by many publishers would be: a Microsoft Word file is imported into InDesign and, after designing and editing, exported to PDF, ready to be printed. After work on the printed edition has been completed, the book may be converted into an electronic version which follows the design of the 'original' as closely as possible. This traditional, print-oriented workflow can be seen as the standard for one-to-one publications (see Fig. 3). Advantages of this workflow are that it is simple and linear, and there are no version branches. You end up with one consolidated manuscript, and WYSIWYG when it comes to design. The main disadvantage of the DTP workflow in 'going electronic' is that it is focused on one single medium, and that the steps to go from there to a digital edition are quite laborious and do not make full use of the possibilities offered by electronic publishing.

3.1.2. From Microsoft Word (.docx) to EPUB

Desktop publishing applications such as InDesign, and WYSIWYG word processors such as Microsoft Word or OpenOffice, are generally not well suited for processing structured text (see section 2.2). Though it is possible to work in a structured manner, for example by using style definitions rather than manually applying formatting, the user is not required to make a distinction between formatting and structure, which is essential in the world of digital publishing. In order to obtain the best possible EPUB file, the .docx file should be formatted using only Word's standard paragraph styles such as 'Normal', 'Title', 'Subtitle', 'Quote' and most importantly 'Heading 1', 'Heading 2', 'Heading 3' for the headings according to their logical hierarchy. For example: 'Heading 1' for chapters, 'Heading 2' for sections, 'Heading 3' for sub-sections. Since the resulting EPUB document will contain a table of contents and document navigation menu based on the 'Heading' hierarchy, a proper structuring of headings is crucial. Word footnotes will appear as linked endnotes in the EPUB, thereby elegantly simplifying an otherwise tedious document redesign task.

Word unfortunately lacks two features that would make it more suitable for hybrid publishing projects:

1. Word does not have a 'strict mode' that would 'force' all writers and editors of a document to use only defined paragraph styles instead of manual formatting.
2. Word provides no automatic or semi-automatic tools for finding manual formatting and replacing it with predefined paragraph styles. The only way to achieve this is to manually review and adjust the whole document.

Often, such inconsistencies in a Word document will only become visible after the EPUB conversion. We recommend two ways of working with Word + Pandoc (an open-source document converter).

1. Conversion from Word to EPUB using Pandoc directly from the Terminal (Mac) / Command Prompt (Windows), or using the browser-based converter developed for this Toolkit [8]. This should only be done only at the very end of the editorial process, because after this conversion no further editorial changes can be applied to the Word document.
2. Conversion from Word to Markdown using Pandoc. Since Pandoc can also convert files *to* the Markdown format, this is often the preferable option, especially for complex publishing projects. The resulting Markdown file can then be used as the 'master file' for conversions to various other file formats (such as EPUB, or HTML for publishing on a website). The advantage of converting to Markdown is that any formatting glitch in the Word document will now become clearly visible. For example, a heading formatted manually will be converted as **heading** while a heading formatted with the proper style definition will be converted as #heading. This makes it much easier to clean up the internal formatting of the document and produce a 'clean' master file for all subsequent document conversions. The EPUB generated from this Markdown file will in most cases be much better structured than an EPUB directly generated from the Word file, making the subsequent work to be performed by the designer much easier.

We would advise against using Pandoc to convert back and forth between Word and EPUB. If the Word document is subject to further editorial changes, then the conversion to EPUB (as in the first scenario) should be done again, as would any work already done by the designer on the previously exported EPUB file. Therefore, if possible the editorial changes should be implemented directly in the EPUB or Markdown file.

3.1.3. Cleaning Up Markdown

Since Markdown is not a word processing application but a document format, it does not provide functions such as automatic renumbering of footnotes and list items. In fact, such numbers don't matter since everything will be renumbered during the document conversion anyway.

However, Pandoc can be used to 'clean up' the Markdown source text; the trick is to convert the document *from* Markdown *to* Markdown. Open a Terminal window (Mac) or Command Prompt window (Windows) and type the following line (be sure to first place the file in the appropriate folder and to navigate to that folder.

```
pandoc beowulf.md -f Markdown -t Markdown -o beowulf_clean.md
```

This means that the program Pandoc is instructed to convert the file 'beowulf.md' file from Markdown (`'-f Markdown'`) to Markdown (`'-t Markdown'`) and save the (cleaned up) result in a new file called 'beowulf_clean.md' (`'-o beowulf_clean.md'`).

3.2. Markdown Workflow

As mentioned above, we recommend the use of the markup language Markdown as part of a hybrid workflow. Though Markdown is not perfect, it is much easier to work with

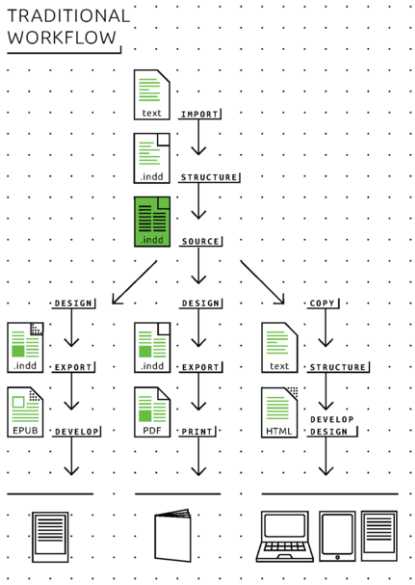


Figure 3. Traditional workflow.

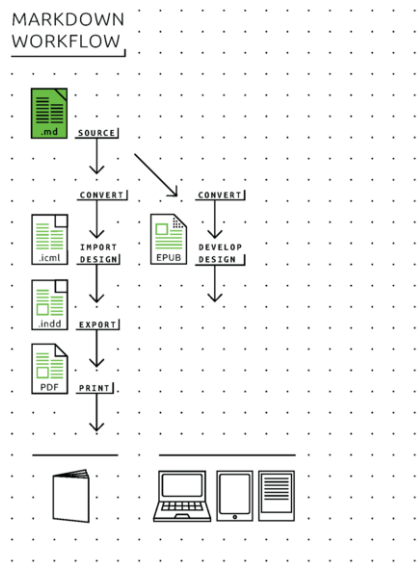


Figure 4. Markdown workflow.

Figure 5. Alice's Adventures in Wonderland.

than, say, the complex markup language XML. Markdown allows for the creation of structured texts, an important requirement in hybrid publishing (see Fig. 4).

3.3. Introduction: Advantages and Limitations

John Gruber, the creator and main developer of Markdown, describes Markdown on his website as follows: 'Markdown allows you to write using an easy-to-read, easy-to-write plain-text format, then convert it to structurally valid XHTML (or HTML).' [6] In other words, Markdown is a way of formatting plain text using human-readable formatting symbols, rather than HTML-style tags such as '<b>' for bold or '<h1>'to define a top-level heading. For example, this is what the beginning of *Alice's Adventures in Wonderland* would look like in Markdown (see Fig. 5).

In this example, the tag '#' defines a top-level heading, '##' a second-level heading, '_' italic text, '**' bold text, and '>' a block quote. Markdown also provides tags for defining lists, embedded images, and links. The popular extended version MultiMarkdown (an extension of the markup language Markdown, with additional support for

footnotes, tables, etc.) provides further support for footnotes, tables, mathematical formulas, cross-references, bibliographies and definition lists. Using simple Open Source conversion applications such as Pandoc, text formatted using Markdown can be automatically converted to well-structured HTML, EPUB, PDF, RTF or other document formats, requiring no manual adjustments.

Markdown is a product of Internet culture. It uses ad-hoc formatting signs commonly used in e-mail and chat platforms, and further popularized on blogging platforms, to provide a standardized, human-readable, user-friendly and well-structured document format, suitable for long-term storage and as a basic source for conversions to contemporary and future document formats. While its formatting *syntax* is simple, it is also both strict and unambiguous enough to allow multiple writers and editors to work on a single document without unnecessary confusion. Another advantage of Markdown is that it can be written and edited in any software application capable of processing basic text: unlike the proprietary file formats of Microsoft Word or other word processors, Markdown can be opened using a simple text editor.

Why do we recommend Markdown in particular? For certain publishing projects – for example, handbooks or books derived from wikis, it may be worth considering alternatives to Markdown, such as reStructuredText. There are many text editing and document conversion programs that support Mardown, such as Mou or MacDown.

However, Markdown/MultiMarkdown is not a magical one-size-fits-all solution. It is particularly well suited for text-oriented documents, but quite limited for creating visually oriented documents, and not really of much use for interactive publishing formats.

Markdown and similar formatting/markup languages are designed for workflows in which there is a clear separation between editorial work (involving writers, translators and editors) and publication design. For publications requiring extensive interaction between writers/editors and visual designers/artists from the very beginning of the authoring process, other tools and workflows are preferable.

3.3.1. Markdown vs. XML

XML is designed for creating structured documents with a clear separation between logical structure and visual formatting. It is the most detailed structuring and formatting language ever developed, and provides the foundation for many other such languages. For example, both HTML and Microsoft Word's .docx are XML-based document formats. So why shouldn't we use XML then? While XML theoretically provides an ideal way of working with single-format files to produce multiple output formats, we do not recommend it for small, independent publishing houses.

The main reason for this is that the broad versatility of XML adds several layers of complexity. Markdown on the other hand can easily be used by non-technical users while still providing good structure and better document conversion into HTML, EPUB and many other formats than Microsoft Word and similar word processing applications. Technically speaking, Markdown provides some of the same features and advantages as XML does, namely separation of content structure from visual layout and painless conversion into multiple output formats. However, unlike XML, it cannot be extended with custom, self-defined markup tags. Still, and particularly for those just getting started with digital or hybrid publishing, Markdown should be more than adequate in almost all circumstances.

4. Showcases, How To Make a Simple EPUB

All showcases underneath are a collaboration with one of the above mentioned publishers, one or more designers and developers from the DPT Collective.

4.1. Showcase BISPublishers

Where other groups of the DPT research group have focused on developing an EPUB, the BISPublishers publication *Sketching, drawing techniques for product designers* [9] [10] has explored the possibilities of a hybrid application as medium. With Phonegap, a mobile development framework, a bridge is made between HTML and a native app, which allows for a multichannel publishing workflow.

4.2. Showcase Valiz

Valiz has developed an EPUB version of Common Skin, a publication in the series Context Without Walls [11]. The multilanguage publications in this series focus on artists from all over the world and contain essays as well as pictures. The transformation from a printed version to an EPUB is not a one-on-one translation, the limitations of graphic design in the EPUB format and the complexities of the production process for a series of publications had to be considered. A publication platform called *EPUBster* has been created to assist with the production of the EPUB [12].

4.3. Showcase nai010 uitgevers

nai010 has developed a digital version of the Stedelijk Museum Highlights Catalogue [13]. Instead of creating an ebook version of the existing catalogue, the group has opted to develop a digital distribution platform accessed via a mobile (web)application. The application allows users to find, filter, search, preview and collect highlights from the Stedelijk Museum collection in preparation for their visit to the museum. Related content (essays, interviews, images, etc.) for the selected highlights may be chosen and are subsequently bundled as a personalised EPUB catalogue, available for generation within the My Highlights application [14]. The generated publication may then be downloaded and viewed (offline) on any device capable of rendering EPUBs.

4.4. Showcase Institute of Network Cultures (INC)

The hybrid workflow INC proposes is based on the need for publishing across different media, while keeping the majority of the work process in-house instead of outsourcing. The research of the INC was aimed at making the transition from a print-centered publication process towards a digital-and-print (hybrid) publication process. The leading question being: how to handle documents in order to publish on different platforms? Creating a workflow that is both structured and flexible enough to cater for different choices is a key step towards an efficient electronic or hybrid publishing strategy. From-scratch development of each publication format is replaced by single source-multi format publishing.

5. Future of Publishing Culture

How will people read in the future? Do people read ebooks from the beginning to the end, or do they casually browse them - or merely keep them as searchable items in a personal electronic library? Do they read PDF versions of books and articles on display screens, or do they first print them on paper? Or will long, complex texts and printed books become fashionable again in a more mature technological environment? It is very possible that in the course of time, users will adjust to the new reading technologies, and have no trouble digesting long and complex texts from the displays of their mobile devices. Already today, increasing numbers of writers and artists self-publish their works. The role of publishers as book producers is rapidly becoming a thing of the past. However, at a certain point of overproduction and oversaturation, publishers may redefine their role to become aggregators and curators.

The electronic publishing models which we have focused on in this Toolkit, using technologies such as EPUB, may in some cases appear counter-intuitive to today's digital media culture: why create what are essentially offline websites in ZIP files, in this age of 'cloud computing' and an 'always-on' culture in which we are constantly connected and networked? However, it is precisely the ephemeral nature of networked media that makes a format like EPUB increasingly attractive. As an offline, stable medium based on World Wide Web technology, it is perfect for everyone who wishes to personally curate, collect and preserve what otherwise may soon be lost.

Faced with on one hand all these new possibilities for self-publishing and self-curating, and on the other hand the rise and consolidation of huge commercial monopolies, the art and craftsmanship of publishing will have no choice but to reinvent and rebuild itself. The changes required may well be greater and more extensive than initially expected.

References

[1] Digital Publishing Toolkit Collective, consists of: Marc de Bruijn, Liz Castro, Florian Cramer, Joost Kircz, Silvio Lorusso, Michael Murtaugh, Pia Pol, Miriam Rasch, Margreet Riphagen, Loes Sikkes and Kimberley Spreeuwenberg.

[2] J. Monk, F. Cramer, M. Rasch, M. Riphagen. & Digital Publishing Collective (2014). *From Print to Ebooks: a Hybrid Publishing Toolkit for the Arts*. Amsterdam, Institute of Network Cultures, 2015.

[3] GitHub [Internet]. [cited 2015 June 2]. Available from: https://github.com/DigitalPublishingToolkit/.

[4] Merriam-Webster [Internet]. [cited 2015 June 2]. Available from: http://www.merriam-webster.com/dictionary/markup%20language.

[5] P. Sloterdijk. & W. Hoban. *You must change your life: On anthropotechnics*. Cambridge, UK: Polity, 2013

[6] John Gruber, *Markdown: Introduction* [Internet]. [cited 2015 June 2]. Available from: daringfireball.net/projects/Markdown/.

[7] Merriam-Webster [Internet]. [cited 2015 June 2]. Available from: http://www.merriam-webster.com/dictionary/wysiwyg.

[8] An overview can be found on the Digital Publishing Toolkit Software Showcase [Internet]. [cited 2015 June 2]. Available from: http://pandoc.networkcultures.org/ or go directly to http://pandoc.networkcultures.org/hybrid.html.

[9] K. Eissen, and R. Steur, *Sketching Drawing Techniques for Product Designers*. Holanda: Bis, 2007.

[10] GitHub Sketching [Internet].[cited 2015 June 2]. Available from: https://github.com/DigitalPublishingToolkit/Sketching

[11] D. Pappers, L. Levy, M. Mihindou, and P. Mason. *Common Skin*. Valiz, 2014.

[12] EPUBster is a web application to create and edit EPUBs, written in CakePHP. Content formatted using Markdown is used as input to generate publications using the EPUB 3 file format [Internet]. [cited 2015 June 2] Available from: https://github.com/DigitalPublishingToolkit/epubster.

[13] H. De Man. *Stedelijk Collection Highlights: 150 Artists from the Collection of the Stedelijk Museum Amsterdam*. Amsterdam: Stedelijk Museum, 2012.

[14] My Highlights application [Internet]. [cited 2015 June 2]. Available from: https://github.com/DigitalPublishingToolkit/My-Highlights.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-58

Open Access and Research Assessment: Dealing with UK Open Access Requirements in Practice

Dominic TATE[a,1]
[a] *Scholarly Communications Manager, University of Edinburgh, United Kingdom*

Abstract. This paper describes research-funder and research assessment policies in the UK and assesses the impact that these policies are having in the transition towards research outputs being made available on an open access basis.

Keywords. Open access, research assessment, research policy, REF, RCUK

1. Introduction

Over recent years there has been a significant increase in the number of institutional and research-funder policies [1] mandating open access (OA) to research results; taking advantage of both green and gold routes. In the United Kingdom, academic institutions and research centres mostly mandate green OA, which is achieved by self-archiving into a repository. Funders have recognised that it is their responsibility not only to fund the original research, but also to ensure the widest possible dissemination of its results. For that reason, some funders do not limit their policies to green OA, but also extend them to gold OA, and take responsibility for covering gold article-processing charges (APCs) when they arise.

In this context, institutional and funding policies can contradict one another and create a level of frustration not only to researchers who need to comply both with their institutions' and funders' policies, but also to the people who advise them on how to comply, such as repository managers and librarians.

2. Background to Open Access at the University of Edinburgh (and in the UK)

The University of Edinburgh has had a long-standing commitment to open access. The University adopted its initial Institutional Repository in 2003, and now has over 34,000 full-text open access research outputs in its systems. This work is facilitated by the library's Scholarly Communications Team and is supported by the University's Research Publications Policy, which strongly endorses OA, with a stated preference for green. Since 2008, the library has managed a fund to pay gold APCs to Wellcome

[1] Dominic Tate, Scholarly Communications Manager, Edinburgh University Library, 30 George Square, Edinburgh EH8 9LJ, United Kingdom; E-mail: dominic.tate@ed.ac.uk.

Trust-funded authors and managed a number of publisher accounts to make best use of these funds.

In the early years, although there were many pockets of enthusiasm, OA did not become part of the fabric of academic life across all disciplines in the way many had hoped.

More recently, additional funder OA policies and initiatives such as those of the Wellcome Trust [2], FP7 [3] and Horizon 2020 [4] have helped to raise awareness of OA issues. However, despite fairly wide publicity since the introduction of the repository, habitual uptake of open access options by University of Edinburgh authors has only really become part of everyday academic practice in some scientific and medical disciplines – a fairly typical scenario in most UK universities.

3. RCUK Open Access Policy

In 2012, Research Councils UK (RCUK) [5] strengthened its existing open access policy [6], effectively requiring that journal articles and conference proceedings arising from research funded by the seven RCUK members are made open access within a maximum 6-24 months from the date of publication. This policy allows both green and gold OA, though rapid access is preferred, and RCUK has provided block grants to 30 research-led universities for gold OA where publisher embargo periods are too long to meet the RCUK requirements.

The University of Edinburgh responded to this policy by beginning an Open Access Implementation Project [7], which paid for some staff time to source and upload repository-appropriate copies of journal articles and conference proceedings to its institutional repository. This approach was successful, and the University achieved OA rates of 64% [8] for RCUK-funded journal articles and conference proceedings in the first year of the policy – a compliance rate fairly typical of research-led Universities in the UK.

4. Research Excellence Framework (REF) Open Access Requirements

The Research Excellence Framework [9] is the current system for assessing the quantity and quality of research undertaken in UK higher education institutions. Following wide consultation, the four UK higher education funding bodies have introduced an open access requirement [10] in the next assessment, (referred to as the post-2014 Research Excellence Framework, likely to take place in 2020). This new requirement comes into effect from 1 April 2016.

The guiding principle of this requirement is that journal articles and conference proceedings must be available in an open access form in order to be eligible for submission to the post-2014 REF. In practice, this means that these outputs must be uploaded to an institutional or subject repository at the point of acceptance for publication. This is a green open access requirement, and even if the author takes a gold route, deposit into a repository must still be made.

The REF assesses the full range of research undertaken at UK universities and research centres and is funder-agnostic. Indeed, in many fields, much of the research assessed is un-funded. The implication of this policy for research-led institutions such as the University of Edinburgh is that all journal articles and conference proceedings

will need to be made OA. The deposit requirements are stringent and auditable (full-text documents must be added to a repository immediately on acceptance by the publisher and made open as soon as the publisher allows). Failure to comply presents significant reputational and financial risks, both for researchers and universities. Accordingly, OA is now considered an institutional priority by university management.

5. How Has REF Changed Open Access in the UK?

The association of OA with research assessment has changed the landscape in the UK in a number of ways:

5.1. Open Access Now Affects Everybody

Until now, there was only really a mandate for those authors in receipt of grant funder for a research project. Large, research-led universities such as Edinburgh aspire to be able to return any member of a staff on a research contract in a REF exercise. In turn, it is important that any member of staff is in a position to be able to select any of his or her publications for such an exercise. In an institution such as the University of Edinburgh, this effectively means that every journal article and conference paper needs to be made OA.

5.2. OA Is No Longer Optional, Even for Un-funded Research

Some researchers who are not in receipt of grant funding for research have argued that they do not need to make papers deriving from their research OA as there is no funder mandate. Linking OA to research assessment has started to break down this argument, with some universities responding that any research conducted during working time or using university facilities is in some way "funded" research, even in absence of a project grant. The argument for tax-payer access to tax-payer funded research still applies.

5.3. OA is Being Discussed

The OA agenda now has the full attention of research directors, administrative support staff and university senior management nationwide. It is on the agenda at many departmental meetings and the issues are being discussed amongst researchers in a way we have not previously seen in the UK.

5.4. OA Can No Longer Be Put Off, or Ignored

Authors must take action immediately on having an article accepted for publication. The OA requirements for REF have been designed in such a way as to prompt the researcher to deposit the author's final peer-reviewed manuscript (postprint, AAM) at the point at which they are most likely to still have that version available. Because it is a requirement to deposit this (even closed-access) immediately on acceptance, authors cannot postpone this task until a later date, or they may risk the paper being ineligible for inclusion in the next REF. Because there is no scope for retro-compliance with the

requirements, appropriate administrative support must be in place to guide authors who are unsure of what actions they must take.

5.5. We Are Starting to See Real Growth in UK Repositories

The relative ease and low cost of making large numbers of research articles available on repositories is highlighting the relative high cost of gold OA, especially with hybrid journals.

6. Preparing for the New Way of Doing Things

The OA requirements for the next REF exercise mean that, in some small way, researchers will have to change what they *do* at the point their papers are accepted by a publisher. What is being asked is small – researchers must simply add a document to a repository (or ask someone to do this for them), but they need to do this within a certain timeframe and they need to get it right. The implementation of the requirements means that awareness must be raised amongst all university research staff in the UK – and adequate support needs to be in place to help answer questions and provide support.

When a new policy such as this is announced, institutions may not be prepared for proactive advocacy and timely compliance for many reasons; they may lack staff, knowledge, or the financial resources. Given the wide scope and broad impact of research funders' OA policies, as well as the differing workflows and approaches of universities across the sector, examples are needed of effective practice that are collaboratively developed but reflect institutional difference within a 'real-world' environment.

Jisc [11] has commissioned a portfolio of Open Access Pathfinder Projects [12] aimed at helping reduce the fragmentation of practice and put in place mechanisms to capture and share lessons quickly and iteratively around the dynamic OA environment.

The University of Edinburgh leads the LOCH Project (Lessons in Open Access Compliance for Higher Education) [13], which is already providing guidance to other institutions on the implementation of both green and gold OA. The project has shared a wide variety of guidance materials to help staff in different universities plan for the implementation of the OA requirements for REF and to provide support and guidance to a range of stakeholders. Project outputs include implementation plans, checklists, text for web pages and a wealth of documentation to help practitioners in the UK with the considerable task ahead.

7. Conclusions

- The UK has been engaged with the OA agenda for over a decade, but progress has undoubtedly been slow.
- The revision of the RCUK policy undoubtedly helped to create an increase in awareness of OA, as well as increased deposits in institutional repositories and demand for gold OA funding.
- Linking OA with research assessment has done more than anything else to get authors interested in and talking about OA.

- It is imperative that we continue to convey a really positive message about the benefits of OA to authors during this time of transition. There is real a danger that authors lose sight of the good things that OA can do for them, and a risk that it becomes perceived as an extra administrative task with a sanction for non-compliance.
- Universities in the UK are working hard to prepare for the REF requirements to 'go live' on April 1st 2016. There is much work to do, and many conversations to be had – but success with this new policy could prove to be a real milestone in the transition towards open access.

References

[1] ROARMAP http://roarmap.eprints.org/
[2] Wellcome Trust Open Access Policy http://www.wellcome.ac.uk/about-us/policy/spotlight-issues/Open-access/index.htm
[3] FP7 Open Access Pilot http://ec.europa.eu/research/science-society/document_library/pdf_06/open-access-pilot_en.pdf
[4] Horizon 2020 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[5] RCUK http://www.rcuk.ac.uk/
[6] RCUK Open Access Policy http://www.rcuk.ac.uk/research/openaccess/
[7] OA Implementation Project http://www.ed.ac.uk/schools-departments/information-services/research-support/publish-research/open-access/oa-imp
[8] University of Edinburgh RCUK Open Access Report 2014 https://www.era.lib.ed.ac.uk/handle/1842/9386
[9] Research Excellence Framework http://www.ref.ac.uk/
[10] REF Open Access Requirements http://www.hefce.ac.uk/whatwedo/rsrch/rinfrastruct/oa/
[11] Jisc http://www.jisc.ac.uk/
[12] Jisc OA Pathfinder Projects http://openaccess.jiscinvolve.org/wp/pathfinder-projects/
[13] LOCH Project http://libraryblogs.is.ed.ac.uk/loch/

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science 63
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-63

Building a Social Semantic Digital Library

Maria NISHEVA-PAVLOVA[a,b,1], Dicho SHUKEROV[a], Pavel PAVLOV[a]
[a] *Faculty of Mathematics and Informatics, Sofia University, Bulgaria*
[b] *Institute of Mathematics and Informatics, Bulgarian Academy of Sciences*

Abstract. The paper analyzes some current trends of research and development in the field of digital libraries. The presentation is focused on the main features of two new generations of digital libraries – the so-called semantic digital libraries and social semantic digital libraries. The design characteristics, principles of functioning and some implementation details of a particular academic digital library have been discussed as an illustration of the suggested ideas.

Keywords. Digital library, semantic technology, ontology, semantic interoperability, search engine, information retrieval, sentiment analysis.

1. Introduction

During the last 2-3 decades, digital libraries are one of the most rapidly developing areas of research and considerable practical results. According to the IFLA/UNESCO Manifesto for Digital Libraries [1], "a digital library is an online collection of digital objects, of assured quality, that are created or collected and managed according to internationally accepted principles for collection development and made accessible in a coherent and sustainable manner, supported by services necessary to allow users to retrieve and exploit the resources". Digital libraries contain electronic copies of valuable books, periodicals, documents, maps, audio archives, etc., and provide convenient tools for comparatively inexpensive access to them. A digital library is an information retrieval system that maintains collections of digital objects along with means for organizing, storing, and retrieving the resources contained in its collections.

Digital libraries should play the role of environments supporting the full life cycle and the best practices of creation, preservation and use of rich digital content. Interoperability and sustainability are the most important principles of building digital libraries able to communicate with each other. The so-called semantic digital libraries have been based on design and implementation standards that are a significant step to providing interoperability at the semantic level.

2. Semantic Digital Libraries

The considerable results in the area of digital libraries during the last two decades played a determinant role in the development of the Digital Library Reference Model [2]. This model has the aim to lead to an agreement between experts with respect to the

[1] Corresponding Author, E-mail: marian@fmi.uni-sofia.bg.

main concepts, structures and activities in digital libraries. Three types of systems play a central and distinct role in the corresponding digital library reference architecture [2]:

- Digital Library – "an organization, which might be virtual, that comprehensively collects, manages and preserves for the long term rich digital content, and offers to its user communities specialized functionality on that content, of measurable quality and according to codified policies";
- Digital Library System – "a software system that is based on a defined (possibly distributed) architecture and provides all functionality required by a particular digital library";
- Digital Library Management System – "a software system that provides the necessary software infrastructure both (i) to produce and administer a digital library system incorporating the suite of functionality considered fundamental for digital libraries and (ii) to integrate additional software offering more refined, specialized or advanced functionality".

The next generation of digital libraries – *semantic digital libraries* – may be considered as digital library systems that apply semantic technologies to achieve their specific goals [3, 4]. They provide new search paradigms for the information space – *intelligent search* (also known as semantic or ontology-based search) [5, 6] and *community-enabled browsing*. The specific technologies of semantic digital libraries make it possible to integrate metadata from various heterogeneous sources. In this way they support the interconnection of different digital library systems.

The utilization of proper *ontologies* is one of the main characteristics of semantic digital libraries. In particular, ontologies play a key role in semantic search. Three types of ontologies have been identified as a support for this type of search [7]: bibliographic ontologies, subject ontologies, and community-aware ontologies. Bibliographic ontologies describe metadata standards. Subject ontologies are useful as knowledge sources which define the meaning of most domain concepts, their hierarchy, properties and relationships. Community-aware ontologies are oriented to the description of the different types of users, their requirements and interactions.

Ontologies are also one of the well-accepted types of resources for achieving semantic interoperability of digital libraries. According to [8], semantic interoperability depends mainly on the existence and use of well-formed and accepted upper and core ontologies, in which the basic concepts and relationships are defined. In addition, the concepts defined in the upper and core ontologies, should be extended by appropriate domain ontologies.

3. Social Semantic Digital Libraries

The critical study of the experience in development and use of digital libraries shows that current semantic digital libraries are not enough from the point of view of their typical end users because [4]:

- digital libraries should not be for scholars and librarians only but mostly for average people;
- they concentrate on delivering content, not on knowledge and opinion sharing within a community of users;
- digital libraries have lost the human part of their predecessors.

The so-called *social semantic digital libraries* [4] are suggested as a solution of these problems. They have the aim to make users (in other words, readers) involved in the content annotation process. They also allow users/readers to share their knowledge within a community. Social semantic digital libraries provide better communication between users in and across communities than the traditional ones.

The rest of the paper discusses the main features of a particular semantic digital library called DjDL [9] and the results of our recent activities directed to its growing into a social semantic digital library.

4. Main Characteristics of DjDL – A Digital Library with Bulgarian Folk Songs

DjDL preserves a collection of over 1000 digital objects representing folk songs from the Thrace region of Bulgaria. This collection constitutes a considerable part of the digitized archive of the distinguished Bulgarian folklorist Prof. Todor Dzhidzhev published in [10]. In particular, the files with metadata and lyrics of songs (in LaTeX format) and the files with the encoded musical notations of songs (in LilyPond format) have been used as original resources in building the repository of DjDL.

The development of the prototype of DjDL was supported by the Bulgarian National Science Fund within a project titled "Information technologies for the presentation of Bulgarian folk songs with music, notes and text in a digital library" [11]. The main characteristics of this prototype were presented at the ELPUB 2012 conference [6]. In this paper we discuss an entirely new version of DjDL that has some essential features of a social semantic digital library.

DjDL has the typical architecture of an academic digital library. Its functional structure is shown in Figure 1. It includes five main components: metadata catalogue, repository, search engine, module implementing the library functionality, interface module. A subject ontology was especially created and has been used to support the full functionality of the search engine.

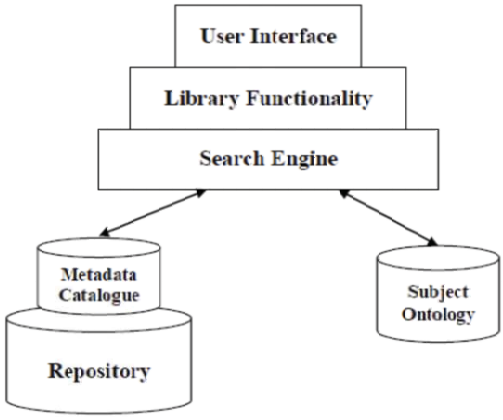


Figure 1. Functional structure of DjDL.

The folk songs treasured in the repository of DjDL have been presented with their notes (musical notations), text (lyrics) and music (digitized versions of their authentic performances).

The subject ontology describes a proper amount of knowledge in several domains, relevant to the content of Bulgarian folk songs. It contains definitions of the main domain concepts, descriptions of their properties and some kinds of relationships between them, as well as a selected set of their representative instances. This subject ontology consists of a set of interrelated subontologies needed by the search engine of DjDL and developed especially for the occasion:

- ontology of folk songs which includes various genre classifications of folk songs (by their thematic focus, by the context of performance, by their cultural functions, etc.);
- ontology of manner of life and family (professions, instruments, clothing, ties of relationship, feasts, traditions and rites, etc.);
- ontology of impressive natural phenomena;
- ontology of social phenomena and relationships;
- ontology of historic events;
- ontology of administrative division – combines the current administrative division of Bulgaria with the one from the beginning of the 20th century.

In addition, a set of natural language-dependent patterns of typical stylistic or thematic constructs, called *concept search patterns*, have been defined and used as domain knowledge aimed at providing satisfactory precision and recall of the search engine.

The purpose of the search engine is to provide adequate access to the variety of resources stored in DjDL. It provides two main types of search: keywords-based and semantic (ontology-based) search. The semantic search tool provides a set of facilities for augmentation and refinement (automatic reformulation according to the available explicit domain knowledge) of the queries for keywords-based search. The augmentation of the user queries is based on proper utilization of the two forms of conceptual knowledge maintained in DjDL – the subject ontology and the set of concept search patterns based on this ontology.

The search engine realizes some additional functionalities enabling the user to combine the search and retrieval of documents kept in the repository of DjDL with a kind of sentiment analysis of their texts. For this purpose some of the subject ontology classes are associated with proper positive or negative numbers which play the role of sentiment estimates of the corresponding concepts. The sentiment estimates of the ontology concepts have been used as default values for their specializations and forms.

The library functionality and the user interface of DjDL are designed in accordance with the expected requirements of its typical users. Three levels of access to the library and the corresponding differentiated roles of users have been defined: *librarian*, *author* and *reader*. Librarians maintain the user accounts and their associated roles. They also may add and register new library resources and upload new versions of the subject ontology. Authors may use a specialized authoring tool developed especially for the purpose. It allows one to create and edit catalogue descriptions and texts of songs (their lyrics and musical notation). Users registered as readers may examine the texts, musical notations and sound recordings of the available folk songs, define and send queries to the search engine, write comments that can be read and

replied by others. In this way users are enabled to participate in the content annotation process, to communicate and to share their impressions and opinions.

5. Semantic Search and Sentiment Analysis of the Lyrics of Songs in DjDL

The semantic search tool of DjDL is aimed at making some kind of pre-processing of the user queries in order to provide better precision and recall of their accomplishment. When the user defines his/her particular query and indicates the search sources (the lyrics of songs or specific metadata), the search engine augments the query so much as possible, in accordance with the available explicit domain knowledge.

The most significant knowledge source for augmentation of the user queries is the taxonomy (the "is-a" hierarchy) of concepts that forms the core of the subject ontology. During the augmentation of the user query, first of all an exhaustive breadth-first search in the graph representing the "is-a" concept hierarchy is performed, starting from the node which corresponds to the original user query. The names of the visited nodes that are in fact the respective more specific concepts described by the ontology, are added to the one formulated by the user. The resulting list of concepts if properly visualized and placed at the user's disposal for optional refinement.

During the next step of query expansion, the search engine adds to the newly constructed set of concepts some derivatives and synonyms of the main terms found as values of their "form" and "synonym" properties in the subject ontology. The corresponding property values from the definitions of all concepts included by that time in the expanded user query and the existing instances of these concepts are added to the query as well. Finally, the values of some properties of the newly included instances that have been explicitly specified as significant for their classes with respect to search purposes, are included in the resulting augmented query. If the search activities have been realized in the lyrics of songs and there is a concept in the augmented query provided with appropriate search patterns, the pattern matching module performs an additional search for each of these patterns.

Thereby the user query is augmented as far as possible in terms of the subject ontology and in fact it has the shape of a disjunction of all included forms of concepts and instance names. In this form the resulting query is ready for further refinement and processing.

As example queries for semantic search being of interest for folklorists, that can be executed by the search engine of DjDL, we could indicate the queries for search and retrieval of:

- songs praising or mentioning significant historic events;
- songs in which typical folk beliefs or rites are described;
- songs in which elements of country work and life are described or mentioned;
- songs in which significant family events are mentioned.

Let us suppose for example that the user defines a query for semantic search in the lyrics of songs which concerns the concept "historic event" ("историческо събитие" in Bulgarian). During the execution of this query, first of all it is augmented and refined with the assistance of the user (see Figure 2). Then a consecutive search in the lyrics of songs follows. As a result, all documents with texts of songs containing phrases that are juxtaposed with at least one element of the augmented query, are extracted. A list with

the titles of the discovered songs is properly visualized on the user screen, as shown in Figure 3.

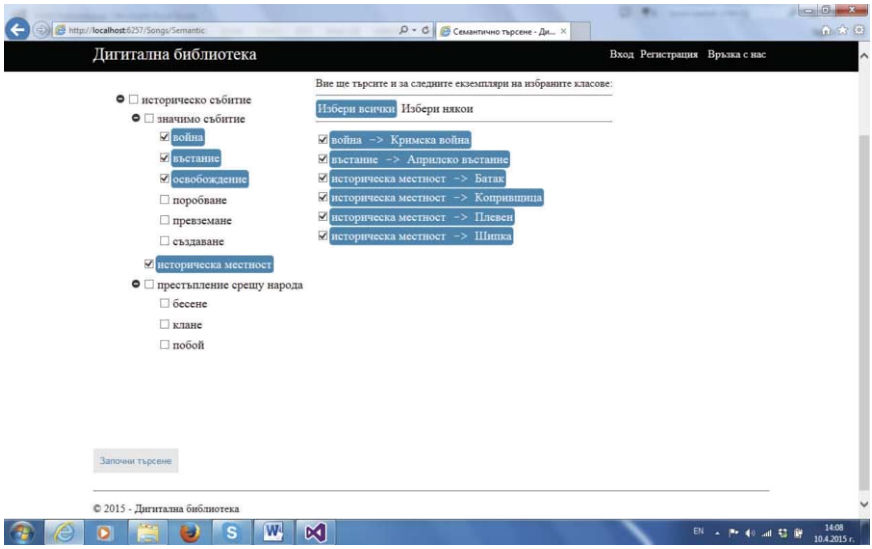


Figure 2. Construction of a user query for semantic search.

When the user clicks on the name of a chosen song satisfying the augmented query, the text of this song is displayed in a new window along with the corresponding metadata. The discovered words and phrases that match the query, are highlighted (see e. g. Figure 4).

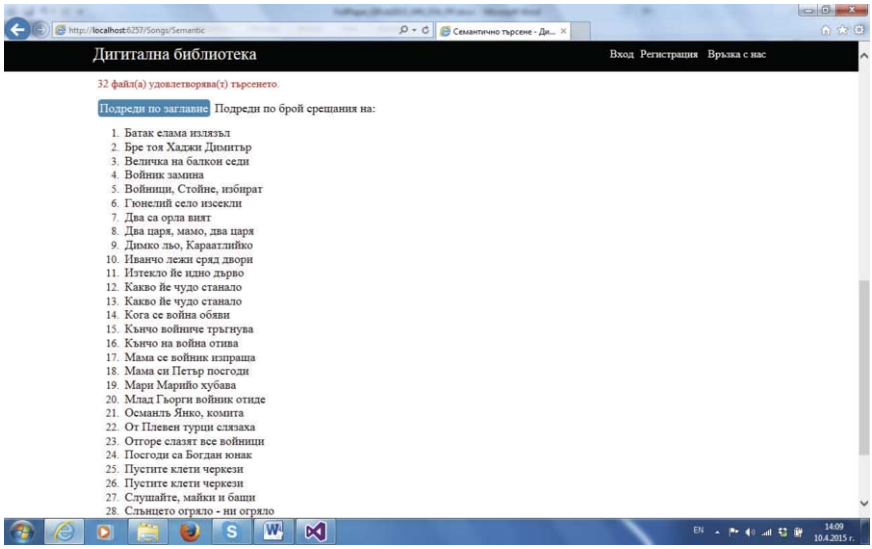


Figure 3. Search results for a user query containing the phrase “historic event” (level 1 – document retrieval).

Figure 5 illustrates some search results for a user query containing the phrase “love infidelity”. A predefined concept search pattern matches the query.

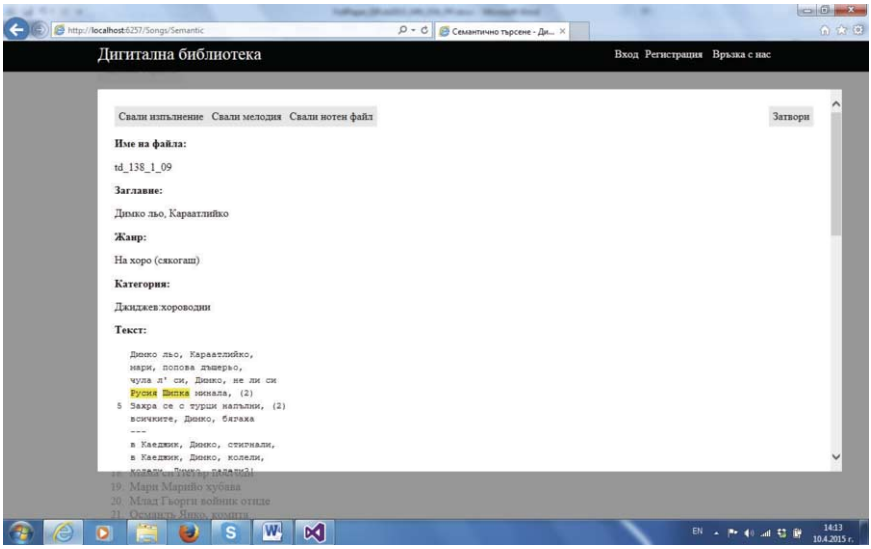


Figure 4. Search results for a user query containing the phrase “historic event” (level 2 – display of lyrics).

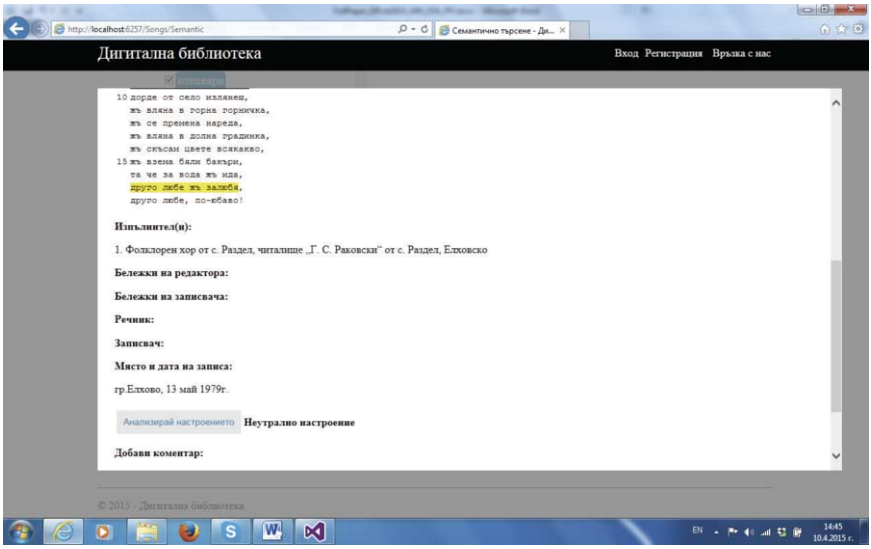


Figure 5. Some results of semantic search in combination with sentiment analysis.

The search engine of DjDL holds up some additional functionalities that enable the user to combine the search and retrieval of documents with a kind of sentiment analysis of their texts (see e. g. Figure 5). The sentiment analysis tool uses the subject ontology

as a source of knowledge about the emotional intensity of its concepts and computes rough estimates of the mood of songs.

More precisely, some of the subject ontology classes are associated with proper positive or negative numbers which play the role of sentiment estimates of the corresponding concepts. The sentiment of a song is currently defined in accordance with the sum of the sentiment estimates of the particular words in the lyrics of this song. Moreover, the specializations of ontology concepts and all their forms and synonyms inherit the sentiment estimates of the main concepts. It other words, the sentiment estimates of the ontology concepts have been used as default values for their specializations and forms.

The results of the experiments carried out with the texts of songs stored in the repository of DjDL indicate that the presented approach is not completely adequate for the domain specificity. For example, the sentiment of a part of the songs has been inferred as "merry" while it may be defined as "sadly" by a human reader. Because of that a new version of the sentiment analysis tool has been under development. Two basic changes have been considered in order to improve its performance. First, the sentiment symbolized by phrases that match existing concept search patterns, will be considered first of all during the sentiment analysis process. A set of new patterns will be defined for the purpose and the concept search patterns will be provided with proper sentiment estimates. Next, the sentiment estimates of some particular forms of a set of distinct ontology concepts will be revalued in accordance with their typical sense and cases of use.

6. Authoring Tools and Social Aspects

The library functionality and the user interface of DjDL are oriented to its three basic types of users and their specific roles: librarian (or library administrator), author and reader (end user). The library administrators add and register new resources in the repository and are responsible for the maintenance of the user accounts and their corresponding roles as well as for any other security and system settings issues.

A set of specialized authoring tools have been developed in order to allow the users with author's role to create and edit metadata, lyrics and musical notations of songs. Authors may also define new and edit existing concept search patterns and define queries for creation of indexes, MIDI files with melodies of songs, etc.

The users of DjDL registered as readers may examine the texts, musical notations and sound recordings of the available folk songs, define and send queries for keywords-based and semantic search and sentiment analysis. They also may write comments that can be read and replied by others – by all users, by administrators or authors only, by all or specific group(s) of readers. These comments could refer specific metadata or the lyrics or music of particular songs, issues concerning the qualities of the search engine or the subject ontology, the performance of the software system of DjDL as well as any other topic of interest to the user.

In this way DjDL is acquiring some characteristics of a social semantic digital library. In particular, its users are enabled to participate in the content annotation process, to communicate and to share their knowledge, impressions and opinions.

7. Implementation Principles

The digital library system of DjDL is a standard client-server application built on the .NET Framework 4.5 and ASP.NET MVC 5 [12]. The tool used for its implementation is Microsoft Visual Studio Ultimate 2012 with additional packages for ASP.NET MVC 5.

A class library called RDFXMLClassLibrary has been especially built for the purpose of automatic conversion of the original files with metadata and texts of songs to the RDF format. It implements the RDF 1.1 XML Syntax standard of W3C.

Another relatively new library used in the project is SignalR. It enables one to add real-time functionality to the software application.

The jQuery library v. 1.10.2 has been used for JavaScript processing.

To generate files with "standard" musical notations and MIDI files with melodies of songs from the original source files, LilyPond [11] should be installed as an external software package on the server.

The software implementation is based on Entity Framework 5 technology in combination with Code First. This enables one to build first of all the data model and then to create the database. The current version of DjDL uses a local database (SqlLocalDB v. 11.0).

Figure 6 shows the software architecture of the digital library system of DjDL.

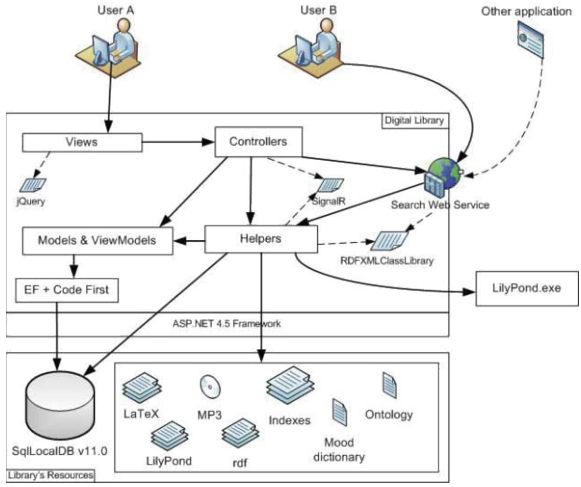


Figure 6. Software architecture of the digital library system of DjDL.

The subject ontology is created using Protégé 4.3. Most concepts of this ontology are constructed as defined OWL 2 classes, by means of necessary and sufficient conditions defined in terms of proper restrictions on certain properties. The current version of DjDL includes one more ontology (named Mood dictionary on Figure 6) which contains some of the subject ontology classes and a "mood_estimation" property. The values of this property are integers that play the role of sentiment estimates of the respective concepts. The Mood dictionary should be considered as a part of the subject ontology and will be merged with it when the values of the "mood_estimation" property will be made sufficiently precise.

8. Conclusion

Semantic digital libraries integrate heterogeneous information resources based on various types of metadata. They provide interoperability at the semantic level with other digital library and information systems and deliver user friendly and adaptive search and document retrieval interfaces. The availability and purposeful use of explicit conceptual knowledge at appropriate level(s) of abstraction in a digital library system may significantly improve the precision and recall of its search engine as well as give it some of the principal characteristics of a social semantic digital library.

References

[1] IFLA/UNESCO Manifesto for Digital Libraries. Available at http://www.ifla.org/publications/iflaunesco-manifesto-for-digital-libraries [Accessed May 2015].
[2] L. Candela, D. Castelli et al., *The DELOS Digital Library Reference Model: Foundations for Digital Libraries*. ISSN 1818-8044, ISTI – CNR, 2007.
[3] M. Nucci, M. Barbera, C. Morbidoni, D. Hahn, A Semantic Web Powered Distributed Digital Library System. In: *Proceedings of ELPUB 2008 Conference on Electronic Publishing*, Toronto, Canada, 2008, 130–139.
[4] S. Kruk, B. McDaniel, Conclusions: The Future of Semantic Digital Libraries. In: S. Kruk, B. McDaniel (Eds.), *Semantic Digital Libraries,* Springer, 2009, 215–222.
[5] R. Guha, R. McCool, E. Miller, Semantic Search. In: *Proceedings of the 12th International World Wide Web Conference,* Budapest, Hungary, 2003, 700–709.
[6] J. Lervik, S. Brygfjeld, Search Engine Technology Applied in Digital Libraries, *ERCIM News* **66** (2006), 18–19.
[7] S. Kruk et al., The Role of Ontologies in Semantic Digital Libraries. In: *Proceedings of the European Networked Knowledge Organization Systems (NKOS) Workshop,* Alicante, Spain, 2006.
[8] N. Guarino, M. Carrara, P. Giaretta, Formalizing Ontological Commitment. In: *Proceedings of the 12th National Conference on Artificial Intelligence AAAI-94,* Seattle, Washington, The AAAI Press, Menlo Park, California, 1994, 560–567.
[9] M. Nisheva-Pavlova, P. Pavlov, Ontology-Based Search and Document Retrieval in a Digital Library with Folk Songs. *Information Services and Use* **31**(3-4) (2011), 157–166.
[10] T. Dzhidzhev, *Folk Songs from Thrace,* L. Peycheva, G. Grigorov, N. Kirov (Eds.), Prof. Marin Drinov Academic Publishing House, Sofia, 2013.
[11] N. Kirov, Digitization of Bulgarian folk songs with music, notes and text, *Review of the National Center for Digitization* **18** (2011), 35–41.
[12] ASP.NET MVC 5 official website: http://www.asp.net/mvc/mvc5 [Accessed May 2015].

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-73

On Key Bespoke Tools to Support Electronic Academic Document Discovery

Fernando Loizides[a,1], George Buchanan[b] and Keti Mavri[a]
[a]*Cyprus University of Technology*
[b]*City University London*

Abstract. Publishing in academic journals and conferences has become faster, and easier with the ability to edit and submit documents electronically. With the increase of publications also come negative effects such as that of information overload and elevated discovery time of relevant resources. An information seeker often wades through several documents in order to find relevant publications having to either select known repositories for their search or utilizing generic search sources which network to several online repositories. Even with the advances in interactive systems, information seekers still carry out a mostly textual search from input to returned results. Several tools have been created by researchers in order to assist the seekers in their visual academic document triage activities but very few have been successfully implemented in actual discovery of electronic publications. With electronic publishing increasing dramatically, we recognize the paramount importance for these tools to be improved and integrated within environments to assist the seekers. In this work, we present an overview of key bespoke tools purpose built for achieving this document selection tasks. Using this work as a reference we hope to encourage structured and novel approaches to creating triage tools and improve the discovery process of electronic academic document publications.

Keywords. Document triage, tools, information seeking

1. Introduction

Electronic publications continue to increase exponentially with the advent of new publishing routes; namely, increasing conferences, journals and online document repositories. Document discovery is becoming more difficult, with the key focus being on the process of publication and retrieval techniques for single database recovery. What is currently under researched and under supported is the process of discovering these publications; a stage which renders the actual publication process void if the documents are never read. Academic repositories are now increasing the tools available to users in order to discover information within documents as well as documents themselves (See Figures 1 and 2).

The purpose of this work is to give the reader a directed primer on the types of tools that have been created over the years to support the document selection and discovery process. We do not aim to present an exhaustive literature review, but rather key findings which represent the under researched field, in order to equip developers, designers and stakeholders with the information needed to guide informed decisions on research and development; crudely speaking, a starting point for the masses. We bring

[1] Corresponding Author. E-mail: fernando.loizides@gmail.com.

together a body of work which is very focused and directed and present suggestions with a hope to encourage more work to be undertaken with the correct incentives in the field of document selection and interfaces to support information seekers.

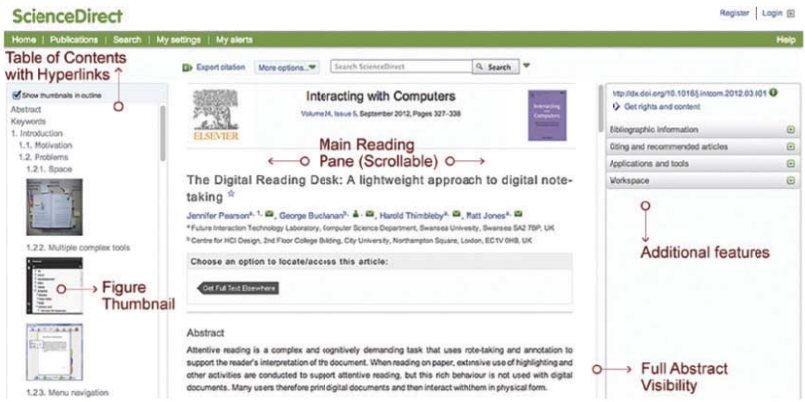


Figure 1. Science Direct website providing tools for faster navigation, also assist information seekers in their triage activities

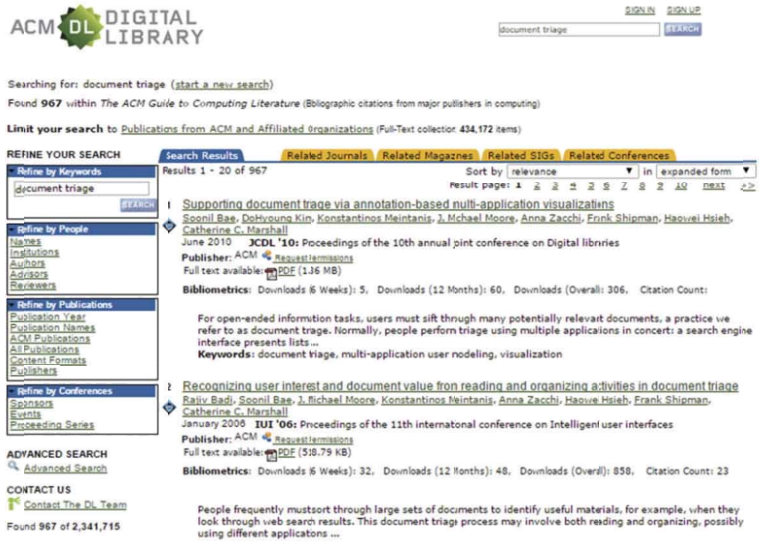


Figure 2. The ACM Digital Library website providing tools for document searching and discovery through facets and snippets

Document Selection is a process undergone by scholars, information professionals and information seekers daily to choose relevant documents on a topic. More recently, the term document triage has been adopted to describe more accurately the document selection process. Document triage is largely a human cognitive process and has not been thoroughly researched, hence this process is not yet fully understood. In order to understand the effect that document triage has on information seeking we focus on the

part of the information seeking process that document triage influences. Information seekers are reported on as making erroneous decisions on the relevance of documents during triage [1]. When influencing factors, such as document features, are altered the behavioural patterns of the users are also likely to change. There are three levels at which the triage process can take place [2], the surrogate stage, the within-document stage and the in depth reading stage.

Most of the work on document triage has been using manual searching, without specialized support from software. Research reports indicate how the libraries themselves lack "better support users' overall information work in context" [3, 4]. Some work, albeit limited, has been carried out to investigate how supportive software can assist users in their information seeking activities. In this work, we take a closer look at key bespoke tools created by researchers and how they affect the document triage process. We begin with an overview of information visualization, which is a technique that is employed in the majority of the individual tools presented. We then present the individual tools themselves. We then discuss the tools as a whole, their potential and limitations. We conclude with future directions that can be taken related to the area of discovering academic publications with tools.

2. Tools and Concepts

2.1. Information Visualization

One of the most challenging problems performing document triage from a results is the sheer amount of documents available. An information seeker is often inundated by more documents that can be possibly looked at. One way that researchers attempt to solve this problem is by using information visualization within their proposed tools. We therefore give a small primer to the reader.

Information visualization has not been restricted to the visual cues alone, but has evolved to include the interactions with the information [5]. Visualizations have, thus far, mostly been effective in a more structured or hierarchical form [6, 7, 8]. Research into query tools, utilizing visualizations to a search a document corpus, has been conducted with positive results [9, 10]. Of course, visualizations are not without their challenges [11], but the results reported are mainly positive and outweigh these issues. Furthermore, advances in information retrieval algorithms (like the TREC conference [12, 13]), based on query terms, indirectly constantly improve the tools that use the results themselves.

2.2. Assistive Tools for Document Selection

In this section we present the tools themselves, outlining the key findings and capabilities of each one. We also present tools which contain common attributes clustered into common themes.

2.2.1. ThemeScapes

Wise et al, implement a visualization technique by employing spatial representations of large document sets [14]. Their aim was to create a visualization that may then be visually browsed and analysed in ways that avoid language processing and that reduce

the analysts' mental load". In their research, they used Themescapes (See Figure 3) "abstract, three dimensional landscapes of information that are constructed from document corpora" and Galaxies "displaying cluster and document interrelatedness" to present the notion of document similarity. Although there was no formal user tests reported on the work provides insights of 'analysts' using the tool giving feedback of reduced time spent looking for relevant material. The users also report using the tool not just for document discovery but also for identifying document relationships (even if this is not a primary function of Themescapes).

2.2.2. VKB

Another interpretation of a collected body of materials is presented by Marshall et al [20]. In this research, a spatial hypertext tool is presented which allows information seekers to interpret results from documents and identify the structure of the document set. This is made feasible by the creation of objects, composites and collections, and allowing relationships to be defined. Building upon this early work, Shipman et al, created the Visual Knowledge Builder (VKB – See Figure 3) [15, 16, 17]. VKB supports the "incremental visual interpretation of information". This tool was thoroughly utilised for collaborative efforts on shared information space. Similarly, a prototype tool called SketchTrieve, was also created to assist information and document triage [18]. SketchTrieve, was based on a conceptual model which followed the pattern: select the services you need, connect them, press Run, and results will be displayed.

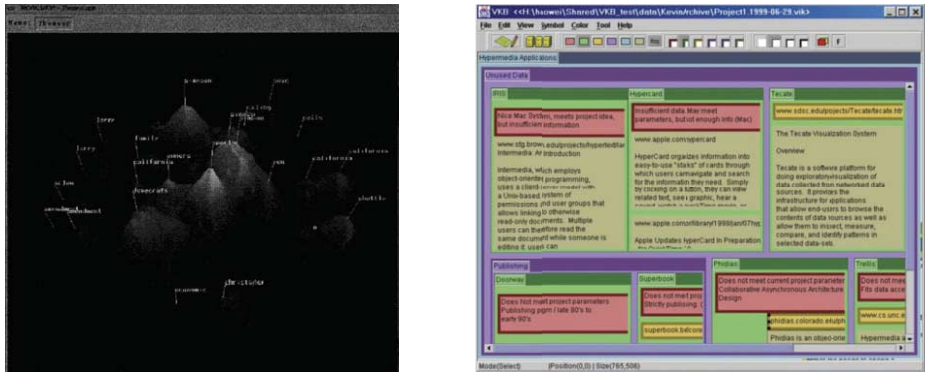


Figure 3. (Left) Themescapes and (Right) Visual Knowledge Builder Tool (VKB)

2.2.3. nSpace and TRIST

Another information visualization tool, created specifically for information triage, is TRIST (The Rapid Information Scanning Tool) [19] (See Figure 4). TRIST is built on the analytical environment nSpace [20] and allows the "rapid scanning over thousands of search results in one display, and includes multiple linked dimensions for result characterization and correlation". TRIST allows for the information seeker to compare queries and find documents that are more tailored to their need. By doing this,

document triage is informed by information that would have otherwise have taken multiple steps to achieve, all within one environment. Matching query terms to document content, like TRIST's attempt is important for information seekers. It helps them to relate their need to potentially relevant parts within a document. It is often hard however to locate the areas of the document which contain the query terms expressed by the user. A search engine will usually utilise the query terms in an information retrieval algorithm. Beyond that, there is usually no connection for the user, between the terms typed and the documents presented. Some users will use the Ctrl-F feature to find their terms within a document, but this is rarely the case [21]. It is evident that a more effective way to communicate the system's relevance decisions to the information seeker is needed. One way is to match up the query terms to areas within the documents.

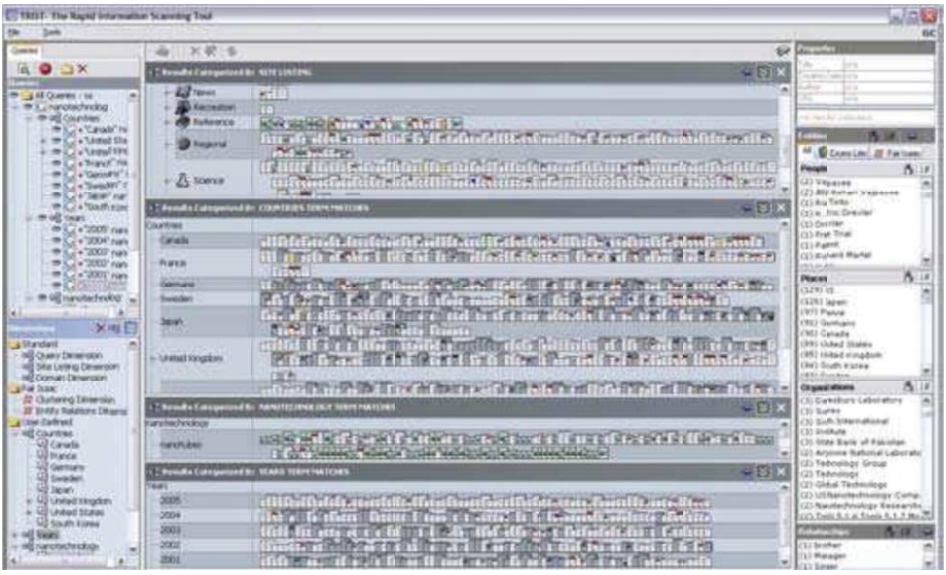


Figure 4. TRIST: The Rapid Information Scanning Tool

2.2.4. Opportunistic Search Tools

Currently, query terms are the established means by which an information seeker can make a request to a search engine. Directed browsing strategies can be assisted by several methods explained above using these terms, or variations of these, formulated by the user. Opportunistic search however, is also a big part of the information seeking process. It requires the triage of information in a less structured way. As it is becoming evident that "keyword and hypertext cannot support all these new tasks well" more opportunistic and exploratory systems are being researched [22]. One such software tool uses Semantic fisheye views (SFEV's) to browser over collections with different metrics [23, 24] (See Figure 5). A similar approach was also implemented by Cockburn et al, this time, using space filling thumbnails with a zooming action to allow better space real estate [25]. Screen real estate is one of the limitations that challenge the above prototypes. The question asked by Bae et al was whether different display types,

would have an effect on the way users perform document triage [15]. In their findings they report how there are more transitions using multiple displays rather than a single display "Additionally, users evaluated documents more by reading their contents and less often relied solely on metadata. Users spent more time reading and interacting with documents that they valued". This corresponds with the finding that reading time correlates with assessing document value in the triage field [26].

Figure 5. Semantic Fisheye Views

2.2.5. Using Task Bars

A common approach to supporting users' triage activities is by enriched visual interfaces using scroll bars (or any bars representing the document length). Two software tools, FindSkim and ProfileSkim (See Figure 6), created visualization in the task bars, to indicate the location of query terms in a document [27, 28, 29]. ProfileSkim, also added bar charts to allow the user to find heavily populated areas, query terms wise, within the document. A similar basis was used by Donald Byrd, who used colour and term highlighting scrollbars [30] and Schwartz et al who used term distribution visualizations [31]. The argument for making use of text structure when retrieving from full text documents, has also been investigated by Marti A. Hearst, and a prototype "a visualization paradigm, called TileBars" is presented to aid the information seeker [32]. The same information and principle as Harper et al was implemented with some additions, such as snippets for reading the results before navigating towards the related area. This method was favourable with participants. Query term matching has also been used in SmartFind (See Figure 6), another hybrid Ctrl/Cmd-F tool which uses Term Frequency x Inverse Document Frequency (TFIDF) algorithms within a document to provide potentially significant document areas to the information seeker [21].

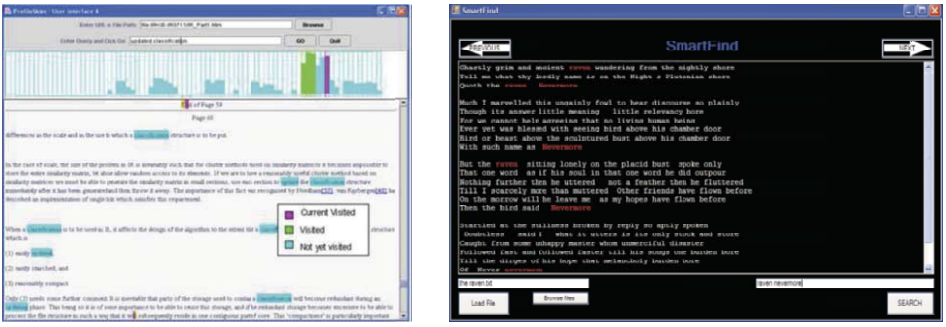


Figure 6. Profile Skim 2 Tool (left) and the SmartFind Tool (right)

2.2.6. TriDoc

TriDoc is a bespoke document triage tool which combines the high level results list view of document results with within-document scanning and information searching [33]. Currently, there are two interfaces supported by TriDoc. Both prototypes are hosted in a single-screen interface that integrates surrogate as well as within-document views; as well as snippets of individual sections of the document, combined with a full-text reading pane (See Figures 7 and 8). This approach follows the research not on visualization presentation but on the visual attention of users; a bottom approach unlike the other presented tools. The interface allows for a 'natural' linear type scanning of the document contents to happen in a non-linear fashion and minimizes scrolling.

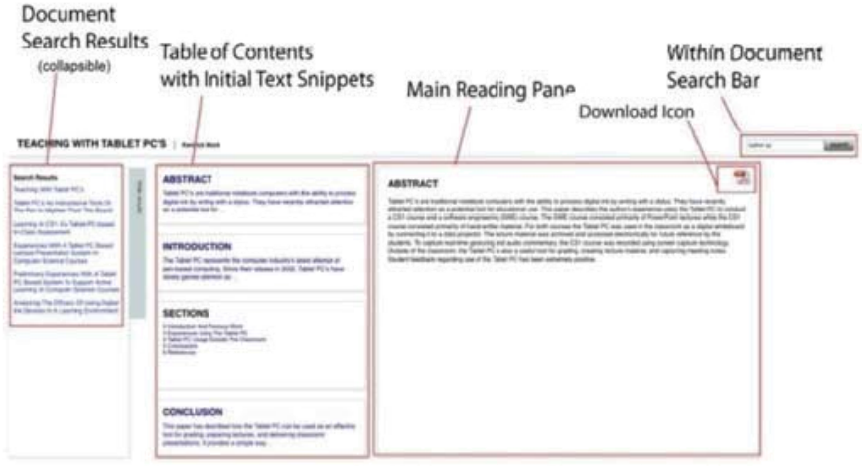


Figure 7. TriDoc Interface 1

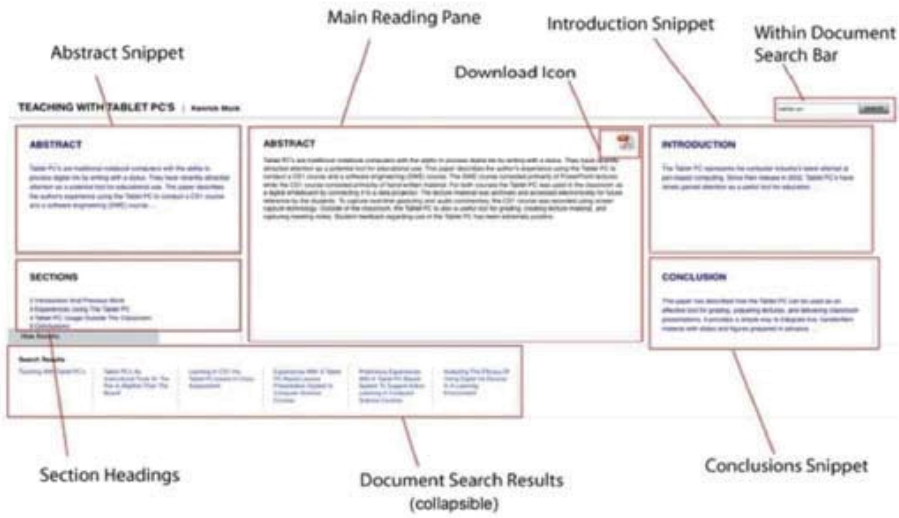


Figure 8. TriDoc Interface 2

TriDoc accommodated faster triage from the users test and received positive feedback from the information seekers (note: one interfaces received a higher rating than the other). TriDoc is an ongoing project which is currently under development (2015), unlike many of the reported tools in this paper.

3. Conclusions and Future Directions

After looking at the individual key tools which were created specifically for document selection / triage purposes we are able to make some inferences in terms of their goals, their similarities and differences. We can begin to deduce the effect on user behavior that these tools have and therefore begin to produce guidelines and understandings for designers and developers on these, similar to those by Mavri et al [34]. We were also are able to comments on the implementation of these tools within commercial repositories for electronic publications.

Most tools reported on for supporting document triage use a visualization approach to present representations of the information, specifically at the highest level of triage; namely, that of the surrogate view. Interestingly, we note that very few of the tools we report on consider the individual document feature important to the users. Furthermore, each tool focuses on matching the search terms inputted by the user rather than providing representations such as document structure. This denotes a reliability on the information retrieval engine, sometimes at the cost of the manual process which occurs after by the information seeker. While, from the reported data, there is a clear improvement regarding triage performance measurements, such as time or accuracy in locating relevant information, we recognize room for further improvement at the post automatic retrieval and presentation stage. There has thus far been limited research into the actual visual attention and processes of information seekers performing within-document triage; the second stage of triage process. Furthermore, most of the findings were taken from subjective feedback rather than empirical quantitative findings. Using

the research as a theoretical foundation, we encourage tools which give emphasis on the visual attention.

Academic document searching has, until now, not been given enough scrutiny in terms of interactive interfaces for document triage in the professional field of academics. We encourage and aim to produce future work which aims to build up knowledge on this topic through the interactive behaviors through an iterative user centered design approach, rather than a waterfall model for development of the tools. A primary goal is to set a foundation for standardising the creation and evaluation of such interfaces.

References

[1] G. Buchanan, F. Loizides, Investigating document triage on paper and electronic media, In *Procs. of the European Conf. on Research and Advanced Technology for Digital Libraries*, **35** (2007), 416–427.

[2] F. Loizides, G. Buchanan, Towards a Framework for Human (Manual) Information Retrieval, In *Multidisciplinary Information Retrieval,* Springer, Berlin Heidelberg, 2013, 87–98.

[3] A. Adams and A. Blandford, Digital libraries in academia: Challenges and changes, In *Proceedings of the 5th International Conference on Asian Digital Libraries: Digital Libraries: People, Knowledge, and Technology*, 2002, 392–403.

[4] A. Adams and A. Blandford, Digital libraries' support for the user's 'information journey', In *Proceedings of the 5th ACM IEEE-CS joint conference on Digital libraries*, 2005, 160–169.

[5] P. R. Keller and M. M. Keller, *Visual Cues: Practical Data Visualization,* IEEE Computer Society Press, Los Alamitos, CA, USA, 1994.

[6] G. G. Robertson, S. K. Card, and J. D. Mackinlay, Information visualization using 3d interactive animation, *Communications of the ACM* **36**(4) (1993), 57–71.

[7] R. Spence, *Information Visualization: Design for Interaction* (2nd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2007.

[8] E. Tufte, *Envisioning information,* Graphics Press, Cheshire, CT, USA, 1990.

[9] R. R. Korfhage. To see, or not to see, is that the query? In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '91, ACM, New York, NY, USA, 1991, 134–141.

[10] A. Spoerri, Infocrystal: a visual tool for information retrieval & management, In *Proceedings of the second international conference on Information and knowledge management*, CIKM '93, ACM, New York, NY, USA, 1993, 11–20.

[11] C. Chen, Visual spatial thinking in digital libraries – top ten problems, In *Joint Conference in Digital Libraries*, 2001.

[12] C. Buckley, G. Salton, J. Allan, and A. Singhal, Automatic Query Expansion Using SMART: TREC 3, In *Third Text REtrieval Conference* (TREC-3), 1994, 69–80.

[13] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, 2005

[14] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, Visualizing the non-visual: spatial analysis and interaction with information from text documents, In *Proceedings of the 1995 IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, USA, 1995, 51–58.

[15] F. Shipman, R. Airhart, H. Hsieh, P. Maloor, J. M. Moore, and D. Shah, Visual and spatial communication and task organization using the visual knowledge builder, In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '01, ACM, New York, NY, USA, 2001, 260–269.

[16] F. Shipman, J. M, Moore, P. Maloor, H. Hsieh, and R. Akkapeddi, Semantics happen: knowledge building in spatial hypertext, In *Proceedings of the thirteenth ACM conference on Hypertext and hypermedia*, HYPERTEXT '02, ACM, New York, NY, USA, 2002, 25-34.

[17] F. M. Shipman, III, H. Hsieh, P. Maloor, and J. M. Moore, The visual knowledge builder: a second generation spatial hypertext, In *Proceedings of the 12th ACM conference on Hypertext and Hypermedia*, HYPERTEXT '01, ACM, New York, NY, USA, 2001, 113-122.

[18] D. G. Hendry and D. J. Harper, An informal information-seeking environment, *J. Am. Soc. Inf. Sci.* **48** (1997), 1036–1048.

[19] D. Jonker, D. Schroh, B. Wright, P. Proulx, and B. Cort, Information triage with trist, In *Conference on Intelligence Analysis*, 2005.
[20] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort, *Advances in nSpace The Sandbox for Analysis*. McLean, VA, 2005.
[21] F. Loizides and G, Buchanan, The myth of find: user behaviour and attitudes towards the basic search feature. In *Joint Conference on Digital Libraries*, 2008, 48–51.
[22] D. Bryan and A. Gershman, Opportunistic exploration of large consumer productspaces. In *Proceedings of the 1st ACM conference on Electronic commerce*, EC '99, 1999, 41–47, New York, NY, USA, ACM.
[23] P. Janecek and P. Pu, Opportunistic search with semantic fisheye views. In *Web Information Systems* WISE 2004, volume 3306, Springer Berlin / Heidelberg, 2004, 668–680.
[24] P. Janecek and P. Pu, An evaluation of semantic fisheye views for opportunistic search in an annotated image collection, *International Journal on Digital Libraries* **5** (2005), 42–56.
[25] A. Cockburn, C. Gutwin, and Jason Alexander, Faster document navigation with spacefilling thumbnails. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, 2006, 1–10, New York, NY, USA, ACM.
[26] P. K. Chan, *A non-invasive learning approach to building web user profiles*, 1999.
[27] D. J. Harper, S. Coulthard, and Sun Yixing, A language modelling approach to relevance profiling for document browsing, In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, JCDL '02, New York, NY, USA, ACM, 2002, 76–83.
[28] D. J. Harper, I. Koychev, and Y. Sun, Query-based document skimming: a user-centred evaluation of relevance profiling, In *Proceedings of the 25th European conference on IR research*, ECIR'03, 200 377-392, Berlin, Heidelberg, Springer-Verlag.
[29] D. J. Harper, I. Koychev, Y. Sun, and I. Pirie, Within-document retrieval: A user-centred evaluation of relevance profiling. *Information Retrieval* **7** (2004), 265–290.
[30] D. Byrd. A scrollbar-based visualization for document navigation. In *Proceedings of the fourth ACM conference on Digital libraries*, DL '99, New York, NY, USA, ACM, 1999, 122–129.
[31] M. Schwartz, C. Hash, and L. M. Liebrock, Term distribution visualizations with focus+context, In *Proceedings of the 2009 ACM symposium on Applied Computing*, 2009, 1792–1799.
[32] M. A. Hearst, Tilebars: visualization of term distribution information in full text information access, In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1995, 59–66.
[33] F. Loizides, Thomas Photiades, Aekaterini Mavri, and Panayiotis Zaphiris, On Interactive Interfaces for Semi-Structured Academic Document Seeking and Relevance Decision Making. *New Rev. Inf. Networking* **19**(2) (2014), 67–95.
[34] A. Mavri, F. Loizides, T. Photiadis, and P. Zaphiris, We have the content… now what?, *Information Design Journal* **20**(3) (2013), 247–265.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-83

Measuring the Usage of Repositories via a National Standards-based Aggregation Service: IRUS-UK

Ross MACINTYRE[a],1, Jo ALCOCK[b], Paul NEEDHAM[c], Jo LAMBERT[a]
[a]*Jisc, United Kingdom*
[b]*Evidence Base, Birmingham City University, United Kingdom*
[c]*Cranfield University, United Kingdom*

Abstract. Many educational institutions have repositories for research outputs. The number of items available through institutional repositories is growing, and is expected to continue to do so due to requirements for outputs from public-funded research to be open access. But how much usage are institutional repositories and their individual items getting? The Jisc-funded service IRUS-UK is designed to help institutions understand more about the usage of their institutional repositories. IRUS-UK collects raw usage data from participating repositories and processes these into COUNTER-compliant statistics. This provides repositories with comparable, authoritative, standards-based data and opportunities for profiling and benchmarking. It enables institutions to run reports at both repository level (e.g. total download figures) and at item level. IRUS-UK utilises a robust, multistage ingest process, validating data, stripping out robot and unusual accesses, and filtering out double clicks, to transform raw usage data into COUNTER-compliant statistics. IRUS-UK currently has data from 83 UK institutional repositories (using Eprints, DSpace and Fedora software) and has recorded over 35 million downloads since July 2012. The data from IRUS-UK can be used to provide information for management reporting, for usage monitoring, and for external reporting. Data can be viewed within the online portal, downloaded for further analysis, or harvested using the SUSHI service (NISO Z39.93). IRUS-UK is also working with and contributing to other groups and initiatives involved in a range of activities relating to usage statistics. These include: the Distributed Usage Logging/CrossRef DOI Event Tracker Working Group, OpenAIRE2020 and COAR Working Group.

Keywords. COUNTER, usage statistics, repositories, benchmarking, altmetrics

1. Introduction

Institutional repositories (IRs) have attracted much attention over the last decade and there has been considerable interest in the growing number of repositories and their contents. However, until now there has been a lack of comprehensive information about usage of the resources hosted by IRs.

Although most IRs provide statistics that purport to show usage, you can't count on them – not entirely. Different types of software – out-of-the-box, add-ons, Google Analytics and other third-party solutions – process raw usage data in different ways,

[1] Corresponding author. Jisc, J14A Sackville Building, The University of Manchester, Sackville Street, Manchester M1 3BB, UK; E-mail: ross.macintyre@jisc.ac.uk.

making it impossible to compare like for like across repositories. There is currently no agreed standard to measure usage across repositories.

IRUS-UK, funded by Jisc, is a national aggregation service that responds to this problem by providing standards-based statistics for all content downloaded from participating UK IRs. The service collects usage data from participating repositories, processes the data into COUNTER-compliant [1] statistics and then presents statistics back to originating repositories to be used in a variety of ways. It provides opportunities for benchmarking at a national level by enabling UK IRs to access and share comprehensive and comparable usage data. Some of the underlying technical principles were taken forward from the substantial work done in LANL's Mesur project [2] and in the European Knowledge Exchange Working Group on Usage Statistics [3].

IRUS-UK now provides a nationwide view of the majority of the UK's institutional repositories use, helping demonstrate the importance and value of IRs. There is also potential for the service to act as an intermediary between UK repositories and other agencies.

IRUS-UK is one of a number of Jisc-funded repository and infrastructure services which aims to increase the cost effectiveness of repositories of open access (OA) literature. The service was developed by a consortium involving Mimas (now part of Jisc itself), Cranfield University and Evidence Base at Birmingham City University. The team is also responsible for the development of the Journal Usage Statistics Portal (JUSP) [4], which provides a 'one-stop shop' for libraries to view, download and analyse their journal usage reports from multiple publishers. Consequently, the team members have significant skills and expertise in managing and developing usage statistics products and services.

2. Background to Development of the Service and PIRUS2

IRUS-UK builds on the work of the successful Jisc-funded PIRUS2 project [5], which demonstrated how COUNTER-compliant article-level usage statistics could be collected and consolidated from publishers and institutional repositories. The primary aims and objectives of PIRUS2 were to assess the feasibility of and develop the technical, organizational and economic models for the recording, reporting and consolidation of usage of journal articles hosted by publishers, institutional repositories and subject repositories.

PIRUS2 achieved its aims by delivering a prototype statistics aggregation service, comprising:

- usage data and statistics from publishers and institutional repositories
- a practical organizational model based on co-operation between data processing suppliers
- data management and auditing services that meet the requirement for an independent, trusted and reliable service
- an economic model that provides a cost-effective service and a logical, transparent basis for allocating costs among the different users of the service.

PIRUS proposed the establishment of a global central clearing house (CCH) to deliver such a service. Unfortunately, it became clear from a survey conducted at the end of the

project that the majority of publishers were not, largely for economic reasons, yet ready to implement or participate in such a service. Nevertheless, this work has been used to inform the development of a COUNTER Code of Practice for Articles. Furthermore, the project found that usage of articles hosted by institutional repositories was substantial. As a result of this, a second set of aims and objectives emerged: to develop the technical, organizational and economic models for the standardized recording and reporting of usage at the individual item level – regardless of content type – for items hosted by institutional repositories and subject repositories (IRUS).

To support these extra objectives, a secondary demonstrator service was developed, which focused solely on repositories. It revealed that significant numbers of other item types (theses, conference papers, reports, etc.) were also being regularly downloaded. This additional work ultimately led to the establishment of IRUS-UK, which adheres to both the COUNTER Codes of Practice (Articles & e-Resources).

3. IRUS-UK Usage Statistics Portal

The IRUS-UK service provides a single gateway for libraries to access statistics relating to usage events recorded within their IR. In particular, it contains COUNTER-compliant usage statistics for each participating UK higher education institution's IR (institutional repository). The service, underpinned by a MySQL database, comprises:

- A web user interface (written in PHP)
- Downloadable reports
- An initial API
- A SUSHI (Z39.93) server

All institutional members of the UK Access Management Federation [6], whether or not their institutional repository is an IRUS-UK participant, can log in to the IRUS-UK portal and view the statistics and reports listed below.

3.1 Summary Reports

IRUS-UK provides a number of summary tables and reports which give an overview of downloads from our participating repositories. You can see:

1. An overall summary of downloads for all participating repositories.
2. Total number of downloads for each individual participating repository.
3. A breakdown of repository participation and number of downloads by selected countries in the UK (England, Scotland, Wales).
4. A breakdown of repository participation and number of downloads by platform used (DSpace, Eprints or Fedora).
5. Numbers of each type of item downloaded and number of downloads of each type of item for all participating repositories.
6. Numbers of each type of item downloaded and the number and percentage for each item type which have DOIs available in the metadata that we harvest.
7. An analysis of the data ingest process for each repository showing raw data, exclusions for robots and double clicks, and the resulting number of downloads showing in IRUS-UK.

3.2 Usage Reports

The usage reports available include:

1. *'Item Report 1'* provides the number of successful item download requests by month and repository identifier for a selected repository.
2. *'Item Report 2'* provides the number of successful item download requests by month and item type for a selected repository.
3. *'Article Report 4'* provides the number of successful article downloads by month for participating repositories. The report can be filtered to limit the results to a selected journal or repository. It can be run for an individual month or over a number of months
4. *'Book Report 1'* provides the number of successful book downloads by month for a selected repository. It can be run for an individual month or over a number of months.
5. *'Book Report 2'* provides the number of successful book section downloads by month for a selected repository. It can be run for an individual month or over a number of months.
6. *'Electronic Thesis or Dissertation Report 1'* provides the number of successful thesis or dissertation download requests by month and repository identifier for a selected repository. For each thesis or dissertation, it shows the item URL, EThOS ID (British Library's Electronic Theses Online Service) [7] if available, title, author and total downloads by month and in total for the period selected. It can be run for an individual month or over a number of months.
7. *'Journal Report 1'* provides the number of successful Full-Text Article Requests by Month and Journal for participating repositories. The report can be filtered to limit the results to a selected journal or repository. It can be run for an individual month or over a number of months.
8. *'Repository Report 1'* enables you to view the number of successful item downloads by month for all participating repositories. The report can be filtered to limit the results to a selected item type, Jisc Band and/or Country.

3.3 Item Type Usage Reports

We map the hundreds of different item types used by our participating repositories to a core set of 25 item types*: Art/Design Item; Article; Audio; Book; Book Section; Conference Papers /Posters; Conference Proceedings; Conference or Workshop Item – Other; Dataset; Exam Paper; Image; Learning Object; Moving Image; Music/Musical Composition; Other; Patent; Performance; Preprint; Report; Show/Exhibition; Text; Thesis or Dissertation; Unknown; Website; Working Paper*.

All the original item types are stored so that items can be subsequently remapped if necessary. The choice of the 25 item types was informed by a major piece of work [8] examining both metadata guidelines for repositories and actual use of item types. In the IRUS-UK portal you can see for each item type the number of items downloaded and the number of downloads.

3.4 Search for Usage of an Individual Item

One can search for words or phrases in the title or author for all repositories or for an individual repository and for all item types or a specified item type. Search results include basic metadata, a link to the item in the host repository, numbers of downloads and additional statistics relating to the item.

3.5 Check Items with DOIs

IRUS-UK extracts DOIs from downloaded item metadata and provides two tables:

1. A summary for each item type of the number and percentage that have DOIs across all repositories.
2. A breakdown of article DOI availability by repository.

3.6 Robot Usage and Double Clicks

In order to produce COUNTER-compliant usage statistics, IRUS-UK excludes downloads by robots and double clicks on individual items. A table provides an analysis of the ingest process for each participating repository. We have a position statement on the treatment of robots and unusual usage [9] and are undertaking further work to refine this process.

3.7 Report Formats

The reports are made available both for human use and direct machine to machine use:

- Each report can be viewed in a web page in the portal or downloaded for use locally as MS-Excel/CSV files.
- The reports are available via the SUSHI protocol for incorporation into local institutional ERMs, or for automatic gathering for use in other national/global services.

3.8 Ingest Scripts

The ingest scripts, based on the original scripts devised by PIRUS2, have been significantly enhanced and refined through several iterations, adding:

- daily granularity instead of the original monthly granularity
- ‘separation of concerns’ to make the ingest more robust, and to simplify development and maintenance of the scripts.
- improved validation of incoming data
- additional filtering of robots and abnormal usage over and above the minimum specified by COUNTER

Data received for participating repositories gets stored in daily log files. The log for any given day is usually processed the following day.

There is currently a three step daily ingest process:

1. A Perl script parses the logs; processes entries from recognised IRs; sorts and filters entries following COUNTER rules to remove robot entries and double-clicks; filters entries using additional IRUS-UK filters; consolidates raw usage data for each item into daily statistics; and outputs to an intermediate file.
2. A Perl script processes the intermediate file output from Step 1; using the OAI identifier associated with item, it looks up each item against the Item Authority table in the IRUS DB to see if it is already known to the system; if a known item, it retrieves the existing IRUS Item Identifier; if the item is new – as yet unknown to IRUS – the script adds a stub-entry to the Item Authority table – minting a new IRUS Item Identifier and adding the repository identifier, platform and OAI identifier to the table with the rest of the metadata set to 'unknown' at this stage; finally, the script adds the download statistics associated with each IRUS Item Identifier to the Daily Statistics table.
3. A Perl script obtains the "unknown" metadata for new items: it queries the DB to find the 'known unknowns' using the OAI identifiers; issues OAI-PMH GetRecord calls to retrieve OAI_DC metadata; parses the OAI records; updates the metadata – Title, Author, Item Type, etc., in the Item Authority Table in the DB; and additionally maps the Item Type, as given by the source repository, to a smaller (more manageable list) of IRUS Item Types.

In addition to the daily ingest scripts, additional scripts are run every few days to add journal information to article records and a further script is run at the end of each month to consolidate the Daily Statistics for that month into a Monthly Statistics table.

3.9 Robots and Unusual Usage

The starting point for eliminating robots and machine accesses from the raw usage data being collected was the COUNTER robots exclusion list. The list contains a set of regular expressions (regexes) of User agents to exclude and is described by COUNTER as a 'minimum' requirement. However, as the service has taken on-board more repositories, it has become obvious that the list is not comprehensive enough to exclude all robots and unusual usage. The problem became most obvious – and acute – when the London School of Economics (LSE) joined IRUS-UK and apparent download figures rose dramatically. At that point, analysis of the data identified a number of further exclusions not in the COUNTER list, including half a dozen user agents and two IP ranges used by Baidu Spider (which User Agent exclusion would not identify). Consequently, the service supplemented the COUNTER exclusions with:

- a set of additional IRUS-UK filters employing the newly identified User Agents and IP ranges.
- an additional check to exclude data where a single IP has exceeded a daily threshold for downloads –unless identified as a legitimate source of high download levels, e.g. an organisational proxy server.

These filters do work reasonably well, but the team was still convinced that more could be done to eliminate even more suspect usage. So, work was commissioned jointly by IRUS-UK and COUNTER to devise an 'adaptive filtering system' – a set of algorithms that will allow the service to dynamically identify and filter out unusual usage/robot activity. The work was undertaken by Information Power Limited, who have supplied a

report and an initial set of scripts of use by IRUS-UK. The results of that work will be assessed, tested, refined and applied to the service in the next phase of development. The information has been shared with the community, via COAR Interest Group 'Usage Data and Beyond' [10] and has led to the formation of a COUNTER Working Group on Robots.

3.10 Updated Tracker Protocol Specification

The specification for this is quite brief and straightforward:

- When a user clicks on a link to (i.e. downloads) a file from a Repository with the tracker protocol in operation, an OpenURL log entry is sent to a remote server for further processing.
- The OpenURL log entry should be based on a subset of the NISO OpenURL 1.0 standard KEV ContextObject Format. The OpenURL string must be URL encoded, with key-value pairs separated by '&'.

The initial specification used by IRUS-UK – based on PIRUS2 work - was designed to work at 'item' level. This is quite adequate for most items which contain a single file; however, there are a proportion of items that may have multiple files associated with them, i.e. the work is divided into chapters or contains appendices or other supplementary materials. In order to accommodate such items and allow reporting at a finer granularity in a future iteration of the service, the team has devised an updated specification containing an extra metadata element – the fileURL – to be transmitted from repositories to the IRUS-UK server (Table 1).

Table 1. Tracker Protocol

Element	OpenURL Key	OpenURL Value (example)	Notes
OpenURL version	url_ver	Z39.88-2004	Identifies data as OpenURL 1.0. String constant: Z39.88-2004 (Mandatory)
Usage event datestamp	url_tim	2010-10-17T03%3A04%3A42Z	Date/time of usage event (Mandatory)
Client IP address	req_id	urn:ip:138.250.13.161	IP Address of the client requesting the article (Mandatory)
UserAgent	req_dat	Mozilla%2F4.0+%28compatible%3B+MSIE+7.0%3B+Windows+NT+5.1%3B+Trident%2F4.0%3B+GoogleT5%3B+.NET+CLR+1.0.3705%3B+.NET+CLR+1.1.4322%3B+Media+Center+PC+4.0%3B+IEMB3%3B+InfoPath.1%3B+.NET+CLR+2.0.50727%3B+IEMB3%29	The UserAgent is used to identify and eliminate, by applying COUNTER rules, accesses by robots/spiders (Mandatory)
Item OAI identifier	rft.artnum	oai:dspace.lib.cranfield.ac.uk:1826/936	(Mandatory)
FileURL	svc_dat	https://dspace.lib.cranfield.ac.uk/bitstream/1826/936/4/Artificial_compressibility_Pt2-2005.pdf	(Mandatory)
HTTP Referer	rfr_dat	http://www.google.co.uk/url?sa=t&rct=j&q=http%20referer&source=web&cd=4&sqi=2&ved=0CeoQFjAD&url=http%3A%2F%2Fwww.whatismyreferer.com%	(Mandatory) The HTTP header field that identifies the address of the webpage (i.e. the URI) that

		2F&ei=zIBCU6fbEoOqhQf67YcwBg&usg=AFQjCNFt-KmqneTZfEb6OxjPZlD4ogiJcQ&sig2=wZJYkoWgNScNjgxRbRs29w&bvm=bv.64125504,d.ZWU	linked to the resource being requested. The 'HTTP Referer' is used to help identify and eliminate accesses by robots/spiders.
Source repository	rfr_id	dspace.lib.cranfield.ac.uk	(Mandatory)

3.11 Tracker Code

An Eprints Tracker plug-in, developed by Eprints Services, for Eprints 3.2.x and 3.3.x. is available from the 'Eprints Bazaar' [11].

Patches are available for DSpace, developed by @mire, for versions 1.8.1, 1.8.2, 1.8.3, 3.1, 3.2, 4.1 and 4.2. The patches are available on request.

Fedora implementations require bespoke code to be developed by the repository implementers. Example implementations exist for Java and RubyGem for Hydra.

The team conducted a small scale trial involving the OAPEN Library [12], which runs on ARNO repository software. The software has been modified to include IRUS Tracker functionality and is successfully transmitting OpenURL messages about e-book downloads – demonstrating the ease of applying the technical solution and its transferability beyond institutional repositories.

Discussions have been initiated with Atira (now part of Elsevier) to see if it is feasible to add Tracker functionality to the PURE Portal software.

4. Benefits of a Shared Service and Community-driven Developments

In theory, every institution could produce its own COUNTER-compliant statistics for its repository. The rules for eliminating robot accesses and double-clicks and for counting downloads are not that difficult to understand or implement. However, there is more to COUNTER compliance than simply following the COUNTER Code of Practice. In order to become truly COUNTER compliant, it is necessary to go through a regular auditing process. By the time of registration, annual membership and report auditing fees are taken into account, this can potentially cost several thousands of pounds per year per IR. By collecting and processing download data into COUNTER statistics on behalf of IRs, IRUS-UK can substantially reduce these costs; in this scenario, only IRUS-UK itself needs to be audited, the individual IRs do not.

Additionally, IRUS-UK is in a position to act as an intermediary between UK IRs and other agencies, such as OpenAIRE [13], which has an interest in obtaining usage statistics for research outputs funded under the European Seventh Framework Programme (FP7). Having a single point of access to FP7 article statistics for the UK will be a lot easier to manage than collecting those statistics from all the relevant individual repositories.

IRUS-UK data can be used for a number of different purposes, some of which have been highlighted in a series of use cases. The use cases are summarised below.

4.1 Reporting to Institutional Managers

It is useful for institutional managers to understand the usage of items in the institutional repository. This might include, for example, obtaining high-level statistics of total downloads from the repository, or gaining an understanding of the items that have higher downloads. The usage statistics within IRUS-UK can be used to report on downloads for institutional managers. A user can find out the total downloads from each repository, or can use the focused reports for more granular information such as downloads by item (to help identify items receiving high numbers of downloads), downloads by item type, downloads of Electronic Theses and Dissertations, and number of downloads from all participating repositories, which can be filtered by item type, Jisc band, or country (or a combination).

4.2 Reporting to Researchers

Researchers are often interested in knowing the usage statistics of their items in the institutional repository; this could be for review purposes, for reporting to Research Councils, or just for curiosity. Additionally, a researcher may be involved in dissemination or publicity (e.g. a conference presentation) that refers to their research, and they may wish to see if this has resulted in an increase in the number of downloads of the item. In the IRUS-UK search a user can search for a specific item to report on. The result shows monthly downloads of the item (since a download was recorded in IRUS-UK) and daily downloads for the last month.

4.3 Benchmarking

Being able to benchmark institutional repository statistics is valuable, both with an institution's own data (to look at trends), and with other institutions (to allow comparisons to be made and to provide a wider context within which to interpret the performance of an institutional repository). The standardised COUNTER-compliant statistics available through IRUS-UK enable reliable benchmarking, both with an institution's own data for longitudinal analysis, and with other institutions within IRUS-UK. Users can view total downloads recorded by IRUS-UK for all participating repositories, or can look at monthly data for all participating repositories to look at trends. Repositories can also be filtered by different groupings (Jisc band or Country) or filter by item type (for example downloads of Articles) and filters can be combined.

4.4 Supporting Advocacy

The data within IRUS-UK can be used to support advocacy by sharing headline download figures from all participating repositories, reporting on an overall total number of downloads from a repository since joining IRUS-UK, showing monthly download figures (and trends), identifying items with high levels of downloads, gathering statistics on downloads of different item types within a repository, or sharing downloads for particular researchers or research areas. These statistics can then be used in a number of different ways including presentations, newsletters, blog posts, reports, social media, meeting updates, etc. These may be focused specifically on the performance of a repository, or, more broadly, on Open Access advocacy.

5. Conclusion

IRUS-UK provides a usage statistics service for UK repositories, based on the COUNTER standard, which enables them to expose credible, authoritative and trustworthy usage figures for item downloads, on the same basis as – and therefore comparable with – the majority of publishers, in an extremely cost-effective manner.

By providing a nationwide view of UK repository usage, it also benefits national organizations such as Jisc and SCONUL, and offers opportunities for benchmarking as well as the ability to act as an intermediary between UK repositories and other agencies. We hope that IRUS-UK will act as a model which can be adopted in other countries and regions around the world.

Finally, it may help to inform the current debate, taking place in the absence of reliable or comprehensive usage data, about the value of repositories and their place and significance in the dissemination of OA research literature.

References

[1] COUNTER Code of Practice: http://www.projectcounter.org/code_practice.html (accessed 18 May 2015).
[2] J. Bollen, H. Van de Sompel, An Architecture for the Aggregation and Analysis of Scholarly Usage Data, http://arxiv.org/abs/cs/0605113 (accessed 18 May 2015).
[3] Knowledge Exchange: http://www.knowledge-exchange.info/ (accessed 18 May 2015)
[4] JUSP: https://www.jusp.mimas.ac.uk/ (accessed 18 May 2015).
[5] PIRUS2 Final Report: http://www.projectcounter.org/News/Pirus2_oct2011.pdf (accessed 18 May 2015).
[6] UK Access Management Federation: http://www.ukfederation.org.uk/ (accessed 18 May 2015)
[7] British Library eTheses Online Service: http://ethos.bl.uk/ (accessed 18 May 2015)
[8] IRUS-UK Item Type Mappings: http://www.irus.mimas.ac.uk/help/toolbox/IRUS_item_type_report_v3.3.pdf (accessed 18 May 2015)
[9] Position statement on the treatment of robots and unusual usage: http://www.irus.mimas.ac.uk/news/IRUS-UK_position_statement_robots_and_unusual_usage_v1_0_Nov_2013.pdf (accessed 18 May 2015)
[10] Confederation of Open Access Repositories Interest Group: Usage Data and Beyond: https://www.coar-repositories.org/activities/repository-interoperability/usage-data-and-beyond/ (accessed 18 May 2015)
[11] Eprints Bazaar http://bazaar.eprints.org/ (accessed 18 May 2015)
[12] OAPEN-UK: http://oapen-uk.jiscebooks.org/ (accessed 18 May 2015).
[13] OpenAIRE: http://www.openaire.eu/ (accessed 18 May 2015).

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-93

Open Access in Scientific Communication: Bulgaria's Current OA Policies within the International Context

Aleksandar DIMCHEV[a,1] and Rosen STEFANOV[a]
[a] *Sofia University "St. Kliment Ohridski"*

Abstract. The following report aims to examine the current tendencies in the field of Open Access (OA) publishing. Ever since its inception in the 1990s, Open Access has been a topic of interest in regards to its potential to be viable alternative to more traditional publishing models for scientific information and communication. This report examines OA within the context of both international practices (with a focus on European experience in the Open Access field), as well as looking at the current tendencies regarding this publishing model in Bulgaria.

Keywords. Open Access, OA, scientific communication, Bulgaria, DRIVER, OpenAIRE, Sofia University "St. Kliment Ohridski", Bulgarian Academy of Sciences

1. Open Access and the International Scientific Community

The matter of communication has been one of ever increasing importance for the scientific community in the past few years. The advent and advancement of the new communication technologies (the World Wide Web, in particular) have enabled the rapid spread and sharing of information throughout the world. This trend also extends to the international scientific community. Through the Web, scientists are able to gain access to a rich vault of peer-reviewed information that can aid them in their respective research fields. In the past this was done mostly through paid electronic journals and databases. This all changed however with the advent of Open Access.

The term "Open Access" (or OA in short) is used to describe methods of information publication that are not linked to paid subscriptions. The definition of what OA is can be found in the comment of Charles Bailey made during the 2001 meeting in Bucharest. He states that *"Open Access online has to be that literature, which scientists provide to everyone without the expectation that they will be paid for it"* [1]. The beginnings of the Open Access movement can be traced back to the early 1990s. In 1991, Paul Ginsparg created the arXiv electronic publications repository which archives e-documents within fields such as nuclear physics, mathematics and more [3]. In the years following the creation of Ginsparg's electronic archive, Open Access became a topic of great interest for both the scientific community as well as librarians. Even so, the early years of scientific communication via the conduit of the Web were focused on paid electronic journals and databases. These platforms and their respective publishers

[1] Corresponding Author. E-Mail: dimchev_uni@abv.bg

were considered the only reliable means of distributing and gaining access to peer-reviewed information. Several factors however helped Open Access gain additional support as an alternative publication model to the traditional paid platforms. These include [1]:

- The fast development of new information technologies.
- Exponential growth of the volume of data and publications in the field of scientific research.
- The constantly rising prices of scientific information (based on data acquired from the Association of Research Libraries in the USA, the prices of scientific journals have increased by 215% over the past 15 years which has had a serious effect on the budgets of research libraries).
- The rapid development of the scientific fields and the increasing need for acceleration of the process of providing access to research outcomes and achievements to scientists (with traditional distribution methods there is a delay in the cycle of research documents publication: a considerable time can pass between the submission of the initial manuscript and the publication of the finalized document).
- Dynamic changes in the publishing sector.
- The continuing trend of digitalization of research literature (over 95% of all renowned scientific journals have an electronic version, over 35% of all scientific monographs are published in an electronic form and so on).
- Changes to the methods of creation, preservation, access and distribution of information resources and products.
- The appearance of new channels that provide researchers with new ways with which they can organize, store and provide access to scientific facts and knowledge.
- The search for solutions that can overcome the "information based isolation inequality in science".
- The public's disagreement to having their taxes be used to pay scientific organizations "two times" (On one hand, they pay for the financing of research activities. On the other, they also pay for the buy-out of the results of said research, subsidized with government funds, in the form of information products and publications distributed through commercial publications and databases).

Given the influence of factors such as those mentioned above, the Open Access method of information distribution has become a viable alternative to more traditional commercial publication platforms. It allows the greater scientific community to freely share research information and results both in between itself, as well as the those members of the public who might be interested in their findings.

There currently exist several initiatives that aim to create commonly accepted frameworks for Open Access. Some of the more noteworthy examples of such initiatives are:

- **The DRIVER project** (Digital Repository Infrastructure for European Research): Initiated by the European University Association in 2008 and financed by the European Commission, the objective of this project is to foster the creation of institutionalized OA repositories in universities throughout Europe that serve as part of a greater network that can provide free scientific information from all possible research fields. In 2014, DRIVER merged with the OpenAIRE initiative (another Open Access repository project funded by the

European Commission) [7]. As of February 2015, DRIVER/OpenAIRE provides access to over 9, 4 million scientific publications (journal articles, dissertations, books, etc.) which are stored in 5,776 datasets derived from 580 repositories and OA journals. These 580 repositories are located in 38 European countries (Bulgaria does participate in DRIVER but does so with only 5 repositories: the digital archive of The Institute for Mathematics and Informatics – Bulgarian Academy of Sciences, Pensoft, New Bulgarian University, Medical University- Sofia and the Burgas Free University).

- **Digital Agenda: more open access to scientific information**: The initiative was started as an idea of the European Commission, Neelie Kroes (vice-president of Digital Agenda) and Máire Geoghegan-Quinn (the European Commissioner for Research, Innovation and Science for the period 2010-2014). The core concept of the project is that European researchers, engineers and entrepreneurs need to have quick and easy access to scientific information so they can stand on an equal footing with their international colleagues. A modern digital infrastructure can play a key role in making access to knowledge easier and can help in the formation of a unified European scientific environment.

The participation of the European Commission in both of the above initiatives shows that the idea of Open Access has become an integral component of the European information policy. In fact, Open Access plays an important role in the new Horizon 2020 research and innovation program of the European Union. This role is reflected in the Open Research Data Pilot which is an initiative with the goal of improving and maximizing access to and re-use of research data [7]. This is further reflected in the fact that Europe has, as of data from 2015, the largest number of registered OA repositories in comparison with the other parts of the world. This is reflected in Table 1.

Table 1. Registered OA repositories per continent (information relevant as of 2015)

Repositories per continent	Percentage of repositories in comparison with the rest of the world	Number of registered repositories
Europe	45.5%	1,241
North America	19.9%	543
Asia	18.4%	503
Other	**16.2%**	**441**
Total	**100%**	**2,728**

The next segment of this report will cover the current developments of the Open Access movement in the context of the Bulgarian academic community and its impact on scientific communication in the country.

2. Bulgaria's Experience in Open Access

It has to be noted that, in comparison with the above presented international experience, the current developments in regards to Open Access in Bulgaria are somewhat lacklustre. That is despite the fact that Open Access, as matter of interest, is the topic of several documents detailing the future developments of the Bulgarian scientific and educational sectors and that *"due attention will be given to creation and development of scientific networks in which scientific information, knowledge and technology can be*

shared freely" [1]. Unfortunately, as of 2015, these documents remain purely theoretical and little has been done to have Open Access become an integral part of Bulgaria's scientific and academic societies [4, 5, 6]. In the few cases where OA has become a practical reality in Bulgaria, this has mostly been thanks to individual projects and initiatives driven by singular organizations rather than through the aid of any government institutions. The Bulgarian Academy of Sciences (BAS) intends to create a network of OA centers with the idea that its Institute of Mathematics and Informatics will be the coordinating body for this nationwide system. The institute aims to provide support to both academic organizations, as well as individual researchers. The aim here is to achieve integration between BAS, the public and industrial sectors, and to have ties between notable scientific centers, universities and other educational institutions in the country be strengthened. These expectations are based around the future creation of Open Access repositories mainly in university libraries and scientific organizations. As of right now though, this network of academic repositories is fairly small (as can be seen in Table 2) as there are currently only 5 Bulgarian OA archives registered in OpenDOAR (Directory of Open Access Repositories).

Table 2. Open Access repositories in Bulgaria registered in OpenDOAR (information relevant as of 09.02.2015).

Organization	Purpose	Profile	Type of archived documents	Language(s) of the documents	Number of archived documents
Burgas Free University	Research documents of the BFU	Inter-disciplinary	Articles, conference papers, books	Bulgarian, English	460
Institute of Mathematics and Informatics- BAS	Research documents of the IMI	Inter-disciplinary	Articles, conference papers, books	Bulgarian, English	2,203
Pensoft Publishers	Repository for documents	Inter-disciplinary	Books	English	447
Medical University of Sofia- Central Medical Library	Research documents of the MU	Healthcare and medicine	Articles, dissertations, books, study materials	Bulgarian, English	574
New Bulgarian University	Research documents of the NBU	Inter-disciplinary	Articles, dissertations, books, study materials	Bulgarian, English	1,634

Another useful source of information regarding OA repositories in Bulgaria can be found in the **OpenAIRE "European Open Access Repositories Landscape"** tool [8]. The following table presents Bulgaria's standing in regards to the total number of Open Access documents published in comparison with several other European countries.

Table 3. OpenAIRE data regarding number of OA repositories/documents per country (information relevant as of 12.02. 2015)

Country	Number of repositories	Number of OA publications
United Kingdom	71	3,050,186
Germany	78	561,665
France	16	477,334
Bulgaria	**12**	**4,998**
Poland	9	27,012
Turkey	11	2,161
Hungary	5	4,076

As can be seen in Table 3, Bulgaria lags considerably behind some of its EU co-members in regards to pure volume of OA information available online. This information source also shows that a greater number of OA repositories do not always correlate with a greater number of OA documents (Poland has a significantly greater volume of published OA documents despite having less OA repositories than Bulgaria). A possible explanation for this can be found in the differences in document publication output rates in the various countries. Bulgaria, for example, has a fairly low publication output which could understandably also affect the number of OA documents published.

As for Open Access journals, Bulgaria ranks as 37th (from a total of 136 countries covered) in the world based on information from DOAJ (Directory of Open Access Journals) relevant as of 2015 [11]. Please note that the following table aims to compare Bulgaria not only in relation to the top ranking countries, but also those who are closer to it in terms of the number of available Open Access journals.

Table 4. Bulgarian ranking in DOAJ with regards to the OA journals available to the public.

Rankings	Country	Number of OA journals
1.	USA	1,237
2.	Brazil	952
3.	Great Britain	664
6.	Germany	341
9.	Romania	307
11.	Turkey	210
13.	France	175
24.	Croatia	93
25.	Serbia	88
30.	Czech Republic	71
32.	Russia	70
37.	**Bulgaria**	**45**
39.	Slovenia	41
40.	Greece	39

The above information provided in Table 4, indicates that the total number of OA journals in Bulgaria is fairly small in comparison with what is found in other countries in the world. Even so, these numbers are comparable with those found in other neighboring countries such as Greece and Slovenia.

There are several factors that can be connected with the current state of Open Access in Bulgaria. Some of the more notable of them are the following [5, 6]:

- The low level of financing of the science and research fields in Bulgaria (roughly 0.5% of the country's GDP, whereas the recommended norm in the rest of the EU is 3%).
- There exists a declared desire to have Open Access become part of the Bulgarian scientific society. Yet this desire is not backed by practical suggestions as to how an OA framework can be created in the country (i.e. the mostly theoretic state of the OA environment in Bulgaria).
- The idea of Open Access is still largely unknown to Bulgarian scientists. There is also their lack of determination to convert to a more digitally focused method of publishing their research results (as opposed to the more traditional printed publication platforms).
- The lack of a common vision for the creation of OA repositories. As stated above, current Open Access initiatives in Bulgaria are handled by individual

institutions and there is distinct lack of cooperation in regards to the creation of a unified OA network.
- An existing crisis in the Bulgarian publishing sector in regards to scientific periodicals and literature that is mostly the result of poor financing.
- The current trend of having Bulgarian scientists focus on having their works be published in renowned foreign peer-reviewed journals and databases with high impact factors as opposed to local publications.
- The lack of qualified specialists who can work towards the implementation of OA in Bulgaria.
- Lack of financing for the development of the Open Access sector in Bulgaria. Despite the OA model being focused on providing free access to scientific information to all who would benefit from it, the actual implementation of a functioning framework nevertheless requires adequate financial resources.

Apart from the above mentioned there are two additional factors that have a significant impact on the current development of the scientific fields in Bulgaria. These factors also have their effect on Open Access and will therefore be mentioned in this report. The first of these is the notable lack of a proper apparatus for scientific critique and review in Bulgarian universities and scientific organizations. The reviewing institution, as of now, has a mostly formal role in Bulgarian academic publishing. In turn, this has led to a decrease of both the quality, as well as the prestige of Bulgarian scientific publications. The second factor is connected with the above mentioned and has to do with the decrease in the quality of the requirements and criteria for the professional development of academic staff in Bulgarian scientific and educational institutions. This has led to several tendencies that have negative impact on the Bulgarian academic society as a whole. These include:
- The lack of encouragement for the creation of higher quality scientific literature.
- The decreased number of documents published per individual researchers.
- The isolation of part of the Bulgarian scientific society from the greater worldwide scientific community and achievements.

The Bulgarian scientific community is a relatively small one in terms of resources and lacks the capabilities to produce a significant quantity of research documents. Even so, despite its shortcomings, the Bulgarian research community has managed to gain some level of international recognition for its scientific achievements and publications. There are notable Bulgarian scientific publications in all 21 major scientific areas covered in the Essential Scientific Indicators. Similarly, in Scopus, Bulgarian scientists have managed to publish important research papers in all 26 core categories [5, 6]. Based on information gained from the above two sources, there are a several scientific sectors in which the Bulgarian scientific community has managed to achieve some level of international renown. Notable among these are publications in the fields of agriculture, chemistry, physics and medicine (and the various sub-fields connected with them). On the other hand though, some scientific sectors such as the social and humanitarian sciences are poorly represented in Bulgarian scientific literature. This could hopefully be mitigated with gradual introduction of Open Access publishing.

In the past several years, another major issue emerged that has had a significant impact on the ability of Bulgaria to participate in the sharing of scientific information with the international research community. This problem is connected with the lack of

resources in Bulgaria's major scientific libraries. This prevents these institutions from sharing information with other scientific and university libraries, as well as other academic organizations from other countries. Thus, since many Bulgarian scientists rely on these libraries for research information, the Bulgarian academic community has some difficulty in gaining access to part of the current scientific output of their international colleagues. The information presented in the current reports of the National Library "St. St. Cyril and Methodius", University Library "St. Kliment Ohridski" and the Central Library of the Bulgarian Academy of Sciences contains some highly troubling statements in regards to the ability of these institutions to maintain effective international book exchange programs. These libraries lack the required financial resources to buy publications that they can include in these exchange programs. What is more, these institutions lack the financial resources required for them to be able to send these publications to their exchange partners (provided, as mentioned above, that they can buy them in the first place) [14, 15]. As mentioned above, these issues manage to partially isolate the Bulgarian academic community from the rest of the scientific world. The creation of electronic information systems such as OA repositories and journals could mitigate some of these issues but the current efforts in regards to this, as mentioned above, have so far been made on an individual basis with a general overlaying framework missing. This leads to lack of coordination in regards to the creation and support of these information systems. Furthermore, not only are the financial resources provided to the Bulgarian scientific and academic institutions scarce but what is provided is fairly poorly managed (lack of an effective spending strategy). All of this means that the Bulgarian scientific community should be looking to create and/or participate in national and inter-institutional initiatives and projects that help create a unified framework for the effective allocation and use of resources. The next segment of the report will examine the current OA initiatives of the Sofia University "St. Kliment Ohridski" as a whole, as well as the Open Access projects of its "Library and Information Sciences" department.

3. Open Access in the Sofia University: Current Tendencies and Projects

Though at a slow pace, the model of Open Access is gaining momentum in the publishing sector of the Sofia University "St. Kliment Ohridski" [2]. The reasons for the slow introduction of OA into the university are very similar to the ones already discussed as issues encountered by the greater scientific community in Bulgaria: the lack of a critical understanding of what Open Access publishing actually is, lack of an adequate technological framework, lack of a unified OA strategy, financial issues and so on. Even so, the Sofia University has several active OA initiatives. These include:

- **The digital library of the SU** (online portal: http://research.uni-sofia.bg) [13]: An online archive which aims to provide access to the results of the various academic projects developed by researchers at the university. What is more, archived materials are indexed in various search engines and information systems such as Google. This increases the visibility of the publications made by SU researchers in the international scientific environment (bypassing the above mentioned issues related to the more traditional book exchange programs). The digital library archives a veriety of documents including articles, conference reports, books, patents and others. What is more, electronic versions of all academic papers defended in the Sofia University

(such as dissertations) are also archived in the digital library. The repository is registered in OpenDOAR [10, 11].

- **Horizons**: An electronic journal for scientific publications which aims to popularize research publications of the SU in regards to fields such as the social and humanitarian sciences (which we mentioned above, as two of the least covered scientific fields in Bulgaria), chemistry, biology, mathematics and others. The journal is maintained by the "Information procurement" department of the university. The publication comes out twice a year and is, as any OA resource should be, completely free of any charges or subscription fees. The journal is accessible through the online portal of the "Scientific research" sector of the SU and ensures that all copyrights are retained by the authors of the archived materials.
- Finally, the SU also maintains and develops a number of specialized digital libraries that provide free access to their documents. These include "Publications of the Sofia University "St. Kliment Ohridski", "Digital library of Bulgarian Slavic studies" and "Chetivo". All of these resources are in the beginning stages of their development and as such provide access to relatively small amount of documents.

In addition to the above initiatives, the individual faculties and departments of the SU are also free to create their own OA resources. An example of this is the "Library and Information Sciences" (LIS) department.

In 2009, the LIS department began to maintain a digital library. It provides free access to the department's yearbook, as well access to the various documents published by the department's academic staff. A major milestone for the department was made in 2012 when LIS entered into an agreement for the purpose of sharing relevant academic information with several international universities located in Germany, Poland, Holland, Latvia and Spain. This agreement was highly beneficial for the department as it provided access to a wealth of scientific information from international colleagues, as well as creating the opportunity for LIS publications to be available to wider research public. The department's staff played major role in promoting Open Access as an alternative publishing method for the "Philosophy" department of the SU (of which the LIS department is part of). Another notable decision made in 2012 is that all research publications (financed through the SU's budget) be deposited in the above mentioned digital library of the university [9]. An analysis of the statistics regarding the use of the electronic versions of the department's yearbook and three other digital publications show that there is a considerable interest in them [12].The example of the Sofia University (and in particular: its LIS department) shows that Open Access can have an integral part to play in Bulgarian academic society.

4. Conclusion

It has been stated numerous times in the span of this report that Open Access can play a key role in the future of academic publishing on both a national, as well as international level. However this change cannot happen on its own. The scientific community must actively strive to create and maintain effective policies and information systems that will enable it to effectively share and provide access to research materials. This also includes Bulgaria. While the current individual OA initiatives have met with some

measure of success it is imperative that unified framework be created. Through OA, Bulgaria's scientific community has the opportunity to engage in a shared dialogue with its international counterparts and improve the overall quality of its academic output.

References

[1] А. Димчев, Демократизиране на достъпа до научната вселена, или „Берлинска стена" към знанието. Кой път ще изберем? *Българско списание за образование*, **1** (2014), 16–60. Available at: http://www.elbook.eu/images/Aleksander-Dimchev.pdf/ [Accessed 26. 02. 2015].

[2] А. Димчев, Библиотечното и информационното осигуряване на образователната и научната дейност във Философски факултет на Софийския университет „Св. Климент Охридски", *Библиотека* **5** (2010), 69–78.

[3] P. Ginsparg, It was twenty years ago today …., *arXiv* (2011). Available at: http://arxiv.org/abs/1108.2700 [Accessed 26.02.2014].

[4] Кохезионната политика укрепва стратегията 2020, *Inforegio – Panorama* **36** (2011), 4-13. Available at: http://ec.europa.eu/regional_policy/sources/docgener/panorama/pdf/mag36/mag36_bg.pdf [Accessed 26.02.2015].

[5] Ministry of Education, Youth and Science, *Национална стратегия за развитие на научните изследвания 2020.* Министерство на образованието, младежта и науката. Стратегически документи (2014). Available at: https://www.mon.bg/?go=page&pageId=74&subpageId=143 [Accessed 26.02.2015].

[6] Ministry of Education, Youth and Science, *Програма за развитие на образованието, науката и младежките политики в Република България (2009 – 2013 г.)*, Министерство на образованието, младежта и науката. Стратегически документи (2009). Available at: http://www.minedu.government.bg/opencms/export/sites/mon/left_menu/documents/strategies/programa_MOMN-2009-2013.pdf [Accessed 26. 02. 2015].

[7] OpenAIRE (2015). Available at: https://www.openaire.eu/ [Accessed 26.02.2015].

[8] OpenAIRE (2015), Overview of Open Access in EU member states. Available at: https://www.openaire.eu/eu-member-states/noads/member-states-overview#european-open-access-repositories-landscape [Accessed 26.02.2015].

[9] Open Access to the Scientific Literature. A Summary of the Issues. LSU Health Sciences Center, Shreveport. Medical Library, available at: http://lib-sh.lsuhsc.edu/openaccess.html [Accessed 26.02.2015]

[10] OpenDOAR (2015), The Directory of Open Access Repositories. Available at: http://opendoar.org/find.php [Accessed 26.02.2015].

[11] OpenDOAR (2015), The Directory of Open Access Repositories. Available at: http://www.opendoar.org/onechart.php? [Accessed 26.02.2015].

[12] Sofia University (2015), Research at Sofia University "St. Kliment Ohridski", available at: http://research.uni-sofia.bg [Accessed 26.02.2015].

[13] National Library, *Национална библиотека „Св.св. Кирил и Методий": Годишен отчет*, НБКМ, София, 2010.

[14] Б, Яврукова, Значението на книгообмена за комплектуване на библиотеките. *Библиотека* **1** (2008), 13–16.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-102

We Should Not Light an Open Access Lamp and then Hide it Under a Bushel!

Santiago CHUMBE[a,1], Roddy MACLEOD[b] and Brian KELLY[c]
[a]*ICBL (Institute for Computer Based Learning), School of Mathematical and Computer Sciences, Heriot Watt University, Edinburgh, UK*
[b]*Former subject librarian at Heriot Watt University. Edinburgh, UK*
[c]*Independent consultant at UK Web Focus*

Abstract. The rapid growth of hybrid journals in the last few years has seen an unfortunate side effect: the majority of Open Access (OA) articles published in those journals cannot be recognized as OA beyond the publishers' websites, or by the discovery services used by researchers to access full-text articles. This reality has been demonstrated in the literature and solutions have been proposed. This paper explains the causes behind the problem, examines each of the proposed solutions, discusses the few implementations made with those solutions, and estimates whether the potential benefits merit the efforts required to implement the available solutions. Each of the solutions is analyzed from standardization and pragmatic perspectives. In particular, we critically analyze the solution proposed by NISO (RP-22-2015), and compare it with the solution offered by the JEMO project, which is based on using metadata elements from namespaces and XML schemas already being used by publishers. The contribution presents a number of case studies which show that research published as OA ends up erroneously being labelled as non-OA on the electronic services used by the end-user, when one of the components of the supply and delivery chain for e-journals fails to include OA information in its metadata. Furthermore, the case studies demonstrate that publishers of hybrid journals should not be the only ones being answerable for the problem. In fact, during the study, some publishers were actually not allowed to enable OA identification, at the article level, by key components of the supply chain. In those case studies, we worked with a sample of publishers that implemented the JEMO solution. From those experiences we draw answers to the main question of this presentation: which solution should be used to enable OA discovery from hybrid journals? What becomes apparent is that publishers are prepared and willing to implement any of the available solutions in their publishing workflow. The paper proposes that the simplest option is the best solution to provide standardized means to identify OA at the article level.

Keywords. Hybrid journals, Open Access articles, e-Publishing platforms, interoperability and integration, web feeds, metadata standards, e-Journal supply chain, discovery services, RSS.

1. Introduction

When researchers see that an article is published in a subscription journal for which they do not have full-text access, there is a high chance that they will give up instead of

[1] Corresponding Author. E-mail: S.Chumbe@hw.ac.uk.

trying to obtain the full-text (the probability is 70% according to S. A. Knowlton et al, [1].) What if the article in question is an OA article which happens to be published in a subscription-based journal? Are OA articles in hybrid journals read and cited less frequently because end-users are not properly made aware of the OA status of such articles? Unfortunately, this is happening and is a real problem. Figuratively, those OA articles are, in fact, being kept under a bushel. Authors wanting to publish OA in a hybrid journal are being penalized with higher APCs (Article Processing Charges) at the same time that the wider community is not being made aware of the availability of those OA articles. Those articles are not labeled as OA either beyond the publishers' websites, or by discovery services used by researchers to access full-text articles. This is a problem that has been noticed and identified by other studies [2, 3, 4, 5, 6, 7]. Accordingly the community has reacted by proposing solutions [8, 9, 10, 11]. However the problem is still unresolved. Our purpose is to describe the problem and to explain why it is still unresolved. Despite flaws in the hybrid model [12] the importance of hybrid journals cannot be disputed as most publishers are producing them. They attract authors who are interested in publishing OA articles in high ranking and well established journals [13]. Some may never become Gold journals [14]. If open access is intended to improve access to and dissemination of knowledge, it is crucial that any type of OA research output is identified as OA to everyone, everywhere, at all times.

The paper is organized as follows. In section 2, we use case studies to introduce the problem and describe its causes. Section 3 presents an analysis of available solutions and gives reasons why we endorse the solution proposed by JEMO. In section 4, through experiments carried out with a sample of hybrid journals, we demonstrate the benefits produced by a simple programmatic OA identification. The final section provides conclusions and recommendations.

2. Why Does the Problem Remain Unresolved?

The JEMO Project received funding from the Engineering and Physical Sciences Research Council (EPSRC) to find out why the problem introduced above remains unresolved. While some business aspects of the problem were beyond our control, the metadata used across the production, discovery and delivery chain of e-journals emerged as an important factor. We concluded that metadata is at the root of the problem of OA articles from hybrid journals being wrongly identified as non-OA articles. Metadata is important; it can enhance the results produced by retrieval and discovery systems and increase the usefulness and value of delivery systems such as link resolvers [15]. It can also enable the development of new services. But what matters is metadata quality [16]. Metadata has to be FAIR (Findable, Accessible, Interoperable and Re-usable)[2] otherwise it can even be harmful or misleading [17]. We will show that the problem is manifested in the diversity and poor quality of the metadata used in the e-journal supply chain, and in the amount of redundant and sometimes conflicting metadata specifications.

Information from 14 publishers is used in the case studies. Five of those publishers were official project partners and the other nine were invited to implement the project recommendations. The following table consolidates the number of Gold OA, Hybrid and Subscription-only journals currently being published by those publishers. Further

[2] https://www.force11.org/node/6062/ (visited on 30 Apr 2015)

detailed analysis removes their individual identifications as some publishers preferred to remain anonymous. Specific identification of the commercial hosting platforms has also been removed from the discussion[3].

Table 1. JEMO Participating Publishers. Showing hosting type (whether they have outsourced the hosting of their content to external e-publishing platforms or not) and type of journals (Hybrid: includes both OA and non-OA articles; Gold: includes OA articles only; Subscript: includes subscription-based articles only.)

Publisher	Hosting Type	Journals	Hybrid	Gold	Subscript.
IUCr	internal	9	7	2	0
BioMed Central	internal	278	0	273	0
Cambridge U Press	internal	447	151	5	291
Edinburgh U Press	outsourced	39	39	0	0
IGI-Global	internal	146	13	0	133
Inderscience Publishers	internal	397	397	0	0
Libertas Academica	internal	86	0	86	0
Maney Publishing	outsourced	201	200	1	0
MDPI	internal	136	0	136	0
Oxford U Press	outsourced	342	301	28	13
Walter de Gruyter	outsourced	678	328	350	0
Taylor & Francis	outsourced	1920	1810	38	72
The Geological Society	outsourced	10	10	0	0
Thieme Medical Publishers	outsourced	158	140	18	0
	Total	4847	3396	932	509

Publishers of hybrid journals know that metadata is important. Yet, some fail to appreciate that what is critically important is that it has to be *fit-for-purpose* metadata. Being fit-for-purpose means providing reusable (interoperable), consistent, accurate and complete information about the article associated with the metadata [18]. The 14 publishers understood the benefits and importance of producing quality metadata. However, in the implementation stages differences started to emerge. Five of the publishers that were using in-house hosting were able to incorporate OA elements in their metadata. The other two publishers chose to wait for the NISO RP-22-2015 recommendations to be released. The situation with publishers using external hosting platforms was contrasting. They faced an additional situation whereby their outsourced content on e-publishing hosting platforms is outwith the control of the original publishers. Despite their willingness to do so, publishers who have outsourced the hosting of their content to external platforms could not implement the required changes as quickly as done or scheduled by the other publishers. Being the bridge between publishers and the rest of the components of the e-journal supply chain, e-publishing hosting platforms play an important role in the transport of OA metadata. A further analysis of those platforms showed that to be cost-effective they cannot implement on-demand software changes on an individual publisher basis. The changes have to be made globally; usually as part of one or two annual software updates. Also, as in most cases the publisher's production system is not integrated with the external platform, the metadata used by the publisher to feed their platforms with new content, needs to be updated to incorporate new OA elements. The hosting platforms were not aware of the importance of those OA elements for hybrid journals. With the support of three publishers, it required discussions with one of the largest platforms to agree to change their metadata to accept OA elements. The metadata is based in the JATS tag suite[4], which is a schema increasingly being adopted by e-Publishing platforms to ingest

[3] Those platforms are Atypon, PublishingTechnology and PubFactory
[4] NISO Z39.96-2012 Journal Article Tag Suite (http://jats.nlm.nih.gov/, visited on 30 Apr 2015)

content from publishers. JATS already has elements to identify OA at the article level[5]. Figure 1 shows an example of using those elements assuming the copyright is retained by the publisher. Figure 2 represents the values for non-OA cases.

For OA articles:

```
<permissions>
<copyright-statement>Copyright © Publication_Year Publisher_Name
</copyright-statement>
<copyright-year>Publication_Year</copyright-year>
<copyright-holder>Publisher_Name</copyright-holder>
<license license-type="open-access"
xlink:href="http://creativecommons.org/licenses/by-nc-nd/4.0" />
<license-p>This is an open-access article distributed under the terms of the
Creative Commons Attribution License, which permits NonCommercial use,
distribution, and reproduction in any medium, provided the original work is
properly cited and you do not distribute the modified material.</license-p>
</license>
</permissions>
```

Figure 1. OA elements included in a JATS file to enable OA identification at the article level.

For non-OA articles:

```
<permissions>
<copyright-statement>Copyright © Publication_Year Publisher_Name
</copyright-statement>
<copyright-year>Publication_Year</copyright-year>
<copyright-holder>Publisher_Name</copyright-holder>
</permissions>
```

Figure 2. Optional copyright elements included in a JATS file for non-OA articles.

Six months after the five publishers were producing metadata with OA elements; we tested the discovery services that researchers are likely to use when trying to access full-text. Were those services taking advantage of the changes made by the publishers to provide OA identification, at the article level? The answer was no.

Discovery services are at the end of the supply chain and are supposed to be the main full-text access points for end-users. The problem with those services is that they can only identify OA at the journal level. Those services have implemented their own solutions. SerialSolutions and ExLibris for example, are addressing the problem using OA packages from OA aggregators that in theory would allow users to discover OA content published in any journal. However, the following example shows that this approach is not working.

Acta Crystallographica Section A is an hybrid journal published by the International Union of Crystallography (IUCr), which was one of the five publishers that immediately implemented OA identification at the article level in their metadata (in March 2014.) The journal published one OA article in its Volume 71, Issue 2 (2015). Figure 3 shows how the article is identified as OA on the journal's website.

[5] JATS will add support for the new NISO RP-22-2015 elements too (http://jats.nlm.nih.gov/1.1d3/)

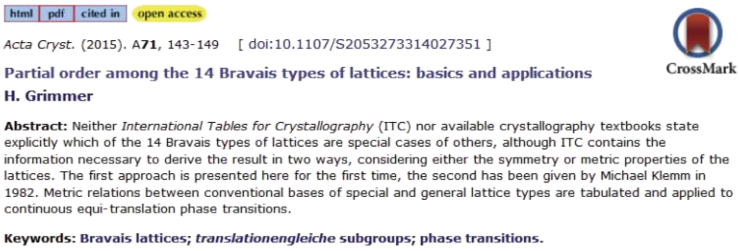


Figure 3. OA article identified on the journal's website. Source: http://scripts.iucr.org/cgi-bin/paper?eo5044 (visited on 1 May 2015)

The Wiley Online Library database also hosts articles of the journal and has no problem identifying the OA article[6] or any OA article published in this hybrid journal. The same could happen with any aggregator or discovery service that supports OA identification at the article level, for example on JournalTOCs[7].

We tried to access this same OA article from Ex-Libris Primo Central, without success. Figure 4 shows that Primo is erroneously labelling this OA article as non-OA ("**No full-text**.") If you select the "Services" link to gain full-text, you will be suggested to use the Inter Loan Library (ILL) service to read this OA article[8]. The first screenshot in Figure 5 is from EbscoHost, where the user is advised to request ILL to be able to read the OA article. The last screenshot comes from Summon, which includes the "**Full Text Online**" link; giving the appearance that through this link you could get full-text access. However the link will send you to the SerialSolutions OpenURL landing page[9], which will point you to the OA Digital Library[10] aggregator from where you will need to start your search again, only to find out at the end that this aggregator doesn't include articles from *Acta Crystallographica Section A*.

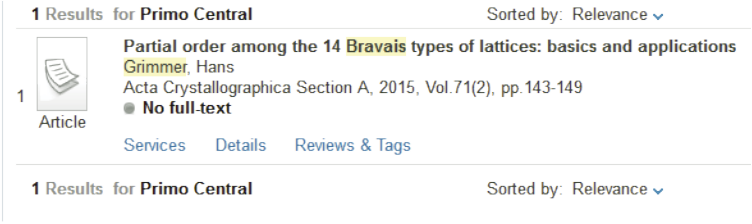


Figure 4. Primo hiding the OA article behind subscription walls. Seen at http://goo.gl/OBvYNN, May 2015

[6] As seen at http://onlinelibrary.wiley.com/doi/10.1111/aya2.2015.71.issue-2/issuetoc on 01 May 2015
[7] As seen at http://www.journaltocs.ac.uk/?issn=2053-2733 on 15 Feb 2015
[8] As seen at http://goo.gl/ajitF8 on 1 May 2015
[9] As seen at http://goo.gl/5xtSZ6 on 1 May 2015
[10] As seen at http://grweb.coalliance.org/oadl/oadl.html on 1 May 2015

Figure 5. EBSCOhost and Summon discovery services showing the OA article erroneously hidden behind subscription walls. Sources: http://goo.gl/bjhscZ and http://goo.gl/CUWsuo (visited on 1 May 2015)

The scenario illustrated by the previous case study is not acceptable. The fact that OA identification is still done at the journal level across the supply chain needs to change. OA articles will continue to be erroneously labelled as non-OA on the electronic services used by the end-user if at least one of the components of the supply chain fails to embed the publication's OA status in the metadata shared across this chain. OA identification at the article level requires cooperation between all parties involved and the use of common and standard metadata elements. This lack of cooperation and interest is one of the underlying causes of why this problem is still unresolved. Furthermore, if the solution passes for embedding OA information in the metadata, it cannot be a responsibility of the publishers or publishing platforms only. We believe that, as long as discovery services don't use metadata with OA elements at the article level, any effort made by the publishers will fail.

As mentioned before, two of the publishers decided to wait until a standard solution was agreed across the publishing industry; specifically the one that NISO was preparing at that time. They didn't want to implement something that may not be interoperable with the other components of the e-journal supply chain. This "insecurity" hints at the second cause of the problem. While existing standards could solve the problem, they were not considered and formulated as a consensual solution to identify OA articles. Consequently, publishers are waiting, implementing their own solutions, or just ignoring the problem. The large number of standards, schemas and namespaces to produce metadata for research resources adds more uncertainty[11]. Well-intentioned machine-readable solutions instigated by publishers, such as the *Open Access Collection* of the Geological Society[12] and the *Get New Open Access Article Feed* of Elsevier[13] as well as using HTML meta-tags [10], shows the publishers' willingness to enable programmatic identification of OA, but they are still far from being efficient solutions for aggregators, databases and discovery services as these services would need to shoulder a greater demand to perform the normalizations and transformations required when dealing with diverse types of feeds and metadata elements.

[11] E.g. http://goo.gl/kD3kSS or http://goo.gl/629OqO (visited on 2 May 2015)
[12] http://cct.highwire.org/svc/getfile?fileId=283&publisherId=gsl (visited on 22 April 2015)
[13] For example http://www.sciencedirect.com/science/journal/15708705 (visited on 22 April 2015)

3. Analysis of Available Metadata Standard Solutions

Regarding standard solutions proposed by the community to resolve this imbalanced situation, M. Van Ballegooie [19] identified two options that are currently available, the NISO RP-22-2015 recommendation [20] and the elements proposed by JEMO [21].

In December 2012, NISO formed a group to recommend a specification for the accessibility of journal articles. In January 2015, the group released the NISO RP-22-2015 recommendation. Initial reactions endorsed an eagerly expected specification, but concerns were also expressed by the community.[14] [15] [16] [17] It was noted that the recommendation does not aim to specifically solve the problem of OA articles published in hybrid journals[18]. Aiming to cover all scenarios, the recommendation prefers to use the "Free to Read" term instead of Open Access. From a pragmatic perspective, NISO created the *free-to-read* and *license_ref* elements and a new namespace[19] specifically designed to support these new elements.

On the other hand, JEMO draws on the wealth of experience provided by the simple yet effective CC (Creative Commons) and DC (Dublin Core) metadata schemas[20] that have been adopted by publishers and are widely used in the e-journal supply chain. The implementation, presented at the NASIG 2014 Conference, aims to resolve the machine-readable or programmatic identification of OA at the article level. It proposes using the *dc:rights* and *cc:license* elements to embed OA information in the metadata already being used by publishers.

The elements proposed by NISO and JEMO are described and assessed from the implementation perspective in Tables 2, 3 and 4.

Table 2. Metadata elements proposed to identify OA at the article level. *free-to-read* and *license_ref* are new elements created by NISO RP-22-2015. *dc.rights* and *cc*:*license* are elements of the DC and CC metadata standards, respectively. Applicable or needed attributes are noticed.

Element	Purpose	Attributes	Namespace
dc.rights	To inform about the ownership of, or rights held in and over, an article	None	DC, implemented since 2000[22]
cc:*license*	To provide a reference to a URI that defines the associated license, indicating the restrictions and how the article may be used and accessed.	*rdf:resource*	CC, de facto license for OA publications[22]
free-to-read	To define whether the article is accessible, without charge or other restriction to read online.	*start_date* *end_date*	New Access and License Indicators (ALI)[21] to be implemented
license_ref	To provide a reference to a URI that carries the license terms specifying how the article may be used.	*start_date*	ALI

[14] http://rioxx.net/guidelines/RIOXX_Metadata_Guidelines_v_3.0.pdf, pp 6-8 (visited on 2 May 2015)

[15] http://goo.gl/ZQkjbL (visited on 2 May 2015)

[16] In some way NISO RP-22-2015 blurs the term Open Access by stating that publishers use the terms Open Access, Increased Access, Public Access and other names to identify their offerings; which is not the case as no publisher or entity of the e-journal supply chain uses "Increased Access" or "Public Access" to name Open Access articles or to identify any type of journal.

[17] http://goo.gl/FkhkV3 (visited on 4 May 2015)

[18] The NISO document tangentially mentions "Open Access" a few times only. In some way, it explains its stand by stating that "*this is a contentious area where political views on modes of access lead to differing interpretations of what constitutes 'open access.*'" [20]

[19] http://www.niso.org/schemas/ali/1.0/ (visited on 2 May 2015)

[20] http://creativecommons.org/ns and http://purl.org/dc/elements/1.1/ (visited on 2 May 2015)

Table 3. Remarks for the *free-to-read, license_ref, dc.rights* and *cc:license* metadata elements that should be considered by implementers.

Remark	dc.rights	cc:license	free-to-read	license_ref
It is part of a mature and widely adopted standard	YES	YES	NO	NO
Can provide information on whether a specific article is Open Access (OA)	NO	YES	Partially	Partially
Can provide information on the restrictions and re-use rights of a specific OA article	NO	YES	NO	YES
It is already being used in the e-journal supply chain.	YES	Partially	NO	NO
Can indicate the period of time when access to an article is delayed	NO	NO	YES	NO
Can indicate how the license's terms change over time	NO	NO	NO	YES
It has already been used to provide copyright metadata	YES	NO	NO	NO

Table 4. Issues particularly relevant for OA articles published in hybrid journals.

Question	Answer
Do OA articles published in hybrid journals have embargo dates?	NO
Do OA articles published in hybrid journals have "moving wall" dates?	NO
Do CC licenses have end or expire dates?	NO
Could the use of end dates inadvertently create gaps between applicable licenses?	YES
Are any of the hybrid journals exclusively using licenses different to CC licenses?	NO
Have publishers been in the past quick in implementing new metadata specifications?	NO
Can an OA article already published as OA using a CC license, become non-OA?	NO

It is noticeable that, from the OA perspective, the function of the *license_ref* element can be provided through use of the *cc:license* element. *cc:license* can state the OA status of an article, plus its associated re-use rights. In the NISO case, *free-to-read* can only tell us whether an article can be freely read or not, but this can be an OA article, a free sample, a temporary promotion, etc. *free-to-read* alone is not enough to know the re-use rights of OA articles. A second new element (*license-ref)* is needed to complete the article's OA status. Consequently, *cc:license* resolves the specific OA problem caused by hybrid journals, while *license-ref* is a general-purpose solution, which needs to be combined/analyzed with *free-to-read* to indicate OA as a function: OA = *f*(free-to-read, re-use rights, embargo-period).

The new NISO elements provide an embargo period; a concept relevant to subscription-based journals but alien to OA. OA means full-text access without any delay, forever. The *start_date* and *end_date* attributes of those two new elements do not apply in OA; if used, they would need to semantically be analyzed by the services trying to identify OA articles and ignored for OA articles, a process that would introduce additional complexity to the handling of terms used in Open Access. In contrast, when the value of *cc:license* is a valid CC URI, the risk of identified a non-OA article as OA is null. CC licenses are not revocable[21]. Therefore, an OA article licensed with any CC license is perpetually OA. Furthermore, an OA article published under any CC license is immediately OA upon publication. The unanimous praxis among OA stakeholders is that OA means immediate open access[22][23]. Certainly, if

[21] https://wiki.creativecommons.org/Frequently_Asked_Questions (visited on 2 May 2015)
[22] "*Delayed Access is neither Green OA nor Gold OA*" (http://goo.gl/gGIxWv visited on 2 May 2015)

publishers use different custom licenses instead of CC licenses, the perpetuity and immediateness concepts associated with OA wouldn't apply and the *cc:license* element wouldn't be enough to identify OA. However, CC is universally accepted by hybrid journals. Data analyzed by the project shows that every hybrid journal accepts CC licenses[24]. Therefore, using *cc:license*, together with *dc:rights*[25], becomes a suitable, less onerous and low-barrier solution to identification of OA articles published in those journals; with the ease of implementation illustrated by the experiments run with JournalTOCs, an aggregator of scholarly journal RSS feeds.

4. Results of Prototyping Programmatic OA Identification

Five participating publishers added the *cc:license* and *dc:rights* elements to their RSS feeds in a matter of weeks. Three of them, whose feeds were already following the CrossRef recommendations for scholarly feeds[26], needed only a week. When NISO RP-22-2015 became available, publishers were given the choice of implementing either NISO or JEMO elements or both. At the end of the project over 20 publishers were using the *cc:licence* in their RSS metadata, including SpringerOne and Biomed Central. No publisher had implemented the new *free-to-read* and *license_ref* elements yet. Once RSS feeds providing OA elements in their metadata became available, JournalTOCs was able to create an API exposing OA articles collected from different gold and hybrid journals[27] and demonstrate the benefits of those new elements. The experience has shown that the maturity of the metadata specifications in question, the level of support from experts and validation services are important factors for adoption. As long as the e-journal supply chain components are unable to parse new elements, metadata providers will use what is easier and convenient for them.

Content providers prefer to provide metadata with the minimal effort possible for them [22, 23]. For example, only 50% of journal TOC RSS feeds use the CrossRef recommendations for RSS feeds published in 2009 [24]. This fact should make us cautious when proposing new metadata elements. As the complexity and number of metadata specification increase, their adoption by metadata providers tends to proportionally decrease. Some publishers were reluctant to enrich their RSS feeds until the suggested metadata had reached a certain level of maturity and acceptance; confirming that new specifications create high barriers to adoption.

The new NISO elements suffer from the same problems affecting other standards that have low or incomplete adoption. They are rich in theory but demanding in practice. There is the over-optimistic assumption that aggregators will know how to fully implement the new specifications (e.g. the NISO recommendations don't provide any technical means of enforcement for its *start_date* and *license* URI attributes of the new NISO elements, leaving the decision to aggregators.) The recommendation of NISO in some way contradicts its own advice that before creating new metadata elements, adapting existing schemas should be considered. "*We use standards to*

[23] Publishing open access makes your work immediately and permanently available online for everyone, worldwide, http://www.springer.com/open+access (visited on 2 May 2015)

[24] https://openjemo.wordpress.com/2015/05/04 (visited on 5 May 2015)

[25] Stating the rights associated with the CC license is recommended because CC licenses are operative only when applied to material in which a copyright exists.

[26] http://oxford.crossref.org/best_practice/rss/ (visited on 9 May 2015)

[27] http://www.journaltocs.ac.uk/api/articles/oa/ (visited on 10 May 2015)

improve interoperability and to reduce unnecessary variation. It is better and easier to adopt something that already exists, is well modelled, and comprehensively supported." [28] What becomes apparent is that the participating publishers were actually prepared and willing to implement the simplest of the available initiatives in their publishing workflow. This response made sense because publishers will normally be more disposed to implement a new specification if it involves using elements with which they are already familiar. The fact that CC and its different licensing flavors are used by practically all the publishers of hybrid journals was an important factor in their quick understanding and adoption of the JEMO CC-based tagging scheme.

5. Conclusion

Open Access articles are being erroneously hidden behind subscription-access walls because the OA status of articles is not embedded in all of its metadata manifestations shared by the multiple databases and discovery services involved in the e-journals delivery chain. The confusing landscape of various standard metadata exchange specifications proposed to cover every free to read possibilities, without giving a particular solution for OA articles, escalates the problem. The JEMO project has shown that using Creative Commons and Dublin Core elements is an easy and effective option for metadata providers (e.g. publishers) and consumers (e.g. discovery services) to programmatically identify OA at the article level. OA identification will eventually fail if OA status is not embedded in all metadata manifestations in the e-journals delivery chain.

Instead of creating new general-purpose specifications, we argue that efforts should be directed to implement elements that are already part of schemas being used by publishers and to enable OA identification at the article level on any online service used to access full-text.

cc:license provides a framework for conveying essential information that addresses common OA use cases. All publishers of hybrid journals offer CC licenses.

The JEMO case studies demonstrate that publishers do not intend to hide their Open Access articles behind subscription walls; it is, rather, a question of whether the e-journals delivery chain is propagating the appropriate forms of access in the right places.

This study has tangentially uncovered some problems with discovery services. We have shown that because discovery services are not using OA elements in their metadata, users are being denied access to OA articles published in subscription journals. Articles that were tagged as OA on the publishers' websites are being kept undiscoverable as OA in discovery services.

Our study has demonstrated that enabling programmatic identification of OA at the article level would enhance current services; hence benefiting both the research community and the OA hybrid business model.

[28] ISO/TC 46/SC11N800R1 Recommendations (as seen at http://goo.gl/Dv5zjp on 5 May 2015)

References

[1] S.A. Knowlton, I. Kristanciuk, M. Jabaily, Spilling Out of the Funnel: How Reliance Upon Interlibrary Loan Affects Access to Information, *Library Resources & Technical Services*. **59**(1) (2015), doi: 10.5860/lrts.59n1.4

[2] C. Bulock and N. Hosburgh. OA in the library collection: The challenges of identifying and managing open access resources. *The Serials Librarian* (In Press)

[3] Fair Prices for Article Processing Charges (APCs) in Hybrid Journals, *UK Open Access Implementation Group*, http://goo.gl/cGBY2l (as of April 2014)

[4] Not all hybrid is equal, *Australian Open Access Support Group* (AOASG) Weblog, http://aoasg.org.au/not-all-hybrid-is-equal/ (as of March 2014)

[5] S. Pinfield, J. Salter and P.A. Bath, The 'total cost of publication' in a hybrid open-access environment: Institutional approaches to funding journal article-processing charges in combination with subscriptions, *J. of the Assoc. for Inf. Science and Tech.* (In Press) http://goo.gl/WW59U4

[6] S.M. Shieber, Equity for Open-Access Journal Publishing, *PLoS Biology* **7**(8) (2009), doi: 10.1371/journal.pbio.1000165

[7] Wellcome Trust calls for greater transparency from journals on open access publishing costs, *Europe PubMed Central Weblog*, http://goo.gl/lBjFyJ (as of December 2013)

[8] C. Xiaotian, Journal Article Retrieval in an Age of Open Access: How Journal Indexes Indicate Open Access Articles, *Journal of Web Librarianship* **7**(3) (2013) doi:10.1080/19322909.2013.795426

[9] S. Chumbe, B. Kelly and R. MacLeod, Hybrid journals: Ensuring systematic and standard discoverability of the latest Open Access articles, *The Serials Librarian* **68**(1-4) (2015), 143–155 doi: 10.1080/0361526X.2015.1016856.

[10] C. Hutchens, Open access metadata: current practices and proposed solutions, *Learned Publishing* **26**(3) (2013), 159–165(7) DOI: 10.1087/20130302.

[11] N. Lagace and G. Tananbaum, NISO Open Access Metadata and Indicators Working Group: Creating a Cross-Audience Solution, *The Serials Librarian* **65**(2) (2013), doi: 10.1080/0361526X.2013.813892.

[12] B.-C. Björk, The hybrid model for open access publication of scholarly articles: A failed experiment? *Journal of the American Society for Information Science and Technology* **63**(8) (2012), 1496–1504, doi: 10.1002/asi.22709.

[13] T. Koler-Povh, P. Južnič and G. Turk, Impact of open access on citation of scholarly publications in the field of civil engineering, *Scientometrics* **98**(2) (2013), 1033–1045.

[14] S. Armato III, C. Baldock and C.G. Orton. "Hybrid gold" is the most appropriate open-access modality for journals like Medical Physics, *Med. Phys.* **42**(1) (2015) doi: 10.1118/1.4895979.

[15] S. Glasser, NISO Webinar: It's Only as Good as the Metadata: Improving OpenURL and Knowledge Base Quality. *Serials Review* **37**(1) (2011), 58–60, doi: 10.1080/00987913.2011.10765348.

[16] A. Tani, L. Candela and D. Castelli, Dealing with metadata quality: The legacy of digital library efforts, *Information Processing & Management* **49**(6) (2013), 1194–1205, doi: 10.1016/j.ipm.2013.05.003.

[17] C.M. Yassera. An Analysis of Problems in Metadata Records. *Journal of Library Metadata*. **11**(2) (2011), 51–62, doi: 10.1080/19386389.2011.570654.

[18] J-R Park. Metadata Quality in Digital Repositories: A Survey of the Current State of the Art, *Cataloging & Classification Quarterly*, **47**(3-4) (2009) 213-228, doi: 10.1080/01639370902737240

[19] M. Van Ballegooiea. Knowledgebases: The Cornerstone of E-Resource Management and Access. *Serials Review*. **40**(4) (2014) pp. 259-266. doi: 10.1080/00987913.2014.977127

[20] Access License and Indicators: A Recommended Practice of the National Information Standards Organization, NISO RP-22-2015, available at: http://www.niso.org/publications/rp/rp-22-2015 (as of January 2015)

[21] Step by Step Guide to enable OA identification from RSS feeds, *JEMO Project* weblog, http://goo.gl/ZXWpPb (as of April 2015)

[22] D. Mietchen, C. Maloney and N.D. Moskopp, Inconsistent XML as a Barrier to Reuse of Open Access Content, *Journal Article Tag Suite Conference (JATS-Con) Proceedings* 2013 http://www.ncbi.nlm.nih.gov/books/NBK159964/

[23] S.D. Shapiro, We are all aggregators (and publishers) now: how discovery tools empower libraries, *Library Hi Tech News* **30**(7) (2013), 7–9, doi: 10.1108/LHTN-07-2013-0041

[24] Recommendations on RSS Feeds for Scholarly Publishers, http://oxford.crossref.org/best_practice/rss/ (as of May 2015)

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-113

Journals' Editorial Policies – An Analysis of the Instructions for Authors of Croatian Open Access Journals

Jadranka STOJANOVSKI[1]
University of Zadar / Rudjer Boskovic Institute

Abstract. The growing number of publications presenting research findings, the pressure on scientists to produce publications in great quantity, and the shift in the business models of many journals increased importance of journals' editorial practices, which are well represented in guidelines for preparing the manuscript for submission. Journals have a special responsibility to protect research integrity and to keep trust in journal publishing. This study looked at information on editorial practices in the instructions for authors of Croatian Open Access journals. 283 instructions for authors from all disciplines were examined according to the broad range of publishing issues grouped in hierarchically organized categories. Mostly addressed issues were manuscript layout (276/283) and journal language (269/283). The most common ethical issues among journals from all disciplines were responsibility of author (73/283), funding (52/283), and accuracy (51/283). There are several ethical issues addressed significantly more often by biomedical journals, like responsibility of authors (14/30), publishing ethics (14/30), conflict of interest (12/30), funding (11/30), and authorship (11/30). In comparison with ethical issues common publishing issues like manuscript layout, manuscript elements, and type of paper were richly represented in journals from all disciplines.

Keywords. Instructions for authors, publishing ethics, Open Access, Croatia

1. Introduction

The value of scientific research and its role in the world prosperity is unquestionable, and publishing is still the main scholarly communication channel. The growth of science, *"publish or perish"* pressure, competition for research funding, more and more profitably scholarly publishing industry, and repeated reports on research misconduct create the need for self-regulation and guidance in the conduct of science and the dissemination of scientific results [1]. One of the first initiatives for *"self-regulation and guidance"* came from a group of journal editors who created the first technical guidelines, known as *"Uniform Requirements for the Submission of Manuscript to Biomedical journal"*. Evolution of *"Uniform Requirements"* inspired other editors, researchers and funders, causing publication ethics to become part of major guidelines and publishing standards. When the UK Committee on Publication Ethics (COPE) was established in 1997 editor of the BMJ Richard Smith stated that COPE *"will serve the editors"* and *"advise on cases brought by editors"* [2] *"Guidelines on good publication practice"* published by COPE in 1999 addressed

[1] Corresponding Author. E-mail: jadranka.stojanovski@irb.hr.

study design and ethical approval, data analysis, authorship, conflicts of interest, peer review, redundant publication, plagiarism, duties of editors, media relations, advertising, and dealing with misconduct [3]. Other guidelines for authors and editors included additional ethical and legal considerations like authorship responsibilities, acknowledgments, duplicate publication, intellectual property, confidentiality, protecting individual rights, and defamation and libel [1].

Ethical issues, often neglected in the small research communities, are actual topics in the area of scholarly publishing. In order to respond properly to potentially low quality submissions, editorial policies of small journals should rely on best practices and guidelines which share the responsibility for research integrity between authors, editors and publishers. Instructions for authors are mirror of editorial policies, summarized in the carefully shaped segments comprising information about journal, manuscript or publishing ethics. Editorial policies reflecting trough the instructions for authors were often neglected in the past. Early study of the instructions for authors proved that beside directions on formatting and style, they often included financial disclosures and policies on coverage, peer review, confidentiality, human experimentation and duplicate submission [4]. Editors of the Croatian journal *Biochemia Medica* stated: *„Whereas our former Instructions to authors have mostly been concerned with recommendations for manuscript preparation and submission, the revised document additionally describes the editorial procedure for all submitted articles and provides exact journal policies towards research integrity, authorship, copyright and conflict of interest."* [5] Several studies have assessed journals' instructions for authors on the reporting of ethical issues, and the majority of them are from the field of biomedicine or related disciplines [6]–[11]. Croatian journals were included in two recent studies. The first study examined publication ethics policies in biomedical journals published in Central and Eastern Europe. These looked for differences between ethical issues addressed in East (EU) and South East European countries [12]. A second larger study examined 197 instructions for authors in English language of the Croatian OA journals. The results suggested that emerging ethical issues are not well addressed in the instructions for authors, and that biomedical journals performed significantly better compared with all journals [13]. Instructions for authors in Croatian language were not yet analysed. Also, there was no published study comparing Croatian journals among different disciplines according broad range of issues, including journal characteristics and manuscript formatting and style.

The primary aim of this study was to answer the research question: "Were ethical issues in current instructions for authors of Croatian Open Access (OA) journals adequately described compared with information about journal and manuscript formatting and style?" The secondary specific aim was to investigate the differences between instructions to authors of Croatian OA journals from different disciplines. The hypothesis was that biomedical journals, where publication ethics is particularly important, adopted ethical requirements in much greater extent compared with journals from other disciplines.

2. Methods

Based on the Croatian repository of Open Access journals HRCAK[2] journals with available instructions for authors written in English or Croatian language were identified. Out of 363 Croatian OA journals available on HRČAK we identified 298 journals with publicly available instructions for authors. Out of 298 journals, 283 journals from all disciplines had instructions stored in a machine-readable format which enabled the text to be coded automatically.

Then we classified all OA journals according to the discipline, which resulted with distribution of 283 journals as follows: sciences (21), biomedicine (30), technical sciences (29), biotechnical sciences (22), social sciences (87) and humanities (94).

In order to find out the presence of publishing issues in journals' instructions for authors, an analysis was performed. The unit of content analysis in this research was the single document containing instructions for authors. We accessed the 'guidelines for authors' or 'instructions on/for/to authors' at HRCAK website during April 2015, and PDF, DOC and TXT versions were stored locally. The Croatian version of the instructions for authors was included in the absence of an English version. Provalis' Research[3] software QDA Miner and WordStat was used as a tool for content analysis. The categorization scheme used in the previous study [3] was modified and manuscript and journal categories were added. The text was coded automatically according to the hierarchically defined categories, subcategories, words, phrases and rules stored in the categorization dictionary (Table 1). Our categorization dictionary contains 3 top-categories, 24 first level, 30 second level, and 7 third level subcategories including 243 words, phrases and rules used for coding. Table 1 is presenting a simplified version of the categorisation dictionary.

Table 1. Simplified version of the categorization dictionary used for content analysis

Category	Subcategory	Terms used for coding
1. Ethical issues		
	Accuracy	accuracy
	Authorship	authorship, contributorship
	Confidentiality	confidentiality, privacy
	Ethics	ethics
	Funding and CoI	funding, grant, project, sponsor, conflict of interest, competing interest
	Misconduct	allegation, fabrication, falsification, fraud, malpractice, manipulation, misconduct
	Plagiarism	plagiarism
	Redundancy	compilation, dual submission, duplicate submission, multiple submission, recycled, redundant
	Reporting	reporting
	Research integrity	research integrity
	Responsibility	author's responsibility, editor's responsibility, publisher's responsibility
	Retraction	expression of concern, retraction, suspicion, withdrawal

[2] hrcak.srce.hr
[3] provalisresearch.com

Category	Subcategory	Terms used for coding
2. Journal		
	Business model	additional charge, article processing charges, fee, free of charge, open access
	Carrier	analogue (paper), digital (electronic, online)
	Copyright transfer	copyright, creative commons, rights transfer
	Research data	dataset, raw data, underlying data
	Language	Croatian, English, French, German, Italian
	Media	audio, graphics, multimedia, text, video
	Peer review	anonymous, blind, open
	Scope	discipline, field, subject, topic
	Timeliness	estimated time, timely
3. Manuscript		
	Manuscript elements	title, author, abstract, key-words, introduction, materials and methods, results, discussion, conclusion, literature, acknowledgement
	Manuscript layout	layout (spacing, margins, header, footer, paragraph), tables&figures (figure, graph, illustration, image, formula, table), typography (font, italic, bold)
	Type of paper	article, book review, preliminary communication, review article, conference paper

Results were expressed as frequencies and percentages for categorical variables or mean ± standard deviation for continuous variables. Associations between discipline and categorical parameters were tested using $\chi 2$-test. The level of significance was set at 0.05.

3. Results and Discussion

At the top level categories there was no significant difference between disciplines on publishing issues (journal and manuscript). Significant difference was present on ethical issues (Table 2). As expected, ethical issues are best represented in the instructions for authors of biomedical journals (76.7%), followed closely by science journals (76.2%) what was not expected. Less than half of the journals from social sciences have present any of ethical issue in their instructions for authors.

Table 2. Publishing and ethical issues in instructions for authors

Category	Subcategory	Sci	Biomed	Techn	Biotech	Soc	Hum	Chi square value	P(Chi square test)
		N=21	N=30	N=29	N=22	N=87	N=94		
1. Ethical issues		**16**	**23**	**20**	**16**	**43**	**52**	**13.141**	**0.022**
	Accuracy	6	9	9	6	11	10	14.266	0.014
	Authorship	6	11	4	3	7	17	15.654	0.008
	Confidentiality	4	5	3		3	5	13.118	0.022

Category	Subcategory	Sci	Biomed	Techn	Biotech	Soc	Hum	Chi square value	P(Chi square test)
	Ethics	4	14	3	2	11	12	23.751	< 0.001
	Funding & CoI	8	13	4	6	21	7	25.203	< 0.001
	Misconduct	6	8	9	8	6	3	36.301	< 0.001
	Plagiarism	4	4	2	1	5	2	11.143	0.049
	Redundancy	5	9	1	1	5	6	23.924	< 0.001
	Reporting	5	9	1	2	3	2	35.463	< 0.001
	Research integrity	8	14	11	6	15	19	15.6	0.008
	Responsibility	1	5		1	1	1	21.51	0.001
	Retraction	4	4	5	5	2	1	26.085	< 0.001
2. Journal		**21**	**29**	**29**	**22**	**86**	**94**	**4.374**	**0.497**
	Business model	14	13	5	5	23	19	24.018	< 0.001
	Carrier	19	29	26	22	80	87	3.125	0.681
	Copyright transfer	10	13	8	9	22	17	14.256	0.014
	Research data	13	15	10	17	28	18	37.228	< 0.001
	Language	21	27	28	22	82	89	4.152	0.528
	Media	20	25	25	22	74	67	15.788	0.007
	Peer review	2	10	2	2	36	33	22.819	< 0.001
	Scope	19	24	21	21	68	63	11.917	0.036
	Timeliness	10	6	5	4	13	11	16.044	0.007
3. Manuscript		**21**	**30**	**29**	**22**	**87**	**93**	**2.018**	**0.847**
	Manuscript elements	21	30	29	22	87	91	6.097	0.297
	Manuscript layout	21	30	29	22	87	87	14.431	0.013
	Type of paper	21	29	28	22	86	86	8.903	0.113

Manuscript issues (layout, manuscript elements and type of paper) were present in almost all journals (Table 2). According to the frequency of coded categories manuscript layout, including instructions for chapters, paragraphs, margins, page size, line spacing, alignment, indentation, headers and footers, makes large part of the instructions for authors. Significant differences among manuscript layout elements were present for tables and figures which are less prominent in the instructions of journals from social sciences and humanities. According to the frequency of coded categories journal editors are often describing manuscript elements: article title, authors, abstract, key-words, introduction, materials and methods, results, discussion, literature and acknowledgement, tracking IMRAD standard for the structure of scientific journal article. Most frequently mentioned were author(s), abstract and literature list, while the presence of discussion and acknowledgement varied across disciplines. The most popular type of papers in all disciplines are article (scientific paper) and conference paper.

Regarding journal information the most addressed media was text presented by PDF format, as expected. Information about journal carrier were equally represented by terms "print" and "electronic", and suggested languages were mostly English and Croatian. An interesting observation was that CD as the carrier for journal or manuscript is still very popular, and mentioned by 38% of all journals. Business models, including fees and charges, are mostly present in journals from science and biomedicine disciplines. The terms "article processing charges" or APC were not mentioned, although a few journals are charging for publishing papers. Open Access was addressed only by 14 journals, even all 283 journals included in the analysis were OA journals. Open Access is a huge privilege for authors and should be mentioned in the instructions for authors. In the instructions for authors, editors are not communicating copyright issues, peer review type and timeliness, all issues of crucial importance for potential authors.

4. Conclusion

An analysis of the author instruction of Croatian OA journals show that ethical issues was the least prominent category in our study. The most frequent ethical issues addressed by Croatian OA journals were responsibility, funding and accuracy. Guidance regarding redundancy, conflict of interest, reporting, retraction, confidentiality, plagiarism, and research integrity was addressed by less then 10% of the journals. There are several issues addressed more often by medical journals, compared with journals from other disciplines, like responsibility, publishing ethics, conflict of interest, funding, and authorship. Ethical issues like retraction, plagiarism, research integrity and confidentiality were represented by few biomedical journals.

It is important to keep in mind that the data presented in the study are the policies of the journals as stated in the instruction for authors. A lack of presence does not mean that particular item is not important for editor. It means that instructions for authors should be revised and improved.

References

[1] M. C. Atlas, Emerging ethical issues in instructions to authors of high-impact biomedical journals, *J. Med. Libr. Assoc.* **91**(4) (2003), 442–449.
[2] R. Smith, Misconduct in research: editors respond, *Br. Med. J.* **315**(7102) (1997), 2001–202.
[3] The COPE Report 1999, Guidelines on good publication practice, *Hum. Reprod.* **16**(8) (2001), 1783–1788.
[4] A. C. Weller, Editorial policy and the assessment of quality among medical journals, *Bull. Med. Libr. Assoc.***75**(4) (1987), 310–316.
[5] A.-M. Šimundić, News at Biochemia Medica: Research integrity corner, updated Guidelines to authors, revised Author statement form and adopted ICMJE Conflict-of-interest form, *Biochem. Medica* **23**(1) (2013), 5–6.
[6] W. Gardner and K. Heck, Ethical Requirements in the Instructions for Authors in Journals Publishing Randomized Clinical Trials, *Res. Ethics* **5**(4) (2009), 131–137.
[7] A. Y. Gasparyan, L. Ayvazyan, S. V. Gorin, and G. D. Kitas, Upgrading instructions for authors of scholarly journals, *Croat. Med. J.* **55**(3) (2014), 271–280.
[8] F. Kunath, H. R. Grobe, G. Rücker, D. Engehausen, G. Antes, B. Wullich, and J. J. Meerpohl, Do journals publishing in the field of urology endorse reporting guidelines? A survey of author instructions, *Urol. Int.* **88**(1) (2012), 54–59.

[9] J. J. Meerpohl, R. F. Wolff, C. M. Niemeyer, and G. Antes, Editorial Policies of Pediatric Journals: Survey of Instructions for Authors, *JAMA Pediatr.* **164**(3) (2010), 268–272.
[10] P. Pitak-Arnnop, U. Bauer, K. Dhanuthai, M. Brückner, C. Herve, J.-P. Meningaud, and A. Hemprich, Ethical issues in instructions to authors of journals in oral-craniomaxillofacial/facial plastic surgery and related specialties, *J. cranio-maxillo-facial Surg.* **38**(8) (2010), 554–559.
[11] D. Strech, C. Metz, and H. Knüppel, Do editorial policies support ethical research? A thematic text analysis of author instructions in psychiatry journals, *PLoS One* **9**(6) (2014), e97492.
[12] M. Broga, G. Mijaljica, M. Waligora, A. Keis, and A. Marusic, Publication ethics in biomedical journals from countries in Central and Eastern Europe, *Sci. Eng. Ethics* **20**(1) (2014), 99–109.
[13] J. Stojanovski, Do Croatian open access journals support ethical research? Content analysis of instructions to authors, *Biochem. Medica* **25**(1) (2015), 12–21.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-120

Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research

Christian HANDKE[a,1], Lucie GUIBAULT[b] and Joan-Josep VALLBÉ[c]
[a]*Erasmus University Rotterdam*
[b]*University of Amsterdam*
[c]*University of Barcelona*

Abstract. With the diffusion of digital information technology, data mining (DM) is widely expected to increase the productivity of all kinds of research activities. Based on bibliometric data, we demonstrate that the share of DM-related research articles in all published academic papers has increased substantially over the last two decades. We develop an ordinal categorization of countries according to essential aspects of the copyright system affecting the costs and benefits of DM research. We demonstrate that countries in which data mining for academic research requires the express consent of rights holders, data mining makes up a significantly smaller share of total research output. To our knowledge, this is the first time that an empirical study identified a significant negative association between copyright protection and innovation. We also show that within countries where DM requires express consent by rights holders, there is an inverse relationship between rule of law indicators and the share of DM related articles in all research articles.

Keywords. Copyright, data mining, research output

1. Introduction

This paper discusses the effect of different copyright arrangements on data mining (DM) by academic researchers.[2] Hand et al. [1] broadly define DM as "the discovery of interesting, unexpected or valuable structures in large datasets." Digital information and communication technology (ICT) reduces the costs of collecting, accessing, combining and jointly analysing large amounts of data. DM is widely expected to increase the productivity of many types of research activities and to produce many valuable new insights. As we will show, conceptual, methodological and applied empirical DM research has accounted for an increasing share of total academic research output over the last two decades.

In particular, this paper is concerned with copyright arrangements that affect academic researchers' costs and benefits when accessing and jointly analysing data from databases or publications produced by others. The familiar expression "standing on the shoulders of giants" alludes to the cumulative nature of much academic research.

[1] Corresponding author. E-mail: handke@eshcc.eur.nl.
[2] This is an abbreviated and preliminary version, as of May 22, 2015.

The principle applies in cases where researchers acquire and analyse data collected and made available by others, say other researchers and publishers, private firms or public institutions.

Where unrestricted IP rights like copyright vest in 'input works' – databases or their content – data mining activities often require the express consent of rights owners to avoid the risk of litigation or other punitive measures. Subject to the transaction costs of clearing rights and any price charged by rights holders, effective copyright protection can thus increase the full economic costs of data mining. On the other hand, copyright could also encourage the supply of valuable input works and thus foster the benefits of data mining. We present empirical evidence regarding the net effect on the output of data mining-related academic research.

The evidence presented in this paper relates to a current policy debate in particular in the European Union (EU). Under current EU legislation (Directive 2001/29/EC on copyright in the information society and Directive 1996/9/EC on the protection of databases), DM requires prior authorization of rights holders even if the potential user has lawful access to the research articles and databases in question. The European Commission is currently considering copyright reforms to allow for data mining without express consent of the rights owner, so that the right to read would be the right to mine (cf. [2]). The USA have generally followed a more permissive attitude towards DM. Other countries like the United Kingdom and Japan have introduced more permissive legislation over recent years. Uncertainty complicates the matter, as neither the law nor the rights holders follow a clear line with respect to data mining. Another important variation is probably the extent to which rights holders or public authorities enforce copyright legislation in practice, and countries differ widely in this respect as well.

This paper exploits the variations in relevant copyright policy to develop evidence on the effect of different copyright regimes relevant for DM. We analyse bibliometric data to establish whether copyright policy and its enforcement affect the application of DM in academic research. We demonstrate that countries in which data mining for academic research requires the express consent of rights holders, data mining makes up a significantly smaller share of total research output.

2. The Empirical Literature

The empirical literature regarding copyright, academic research and DM in particular is limited to descriptive analyses. Regarding the supply of academic work, the paper by Tsai [3] contains recent bibliometric data on DM. Tsai uses information from the Social Science Citation Index, a section of a database called Thomson Reuter's Web of Science (WoS). He finds 1,181 academic publications between 1998 and 2009 with the topic "data mining". The vast majority of these articles, over 97%, were in English. Relevant articles are spread over a great number of academic fields.

The data presented by Tsai [3] illustrates rapid, approximately exponential growth in the number of DM-related publications and their citation counts between 1992 and 2009. For all the difficulties in predicting technological change, this makes rapid further growth likely. [3] also contains data on the share of various countries in DM related, academic publications. The U.S.A. accounts for almost 47% of all publications featuring DM in subject headers. Over 11% of the DM publications came from the U.K. and the other five largest EU economies accounted for just below 10% (Germany,

France, Italy, Spain, and the Netherlands).[3] Tsai [3] does not control for the size of countries and their domestic research output nor does he relate these findings to copyright policy. Filippov [4] contains an update of Tsai, confirming continued growth in the number of articles with "data mining" in the title up to 2013.

3. Empirical Strategy in this Paper

This paper makes greater use of the rich data available on academic research output than the preceding literature. Main factors driving the output of academic publications of any type should be (1) the means available for academic research, in particular labour and capital, and (2) the productivity of researchers, as measured by the number and quality of research output relatively to the resources used. To control for the size and productivity of academic research, we use the ratio between DM-related research output and total research output per country as the dependent variable.

The main independent variables of interest derive from a categorization of countries according to relevant copyright law and practice in each jurisdiction. Copyright protection has ambiguous effects according to economic theory. On the one hand, it should increase the number and quality of potential input works for DM applications made available. On the other hand, holding other things equal, stronger copyright protection increases the price of such works and the transaction costs compared to a situation where input works that researchers can acquire are available without an explicit contract with any rights holders.

There is often a gap between the provisions of IP law and social practice, since IP is hard to enforce. In our analysis, we thus consider relevant indicators of the rule of law within countries. The share of DM-related research output should also be affected by a number of further factors, for instance: (1) the supply of potential input works relevant to domestic academic researchers independent of copyright policy; (2) inter-country differences in academic cultures and incentive schemes that would affect the propensity to publish DM-related articles relative to other research output; (3) differences in the age structure of academic researchers that could affect adoption of DM, assuming that younger researchers may be more likely to adopt novel data collection and analysis methods; (4) targeted funding for DM; (5) learning curves as researchers improve their DM related skills with practice. However, no valid measures on these factors are available from a sufficient number of countries. We estimate a multilevel model to account for unobserved, constant country differences.

4. Data

4.1. Dependent Variable: Data Mining Research Output

One important measure of research output is the number of academic journal articles published. We collected data from Thomson Reuter's Web of Science (WoS). This is a relatively comprehensive database of academic publications, which features items from thousands of journals with a strong international reputation. We used the entire WoS

[3] The figure for the UK is the sum of England, Scotland and Wales reported separately on SSCI and in Tsai [3].

Core Collection Database including the so-called Science Citation Index Expanded, Social Science Citation Index and Art & Humanities Citation Index.

To identify the research output of interest, we extracted the number of all published research articles from a number of countries that contained the exact expression "data mining". Our panel includes the 15 largest EU member states, as well as the 27 largest other economies based on national GDP in 2013 according to the World Bank. The data covers the years 1992 to 2014. WoS includes articles published since 1975. It contains no articles on DM published before 1992. We thus have 966 country-year observations. In the data analysis, we exclude some countries for years on which no data on control variables are available.

The Boolean searches on the WoS database were defined by three simultaneous restrictions: (1) "data mining" entered in inverted commas in the field 'Topic'; (2) a country name according to the format used on WoS in the field author's 'Address', which relates to the country of residence of the first or main author; (3) a year of publication in the field 'Year Published'. Search results were further restricted by ticking the option 'Articles' in the user interface of WoS, so that results only contain academic journal articles rather than conference proceedings, book reviews and the like. For each country and year, we recorded the number of different items in the WoS database that fulfill these search criteria.

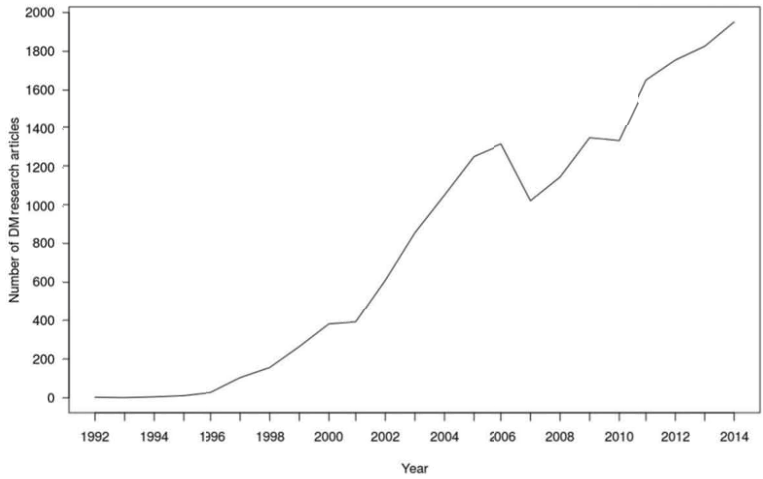


Figure 1. Absolute number of DM research articles published per year (42 countries, 1992 to 2014). *Source*: Own calculations based on search results on the WoS database.

The articles featured in the search results contain DM applications as well as related conceptual and methodological work. Among the 40 countries covered, searches on WoS brought up 18,441 DM-related articles between 1993 and 2014. We also collected data on the total number of research articles published for the same set of countries and years. Search parameters were the same as reported above, except that no 'Topic' was specified. This brought up 23,802,650 articles for the entire panel. That is for all countries and entire time period covered, 0.7‰ had DM as a topic. Starting from a low

base, there has been quite a rapid expansion of the DM share in total research output. See Figure 1 for an illustration.

In our empirical analysis, for each country and year we used the ratio of the number of DM-related research and total number of articles as the dependent variable. This variable is referred to as 'DM share' below.

4.2. Main Independent Variable: Copyright

Despite international and regional harmonization efforts, copyright protection is determined at the national level [5]. Different elements of the copyright regime may have an impact on the lawfulness of DM activities. The two features of the copyright system that bear the most on DM are the scope of rights granted on compilations of articles and other data, and the exceptions on these rights recognized in the various jurisdictions. Worldwide copyright laws can roughly be divided in two main traditions: first, the author's rights tradition, existing in countries of Continental Europe and countries that were inspired at some point in their history by the laws of one of these countries; second, the copyright tradition, reflected in the legislation following the Anglo-American legal system. Because the theoretical foundations of both regimes diverge, they are considered to follow a different approach with respect to the scope of rights and exceptions. In some countries, exceptions to copyright expressly allow DM activities to take place for research purposes without the authorization of the rights holder, while in others such activities are only lawful with a specific permission of the rights holder.

The laws of the countries examined in this paper are classified according to the possibility for academic researchers to engage in DM activities for research purposes without the need to obtain prior permission from the rights holder. Our assessment of the state of the copyright rules in each jurisdiction is based on a reading of the current legislative provisions, as well as the scholarly commentaries and the judicial interpretation, when available. In the following, we classify a number of other countries according to whether DM by academic researchers, who have lawful access to data, is either definitely 'not allowed', 'probably not allowed', 'probably allowed', or definitely 'allowed'.

Among the countries examined, sixteen belong to the European Union (EU) or the European Economic Area (EEA). Directive 2001/29/EC confers rights owners with the exclusive right to reproduce, communicate to the public and distribute their works. Directive 1996/9/EC grants protection with respect to non-original databases if they show a substantial investment in the obtaining, verification or presentation of the data. The rights granted under Directive 2001/29/EC and Directive 1996/9/EC have traditionally received a broad interpretation from the European Court of Justice (ECJ) [6, 7]. Directive 2001/29/EC contains a list of exceptions on these exclusive rights, the most relevant of which in the context of data mining allows Member States to provide for exceptions in the case of 'use for the sole purpose of illustration for teaching or scientific research. This exception is optional; Member States may decide whether to implement it or not [8]. As a result the research exception is generally vague and unevenly implemented at national level [7]. The same holds for the database right, where Member States are free to adopt an exception allowing the substantial extraction of the content of a database for research purposes, but not the re-utilization. Most EU and EEA Member states are thus 6 classified as academic DM 'not allowed' without express consent. The UK differs from the rest of the EU/EEA, as the legislator adopted

a specific copyright exception in the course of 2013, allowing DM activities for non-commercial research purposes to take place without the need to obtain prior authorization from the rights holders. Since 2014, the UK is thus classified as 'allowed'. Before, the situation in the UK was similar to that of the rest of Europe ('not allowed').

Switzerland, not being an EU/EEA country, is not bound by the European legal framework. The Swiss Copyright Act grants authors of original works a number of exclusive rights including that of reproduction, retransmission and making available. Among the several exceptions listed in the act, none would seem to cover acts of mining for purposes of research, beyond the right to make a copy for private use. Switzerland does not protect databases separately. Nevertheless, DM activities are most likely 'not allowed'. The same holds for Russia and Turkey. In these countries, DM activities are most likely 'not allowed'.

The laws of most countries with a colonial past have been influenced by those of the colonial power, most often a European country. The copyright laws of Latin American countries[4] find their root in the legislation of Continental Europe, where the rights and exceptions recognised show similar features to their European counterparts. As a result, acts of DM are most likely 'not allowed' in these countries.

While the modern Japanese Copyright Act was strongly influenced by the German Copyright Act of 1965, the recent development of the Japanese Act pursued its own course, under a greater influence from the USA. Up to 2009, DM activities were 'probably not allowed' in Japan. In that year, Japan is reported to have been the first country in the world to adopt a specific copyright exception allowing the 'analysis of in copyright works using computers in order to extract statistics and information', and come up with new ideas [9]. At the current state of our information, Japan is classified as 'allowed' since 2010.[5]

In the countries adhering to the Anglo-American copyright system, there are two different approaches with respect to exceptions on copyright: some countries recognize a 'fair dealing' exception, while others recognize a 'fair use' defense. The countries of the British Commonwealth[6] commonly recognize 'fair dealing' exceptions for different purposes, including for criticism and comment, private use and research. To fall under the fair dealing exception, the purpose of the dealing must qualify as one of the allowable purposes under the copyright act, and the dealing must be fair. The fair dealing exception generally receives a restrictive interpretation, which lets us conclude that DM activities are 'probably not allowed' in most 'fair dealing' countries. Compared to other 'fair dealing' countries, Canada has followed in recent years a more flexible approach: not only has the Supreme Court ruled twice in favour of fair dealing for research purposes, but the Copyright Act was amended in 2012, to expand the allowable fair dealing purposes. Since 2012, DM activities in Canada are 'probably allowed'. The same development can be observed for Singapore where acts of DM are today 'probably allowed'.

The 'fair dealing' exception differs from the 'fair use' defense primarily in the fact that the latter is characterized by an open-ended list of purposes for which the use of a work may be regarded as fair, marked by the words 'such as'. The fair use defense was first developed at the beginning of the 20th Century in the USA as a judicial doctrine before being codified in § 107 of the Copyright Act 1976. The assessment of whether a

[4] Argentina, Brazil, Colombia, Mexico, Venezuela.

[5] The classification of Japan is not straightforward, since this provision excludes its application to databases that are precisely made for data analysis.

[6] Australia, Canada, India, Ireland, Malaysia, Nigeria, Singapore, South Africa, United Kingdom.

particular use is fair is done by the judge according to four factors: the purpose and character of your use; the nature of the copyrighted work; the amount and substantiality of the portion taken, and the effect of the use upon the potential market. The USA is classified as 'probably allowed'.[7]

For a long time, the fair use doctrine was a unique feature of the American copyright regime. The copyright acts of countries, like Israel and the Republic of Korea, contained a list of specific exceptions, which were too narrow to cover acts of DM. However, following the conclusion of a bilateral trade agreement with the U.S.A., both Israel (2008) and the Republic of Korea (2012) introduced a fair use defense in their copyright legislation in addition to a list of specific exceptions. Since the legislative amendment introducing the fair use defense, acts of DM possibly shifted from 'probably not allowed' to 'probably allowed'.

The People's Republic of China only adopted Berne Convention compliant copyright norms in 2007, upon its accession to the TRIPS Agreement. Before that time, copyright protection on the Chinese territory was below the Berne standard, meaning that before 2007 the existence and enforcement of copyright rules was not a priority. The Chinese Copyright Act of 2007 lists the permissible exceptions, including for use of a published work for the purposes of the user's own private study, research or self-entertainment. Literal interpretation of this provision would not permit acts of data mining. However, Geller and Nimmer [10] report that the Supreme People's Court of China issued a policy document at the end of 2011, according to which in circumstances necessary to stimulate technical innovation and commercial development, an act that would neither conflict with the normal use of the work nor unreasonably prejudice the legitimate interest of the author could be deemed "fair use". This policy document was followed in a 2014 case. Since that time, acts of DM are 'probably allowed' in China while they were 'probably not allowed' between 2007 and 2012.

Taiwan Copyright Law has a long history, having first been enacted in 1928. Taiwan's modern Copyright Act was adopted in 1992 and contained a list of exceptions, none of which was broad enough to encompass DM activities. In recent years, Taiwan copyright law has been marked by American influence. In 2003 the list of exceptions was complemented by a fair use provision, which must be applied in conjunction with the specific exceptions. Since that time, acts of DM are 'probably allowed' in Taiwan, provided that they meet the four factors used to evaluate fair use in any of its many enumerated circumstances. By contrast, the law of Thailand contains a more restrictive provision according to which DM is 'probably not allowed'.

Although most Muslim countries included in the sample are members of the Berne Convention, finding specific information on the scope of the exceptions in the laws of Iran, Indonesia, Saudi Arabia and the United Arab Emirates proves difficult. They are thus excluded in the data analysis. See Table 1 for the average DM share among the four main copyright categories. By far the largest number of observations is available for the category 'not allowed', and the average DM share for this category is lower than for all other categories. The category 'allowed' only contains six observations, so that it is hardly suited for a statistical analysis. The average DM share for this category is

[7] A very recent ruling in 2014 may result in a status change of the USA to 'allowed'. In the Authors' Guild of America vs. Hathitrust case, the Court of Appeal for the Second Circuit ruled in 2014 that the digitization of books held by the Libraries for the purpose of allowing full-text searches is permissible under all four fair use factors. United States Court of Appeal for the Second Circuit, June 10, 2014 (Authors' Guild of America vs. Hathitrust), No. 12-4547-cv.

relatively low. These observations come from very recent changes, however. Chances are that the full effect of changing to 'allowed' transpire over a longer period than covered by our data.

Table 1. Categorization of countries according to their level of copyright restriction to data mining research.

Country	Not allowed	Probably not allowed	Probably allowed
Not allowed	0.54	0.54	528
Probably not allowed	0.67	0.66	162
Probably allowed	1.64	1.37	71
Allowed	0.60	0.18	6

4.3. Other Control Variables

Besides the total research output of countries, we use several control variables: (1) GDP per capita as reported by the World Bank World Development Indicators [11], with complete data for the 1992-2013 period; (2) country population size, also from official World Bank data [11], also complete from 1992 until 2013; (3) and the level rule of law as reported by the Worldwide Governance Indicators Project [12]. The level of rule of law is captured by one of the six dimensions of governance of the WGI indicators, and is defined as "the extent to which agents have confidence in and abide by the rules of society" [12], including the quality of contract enforcement and property rights. We use it as a proxy to measure the level of enforcement of property rights. Data availability for this indicator begins in 1996 and last estimates are from 2013.

5. Main Empirical Results

Due to the panel structure of our data and the low temporal variation of copyright legal arrangements within countries, we fit a multilevel linear regression model with varying intercepts by group (i.e. country), also known as a random effects model. The dependent variable is the share of articles on DM in the total number of articles published (DM share) per country and year. The main predictor is each country's copyright category, with 'not allowed' as the reference category. Table 2 presents the results of four different specifications of the model.

Model 1 only contains the main predictor. As expected, the 'allowed' category does not yield significant coefficients: it contains only six observations and we only report it for completeness.[8] There is a significant positive coefficient for the category 'probably allowed', which suggests that a more permissive copyright framework is associated with more DM research. The specification in Model 2 tests the effect of copyright categories controlling for GDP per capita, population size, and the rule of law. In Model 3 we also control for the total number of research articles published to test whether changes in DM share are confounded by changes in total research output. (We discuss Model 4 with interaction terms separately below.) The number of observations is reduced in models with control variables, since no data are available on the 'rule of law' before 1996. The control variables improve model fit considerably

[8] Furthermore, the effects of introducing permissive copyright regulations on DM share should be gradual, so that the full effect of recent changes in Japan and the UK may not have transpired.

compared to Model 1. In all specifications, we find significant positive coefficients for the category 'probably allowed' (p<.01). For the category 'probably not allowed', results are less stable. Coefficients for the category 'allowed' are generally positive but not significant with a very low number of observations. The coefficient for 'probably allowed' is consistently larger than for 'probably not allowed' in all specifications. This is in line with our ordinal categorization: there is a stronger and more reliably significant coefficient for the category that differs more from the reference category 'not allowed'. Overall, there is extensive evidence that DM share is greater in countries with more permissive DM-related copyright than in the 'not allowed' category of countries.

Table 2. Results of the multilevel regressions (varying intercept, random effects) with DM share as dependent variable.

	Model 1	**Model 2**	**Model 3**	**Model 4**
(Intercept)	0.542*** (0.075)	0.559 (1.072)	3.550*** (1.120)	4.388*** (1.009)
Copyright [Ref. *Not allowed*]				
Allowed	0.245 (0.283)	0.440 (0.355)	0.362 (0.331)	9.383 (26.470)
Probably allowed	1.407*** (0.160)	1.653*** (0.191)	1.465*** (0.188)	1.419*** (0.327)
Probably not allowed	-0.005 (0.137)	0.403*** (0.164)	0.264 (0.164)	0.503** (0.229)
GDP/capita ($1,000)		0.046*** (0.006)	0.013*** (0.007)	0.007 (0.006)
Population (log)		-0.028 (0.059)	-0.478*** (0.076)	-0.512*** (0.072)
Rule of Law		-0.761*** (0.126)	-0.701*** (0.120)	-0.639*** (0.116)
Total research output (log)			0.604*** (0.064)	0.600*** (0.065)
*Definitely allowed**Rule of Law				-6.922 (20.131)
*Probably allowed**Rule of Law				0.036 (0.266)
*Probably not allowed**Rule of Law				-0.258 (0.187)
R^2	0.144	0.233	0.351	0.333
F	42.820***	28.208***	42.948***	27.564***
N	767	564	564	564

*p<0.1; **p<0.05; ***p<0.01

In Models 3 and 4, 'total research output' has a positive and significant coefficient. Countries with a high share of data mining articles in total research output also tend to have larger total research output. There is thus no indication that DM would reduce incentives for other types of research within the same country. With the control for total research output in Model 3, the coefficient for 'probably allowed' changes little compared to Model 2.

For the entire panel, the rule of law consistently has a significant, negative coefficient. The rule of law should make the enforcement of copyright law more effective, and thus have a stronger negative effect on DM share in countries with

stronger copyright. To test for this, Model 4 includes a multiplicative interaction between copyright categories and the rule of law indicator. In this model, the coefficients of the variables that constitute the interaction (the categories of copyright regulation and rule of law in Model 4) are no longer to be interpreted as unconditional marginal effects. For instance, the coefficient of the 'probably allowed' constitutive term (1.419) represents the effect of this type of copyright regulation only when the rule of law is zero.[9] The coefficients for the multiplicative interaction terms (copyright*rule of law) are not significant, suggesting that for less restrictive countries, different levels of rule of law do not affect DM share. However, the coefficient for the constitutive term of the rule of law – that represents the effect of the rule of law for countries in the category 'not allowed' – is negative and significant. This suggests that in particular the combination of strong copyright law and strong enforcement (and/or a cultural propensity to adhere to legal norms) reduces academic researchers' data mining performance.

6. Conclusions

In most EU/EEA member states, DM-related copyright protection is comparatively strong. Our results suggest that the net effect is a weaker performance of domestic academic researchers in this increasingly important type of research. To our knowledge, this is the first time that an empirical study identified a significant negative association between copyright protection and the supply of new copyright works of any type. The academic culture and incentive scheme (based on public subsidies and the ideal of producing public goods) sits uncomfortably with academic publishing, which is operated by for-profit firms. One of the battle lines regarding DM is between for-profit academic publishers and representatives of some academics, universities and libraries. At least in Europe, publishers tend to favour restrictions on DM so that there is greater potential to sell rights to DM, whereas many representatives of the "academic community" favour a situation in which academic researchers, who have lawful access to input works, are generally allowed to conduct DM on these works. Our results suggest that in the case of academic research and DM, the adverse consequences of copyright protection on the creation of new information goods are greater than the benefits. As a rule, DM research draws heavily on input works to which others may hold copyrights. Copyright exemptions or limitations could promote this type of research, at least to enable DM of input works that have been publicly financed.

References

[1] D. J. Hand, H. Mannilla, P. Smyth, *Principles of Data Mining*, The MIT Press, Cambridge, 2001.
[2] P. Murray-Rust, *Open content mining*, Working Paper, Cambridge University, 2012. Online: http://www.dspace.cam.ac.uk/handle/1810/243749
[3] H.-H. Tsai, Global data mining: An empirical study of current trends, future forecasts and technology diffusions, *Expert Systems with Applications* **39** (2012), 8172–8181.
[4] S. Filippov, *Mapping text and data mining in academic and research communities in Europe*, The Lisbon Council, Brussels, 2014.

[9] In our panel, there are only four observations in the data with rule of law between -0.01 and 0.01 (there are no exact zero matches), which are South Africa in 1996, Argentina in 1997, and Brazil in 2010 and 2011.

[5] P. Goldstein, P. B. Hugenholtz, *International copyright*, Oxford University Press, Oxford, 2012.
[6] I. Hargreaves, L. Guibault, C. Handke, P. Valcke, B. Martens, *Report from the expert group on standardisation in the area of innovation and technological development, notably in the field of text and data mining*, Publications Office of the European Union, Luxembourg, 2014.
[7] J.-P. Triaille, S. Dusollier, S. Depreeuw, J.-B. Hubin, A. De Francquen, *Study on the application of Directive 2001/29/EC on copyright and related rights in the information society*, European Commission, Brussels, 2013.
[8] L. Guibault, Why cherry picking never leads to harmonisation: The case of the limitations on copyright under Directive 2001/29/EC, *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* **1** (2010), 55-66.
[9] J.-P. Triaille, *Study of the legal framework of text and data mining (TDM)*, European Commission, Brussels, 2014, p. 10.
[10] P. E. Geller, M. B. Nimmer, *International copyright law and practice*, Matthew Bender, Los Angeles, CHI-72, 2015.
[11] World Bank, *World Development Indicators*, Data, 2015.
[12] World Bank, *Worldwide Governance Indicators (WGI) Project*, Data, 2015.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-131

Finding the Law for Sharing Data in Academia

Esther HOORN[a,1] and Marlon DOMINGUS[b,2]
[a] *Rijksuniversiteit Groningen, The Netherlands*
[b] *Erasmus University Rotterdam, The Netherlands*

Abstract. How can universities provide good advice about the legal aspects of research data management? At the same time, how can universities prevent that perceived legal risks become barriers to: conducting research, sharing research data, valorisation of research data, and control mechanisms for the purpose of scientific integrity? A Dutch expert group developed a creative approach based on some core ideas[3] about regulation in the field of academic research.

Keywords. Research, hard law, soft law, code of conduct, guidelines, model contracts, wiki

1. From Self-funded Science to Publicly Funded Academia

History shows that the funding of research has an impact on research itself. From a Renaissance culture with its roots in self-funded science, via noble and religious patronage, to government funding, military funding, patent profits, corporate sponsorship, and private philanthropists, researchers have found ways to contribute to science or scholarship as well as meet the requirements of their research grant providers.

Today, Academia[4] in the Netherlands is mainly publicly funded[5] but in recent years the so called *third flow of funds* has seen a substantial increase.[6] This refers to revenues based on contract research and funds from Dutch ministries and the European Union (FP7 and Horizon 2020).

With the third flow of funds comes the obligation to meet the new funder's requirements. The EU grants, for instance, are aimed at fostering interdisciplinary re-

[1] Corresponding Author. E-mail: e.hoorn@rug.nl
[2] Corresponding Author. E-mail: domingus@ubib.eur.nl
[3] The views expressed here are those of the authors and do not necessarily reflect those of the Rijksuniversiteit Groningen or Erasmus University Rotterdam.
[4] Used as an equivalent to the 14 research universities in the Netherlands. The case for private universities and university colleges is not investigated here.
[5] Directly and indirectly funded by the government. Direct funding from the Ministry of Edu cation, Culture and Science, and Wageningen University's funding comes from the Ministry of Eco nomic Affairs (the so called *first flow of funds*). Indirectly funded by the government: grants from the Dutch Organisation for Scientific Research (NWO) and the Royal Netherlands Academy of Arts and Sciences (KNAW) (the *second flow of funds*). See also: http://www.vsnu.nl/funding-en.html and http://www.rathenau.nl/fileadmin/user_upload/rathenau/De_Nederlandse_Wetenschap/Facts_and_Figures-Dutch_universities_2012__01.pdf
[6] From 22% (2004) to 26% (2013) of the gross income of Dutch research universities. Online source: http://www.vsnu.nl/funding-en.html

search, international collaborations as well as Industry-Academia Partnerships. And the EU runs a pilot on open data:

"*The European Commission's vision is that information already paid for by the public purse should not be paid for again each time it is accessed or used, and that it should benefit European companies and citizens to the full.*" [1]

So Academia's move to the public domain for research funding has resulted in a new Academia ecosystem in which new (consortium) agreements and understandings are to be formulated.

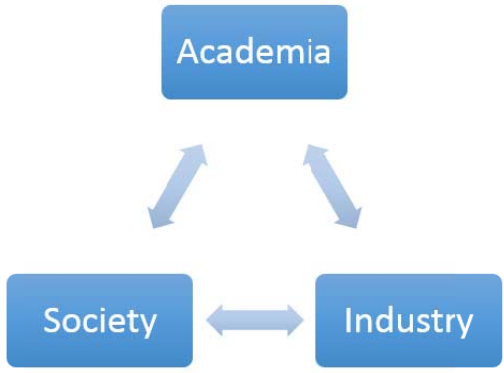


Figure 1. Academia Ecosystem.

Signs of the new ecosystem are, for instance, Academia's focus on 'societal impact' and 'societal relevance' [2]. In the new ecosystem, the moral pressure to make publicly available what was publicly funded, relates to both research publications and the underlying research data. Dissemination of research data is demanded from the researcher rather than exploitation of the data.

The parallel with a company's shareholders is clear: government invests in Academia with citizens as the Academia shareholders, advocating citizen science. Academia's capital is arguably the collective research data, which combined with government's open data and the industry's data open all kinds of new possibilities for all in the ecosystem.

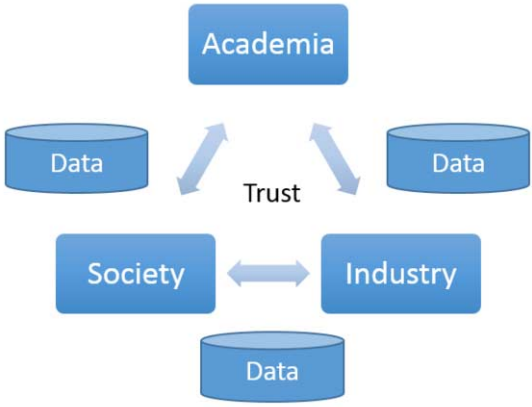


Figure 2. Data sharing in the ecosystem.

In this new ecosystem new rules and agreements are required, new responsibilities assigned for the ecosystem to find its balance. Data exchange seems only possible based on mutual trust. How can universities best provide good advice to their researchers about the new legal and ethical aspects?

2. Hard Law, Soft Law and Ethics

Whereas in some fields of law norms and procedures are crystal clear, in other fields of law open norms leave bargaining power to participants. When considering a good approach for raising awareness on legal aspects of sharing research data, this fuzzy approach to law seems to match the nature of law in the field of academic research. In addition to the criteria of research funders, a broad range of hard law, like privacy regulation and contract law and soft law (opinion juris), like research codes and discipline-specific norms are applicable in decisions about openness and involvement of citizen in science. A code of conduct is not a body of law, but a canon for self-regulation, based upon ethical principles [3]. For Academia these principles are summarized as: Responsible Research and Innovation, for Society: Good Citizenship and for Industry: Corporate Social Responsibility. When research is considered, we distinguish different ethical dimensions [3]: with regards to the *context of research*[7] and the *responsible conduct*

[7] ALLEA, p. 10: "Could the research result in harm for people, nature or society, or be in conflict with basic human values?" This aspect, however, is ignored in the ALLEA Code. R. von Schomberg, European Commission-DG Research and Innovation, introduces the following normative anchor points for this ethical dimension: 1. Compliant with fundamental rights. 2. Sustainable and 3. Socially desirable. See: Von Schomberg, Prospects for Technology Assessment in a framework of responsible research and innovation, M. Dusseldorp and R. Beecroft (eds). *Technikfolgen abschätzen lehren: Bildungspotenziale transdisziplinärer Methoden* (2011),Wiesbaden: Vs Verlag.

of research.[8] In conclusion: good advice to researchers addresses hard law as well as soft law and takes into account the two ethical[9] dimensions.

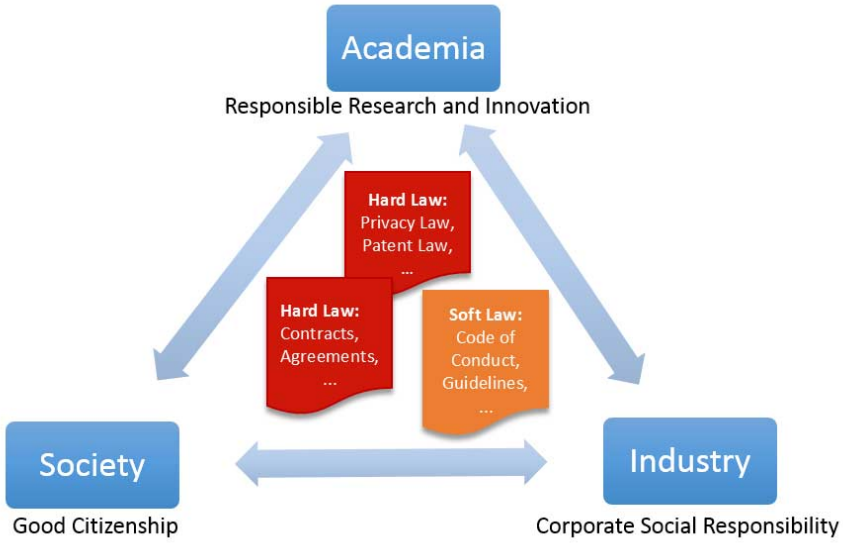


Figure 3. Hard law and soft law in the ecosystem.

At the same time, an approach that depends on specific national or local regulations (hard law) is not considered fruitful, especially since research practices tend to transcend borders, borders between different legal systems with respect to legal aspects of data management. For instance, in the Netherlands ownership of data is not well defined in law. This might be a blessing in disguise to stimulate discussions about data stewardship [4] and shared responsibilities in data management practices. For legal practitioners, however, new solutions are sought within the system of law starting from the relevant and vigorous field of IP law in which the institution is the rights holder.

It's interesting to observe here that principles of scientific integrity are perceived to have a universal character, whereas the different formal legal systems as well as the different good practice rules have national boundaries.[10]

[8] ALLEA, p. 10. This amounts to definitions of proper scientific practice and of scientific misconduct, based upon principles of scientific integrity, and guidelines for good practice rules.

[9] Von Schomberg suggests to regard ethics as a "design" factor of technology and increase social-ethical reflexivity in research practices by incorporating ethical principles in the design process of technology (privacy by design as an example), which can lead to well accepted technological advances. See p. 15: Von Schomberg, Prospects for Technology Assessment in a framework of responsible research and innovation, M. Dusseldorp and R. Beecroft (eds), *Technikfolgen abschätzen lehren: Bildungspotenziale transdisziplinärer Methoden* (2011),Wiesbaden: Vs Verlag,.

[10] See ALLEA: *The European Code of Conduct for Research Integrity*, p. 9.

3. Dimensions in Data Openness

Data openness is yet another dimension which needs to be addressed here. In our ecosystem we would expect there to be data sharing between Academia, Society and Industry, based on a mutual trust, once agreement has been reached on relevant aspects of hard law and soft law. This data openness should be regulated and well defined, but open. We see, however, different dimensions in data openness.

The European Commission supports open data. Open data refers to the idea that certain data should be freely available for use and re-use [5]. More precisely, open data is the engine for innovation, growth and transparent governance [6]. Furthermore, open data is supported by the European Commission in the context of open science[11] and citizen science[12].

Similarly the Netherlands Organization for Scientific Research (NWO) holds [7] that research results paid for by public funds should be freely accessible worldwide. This applies to both scientific publications and other forms of scientific output. In principle, it should be possible to share the research data with others as well. In this way, valuable knowledge can be utilised by researchers, businesses and civil society organisations.

From the point of view of Academia[13] however, openness seems more restricted to fellow researchers (interested colleagues) from whom the general public benefits through their publication of research findings, thus contributing to public knowledge. For this reason, research data should be available to colleagues who want to replicate the study or elaborate on its findings.[14] To be realistic: within Academia the role of the individual researcher and his / her motives for not creating open data should not be underestimated. Within many disciplines, sharing too much data too soon[15], could endanger an academic career.

[11] See the EC policy on Open Science: The European Commission has promoted an approach to research and innovation in which all societal actors (researchers, citizens, policy makers, businesses, civil society organisations, etc.) work together during the whole Research and Innovation process, with the aim to better align research and innovation outcomes with societal values needs and aspirations. It has referred to this approach as Responsible Research and Innovation (RRI). Source: http://ec.europa.eu/research/swafs/index.cfm?pg=policy&lib=science

[12] See the definition in the 2014 EC Green Paper on Citizen Science [*Citizen Science for Europe. Towards a better society of empowered citizens and enhanced research*]: "Citizen Science refers to the general public engagement in scientific research activities when citizens actively contribute to science either with their intellectual effort or surrounding knowledge or with their tools and resources. Participants provide experimental data and facilities for researchers, raise new questions and co-create a new scientific culture. While adding value, volunteers acquire new learning and skills, and deeper understanding of the scientific work in an appealing way. As a result of this open, networked and trans-disciplinary scenario, science-society-policy interactions are improved leading to a more democratic research based on evidence-informed decision making." Online source: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=4121

[13] See ALLEA (p. 10, §2.2.3.) the: *Open Communication* principle.

[14] See ALLEA (p. 13, §2.3.) *Good data practices: availability and access*.

[15] Sometimes it can *never* be the case that research data comes available; to protect the privacy of patients and / or to protect the commercial interests of Industry. This is the case with most medical research, for example in the field of epidemiology.

4. A Data Case

In the case of the recent NWO data contract (April 2015) NWO states that researchers are in some cases (depending on field of science and type of the awarded NWO grant) obliged to enter into a data contract with Data and Archiving Services (DANS), the NWO service provider for research data archiving. The data contract is intended to guarantee accessibility to the data as well as digital sustainability of the data for additional scientific research [8]. As we have learned from NWO's open data position, however, the data should in principle be available not only to researchers, but also to businesses and civil society organisations.

At the same time the granting conditions of NWO recognise the research funder's and the research institution's shared ownership of the research data. It is evident that the data management of a project must also be in line with the institutional policies and responsibilities regarding research data.

Moreover, the interests of the partners involved in the project also need to be taken into account. In this respect, ethical commissions play an important role by shaping self-regulation for research integrity via peer assessments. The institution has a legal obligation [9] to ensure that a proper and independent assessment framework is in place to assure recommendations are followed up. Finally, when personal data is involved, the institution is responsible for technical and organisational measures (privacy by design) to ensure privacy of participants, also during and after the project.

Analysing the contract from a lawyer's perspective the contract seems to imply that the project-leader can bind the institution to sign a license after the project. This is not the case. It would be good practice to make this explicit in the contract.

5. Approach: Legal Research Support

So, the research data landscape is altogether a complex one. In any given case, ethical, legal and social implications can be identified, as we have seen. These implications may be perceived as / or may actually be barriers to:

- conducting research,
- sharing research data,
- valorisation of research data and
- control mechanisms for the purpose of scientific integrity.

How can we lift these barriers? In essence: all those concerned need to have a suitable understanding of the matters. Some, however, should acquire expert's[16] knowledge on these issues and should act as the go-to person for identifying what the relevant aspects are, what the relevant ruling is, what the course of action should be and what agreements, contracts or otherwise need to be formulated.

Currently in Academia in the Netherlands, research support services is a joint effort of staff from Faculty, ICT, Library, Legal Affairs, Academic Affairs, Valorisation Offices and Patent Offices in many different roles. As is the case with most hot topics, there are many perspectives, opinions and interests. This is immediately clear once you

[16] Rob Posthumus suggested the term 'consciously incompetent' in this respect, derived from the psychological "conscious competence" learning model, usually attributed to Abraham Maslow. It is the expert's task to be 'consciously competent'.

engage in a discussion on the topic of research data management. But it is also a topic no single person is likely to get their head around as it requires co-operation.

We propose not so much a staff / responsibilities matrix, but rather suggest that within a university the following basic steps should be taken by experts:

1. Identify barriers and pitfalls, for instance in a research project plan,
2. Acquire accurate knowledge of the rules and requirements regarding to research data,
3. Take proper legal advice and applying this legal knowledge correctly and timely, for instance by writing tailored paragraphs in a Consortium Agreement.

Even within a single university this is a challenging task. The co-operation between universities to collectively build a body of knowledge, best practices and model agreements looks promising. From a researcher's perspective, it should be clear whom to turn to for support related to these matters. We suggest implementing an awareness program in which the researcher is offered an overview of the research support services and choose[17] from them when specialized support is needed.

From the expert point of view, one would expect an ongoing process of creating a structured body of knowledge, resulting in a detailed analysis of the legal and ethical requirements and the corresponding best practices, template paragraphs and model agreements.[18] On a research data management services level, this list could be considered as a basic functional roadmap to ensure that what is agreed within a research project, is actually executed as promised.

6. The Wiki and the Mood Board

In a seminar [10] addressing these matters, a Dutch expert group presented a wiki [11], in which the three basic steps as described above form *la ligne rouge*.

As a tool to further discussions about legal aspects of data management, the diversity of perspectives and approaches to regulation is visualised on a map. This Mood Board [12] is also a playful way to identify domain specific legal and ethical barriers and pitfalls.

The group is now setting up a broader network with legal practitioners of the institutions and hopes to set, with you, an agenda for the development of helpful support material.

[17] And in this respect following *The Netherlands Code of Conduct for Academic Practice. Principles of good academic teaching and research.* Association of Universities in the Netherlands (VSNU), 2014. Online source:
http://www.vsnu.nl/files/documenten/Domeinen/Onderzoek/The%20Netherlands%20Code%20of%20Conduct%20for%20Academic%20Practice%202004%20(version%202014).pdf

[18] The wiki mentioned in the next paragraph is, in our view, a candidate of a platform of such a body of knowledge. See for instance the section: *The landscape of present rules and requirements regarding to research data, for instance in the recently revised Code of Conduct for Scientific Practice and in the regulations applied by research funding bodies*: https://wiki.surfnet.nl/pages/viewpage.action?pageId=47449662

Figure 4. Mood Board.

Acknowledgements

The authors wish to thank the members of the expert group for fruitful discussions and insights:

- John Doove (SURF - Collaborative organisation for ICT in Dutch higher education and research),
- Rob Grim (Radboud University, Open Data Foundation),
- Theo Hoksbergen (Wageningen UR),
- Kim Huijpen (VSNU – Association of universities in the Netherlands),
- Ana van Meegen Silva (VU University Amsterdam),
- Heiko Tjalsma (Data Archiving and Networked Services),
- Juliën Visser (Erasmus University Rotterdam) and
- René Winter (Erasmus University Rotterdam).

A special thanks to Rob Posthumus (Erasmus University Rotterdam).

Many thanks Heather Boet-Foley (editor) for your professionalism and consideration.

References

[1] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, Version 1.0. 11 December 2013. Online source: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

[2] R. Owen, Responsible research and innovation: From science in society to science for society, with society, *Science and Public Policy* **39**(6) (2012), 751–760.

[3] All European Academies (ALLEA), The European Code of Conduct for Research Integrity, 2011. p. 8 Online source: http://www.esf.org/fileadmin/Public_documents/Publications/Code_Conduct_ResearchIntegrity.pdf

[4] D. Kleppner, Ensuring the integrity, accessibility, and stewardship of research data in the digital age, *International Association of Scientific and Technological University Libraries, 31st Annual Conference*, (2010). Online Source: http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1023&context=iatul2010

[5] http://ec.europa.eu/digital-agenda/public-sector-information-raw-data-new-services-and-products

[6] European Commission, Open data. An engine for innovation, growth and transparent governance, 2011. http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF

[7] Online source: http://www.nwo.nl/en/policies/open+science

[8] Online source: http://www.dans.knaw.nl/en/deposit/information-about-depositing-data/datacontract/data-contract?set_language=en

[9] Online source: Artikel 1.7, Wet op het hoger onderwijs en wetenschappelijk onderzoek. Online source (Dutch only): http://wetten.overheid.nl/BWBR0005682/Hoofdstuk1/Titel1a/Artikel17a/

[10] A. van de Wijngaart, *Mapping ownership in the data landscape*. Online source: https://wiki.surfnet.nl/download/attachments/47449647/Seminar%20Report%20RDO.pdf

[11] Online source: https://wiki.surfnet.nl/display/RD/WIKI+Research+Data+Ownership

[12] Online source: https://wiki.surfnet.nl/pages/viewpage.action?pageId=47449973

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-140

Open Data in Global Environmental Research: Findings from the Community

Birgit SCHMIDT[a,1], Birgit GEMEINHOLZER[b] and Andrew TRELOAR[c]
[a] *University of Göttingen, State and University Library*
[b] *University of Giessen*
[c] *Australian National Data Service*

Abstract. This paper presents selected findings of the Belmont Forum's survey on open data which targeted the global environmental research and data infrastructure community. It highlights users' perceptions of the term "open data", expectations of infrastructure functionalities, and barriers and enablers for the sharing of data. Respondents also pointed out a wide range of good practice examples and a desire for enhancement and consolidation.

Keywords. Open data, sharing, e-infrastructures, global environmental change

1. Introduction

The Belmont Forum [1], a group of high-level representatives from major funding agencies across the globe, coordinates funding for collaborative research to address the challenges and opportunities of global environmental change. In the course of this, the Belmont Forum E-Infrastructures and Data Management Collaborative Research Action [2] was initiated in 2013, to survey the state of the art and establish recommendations on how the Belmont Forum can leverage existing resources and investments to better foster a more coordinated, holistic, and sustainable approach to the funding of global environmental change research.

Experts from more than 14 countries collectively assessed existing international e-infrastructure capabilities, gaps and overlaps. They prioritized challenges, and provided recommendations for developing and sustaining human and technical international data infrastructures.

In the context of the working group on open data (one of six working groups), a survey invited researchers from various science communities, interested laypersons, government employees, and others who are providing and/or using open data in the scope of environmental change, or are planning/interested in doing so in the future, to share their views and experiences on data publishing, access and (re)use.

The main aim of the survey was to learn more about key open data initiatives of relevance for global environmental change from a data user/provider/manager perspective; areas where users' desire to share could be enhanced by new/other developments; and to detect barriers to "open data sharing" from a user perspective.

[1] Corresponding Author. E-mail: bschmidt@sub.uni-goettingen.de.

2. Methods and Results

From September to November 2014, the survey collected 1,330 responses through a web form [3]. Of these, 1,253 qualified as valid responses. The survey was distributed to about 20 disciplinary and professional mailing lists, and to all the authors of a well-renowned open access publisher, central to the research area (Copernicus Publications). A potential bias in the responses should be taken into account as the participants of the survey might not be representative of the community, and might also be more positive towards the topic of "open data" than the average researcher. Note that "open data" was not defined in the survey – respondents were free to respond on the basis of their own understanding.

Table 1 indicates the regional distribution of the collected data.

Table 1. Population of the survey

	Frequency (N=1248)	Percentage
Germany	205	16.4
United States	184	14.7
Italy	117	9.4
United Kingdom	88	7.1
France	68	5.5
Australia	45	3.6
Spain	43	3.4
China	39	3.1
Other countries (76)	459	36.8

As expected, the majority of respondents belonged (multiple answers were allowed) to earth and environmental sciences (67.5%, 846 answers) as well as climate and atmospheric sciences (30.8%, 386 answers). In addition, there were 50 or more answers from the biological sciences (20.6%, 258 answers), physical sciences (12.9%, 162 answers), engineering (7%, 88 answers), computer sciences (6.8%, 85 answers), social sciences (5.3%, 66 answers), agricultural and veterinary sciences (4.2%, 53 answers) and chemical sciences (4%, 50 answers).

The survey results highlight the users' perspective about the term "open data", such as the importance of information that enables users to assess the quality of data (82% very important), to select data based on metadata (78%), and to easily access (76%) and reuse the data (70%). The provision of unrestricted data was considered as very important by 2 out of 3 respondents. Open data already seems to be of substantial relevance for the global environmental change research community as more than 4 out of 5 respondents considered open data as important for advancing research. Half of the respondents saw open data as important for supporting applications to societal problems.

Moreover, motivators and barriers to publish data as open data were analyzed. The desire to publish data as open data was mainly linked to research-intrinsic motives ranging from general considerations, i.e. the acceleration of scientific research and applications, to personal motivations, i.e. dissemination and recognition of research results, personal commitment to open data and requests from data users (see Fig. 1).

From all the policies, funder policies seem to be the most important motivator, supporting the conclusion that stronger mandates will likely further strengthen the case for data sharing.

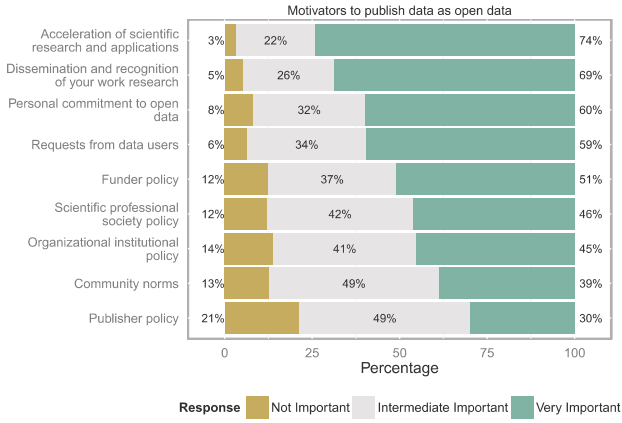


Figure 1. Motivators to publish data as open data.

The most important barriers (see Fig. 2) were the desire to publish results before releasing data (54% very important), legal constraints (47%), loss of credit and recognition (41%) and possible misinterpretation or misuse (37%). The ranking of these perceived barriers varied across fields, e.g. legal constraints were the most important barrier in economics, computer sciences and engineering. As expected, the desire to publish results before releasing data was most prevalent at early stages of a research career, i.e. the age from early to mid 30s, and was perceived as a major barrier by 69% of all respondents.

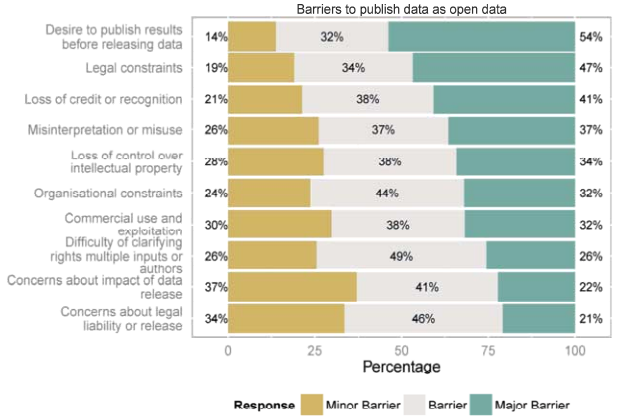


Figure 2. Barriers to publish data as open data.

When it comes to expectations towards infrastructure, the most important functionalities are that authorship and attribution are highlighted (75% most important, 23% intermediate important), data are citable via persistent identifiers (73% and 25% resp.), links to publications are provided (63% and 35% resp.) and restrictions, conditions and/or licensing information is communicated (61% and 36% resp.). In addition, a wide range of good practice examples was pointed out by respondents,

which demonstrates a substantial uptake of data sharing and reuse through data e-infrastructures in the global environmental change community.

Lead examples of repositories that enabled the sharing of data included Dryad and Pangaea, Figshare, a range of repositories managed by NASA and the National Oceanic and Atmospheric Administration (NOAA), the Global Biodiversity Information Facility (GBIF, a network of repositories), Genbank, government data, and the Coupled Model Intercomparison Project (CMIP, a multi-phase project that provides infrastructure in support of climate modeling, documentation and data access) (see fig. 3) [4]. Comments highlighted the importance of the reputation of the managing institution, the repository's well-defined quality information, metadata and documentation as well as portals' ability to support the discovery and visualization of data.

Figure 3. Repositories to find and use data.

A need for further enhancement and consolidation can be derived from respondents' expectations about functionalities of infrastructures and desires expressed about access to specific types of data, e.g. environmental data from several countries which currently do not share these data, private-sector data and high-resolution satellite data. General comments highlighted the importance of long-tail research datasets, the need to increase access to heritage data (if still in analogue form only), the lack of community standards and the importance to openly publish source code, scripts of simulation programs and analytical solutions.

3. Discussion and Conclusions

Over the last decade several authors have studied researchers' attitudes towards data sharing, barriers and motivators as well as citation impact (e.g. [5] and [6], and the literature review [8]), some with a disciplinary focus (e.g. [7]).

Van der Eyden and Bishop's recent study elaborates on researchers' motivations for sharing data which was based on selected disciplinary case studies. They found that in certain fields sharing data is an essential part of the research process or the research depends entirely of data sharing. Other motivations they identified were to enhance the researcher's career, e.g. visibility via publications, and extrinsic incentives such as funder or publisher requirements ([9], p. 27f).

Our findings confirm that accelerating research and applications and scientific merits are the main motivators for publishing data as open data. The high number of examples provided and the wishes expressed by respondents also seem to indicate that global environmental research highly depends on data sharing.

Among all policies, funder policies were ranked first as a motivator, and 88% of all respondents acknowledged their importance. Therefore it seems that acceptance of open data could be further enhanced by making open data archiving mandatory which is currently not the case for funders in several countries. However, this will be one recommendation of the Belmont Forum E-Infrastructures and Data Management Collaborative Research group to the Belmont Forum to support global environmental change research. Although ranked lowest among policies as a motivator, references in journals were the top route for the discovery of data (followed by search engines and data repositories and other discovery routes), and were also mentioned several times in the free text comments. This however should not lead to an encouragement of publishers to establish commercial databases for data storage, as paying for data access was not well perceived by the respondents.

Based on the findings of the survey, we have made the following recommendations to the Belmont Forum:

- that funders should make open data archiving mandatory, to take into account the main motivators revealed by the survey,
- to strengthen support and training activities,
- to further facilitate interoperability between data infrastructures, and
- to support the long-term sustainability of archives and data infrastructures.

Acknowledgements. The authors would like to thank Jonathan Hodge and André Santachè for their contribution to the design of the survey, Kim Oakley for collecting the data, and all survey respondents for their input.

References

[1] Belmont Forum, available at: https://igfagcr.org/ [Accessed 14 May 2015].
[2] Belmont Forum E-Infrastructures and Data Management Collaborative Research Action, available at: http://www.bfe-inf.org/ [Accessed 14 May 2015].
[3] B. Schmidt, B. Gemeinholzer, A. Treloar, J Hodge, A. Santanchè, K. Oakley (2015). Belmont Forum Open Data Survey 2014. Zenodo. doi:10.5281/zenodo.16384.
[4] S. Sinclair, G. Rockwell, Cirrus, *Voyant*, 2015. Available at: http://voyant-tools.org/tool/Cirrus/ [Accessed 14 May 2015].
[5] H.A. Piwowar, R.S. Day, D.B. Fridsma. Sharing detailed research data is associated with increased citation rate. *PLoS ONE* **2**(3) (2007), e308 (2007). doi:10.1371/journal.pone.0000308
[6] C. Tenopir, S. Allard, K. Douglass, A.U. Aydinoglu, L. Wu et al., Data sharing by scientists: practices and perceptions, *PLoS ONE* **6**(6) (2011), e21101. doi:10.1371/journal.pone.0021101.
[7] N. Enke, A. Thessen, K. Bach, J. Bendix, B. Seeger, B. Gemeinholzer (2012), The user's view on biodiversity data sharing – Investigating facts of acceptance and requirements to realize a sustainable use of research data, *Ecological Informatics* **11** (2012), 25–33. doi:10.1016/j.ecoinf.2012.03.004
[8] B. Fechner, S. Friesike, M. Hebing. What drives academic data sharing? *PLoS ONE* **10**(2) (2015), e0118053. doi:10.1371/journal.pone.0118053.
[9] V. Van den Eynden, V., L. Bishop, Incentives and motivations for sharing research data, a researcher's perspective, 2014. A Knowledge Exchange Report, available at: http://knowledge-exchange.info/Default.aspx?ID=733 [Accessed 14 May 2015].

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-145

Data Policies and Data Archives: A New Paradigm for Academic Publishing in Economic Sciences?

Sven VLAEMINCK[a,1], Lisa-Kristin HERRMANN[a]
[a] *ZBW – Leibniz Information Centre for Economics, Hamburg*

Abstract. In our paper we summarise the findings of an empirical study in which a sample of 346 journals in economics and business studies were examined. We regard both the extent and the quality of journals' data policies, which should facilitate replications of published empirical research. The paper presents some characteristics of journals equipped with data policies and gives some recommendations for suitable data policies in economics and business sciences journals. In addition, we also evaluate the journals' data archives to roughly estimate whether these journals really enforce data availability. Our key finding is that we are currently not able to determine a new publishing paradigm for journals in economic sciences.

Keywords: Reproducibility, replication, economics, business studies, social sciences, academic publishing, data policies, data archives.

1. Introduction

In economic sciences, empirically based studies have become increasingly important: The number of contributions to journals in which authors utilised self-collected or externally produced datasets for statistical analyses have massively increased [1].

With the growing relevance of publications based on empirical research, new questions and challenges for academic publishing emerge. Issues like integrating research data and scripts to run a data model in the broader context of a published article to foster replicable research and validation of scientific results are becoming increasingly important for both researchers and editors of scholarly journals.

This growing importance of research data and its integration in the academic publishing process is also reflected in numerous statements and partially also in requirements of funding agencies and scientific and political bodies in Europe and abroad. For instance, the European Commission (EC) recommends that EU member states should implement policies to ensure that "datasets are made easily identifiable and can be linked to other datasets and publications through appropriate mechanisms" [2]. This is also reflected in the goals of the 8th research framework programme of the EC, better known as Horizon 2020, inter alia the open research data pilot, which aims to improve and maximise access to and re-use of research data generated by funded projects [3].

[1] Corresponding Author, ZBW - Leibniz Information Centre for Economics, Neuer Jungfernstieg 21, D-20354 Hamburg, Germany; E-Mail: s.vlaeminck@zbw.eu.

To date, there exist only few means to replicate the results of economic research within the framework of published journal articles and to verify the results claimed in such a paper. This is unsatisfactory from a scientific point of view, because replicability is a cornerstone of the scientific method. The US economist B.D. McCullough outlined the importance of replicable research: "*[…] replication ensures that the method used to produce the results is known. Whether the results are correct or not is another matter, but unless everyone knows how the results were produced, their correctness cannot be assessed. Replicable research is subject to the scientific principle of verification; non-replicable research cannot be verified. Second, and more importantly, replicable research speeds scientific progress. […] Third, researchers will have an incentive to avoid sloppiness. […] Fourth, the incidence of fraud will decrease"*[4].

Especially for a scientific discipline like economic sciences, the effects of flawed research might have a huge impact on society, as the prominent example of the US economists Reinhart and Rogoff illustrated: The two top economists had published a paper [5] on the interrelation of economic growth and public debt in 2010 that attracted much attention: US vice presidential candidate Paul Ryan and also EU monetary affairs commissioner Olli Rehn used the findings claimed in the paper to justify austerity policy [6].

In 2013, the two authors provided the Excel sheet of their calculations to a student for teaching purposes. This student discovered that the Excel sheet contained faulty calculations and selectively omitted data [7], which casted massive doubts on Reinhart's and Rogoff's findings. This example clearly illustrates the necessity for economic research to be replicable.

One possible way to facilitate replications of published research is to implement strict research data policies for journals and also to implement data archives for code and data associated with articles published in scientific journals. But currently, especially professionally maintained data archives for publication-related research data in the social and economic sciences are not widespread. Often such data centres exist only for large surveys [23]. Therefore, a growing number of organisations and initiatives from all over the world have started to offer suitable services. Examples include the Interuniversity Consortium for Political and Social Research (ICPSR) (US)[2], GESIS-datorium (DE)[3], DANS-EASY (NL)[4] and the UK Data Service ReShare (UK)[5].

For our project[6] the question arose, how many journals in economic sciences currently are equipped with policies which facilitate access to underlying research data and the code of computation, which supports replications of applied economic research. To evaluate the current status quo in economic sciences and to clarify whether the implementation of data policies and related data archives tend to be a new paradigm for economic research, our project conducted a broad evaluation of 346 journals in economics and business studies. In addition, we also examined the way in which journals provide researchers and interested readers with research data and other materials.

[2] Cf. http://www.icpsr.umich.edu/icpsrweb/deposit/

[3] Cf. https://datorium.gesis.org/xmlui/?locale-attribute=en

[4] Cf. https://easy.dans.knaw.nl/ui/deposit

[5] Cf. http://reshare.ukdataservice.ac.uk/

[6] These project results have been developed in the EDaWaX project (European Data Watch Extended, http://www.edawax.de). EDaWaX is financed by the German Research Foundation (http://www.dfg.de).

2. Literature Review

To date, not many economists have dealt with the topic of data policies, despite the fact that discussion around replicable research in the discipline has been ongoing for several decades now. In 1986, a broadly noticed paper [8] reported the findings of a two-year study that collected programs and data from authors and attempted to replicate their published results. Ultimately, the authors were able to replicate only 2 of 54 papers – 3.7%.

Data policies of journals, especially the data policy of the Journal of Money, Credit and Banking (JMCB), which in 1982 was one of the first journals to introduce a data policy, were already discussed in the paper.

Almost 20 years later, the US economist B.D. McCullough published remarkable articles on data policies and data availability in economic journals. McCullough analysed the data policies of selected journals [9] and their data archives [10]. In 2008, he broadened his analyses and checked the data policies and data archives of journals in regard to their functionality for replication purposes [11]. One year later, he recapped his findings and also discussed the open access question for economic research. In total, he was not able to find more than 11 journals equipped with a mandatory data and code archive within the top 50 economics journals [4].

For our project the question arose, how the "market for replicable economic research" has developed since 2009. A first attempt was published in 2013, using a sample of 141 economics journals [12]. We found a total of 40 journals equipped with a data policy. 29 journals had a data availability policy[7], another 11 held weak policies which ask authors to cooperate with researchers in case of future request for data. Therefore, we name these policies author responsibility policies (in the following abbreviated with "ARP"). In addition, we found that journals with a data availability policy (in the following abbreviated with "DAP") are much better rated than journals without such policy. These findings are in line with other disciplinary and interdisciplinary studies [13, 14].

In addition, our project published some insights on the current status quo in providing research data in economics journals and the extent to which journals enforce data availability and replicable research. Our findings suggest that the main way to provide research data and associated materials is via the journals' websites. But we also noticed that journals obviously do not really enforce data availability: Only eight out of 29 journals had more than 50% of all articles in two issues checked accompanied by research data; 10 out of 29 journals with a DAP did not even have a single article in their archive supplemented by research data [12].

3. Study Methodology and Characteristics of the Research Sample

To compile a sample for our analyses, we used several lists of academic journals assembled by German economic associations. For instance, we included the

[7] We distinguish two types of data policies: An "author responsibility policy" requires authors to provide data (and sometimes code and other materials, too) to would-be replicators. In contrast, a data availability policy asks or mandates authors to provide research data (and partially code and other associated materials) to the journal. The journal provides this information to would-be replicators by attaching the data and other materials to the article (often in the "supplementary information" section). Cf. McCullough, McGeary & Harrison (2008) [11].

JOURQUAL2.1 list [15], maintained by the German Academic Association for Business Research (VHB), for journals in business studies. In addition, we included a sample of journals used by Bräuninger, Haucap and Muck [16], which primarily focuses on journals in economics. Both lists of journals have been used to evaluate the quality and relevance of the included journals from the point of view of German economists.

Because the JOURQUAL list contains 838 journals, we had to select a subsample. Therefore, we chose all journals from the JOURQUAL list ranked A+, A or B. This selection criterion is based on the results of our analyses in project phase 1, during which we found that primarily high-ranked journals are equipped with DAPs [12].

Using this approach, 258 out of 838 journals remained in the sample. Additionally, we randomly selected 60 journals rated C, D or E. With the aid of this subsample, we again wanted to check whether our assumption regarding the interrelation of highly ranked journals and the existence of data policies is correct. The entire sample used by Bräuninger, Haucap and Muck was also added to our research sample. In the next step, we removed double entries (some journals in the Bräuninger, Haucap and Muck sample are also included in the JOURQUAL list) and carefully checked the "aims and scope" section of each journal to find out whether the particular journal generally publishes empirically based studies and research papers. Journals publishing only theoretical papers or papers based on policy debates were removed from our sample.

Due to the outcome of this examination, the sample's size slightly decreased: In total, our database contains 346 journals, which is still quite a big sample compared to similar analyses.

Subsequently, we determined the primary scope of all journals in our sample. With such a classification, we were able to differentiate the results of our study by the subdomains of economic research. The lists of journals provided by professional associations are not sufficient for this purpose, because they do not distinguish accurately among subject categories. Therefore, we employed the subject categories used by the Thomson Reuters Journal Citation Report (JCR). In the event of more than one subject category being listed in the JCR, we used all of those, as long as all subject categories are derived from the broader field of economic research (we named this category 'economics & business studies in equal parts'). In the event of only one of the categories being dedicated to the field of economic research, we only used this subject category (either 'primarily economics' or 'primarily business studies'). In the case of none of the subject categories being primarily dedicated to economic research, we assigned the journal to a group called 'other'. For journals not listed in the JCR, we employed the indexing guidelines of the ZBW (German National Library of Economics/Leibniz Information Centre for Economics) to determine the subject category.

Beyond this, we also collected further information on the journals in our sample. For instance, we collected the impact factor of these journals (if available) and the rating both in the JOURQUAL2.1 and in the Handelsblatt ranking [17] – the latter being an important ranking for German economists.

Subsequently, we checked the websites of the journals (in some cases there are two websites for a single journal – the publisher's website and a website maintained by the editors) for existing data policies. In cases where we found such a guideline, we carefully analysed the wording of each policy and checked whether the policy complies with the criteria listed below. These criteria have been derived from previous studies in the field of replicable economic research [4, 8, 9, 10, 18]:

- A data policy must be mandatory.
- A data policy should not only require authors to provide the datasets used, but also the code of computation (syntax), self-compiled software components (e.g. in Fortran) and detailed descriptions of the data (data dictionary or codebook). In addition, such a policy should mandate authors to submit the original data from which the final dataset is derived and all instructions/codes necessary to achieve the final results of computation. Also, a README file should list all submitted files with a description of each and indicate which programs correspond to which findings in the paper.
- The data policy should require authors of empirically based articles to provide data and other materials listed above to the editorial office prior to the publication of an article.
- All submitted data and files (apart from confidential or proprietary datasets) must be made publicly available by the journal to interested researchers.
- A data policy has to have a procedure in place which allows interested readers to replicate research based on proprietary or confidential datasets in principle, even if the raw dataset cannot be submitted to the journal due to juridical reasons.

In addition, a journal should have a replication section or publish positive and negative replications. Furthermore, journals should encourage their readers to use the replication section (if available) to conduct replications of previously published research. This will encourage authors to scrutinise their data; submission of poorly documented data or even junk will most likely be prevented.

Subsequently to the analyses of data policies, we also checked two other aspects: On the one hand, we analysed in which way journals provided research data and other materials to interested readers and possible replicators. For this purpose, we carefully examined both the websites and the data policies for hints on how these journals provide research data. On the other hand, we selected four issues of each journal equipped with a data availability policy and checked how many of the articles of each issue contain additional materials like datasets, code and descriptions of the data and the analyses.

3.1. Some Characteristics of the Sample

Based on this sketched approach, we were able to determine that 46.2% (160) of all journals in our sample primarily belong to the subject category of business studies and 38.2% (132) to economics. 9.8% (34) of all journals in our sample are open to submissions from both economics and business studies in equal parts. 5.8% (20) are primarily associated with other subject categories (for example psychology, mathematics or sociology).

When we had a look at the major publishers in our sample, we were able to determine the three biggest publishers in our sample: 19.7% (68) of all journals in our sample are published by Wiley-Blackwell, and the same percentage is published by Elsevier. In third place, Springer follows with 12.4% (43).

When we examined the statistical distribution of the journals in our sample, we noticed that the biggest group is rated with a 0.5 (mode) in the Handelsblatt ranking.

In total, more than 50% of all journals are among the three best-rated groups. Hence, better-rated journals are disproportionally represented. Nevertheless,

approximately 35% of the journals are among the three lower-rated groups. Moreover, 21 journals in our sample are not considered in the Handelsblatt ranking. The likely reason is that these journals do not appear important enough to be indexed. When we take these journals into account, the extent of lower-ranked journals in our sample is around 38%.

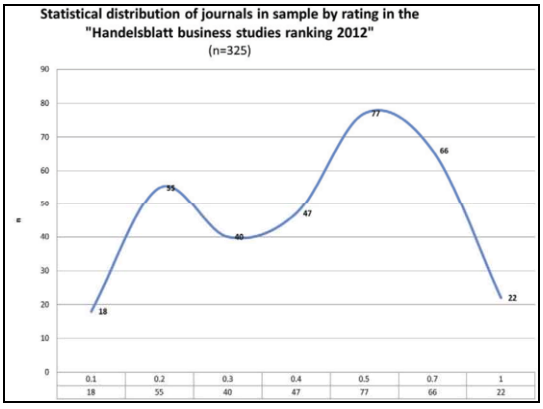


Figure 1. Statistical distribution of research sample in the Handelsblatt ranking 2012.

4. Findings of the Study

In the following paragraphs, we subsume the empirical findings of our study. First, we describe the outcome of the analyses on data policies, followed by results on current modes of journals in economic sciences to provide research data. To conclude, we appraise the degree in which journals enforce data availability by an investigation of the journals' data archives. We present the findings we obtained while checking four issues of each journal with regard to available research data.

4.1. Data Policies of Journals in Economic Sciences

Based on our approach described above, we were able to identify a total of 71 journals which have a data policy (20.5% of the total sample). 49 journals held a policy we classified as a data availability policy (14.2%).

Among the different subsamples (subject categories) of journals, we were able to determine important differences: 34 of the 49 journals belong to journals primarily publishing economics research (this equates to 25.8% of all economics journals in the sample), whereas only nine journals from the field of business studies have such policies (which equates to 5.6% of all business studies journals in the sample).

Another 22 journals (6.4% of the full sample) were equipped with a policy that relies on the author's willingness to provide research data (and sometimes code), even though some journals mandate their authors to do so.

Nevertheless, this latter type of policy does not work in practise: Feigenbaum and Levy [19] and Mirowski and Sklivas [20] have shown the disincentives for economists to participate in the replication of their work. Their theoretical work was underpinned

by McCullough's and Vinod's experiences when they tried to replicate all empirically based articles in a single issue of the American Economic Review (AER): "*Though the policy of the AER requires that "Details of computations sufficient to permit replication must be provided," we found that fully half of the authors would not honor the replication policy.*" [22].

Data sharing and helping to replicate one's own work does not comply with the common incentive schema. Therefore, such policies can be considered to be weak policies.

Among the remaining 49 journals equipped with a DAP, we found highs and lows. While some of the journals hold strong data availability policies, other policies merely appear to be window dressing: Only 61.2% of all data availability policies are mandatory. Against the background that data sharing is not widespread among economists – Andreoli-Versbach and Mueller-Langer found that roughly 2.5% of 488 applied economists regularly share their data [21] – and current practices on how to obtain credit in science do not incentivise documenting and sharing data, it is crucial to mandate the submission of underlying datasets and other materials.

77.6% (38) of all DAPs require the authors to provide the code of computation, 53% (26) also require researchers to submit self-compiled software components and another 71.4% (35) want their authors to provide descriptions of submitted datasets and other materials. While 69.4% (34) of all DAPs offer exemptions (e.g. for proprietary or confidential datasets) to the policy, another 24.5% (12) did not state whether such exemptions exist. Normally, such exemptions are granted by journals, so we conclude that 93.9% (46) of all journals with DAP seem to allow exemptions. On the other hand, only 52.2% (24) of these journals have a procedure, normally a requirement to post the code of computation in addition to other information like a contact address and the version and name of the dataset used, which would principally allow interested researchers to replicate even research based on proprietary or confidential datasets. This obviously is not a good result, because research based on such data is not replicable in almost every second case.

37 out of 49 journals (75.5%) with a data availability policy require their authors to provide the data and other materials with the initial submission or prior to publication – a good result. The editorial offices seem to have recognised the importance of the timely submission of data and associated materials.

Table 1. Requirements for data availability policies which facilitate replications (n=49)

criterion	yes	no	not stated
mandatory policies	30 (61.2%)	19 (38.8%)	-
code of computation	38 (77.6%)	-	11 (22.5%)
descriptions of data	35 (71.4%)	-	14 (28.6%)
self-compiled software	26 (53.1%)	-	23 (46.9%)
exemptions allowed	34 (69.4%)	3 (6.1%)*	12 (24.5%)
procedure for prop. data	24 (52.2%)**	-	22 (47.8%)**
public data disclosure	45 (91.8%)	1 (2%)	3 (6.1%)

* These journals "discouraged" the use of proprietary or confidential datasets.

** Due to three journals which discouraged the use of proprietary data, the sample size was reduced to 46.

Replications, in sharp contrast, are published by only five of the journals investigated, even though a few journals claimed to support the publication of replication studies (both positive and negative). Journals using Dataverse or similar software components offer readers the possibility to comment on data and code submitted and to give feedback on data quality and success or failure in replication attempts.

When examining some other characteristics of journals with data policies, we found that journals with DAPs are better rated than journals without such policies. 95.9% (47 out of 49) of journals with a DAP possess an Impact Factor, compared to only 57.2% (198 out of 346) for journals without any data policy. In the Handelsblatt ranking, journals with DAPs are rated 0.1 points better compared to journals with an ARP and to journals without any data policy (median). The same goes for the Impact Factor: Journals with DAPs are rated 0.63 points better than journals with an ARP and 0.69 points higher than journals without any data policy (median). Also, in the JOURQUAL ranking, journals with a DAP are rated 0.54 points better compared to journals without a data policy and 0.56 points better than those with ARPs (median). We also found more than three quarters of all journals with a DAP are among the three best-rated groups of journals in the Handelsblatt ranking.

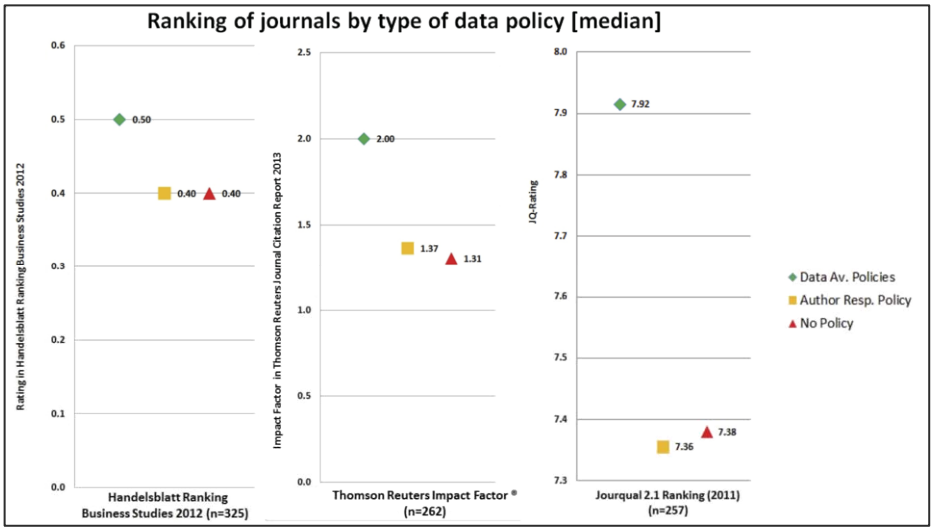


Figure 2. Ranking of journals in our sample by type of data policy (median).

4.2. Infrastructure Used for Journals' Data Archives

When we were able to find a DAP, we also checked in which mode research data and associated materials are made available for would-be replicators. On the one hand, we checked the content of the data policy, and on the other hand, we checked the journal websites to find out which infrastructure component is used to provide these additional materials.

Our findings suggest that most often research data is provided by the journals' websites: 83.7% of all journals with a DAP choose this mode to provide research data and other materials, 14.3% use special software for this purpose or suggest the use of external repositories to their authors.

Table 2. Provision mode for research data in journals equipped with a DAP (n=49; multiple modes possible)

Website	Author's Website	Repositories/ special software	No publication	Not stated
41 (83.7%)	2 (4.1%)	7 (14.3%)	3 (6.1%)	1 (2%)

The major problem in providing research data via websites from the viewpoint of scientific infrastructure providers is that there is no additional metadata for the supplements. Therefore, these datasets can neither be cited adequately, nor is it possible to reuse the datasets in any context other than the original article's – simply because these datasets are not findable. Though there are useful and easy to use solutions (e.g. Dataverse), only a small minority of journals in economic sciences apply these solutions: In total, only four journals with a DAP and a focus on economics research (i.e. 12.5% of all economics journals equipped with a DAP), and three with an 'other' classification (i.e. 60% of all 'other' journals with a DAP) used specialised software or employed external research data repositories. Not a single journal with a focus on business studies chose this way to provide readers with data and code of empirically based research.

4.3. Do Journals Enforce Data Availability?

In the course of our study, we also checked the data archives of all journals equipped with a data policy. We investigated four issues of each journal to determine how many articles are supplemented by research data and other materials.

The results we obtained suggest that data availability and replicable research are not among the top priorities of many of the journals surveyed. For instance, we found 10 journals (i.e. 20.4% of all journals with such policies) where not a single article was equipped with the underlying research data. But even beyond these journals, many editorial offices do not really enforce data availability: There was only a single journal (American Economic Journal: Applied Economics) which has data and code available for every article in the four issues.

5. Discussion

With the results we obtained, we are currently not able to determine a new publishing paradigm for journals in economic sciences. But there are differences among the subdomains of economic research: Especially economics journals with DAPs are slowly but steadily increasing: While McCullough [4] in 2009 was able to find only 10 journals equipped with such policies, Vlaeminck [12] was able to find a total of 29 journals with DAPs. Two years later, we identified 49 economics journals outfitted with such policies. These editorial offices seem to reflect the recommendations of scientific and political bodies to foster replicability of published research.

But we also found great discrepancies among the different subsamples in our sample: While journals focusing on economics research frequently have much more suitable data policies, DAPs are rare for journals in business studies. To explain these differences, we should keep in mind that research data in economics and business

studies is not identical. For instance, research data in business studies often consists of proprietary or even confidential data. Potentially the nature of this data leads the editorial offices of journals in business studies not to implement strong data policies, because they do not believe they would receive a noteworthy amount of data. Because developing and implementing data policies and related workflows is time and cost consuming, journals in business studies seem to be reluctant to enact such guidelines and processes.

Another assumption also provided true: In most cases, journals with strong DAPs are among the profession's top journals. Editors often mention that such journals can afford to implement such guidelines, because everyone would like to publish a paper in such a journal and is willing to submit datasets and other requested files, while a medium or low-ranked journal planning to implement a DAP could see a reduction in the amount of submissions it receives. However, we were able to identify a few lower-ranked journals which nevertheless are equipped with a strong data policy.

Relating to the requirements of the DAPs in our sample to foster replicable research, there is still room for improvements for many policies. The fact that a large portion of the guidelines are not mandatory is one such aspect. The failure to require self-compiled programs in many policies is another. Also, the absence of clear rules in cases in which proprietary data was used to obtain results in empirically based papers is an aspect that should be improved.

But even the best policy is meaningless if it is not enforced – and obviously many journals do not treat data availability as an important issue. With more than 20% of all journals in our sample clearly not putting their policies into action, there is a serious problem in terms of replicable research.

There are several aspects where research libraries and organisations dealing with research data might help publishers and editorial offices in lowering the burdens of implementing research data policies: One of these aspects is to advise editors how to develop suitable data policies. Another is to develop and to implement –powerful and lightweight software components which would reduce the cost and effort of managing data from empirically based articles. The fact that most journals still provide research data and other materials as a zip file on the publisher's or editor's website shows that there is an urgent need for such technical solutions.

References

[1] D.S. Hamermesh, Six Decades of Top Economics Publishing: Who and How?, *National Bureau of Economic Research*, Working Paper 18635 (2012).

[2] European Commission, COMMISSION RECOMMENDATION of 17.7.2012 on access to and preservation of scientific information (No. {SWD(2012) 221 final}{SWD(2012) 222 final}). Brussels (2012), available at: http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf [Accessed 5 June 2015].

[3] European Commission, HORIZON 2020 WORK PROGRAMME 2014 – 2015, (European Commission Decision C (2014)4995 of 22 July 2014), Brussels (2014), available at: http://ec.europa.eu/research/participants/portal/doc/call/h2020/common/1617601-part_1_introduction_v2.0_en.pdf [Accessed 5 June 2015].

[4] B.D. McCullough, Open Access Economics Journals and the Market for Reproducible Economic Research, *Economic Analysis and Policy* **39**(1) (2009), 117–126.

[5] C.M. Reinhart & K.S. Rogoff, Growth in a Time of Debt, *American Economic Review* **100** (2010), 573–578.

[6] P. Ryan, The Path to Prosperity: A Blueprint for American Renewal. Fiscal Year 2013 Budget Resolution, *House Budget Committee* (2013), available at http://budget.house.gov/uploadedfiles/pathtoprosperity2013.pdf [Accessed 5 June 2015].
[7] T. Herndon, M. Ash & R. Pollin, Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff, *Political Economy Research Institute*, (2013), available at http://www.peri.umass.edu/fileadmin/pdf/working_papers/working_papers_301-350/WP322.pdf [Accessed 5 June 2015].
[8] W.G. Dewald, J.G. Thursby & R.G. Anderson, Replication in Empirical Economics: The Journal of Money, Credit and Banking Project, *American Economic Review* **76**(4) (1986), 587–603.
[9] B.D. McCullough, K.A. McGeary & T.D. Harrison, Lessons from the JMCB archive, *Journal of Money, Credit, and Banking* **38** (2006), 1093–1107.
[10] B.D. McCullough, Got Replicability? The Journal of Money, Credit and Banking Archive. *Econ Journal Watch: Scholarly Comments on Academic Economics*, **4**(3) (2007), 326–337.
[11] B.D. McCullough, K.A. McGeary & T.D. Harrison, Do economics journal archives promote replicable research? *Canadian Journal of Economics* **41** (2008), 1406–1420.
[12] S. Vlaeminck, Data management in scholarly journals and possible roles for libraries – Some insights from EDaWaX, *LIBER Quarterly* **23**(1) (2013), URN:NBN:NL:UI:10-1-114595
[13] P. Sturges, M. Bamkin, J. Anders, & A. Hussain, Access to Research Data: Addressing the Problem through Journal Data Sharing Policies, *Proceedings of the IATUL Conferences*, Paper 3 (2014), available at http://docs.lib.purdue.edu/iatul/2014/openaccess/3/ [Accessed 5 June 2015].
[14] H.A. Piwowar & W.W. Chapman, A review of journal policies for sharing research data, 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* (2008), 1-14, available at http://elpub.scix.net/cgi-bin/works/Show?_id=001_elpub2008&sort=DEFAULT&search=%22ELPUB%3a2008%22&hits=52, http://www.bbc.com/news/magazine-22223190 [Accessed 5 June 2015].
[15] JOURQUAL 2.1 journal list (2011), available at http://vhbonline.org/service/jourqual/vhb-jourqual-21-2011/jq21/ [Accessed 5 June 2015].
[16] M. Bräuninger, J. Haucap & J. Muck, Was lesen und schätzen Ökonomen im Jahr 2011?, *DICE - Ordnungspolitische Perspektiven* **18** (2011), available at http://hdl.handle.net/10419/49023 [Accessed 5 June 2015].
[17] Handelsblatt Ranking BWL 2012 – Zeitschriftenliste – formatted (2012), Available at: https://docs.google.com/spreadsheet/pub?key=0AuEtgCUuVBDUdGVpTzE3TEp6QWNTaU43SjZWT2tDVFE&output=html [Accessed 5 June 2015]. For the methodology used please consult: J. Schläpfer & O. Storbeck, Methodik und Zeitschriftenliste für das Handelsblatt-BWL-Ranking 2012 (2012), available at http://htmldb-hosting.net/pls/htmldb/FMONITORING.download_my_file?p_file=801 [Accessed 5 June 2015].
[18] G. King, Replication, Replication, *PS: Political Science and Politics* **28** (1995), 443–499.
[19] S. Feigenbaum & D. M. Levy, The market for (ir)reproducible econometrics, *Social Epistemology* **7**(3) (1993), 215-232.
[20] P. Mirowski. & S. Sklivas, Why econometricians don't replicate (although they do reproduce), *Review of Political Economy* ***3***(2) (1991), 146–163.
[21] P. Andreoli-Versbach & F. Mueller-Langer, Open access to data: An ideal professed but not practised, *Research Policy* **43**(9) (2014), 1621–1633.
[22] B.D. McCullough & H.D. Vinod, Verifying the solution from a nonlinear solver: a case study, *American Economic Review* **93**(3) (2003), 873-892.
[23] S. Vlaeminck & G.G. Wagner, On the role of research data centres in the management of publication-related research data, *LIBER Quarterly,* ***23***(4) (2014), 336–357, available at http://liber.library.uu.nl/index.php/lq/article/view/9356 [Accessed 5 June 2015].

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-156

A New Platform for Editing Digital Multimedia: The eTalks

Claire CLIVAZ[a,1], Marion RIVOAL[a,b] and Martial SANKAR[b]
[a]*University of Lausanne (CH)*
[b]*Vital-IT SIB (CH)*

Abstract The eTalks are a new digital multimedia editing plaform developed at the University of Lausanne: their application is implemented via an easy-to-use editor interface, designed for the use of researchers themselves, to create and edit original eTalks. This permits the linking together of images, sounds and textual materials with hyperlinks, enriching it with relevant information. The final release of eTalks allows complete 'citability' of its contents: each and every portion of the researchers' talks can be precisely referred to and thus cited with a specific identifier, just like any traditional, paper-based scientific publication but with all the potential for plural literacies. It is openly accessible and the code is open source, including guidelines to install the eTalks. It contributes to the development of multiliteracies in the digital academic production of knowledge.

Keywords. Multiliteracies, digital edition, eTalks, electronic publishing, enhanced talks.

1. Introduction: Academic Communication and Digital Multiliteracies

Printed monographs, collected essays and articles have been cornerstones of modern academic communication for decades. If the use and teaching of rhetoric was expunged from German and French universities at the end of the 19th century [1], today, the strong presence of orality in digital publications invites one to reconsider the place of rhetoric in academic communication. As Kress argued in 1998, digital culture leads to the emergence of plural literacies, or multiliteracies [2]. The preoccupation with multiliteracies in academic publications and education began before the expression "digital humanities" came into being, and outside of the Humanities and Computing field. In 1996, the *Harvard Educational Review* (*HER*) published an article illustrating how literacy pedagogy in the digital age can reflect societal changes such as globalization, technology and increasing cultural and social diversity [3]. Before this junction within digital culture, the term 'literacy', born in the middle of the 19th century at the height of printed culture [4], was first changed to plural form by ethnologists and anthropologists [5], then by Ancient Classicists [6]. Naturally, this plural term has come into contact with present Western culture in the digital age.

In a 2012 article, Tanya Clement draws a picture of digital academic communication and education, putting forth the core term "multiliteracies" in her definition thereof. She discusses diverse DH pedagogies, such as new media studies and game studies, by looking at multiliteracies "that are engaged within undergraduate humanities curricula through general skills, principles and habits of mind that allow

[1] Corresponding author; email: claire.clivaz@unil.ch

students to progress within and engage society in the twenty-first century" [7]. The current students in classrooms have still been trained at school in a quasi-unique literacy mode: printed literacy. However, they are living in a cultural world that has already switched to multimodal literacies. These same students will be the future scholars who produce knowledge in multimodal digital ways. Convinced that textuality, images and sounds have to be used together in Humanist digital academic publications [8] and that digital multiliteracies have to meet high editorial requirements, we have built a new editorial form: the eTalk, based first and foremost on speeches of scholars.

2. The eTalks

Simple videos or MP3 recordings of lectures may prove insufficient to many researchers since they are not quotable in detail and they do not offer the possibility of being combined with textuality, images, hyperlinks, and references. Consequently, the eTalks application implements an easy-to-use editor interface, designed for the use by researchers themselves, to create and edit original enhanced talks. This permits the linking together of images, sounds and textual materials with hyperlinks, thereby enriching the content with relevant information. The result of the edition is displayed through a viewer interface, allowing one to experiment with the entire eTalk or to actively navigate, scroll and search inside its content. After having recorded the speech of the scholar, the software, Audacity, allows for the splitting of the speech in pieces of 2-3 sentences. Each piece of speech can be associated with its written version, a slide, images, or hyperlinks and so forth. Each piece is also quotable with a specific URL: a new kind of reference (see a presentation video[2]). Thus, the final release of eTalks allows for the complete 'citability' of its contents: each and every portion of the researchers' talks can be precisely referred to and therefore cited, just like any traditional, paper-based scientific publication but with all the potential for plural literacies.

The concept of the eTalk was developed by Claire Clivaz (UNIL) and Frédéric Kaplan (EPFL) in conjunction with an interdisciplinary team of colleagues. The core of the eTalk engine was developed in JavaScript by Frédéric Kaplan and Cyril Bornet (EPFL). The code is now available as open source on Github as a free application for further development[3]; we will soon provide the relevant guidelines to interested users[4]. The eTalks are currently being developed and disseminated further by an interdisciplinary team of researchers in Digital Humanities and bioinformatics, at the University of Lausanne (CH)[5]. As of now, three series of eTalks have been published as openly accessible: twelve on funerary rituals, two on the institutional biobank of Lausanne, and one on a DH2014 lecture.[6] A new series is being prepared regarding the topic of the enhanced human. The eTalks are now in development by institutional and research

[2] https://www.youtube.com/watch?v=NHqX-DVoBb8; all the hyperlinks have been last accessed on 05/15/2015.

[3] https://github.com/OZWE/eTalk

[4] Please contact claire.clivaz@unil.ch for the guidelines.

[5] Swiss Institute of Bioinformatics (SIB/VITAL-IT) and LADHUL (Laboratory of Digital Humanities and Cultures), at the University of Lausanne (CH): Claire Clivaz, Cécile Pache, Marion Rivoal and Martial Sankar.

[6] etalk.vital-it.ch/rites-funeraires; etalk.vital-it.ch/mooser; etalk.vital-it.ch/dh

collaborations, notably: colleagues from the Pedagogical High School of Lausanne (HEPVaud)[7] and the ERASMUS+ dariahTeach project[8], whose purpose is to offer a webportal in 2017 that will include digital teaching modules. The following pattern summarizes how an eTalk is built.

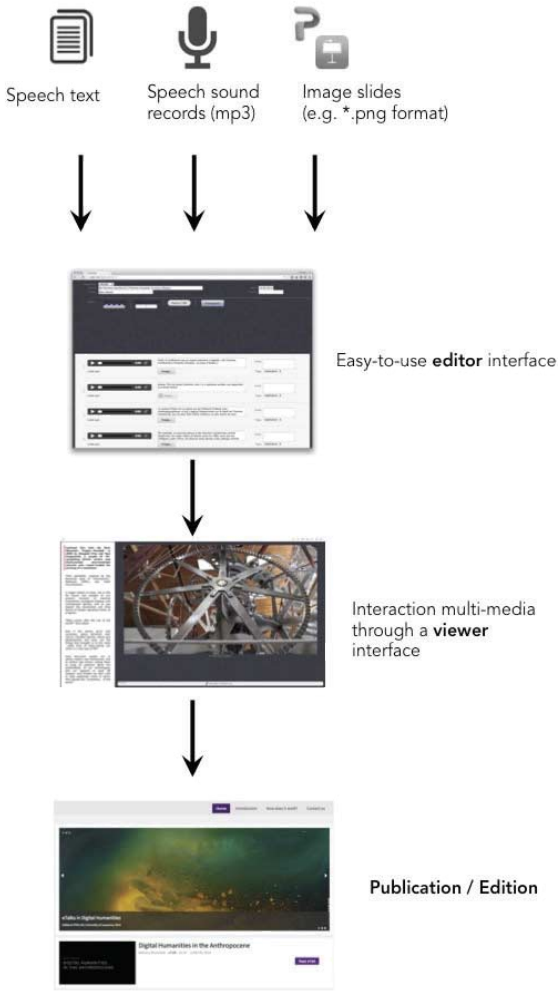


Figure 1. The eTalks editing process

[7] Prof. Nicole Durisch Gauthier and Prof. Christine Fawer Caputo.
[8] On twitter: @dariahTeach

3. eTalk Software Improvements and Future Directions

With respect to improvements and the future direction of eTalks, we will soon provide guidelines for future authors aiming to record an eTalk of their work. Those guidelines will permit a more efficient exchange between our team and the author during the creation process, such as providing image format and size requirements. With regards to the eTalk software specifically, several major improvements will be considered. Firstly, the eTalk application is currently only accessible through the Chrome, Safari and Internet Explorer browsers. Making allowances for multi-browser compatibility will be one of our first tasks, Mozilla/Firefox in particular. A user-rights management system also has to be established: it will allow for the coping with of a steady increase in eTalk projects and their authors' expectations in terms of privacy and diffusion.

We also aim to extend the range of media supported, such as enabling the citation of short films in eTalks, as well as the use of a stable URL system for quotation[9]. Moreover, sharing options have to be taken into consideration in order to take full advantage of social media sharing possibilities and the increase of authors' eTalk visibility. A top-level web application has to be designed, including the implementation of a search engine. This would permit users to search for specific eTalks, authors or topics inside the eTalk library. However, we will also continue to explore collaborations with the usual publishers and to examine the diffusion of eTalks with their help. Finally, in collaboration with the ERASMUS+ dariahTeach project, we will develop the most efficient way of recording and editing an eTalk by oneself with some editorial help from our team. Our purpose is to encourage academics to use this method of editing to quickly make their recent talks and slides available online.

References

[1] B. Belhoste, L'enseignement secondaire français et les sciences au début du XXe siècle. La réforme de 1902 des plans d'études et des programmes, *Revue d'histoire des sciences,* **43** (1990), 371– 400.

[2] G. Kress, Visual and Verbal Modes of Representation in Electronically Mediated Communication: the potentials of New Forms of Text, In *Page to Screen. Taking Literacy into the Electronic Era*, Ilana Snyder (ed.), Routledge, London/New York, 1998, 53– 79.

[3] New London Group, A Pedagogy of Multiliteracies: Designing Social Futures, *Harvard Educational Review* **66**(1) (1996), 60– 92.

[4] D. Barnton, *Literacy. An Introduction to the Ecology of Written Language*, Blackwell Publishing, Malden (MA), Oxford and Victoria, 2007.

[5] C. Clivaz, Common Era 2.0. Mapping the Digital Era from Antiquity and Modernity, In *Reading Tomorrow. From Ancient Manuscripts to the Digital Era / Lire Demain. Des manuscrits antiques à l'ère digitale*, C. Clivaz, J. Meizoz, F. Vallotton and J. Verheyden (eds.), with B. Bertho, PPUR, Lausanne, 2012, 23– 60.

[6] W. Johnson and H. Parker (eds.), *Ancient literacies: the culture of reading in Greece and Rome,* Oxford University Press, New York, 2009.

[7] T. Clement, Multiliteracies in the Undergraduate Digital Humanities Curriculum: Skills, Principles, and Habits of Mind, In *Digital Humanities Pedagogy: Practices, Principles and Politics*, B. Hirsch (ed.), Open Book Publishers, Cambridge, UK, 2012, 365– 388.

[8] C. Clivaz, D. Vinck, Introduction. Des humanités délivrées pour une littérature plurielle, *Les Cahiers du numérique* **10**(3) (2014), 9-16. DOI:10.3166.

[9] See for example www.purl.org.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-160

Towards Privacy Aware Social Semantic Digital Libraries

Owen SACCO [a] and John BRESLIN [b]
[a] *University of Malta - owensacco@gmail.com*
[b] *National University of Ireland, Galway - john.breslin@nuigalway.ie*

Keywords. privacy, social semantic web, semantic digital libraries, ppo, ppm

Digital Library systems provide an effective means to share digital information. These systems provide repositories of digital objects such as catalogues, books, articles, documents, images, video material and audio material [1]. Most digital library systems also provide users to create profiles that include personal information such as their topics of interest, publications, number of citations, subscribed material, history of viewed material, connections with co-authors, and other personal information. Therefore, digital libraries can be seen as implicit social networks since connections amongst users can implicitly be formed from the personal information of users created and gathered within these systems [3]. Moreover, several activities carried out in social networks are also carried out in digital libraries such as social bookmarking. Utilising personal information and the connections between users could enhance personalised services that digital libraries could offer, such as personalised search based on the user's interests and the interests of the user's connections.

Apart from utilising personal information created within digital library systems, these platforms could also be integrated with social network platforms in order to improve personalised services [3]. For instance, recommending or searching digital objects can be based on social network interactions between the user and his/her connections (i.e. "friends"). The digital library platform can recommend articles ranked according to the number of "likes" an article has and whether the topic of the article is similar to the user's interests or similar to the topic of the user's recent publications. Recommending articles can also take into consideration trust measures. For example, the profile similarity between profiles of the user and the user's friends could provide better ranking values especially when comparing the interests between the user's interests and the user's friends interests [2] – those friends that have the same interests as the user and have liked a particular article, then that article will have a higher ranking. Other measures could also be taken into consideration such as the number of interactions amongst users. These trust measures will eventually form a trusted social network which digital libraries could benefit from for providing personalised services. Moreover, digital libraries could be integrated with multiple social network platforms and also with other library platforms in order to take advantage of the personal information about users. This integration will not only enhance searching and recommending articles but also other services such as personal information management within digital libraries.

Digital libraries however pose several challenges: first, most digital libraries are data silos – most platforms structure their data using system specific schema which create a walled garden effect such that digital library data sources are not interoperable with each other and make it hard to link data elements amongst each system. Second, most digital libraries do not make use of current social data residing in social network platforms or in other digital library systems but provide the user to re-create his/her personal data. Digital libraries should utilise social data by aligning and aggregating user's personal information with the user's personal data residing in social network platforms and also residing in other digital library systems. Third, digital libraries lack privacy measures to provide users ways how they want to share their personal information.

The Semantic Web, which evolves from the conventional Web, provides techniques to markup data with meaning which can be processed by machines to offer enhanced services for data sharing and interoperability amongst different data sources. Such structured data is increasing as developers are becoming more aware of the advantages that Semantic Web technologies have to offer. However, the meta-formats which the Semantic Web provides are difficult for non-technical users to grasp in order to structure their data. Therefore, other formats emerged such as microformats[1] which are structured on current standards and provide easy to use formats to markup content with semantics (meaning) in Web documents. All of these formats have an underlying goal: to add structure to Web content in a *graph model*. The aim is to use identifiers, Uniform Resource Identifiers (URIs), to uniquely identify things (also known as resources) such as people, events, blog posts, reviews and tags published in Web documents. Therefore, each resource can link to other resources by referring to the URI of the specific resource to link to. Resources can be depicted as nodes in graphs and the edges between nodes illustrate the links between them. The advantage of linking resources is that different datasets can be linked, and hence create the *Web of Data*.

Content stored in digital libraries and social network platforms can be standardised and represented using various vocabularies such as Friend-of-a-Friend (FOAF)[2] for describing basic personal information, the Relationship Ontology[3] for describing relationship types with other users, the Description-of-a-Career (DOAC)[4] for describing career related information and Semantically Interlinked Online Communities (SIOC)[5] for describing activities. In order to disambiguate terms such as user's interests, DBPedia[6] concepts are used to describe such terms. Standardising how data is represented in this way would enable digital library systems to interoperate amongst each other and also with social networks that would create social semantic digital library platforms. However, this data is easily accessible and open since no access control mechanisms are in place for the Web of Data.

Our work, the *Privacy Preference Framework*, provides an attribute-based access control (ABAC) approach which allows expressing access control restrictions based on attributes which the requester and the restricted data must satisfy. Considering that no vocabulary provides fine-grained access control mechanisms for the *Web of Data*, our

[1] microformats – `http://microformats.org/`
[2] FOAF – `http://www.foaf-project.org`
[3] Relationship – `http://vocab.org/relationship/.html`
[4] DOAC – `http://ramonantonio.net/doac/0.1/`
[5] SIOC – `http://sioc-project.org/`
[6] DBPedia – `http://dbpedia.org/`

work provides a vocabulary for describing privacy settings and a manager to filter data based on these settings.

The *Privacy Preference Ontology (PPO)* [5] - `http://vocab.deri.ie/ppo#` – is a light-weight Attribute-based Access Control (ABAC) vocabulary that allows people to describe fine-grained privacy preferences for restricting or granting access to specific Linked Data. Among other use-cases, PPO can be used to restrict part of a user's digital library records only to users that have specific attributes. It provides a machine-readable way to define settings such as "Provide my list of publications only to those who have published articles of the same topic" or "Grant access to my personal contact details only to my co-authors".

As PPO deals with RDF(S)/OWL data, a privacy preference, defines: (1) the resource, statement, named graph, dataset or context it must restrict access to; (2) the conditions refining what to restrict; (3) the access control type; and (4) a SPARQL query, (`AccessSpace`) *i.e.* a graph pattern representing what must be satisfied by the user requesting information.

The *Privacy Preference Manager (PPM)* [4], is a privacy preference manager for the Web of Data. It allows users to manage their privacy preferences and also grants or denies access to user's information when requested by others. Using it, users can (1) authenticate to their instance and create privacy preferences for their digital library data and social data; and (2) authenticate to other user's instance and access the filtered digital data of these users. Moreover, the *Privacy Preference Manager* provides an API that can be used by Web systems to take advantage of incorporating privacy preferences enforcement within their system.

In this work, we describe models that provide a standard format for structuring digital library data and social Web data. Among other applications, these models could be used: (1) to define meta-structures for characterising and representing digital library data abstractly that could then be re-used on the Web; (2) to integrate Social Web data or other information from the Web within digital libraries. The latter could lead to a new kind of digital library experience which creates social networks for digital libraries. Furthermore, this work also provides an approach for defining fine-grained privacy preferences to social data and digital library data enabling users to control who can access their information.

References

[1] V. S. Chooralil. *Semantic Digital Library*. PhD thesis, Cochin University of Science and Technology, 2010.

[2] J. Golbeck. Trust and nuanced profile similarity in online social networks. *ACM Transactions on the Web*, Sept. 2009.

[3] S. Kruk, E. Kruk, and K. Stankiewicz. Evaluation of semantic and social technologies for digital libraries. In S. Kruk and B. McDaniel, editors, *Semantic Digital Libraries*, pages 203–214. Springer Berlin Heidelberg, 2009.

[4] O. Sacco and A. Passant. A Privacy Preference Manager for the Social Semantic Web. In *Proceedings of the 2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation, SPIM2011*, 2011.

[5] O. Sacco and A. Passant. A Privacy Preference Ontology (PPO) for Linked Data. In *Proceedings of the Linked Data on the Web Workshop, LDOW2011*, 2011.

Posters

This page intentionally left blank

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-165

Reaching Out to Global Interoperability through Aligning Repository Networks

Kathleen SHEARER[a,1], Katharina MUELLER[a,b], Maxie GOTTSCHLING[a,b]
[a]*Confederation of Open Access Repositories, Germany*
[b]*University of Göttingen, State and University Library*

Abstract. COAR is working towards greater interoperability of systems in a number of areas, with an emphasis on open access metadata elements and vocabularies. This paper presents the outcomes of the recently published roadmap on future directions for repository interoperability as well as an overview of the current status of the associations' work related to this topic, i.e. the initiative "Aligning Repository Networks".

Keywords. Repositories, interoperability, standards

1. Introduction

Scholarly communication is undergoing fundamental changes, in particular with new requirements for open access to research outputs, new forms of peer-review, and alternative methods for measuring impact. In parallel, technical developments, especially in communication and interface technologies facilitate bi-directional data exchange across related applications and repository and research information systems [1].

In the past few years, Open Access repositories and their associated services have become important components of the global e-research infrastructure [2]. Repositories can be connected into networks (e.g. at the national or regional level) to support unified access to an open, aggregated collection of scholarship and related materials that machines can mine enabling researchers to work with content in new ways and allowing funders and institutions to track research outputs [3]. In addition, repositories are increasingly being integrated with other systems, such as research administrative systems and with research data repositories, with the aim of providing a more integrated and seamless suite of services to various communities [4].

Interoperability between repositories, repository networks and across systems is absolutely critical to ensure the development of a global knowledge infrastructure. It will enable the exchange of data between repositories and support the creation of new services such as disciplinary portals or text mining. In addition it will enable networks to learn from each other allowing the global community to progress more quickly leading to cost synergies by preventing duplication of work across networks.

COAR, the Confederation of Open Access Repositories (an international membership-based organization based in Germany), acts as forum to learn about new

[1] Corresponding Author. COAR e. V. Office, c/o State and University Library Goettingen, Platz der Goettinger Sieben 1, 37073 Goettingen, Germany; E-Mail: kathleen.shearer@coar-repositories.org

trends; engage with international colleagues in the repository community; and ensure repositories adhere to global best practices and interoperability standards. COAR has been working to improve the interoperability across repository networks and between repository networks and other systems. In March 2014, COAR launched the Aligning Repository Networks initiative to work towards more harmonized standards and common vocabularies.[2]

2. Aligning across Repositories

As open access repositories become key instruments in research infrastructure, many regions around the world are investing in the development of repository networks. These networks have evolved in their specific local contexts and currently differ in a number of ways. However, the real value of repositories is when they are interconnected to provide unified access to research materials for researchers around the world [3].

To achieve some level of alignment across repository networks COAR initiated a meeting of representatives of globally important repository networks in 2014. At this meeting, delegates from Australia, Canada, China, Europe, Latin America and the United States identified several key elements for alignment [3] and made first steps to establish a mechanism for ongoing dialogue between repository networks. This initiative will give the repository community a stronger global voice and raise the visibility of the role of repositories as critical research infrastructure. At the practical level, this activity will allow repository networks to discuss and adopt best practices for metadata standards, vocabularies and services [3].

A number of follow-up activities were undertaken. At the strategic level, the group identified the fact that publishers are lengthening embargo periods, making it challenging for repositories to provide open access to articles. COAR, in conjunction with SPARC US[3] and other organizations crafted a statement against embargo periods which emphasizes that immediate access should be considered best practice and if embargos are to be imposed they should be a maximum of 6-12 months (depending on discipline). Numerous institutions and individuals have signed the statement.[4]

At the technical level, a working group was launched in April 2014 with representatives from major regional repository networks[5] seeking to harmonize open access elements and metadata schemas, and improved visibility of repository networks worldwide. The aim is to discuss how to express new elements needed in order to track open access outputs such as funder name, open access status, embargo periods, and licenses.[6] The regional Aligning committee met for a second time in April 2015 in

[2] See https://www.coar-repositories.org/activities/advocacy-leadership/aligning-repository-networks-across-regions/

[3] The Scholarly Publishing and Academic Resources Coalition: http://www.sparc.arl.org/

[4] https://www.coar-repositories.org/activities/advocacy-leadership/aligning-repository-networks-across-regions/statement-about-embargo-periods/

[5] Australia, LA Referencia (http://lareferencia.redclara.net/rfr/), OpenAIRE (http://www.openaire.eu) and SHARE (SHared Access Research Ecosystem: http://www.share-research.org/), CASRAI (Consortia Advancing Standards in Research Administration Information: http://casrai.org/) and EuroCRIS (Current Research Information Systems: http://www.eurocris.org/).

[6] Working Group – developing a blueprint for global interoperability of open access repository networks, https://www.coar-repositories.org/activities/advocacy-leadership/working-group-global-interoperability-of-oa-repository-networks/

order to review priorities and identify new areas in which they can further align their expanding repository networks. Amidst the intensifying global debate about the most sustainable ways to implement open access and research infrastructures, meeting participants reinforced their aim to foster solutions that reflect the diversity of approaches and capacities across different regions.[7] COAR will work with the community to accomplish these activities in the coming year.

3. Interoperability

The interoperability of repositories and between repositories and other systems has been an important topic for the COAR community since the association was established in 2009. With the papers "The case for interoperability for open access repositories" (2011) [6] and the "Current State of Repository Interoperability Report" (2012) [7] the Working Group Repository Interoperability built the foundation for the establishment of the interoperability roadmap. This roadmap presents the results of a community wide consultation to identify the important issues around repository interoperability. The issues and priorities outlined in the roadmap were derived based on input from an Expert Advisory Panel representing knowledgeable experts from around the world. Through this process, 35 interoperability issues were identified, 9 of which are considered "priority" issues based on their immediate relevance, with varying levels of complexity. These issues can be viewed as represent the most pressing priorities for efforts around interoperability [1]:

1. Exposing citation formats
2. Supporting data export functions
3. Supporting author identification systems
4. Exposing publication lists
5. Exposing bibliometric information
6. Exposing usage statistics
7. Supporting additional metadata format(s)
8. Integrating different persistent identifiers
9. Supporting search engine optimization (SEO)[8]

Many of these issues involve some level of standardization across vocabularies, metadata and indicators, both within the repository environment as well as with other systems. Interoperability in these areas, therefore, will require collaboration across countries and regions as well as with other systems developed by different communities. In addition, ensuring local implementation of guidelines and standards at the level of individual repositories is very difficult and often requires significant community outreach to raise awareness of the benefits of adopting standards. One strategy is to work with the repository platform developers to have the standards implemented into repository software systems. In parallel the available interfaces of repositories and the corresponding systems should be open to enable bi-directional communication and information channels in order to allow concrete system interoperability [1].

[7] https://www.coar-repositories.org/news-media/communique-international-repository-networks-reinforce-their-aim-to-develop-a-global-open-access-knowledge-commons/

[8] For the full table see https://www.coar-repositories.org/files/Roadmap_final_formatted_20150203.pdf

COAR is already working to advance interoperability in several of the priority areas including author identification systems, standard vocabularies, persistent identifiers, and usage statistics [1]. In particular, the interest group "Controlled vocabularies for repository assets"[9] is working on developing a controlled vocabulary for open access repositories that will be used globally and connecting different languages through linked-data techniques.[10] The COAR interest group "Usage Data and Beyond"[11] reviews existing and emerging article level metrics with the aim of identifying a common set of metrics that can be adopted across repositories, enabling the community to compare statistics across repositories.

4. Conclusions

Interoperability is a core issue for repositories and will remain so as the roles of repositories evolve and expand. The success of future repository services depends on the ability to offer seamless services across repository networks and in conjunction with other stakeholders at the local, national and international community level. In order to achieve this, the repository community must work with and engage in ongoing dialogue with these other communities. COAR, with its vision of a global network of open access repositories, will continue to work towards greater interoperability both within the repository community as well as with other players in the scholarly communication system. In addition, ensuring local implementation of guidelines and standards at the level of individual repositories is very difficult and often requires significant community outreach to raise awareness of the benefits of adopting standards.

References

[1] COAR Working Group 2: Repository Interoperability, COAR Roadmap Future Directions for Repository Interoperability, 2015, https://www.coarepositories.org/files/Roadmap_final_formatted_20150203.pdf.
[2] Directory of Open Access Repositories (OpenDOAR) http://www.opendoar.org/onechart.phpZ?cID=&ctID=&rtID=&clID=&lID=&potID=&rSoftWareName=&search=&groupby=c.cContinent&orderby=Tally%20DESC&charttype=pie&width=600&height=300&caption=Proportion%20of%20Repositories%20by%20Continent%20-%20Worldwide
[3] Kathleen Shearer, Towards a Seamless Global Research Infrastructure, Report of the Aligning Repository Networks Meeting, 2014, https://www.coar-repositories.org/files/Aligning-Repository-Networks-Meeting-Report.pdf
[4] COAR Repository Observatory, Third Edition, IR and CRIS Interoperability, https://www.coar-repositories.org/activities/repository-observatory/third-edition-ir-and-cris/
[5] COAR Strategy 2012 – 2015, https://www.coar-repositories.org/files/COAR-Strategy-2012-2015-Workplan-2012-2013.pdf
[6] COAR Working Group 2: Repository Interoperability, The case for interoperability for open access repositories, 2011, https://www.coar-repositories.org/files/A-Case-for-Interoperability-Final-Version.pdf
[7] COAR Working Group 2: Repository Interoperability, Current State of Repository Interoperability, October 2012, https://www.coar-repositories.org/files/COAR-Current-State-of-Open-Access-Repository-Interoperability-26-10-2012.pdf

[9] See https://www.coar-repositories.org/activities/repository-interoperability/ig-controlled-vocabularies-for-repository-assets/
[10] See http://wiki.surf.nl/display/standards/info-eu-repo
[11] See https://www.coar-repositories.org/activities/repository-interoperability/usage-data-and-beyond/

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-169

Social Reading and eBooks

Harri HEIKKILÄ[1]
Aalto University School of Art and Design

Abstract. The success and mainstreaming of e-books is transforming not only the traditional/Gutenbergian idea of the book but also the previous idea of an e-book as mainly an enriched print book. In the new e-book concept, the nature of a book as an artifact is diminishing and disposition as a networked interface to the knowledge is rising. One of the most important emerging concepts is the social reading, which means reading acts while connected to the other people. Social reading is a new and not very well defined area of reading practices. In addition to the traditional reading together and discussing books person to person, social reading includes a large number of networked functions like sharing and receiving shared information. Research of this new phenomena is almost non existent, yet it is expected to be the next big thing in reading and in e-books. This study provides an overview of the history of social reading of printed books and then defines parallel features in the new digital reading activities. Research material consists of popular e-book software and services. The proposed categorization of social reading is based on content analysis of properties that were found in those services. This report claims that social reading functionalities are manifestations of the social needs that have existed during and even before the paper book; digital time enables re-emerging of some of those features, but in a different manner.

Keywords. Social reading, electronic books, future of books

1. Introduction

What we see in people's practices and orientations towards e-books is a shift in emphasis from the book as an artifact to a set of activities associated with reading. We are facing a process of transformation, from the book as text container to a shared interface in a networked environment [1, 34]. One of the most salient and logical consequences of this is the rising of social reading – the act of reading while connected to others. This phenomenon has been described as the next big thing in reading, or even as the "future of books" [2]. This paper explores this emerging culture within e-books.

Sometimes the term social reading is used only to refer to the digital version of book clubs and social media, how ever social reading can be seen as a broader term, embracing a vast number of functions that follow the logic of networked media in general, like sharing, recommending and commenting.

This poster argues that social reading is noteworthy, because it is based on social needs that have always been present, but that have been channeled differently, following the technological prerequisites of distinct times.

Research questions are:

- What are the social dimensions of e-reading?
- How to categorize social reading functions in eBooks?

[1] Corresponding Author. E-Mail: harri.heikkila@aalto.fi

In order to answer these questions, we attempt to unravel social relations and needs that existed before and during the Gutenbergian era of reading, and then reflect what could be their digital counterparts in the future. Finally, we move on to conceptualize social reading according to practices discovered in present-day services.

2. A Tiny History of Social Reading

Social reading, meaning the act of reading while connected with each other, has a longer history than solitary reading. Book historians agree that reading was originally done in groups and by reading aloud [3, 7]. The Gutenbergian time changed reading by making books widely available, but it is often forgotten that it also solidified the practice of solitary and quiet reading, reading with oneself. Similarly, it is rarely noted that also many other practices of social reading existed, for example the culture of annotating to margins of shared volumes, which also faded away with the Gutenberg era. The scribes who copied manuscripts often copied annotations to new versions, and thus knowledge was accumulated socially. The printing press and movable type changed the role of reader as co-author and member of a community engaged in a collaborative search for meaning to a largely private activity [4].

Another example of social reading, which has been faded in history, is the culture of "Commonplace books", which were kind of semi-social clipart-books, personalized encyclopaedias where authors re-organized texts, like quotes and passages from different sources and annotated them. Liz Danzico [5] has described this everyday marginalia as a 300 years old "slow-motion Twitter or Face-book".

The famous main point of McLuhan is a valid foundation for e-reading research: when media changes we change. Media changes our habits and extends different kind of elements to our senses, which in turn affects our choices within the media – and gives birth to new paradigms.

McLuhan claimed that the era of "hot media" (media that favours single sense and low-participation) like print and books, will be replaced by multi-sensory "cool media" with high participation. In a way – McLuhan argues – this is returning to the time before Gutenberg, to time of discussion, non-linearity and non-fixed "cool media" [6].

Bob Stein, a pioneer developer of social reading software, argues that reading and writing have always been social; the paper-medium has just covered that. Stein sees an inevitable development, where we will confront "many levels of reader engagement from the simple acknowledgement of the presence of others to a very active engagement with authors and fellow readers" – because of the Internet [7].

3. Methods

In our research, the original sample of e-reading apps was constructed by finding the 100 most popular e-reading-category applications from the Apple app store (US) and then discarding those that consisted only of a single book (we were interested apps that could host several books, because there can not exist, for example, archiving or rating of other books etc. in a singe book-app) and those that did not offer access to functions without membership. This left us with 22 iOS-programs. In addition, we scanned available English-language social reading web services. The base list of services was constructed with the help of Huffington Post article "Best social reading sites," with

nine services. Additionally, three that had been emerged in the preliminary research. Thus, the whole sample consisted of 33 services and programs.

Using the standard methods of content analysis, the available functions were first listed, the meaningful ones from the point of view of social reading chosen (reduction) from the similar functions combined to classes (clustering) to construct categories (abstraction).

4. Results

We can conceptualize reading actions according to functions that are taking place within the software: First, there is an act of reading itself (reading-category); then, missions that deal with organizing and archiving my readings (bookshelf-category); then, a category of annotations that combines all marginalia-functions, like highlighting, notes and comments; Then, there is the obvious category of ratings, where the reader assesses the book, usually by giving points on some scale, like three stars out of five; And finally, a review-category which is about expressing one's opinion of the book in words. Reviews can be short or literature critique-like lengthy writings.

Furthermore, functions can be categorized according to sharing and sharing direction. The first group consists of actions primarily for myself, usually digital versions of something people do traditionally with paper books, we call these "Book 1.0 acts". The second and third categories represent sharing these same functions or receiving them – note that the sharing function can have two directions. The fourth group is discussion together. We call these social reading categories "Book 2.0 acts".

After cross tabulating these categories and actions, we get the following detailed table of available functions as measurable categories. For example, the category on Reading to oneself, breaks down into "Share what I read now" and "Follow what others read now" in Book 2.0 -category. Similarly, the Bookshelf-class is divided into "Share history of readings or intentions to read" and "See others' history and intentions to read" in the Book 2.0 -category.

Table 1. Social reading operationalized: classified functions and action in e-reading programs and services

	Book 1.0 Actions	**Book 2.0 Actions "From me"**	**Book 2.0 Actions "To me"**	**Book 2.0 Actions "US"**
READING	Reading to my self	Share what I read now	Follow what others read now	Reading together
BOOKSHELF	Archiving my books	Share history of readings, intentions to read	See others' history and intentions	Discussing
ANNOTATING	My annotations to myself	Share annotations (Highlights, notes, quotations, pictures)	See annotations of others	Discussing
RATING	My ratings for myself	Publish a rating	Review ratings	Discussing
REVIEWING	Review for me	Publish a review	Read reviews	Discussing

5. Discussion

New social actions in e-reading can be conceptualized into different categories according to their level and direction of sociality as well as what an action is intended to do.

Since most of the described "Book 2.0." -acts are basically shared versions of existing "Book 1.0" -functions (highlighting, note making, bookmarking, rating, archiving), one could contemplate that these functions should be first broadly and easily available in e-books, before social versions of them can become mainstream. When they become mainstream, one could expect that popularity of social functions will follow the same pattern as found on the net: most of the readers remain passive, only small percentage is willing to produce content him or herself (like reviews) but a vastly larger amount of readers are interested in following others activity by reading given reviews, receiving recommendations and ratings, and following discussions.

Social reading is likely to become more common as the evolution of the network-culture progresses and e-reading and e-books become more mainstream outside US. Books are going to be more retrievable and their content more connected.

Since social reading is more interesting the more participants are involved, this poses a challenge for the smaller systems. This is especially true in the small language regions: it is difficult to achieve the required user base in one service. Public libraries will have an interesting possibility here.

The full version of this paper includes research on reader's preferences in social reading categories. The Finnish project Textbook2020 will continue this research in the field of textbooks.

References

[1] J.-A. Cordón-Garcia, J. Alonso-Arévalo, R. Gómez-Díaz & D. Linder, *Social reading – platforms, applications, clouds and tags,* Chandos Publishing, Oxford, 2011.
[2] S. Prpick, *'Social reading' the next phase of e-book revolution* (2013), CBC http://www.cbc.ca/news/canada/social-reading-the-next-phase-of-e-book-revolution-1.1339149
[3] F.G. Kilgour, *The evolution of the book,* Oxford University Press, New York, 1998.
[4] D. Lebow, D. Lick & H. Hartman, *New Technology for Empowering Virtual Communities.* In: M. Pagani (ed) *Encyclopedia of Multimedia Technology and Networking*, Second Edition. London, IGI Blobal, 2008.
[5] L. Danzico, The Social Life of Marginalia, *Interactions* **18** (2011), 12–13.
[6] M. McLuhan, *Understanding media: the extensions of man* (1994). MIT Press, Cambridge, Mass.
[7] B. Stein, *A unified field theory of publishing in the networked era* (2008). If:Book. http://futureofthebook.org/blog/2008/09/04/a_unified_field_theory_of_publ_1/

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-173

Researchers and Open Data – Attitudes and Culture at Blekinge Institute of Technology

Peter LINDE[1], Eva NORLING, Anette PETTERSSON, Lena PETERSSON, Kent PETTERSSON, Anna STOCKMANN, Sofia SWARTZ
Blekinge Institute of Technology, the Library, Sweden

Abstract. During March 2015, the Blekinge Institute of Technology library carried out an interview survey comprising around 36 senior researchers and postdocs mainly in engineering sciences, with the objective to get a picture of how research data is managed at BTH and to find out what the researcher attitudes are to sharing data. The survey showed that most researchers in the study were positive to sharing research data but lacked any experience of making data management plans and had little or no knowledge of data preservation or of sharing open data. Uncertainties about data ownership are also an issue.

Keywords. Open Research Data, Open Access, Data Management Plan, DMP, Research Data Management, RDM, survey, attitude

1. Introduction

There now seems to be consensus among research funders and policy makers that open research data increases economic growth; the quality and transparency of research; the growth rate of innovation, and that it enriches the civil society. It is not unusual anymore that research funders mandate so called data management plans (DMPs) where researchers need to document how their research data are to be managed, disseminated, shared and preserved. More and more universities are therefore trying to create good environments for handling data by developing plans for research data management (RDM). The Swedish Research Council recently submitted a government commissioned proposal for a national policy for open research data [1]. Behind this activity lies the EU commission recommendation from 2012 inviting member states to make all scientific articles and data produced by tax funds open access [2]. By 2016 it is anticipated that the Swedish government will approve new national guidelines for open access based on the Research Council's proposal. The original proposition is that all research data produced either wholly or partly with public funding is to be made openly accessible by 2025.

The Swedish Research Council recommends that higher education institutions now can be pro-active and prepare for research data management by:

- working actively on the issue of archiving and long-term preservation of research data;
- allocating funds for archiving and long-term preservation of research data;

[1] Corresponding Author. peter.linde@bth.se.

- collaborating on planning of technical solutions, processes and guidelines for researchers.

2. Method

During March 2015, the Blekinge Institute of Technology (BTH) library carried out a semi- structured qualitative interview survey, choosing 40 senior researchers and postdocs out of the ca. 200 employed researchers. The researchers were selected to reflect all BTH research areas, in order to get a fair picture of how research data is managed at BTH and what the researcher attitudes are to sharing data.

We used the BTH institutional repository to select 40 of the most productive researchers during the last three years. The researchers were then contacted by e-mail.

To our surprise a majority of the researchers (36), within a few days, replied positively to our request. During the interviews it was clear that many of them took this as an opportunity to get informed about the topic of research data management and discussions, questions and diversions were common. So instead of running for 10 minutes the interviews generally took around 20 minutes or longer.

We constructed a set of 9 questions specifically targeted to understand research data management and awareness.

The interviews were performed by subject librarians at the BTH library. The answers were entered into a web-based google form and finally the data was saved into an Excel spreadsheet and analyzed.

3. Results

Question 1: What are your main areas of research?

BTH is one of the smallest institutions of higher education in Sweden but with a relatively large share of research focused on computer and information science, electrical engineering, nursing and spatial planning. The research areas of the interviewed scientists reflect this structure and typical research areas mentioned in the survey are telecommunication, computer security, nursing, physical acoustics, software engineering, computer systems, climate aspects in planning, mathematic modelling, and mechanical engineering.

Question 2: What are the most common type data sets that you use or produce?

The answers here reflected a very fragmented picture of the use and produce of research data at the institute. Researchers in health science, sustainable development and planning mainly used qualitative observations of processes or surveys or interviews recorded or transcribed. The computer scientists notably used data collected from commercial companies. This data is then used for scanning financial transactions, for performance factors, measuring traffic or other processes. With some exceptions BTH researchers use or produce rather small or medium-sized data files in Word (doc), Excel (xls) or picture (jpg) formats. Files processed and created in software like SPSS, MATLAB, SPLASH, SPEC, PARSEC, STELLA etc. are also common.

Question 3: Who are your main research funders?

Without a doubt the Knowledge Foundation, a university research funder with the task of strengthening Swedish competitiveness, is the major funder of research at BTH.

Also very important is the Swedish innovation agency VINNOVA as are commercial companies, government agencies, municipalities and county councils. The Swedish Research Foundation, the major research funder in Sweden and at the same time the only funder that mandates Data Management Plans along with the European Commission, is only mentioned by 5 of the researchers as a source for funding. Seventeen of the researchers mentioned the Knowledge Foundation; 11 mentioned VINNOVA; 12 mentioned commercial companies and 16 mentioned either government agencies, municipalities or county councils as important sources for funding.

Question 4: Do you have any experience of using open research data or making your data openly available?

Twenty-two (61%) of the researchers did not have any experience of using open research data or making data openly available. Six (17%) said they used open data but never published open data themselves. Eight (22%) answered the question positively but the majority added that they may have used open data on several occasions but had shared only on single occasions.

Question 5: Do you have any experience of writing Data Management Plans?

Thirty-four (94%) answered no. Two (6%) answered yes. For several researchers Data Management Plans was a new concept. One researcher explained that he usually only saves the data and sometimes describes the data shortly and that usually is enough to satisfy funders and research managers.

Question 6: Who owns your data?

It is common that BTH researchers do projects with commercial companies. When this is the case an IPR (Intellectual Property Rights) agreement is created where it is stated who owns or can use the data. A large majority of the researchers answered that it is either the company or the researcher that owns the data. Only four of the researchers suggested it was the university or the state that owned the data. Six researcher answered they did not know.

Question 7: Are the data sets preserved in any way after the project is completed? If yes, how is the data preserved and who is responsible for it?

The majority of the researchers said they were responsible for the data and that they usually save their data between one and ten years. Most often the data is saved on computer hard drives. It seems that in many cases the data follows the researcher. Other means for safekeeping data was on servers, USB-sticks, CDs, tape, Cloud services and even in binders. In a few cases data sets were not saved at all.

Question 8: Would you consider sharing your data in an open archive? If not, why?

A majority of researchers were positive to sharing data if it is ethically and legally ok and if commercial companies and other partners agree to it. Four of the researchers were negative and meant that no one else could possibly have any use for this specific data or were reluctant since they did not have control over what others might do with the data, or said that sharing data would just create more work and problems.

Question 9: Do you use or know any specific archive for research data? If yes, which ones?

Twenty-seven (75%) of the researchers did not know of any archives dedicated for research data. The ones who named particular archives either only knew of them or had used them as data sources but had not deposited data there.

4. Discussion & Conclusion

Even though BTH has a focus on research in technical areas like computer science and electrical engineering, the data sets produced present a very varied and fragmented picture.

In order to support researchers writing DMPs, the library needs to build a competence not only around funder mandates for DMPs, but also has to understand the varied palette of file formats. Also important is to start cooperation with institutions that have curation skills. Smaller and more uncomplicated data sets could possibly be archived within the BTH institutional repository system DiVA (with a few systems upgrades) but the "big data" sets and the ones that need more sophisticated curating must go to adequate special subject repositories or data centers. The library could also be supportive with information on how to cite data sets.

So far in Sweden only the major research funder, the Swedish Research Council, has mandated DMPs in their applications. But since they are the leading research funder one can suspect that this practice will trickle down to other state- or semi-state – funded research agencies like the Knowledge Foundation and VINNOVA. Also the expected national policy guidelines for open access to scientific information will have a positive leverage on other research funder attitudes towards open data.

At BTH only a small minority of the researchers are funded by the Swedish Research Council. The majority of researchers receive funds from the Knowledge Foundation or VINNOVA. Commercial companies are also, to a great extent, involved in funding or as partners in research projects. This can partly explain why researchers at BTH do not have that much experience of DMPs or any experience of making their data openly available.

Ownership of research data is a tricky question. Many researchers in our study believe they own the data if it is not owned by a commercial partner.

So who owns the data really?

In Sweden it is actually the university, as a government agency, that owns the data. Official documents held by Swedish agencies are in principle public [1, 3, 4]. The idea that individual researchers own the data has no legal support in Sweden. Here is clearly a need for clarifications and information from both university and government level on what the case is and what the exceptions are regarding ownership. When is data state property and when is it not?

From believing that you own the data follows the notion that you can do with it as you please. And the survey reflects this attitude – the data follows the researcher and therefore there is seldom any preservation of the data. It is either disposed of or saved on hard drives, USB sticks or servers for as long as it pleases the researchers or project managers.

BTH has since 2006 an open access mandate for research publications and the awareness of this is quite high. The attitude towards open access publishing is also generally very positive. This attitude is reflected in the answers to question #9 where 32 out of 36 researchers said that they would share their data if they could legally and ethically do this. At BTH it is common practice to cooperate with commercial companies and this fact complicates the process of making researchers share data.

The survey has shown that even if there is a positive attitude towards sharing data the need for information about and support for RDM and DMPs among BTH researchers is widespread. Information and clarification about data ownership issues

are also acute. The library therefore will start a dialogue with the Vice-Chancellor and the deans about preparing and supplying part of this support.

References

[1] Vetenskapsrådet, *Förslag till Nationella riktlinjer för öppen tillgång till vetenskaplig information* [*Proposal for national guidelines for open access to scientific information*], Vetenskapsrådets rapporter, Stockholm, 2015. https://publikationer.vr.se/produkt/forslag-till-nationella-riktlinjer-for-oppen-tillgang-till-vetenskaplig-information/

[2] European Commission, *Commission recommendation of 17.7.2012 on access to and preservation of scientific information*, Brussels, 2012. http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

[3] Swedish government, *Freedom of the Press Act* (SFS 1949:105). http://riksdagen.se/en/documents-and-laws/Laws/The-constitution/

[4] Alf Bohlin, *Offentlighet och sekretess i myndighets forskningsverksamhet [Public disclosure and secrecy in government research]*, Rapport Riksarkivet 1997:2, Stockholm, 1997.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-178

Exploration of Professional Social Networks and Opinions about Scholarly Communication Tools among Italian Astrophysicists

Monica MARRA[a,1]
[a] *Italian National Institute for Astrophysics (INAF) - Bologna Astronomical Observatory*

Abstract. The poster conveys the first results of a survey conducted among astrophysicists working at INAF. Just under 120 respondents made it possible to investigate their behaviour and opinions with regard to use of some major professional social networks and preferences about some aspects of scholarly communication and evaluation.

Keywords. Survey, astrophysics, social networks, scholarly communication

1. Introduction

In recent years, studies have started examining the relationship between scholars and the social media also by means of dedicated surveys. Most of the authors have timely aimed at enlightening whether researchers' involvement in Web 2.0 practices could already be changing some modes of the scholarly communication pathway.

2. Main Results

Basically the same hypothesis led to the drawing up of a questionnaire which was submitted to the personnel working at INAF (Italian National Institute for Astrophysics) through one of the institute's mailing lists in October 2014. The questionnaire was made up of nine questions (plus a question for respondents' age tier and profile self-classification) concerning opinions about the peer-review system, some mainstream scholarly communication practices within the discipline and respondents' use of some major professional social networks for research purposes. The questionnaire was kindly filled in by 122 among researchers and technologists (in fact, almost entirely researchers, technologists being 3,84% of the total respondents), on a total of 577 open-ended contract researchers and technologists included in INAF's most recent available statistics [1]. After elaborating the age data available on INAF's website cited above for the institute's personnel on one side, and respondents' self-classification in the three age tiers provided (25-44; 45-55; >55) on the other side, the 117 researchers who answered the questionnaire proved to be a representative sample of INAF researchers' age tiers due to a sufficiently satisfying correspondence between

[1] Corresponding Author. E-Mail: monica.marra@oabo.inaf.it.

the three couples of values. In particular, the widest age group both according to INAF's data and among the present questionnaire's respondents is the one between 46 and 55 years (44,75% and 47,86% respectively). Following in quantitative order, we find the 25-45 years-old age tier and the >55 age group.

Such a sample (117 researchers) is precious in order to aim at an adequate view of the opinions and behaviour of the institution's personnel, and to check the correspondence of the present questionnaire's results to those of the previous surveys.

According to [2], researchers in physics are the third disciplinary group by intensity of use of the social media in general, with their 88,6% of researchers using these tools. The CIBER study identifies a trend with researchers in the natural sciences being more active than scholars in other disciplines with the social media. Although other studies have shown the low use of single social media tools such as Twitter among astrophysicists [3], the results of the present questionnaire, which was targeted at "proper" social networks' use only, show the inclination towards these specific tools by researchers in astrophysics. The present survey has considered the following social networks (hence also: SN): LinkedIn, ResearchGate, Academia. Now, our questionnaire's respondents who declare they do not have a profile at all on a social network for professional use are less than 35%: more than 65% result to be owners of at least one profile on the professional social networks mentioned above (both percentages are the average among the three different age tiers). Again, the average percentage of astrophysicists with a profile on one or more of these professional SN is much higher than it had been verified by some of the previous surveys addressed at a multi-disciplinary audience (e.g. [4] ; the detail would require a more extended exposition).

Predictably, and in line with the previous literature, respondents who say they don't have a professional SN profile are more numerous among researchers >55 years old; surprisingly, instead, among the remaining researchers it's the youngest age tier that scores second among non-users (32,43%), whereas astrophysicists aged between 45 and 55 appear to be more inclined to testing the three SN here considered (more than 70% have a profile).

Overall, the most popular SN among the ones taken into account is ResearchGate, alone or combined; this confirms the results of a recent study. Less than 10% lower than RG's percentage comes LinkedIn, alone or combined; Academia occupies a very small niche made up exclusively by researchers between 45 and 55 years of age.

When considering researchers with a single SN profile, the most popular, again, results to be RG, which has the highest number of single-profile subscribers among the youngest age group (32%). LinkedIn has the highest percentage of single-profile owners among 45-55 agers. Interestingly, researchers older than 55 years have by far the highest percentage of profiles both on LinkedIn and on ResearchGate: almost 65%.

The widely shared view about ArXiv being astrophysicists' mainstream channel for opinions exchange about (mostly preprint) publications, seems to need reconsideration to some extent. For almost 65% of respondents, and on average among age tiers, papers self-archived by researchers on the invaluable USA-based database receive comments in only 0-20% of cases (n.b.: the percentage varies considerably, and interestingly, according to age groups). And, for the majority of respondents, comments on single aspects of their papers have a noteworthy importance (on average: they are "very important" for 31,4%; "rather important" for ~ 50%. In fact, anyway, it's still complete peer-review that holds the first place in researchers' appraisal of how their production should ideally be assessed: it's "very important" for > 70% of the

astrophysicists in our sample (average of age tiers), although here again age tiers play a role. Rather consistently, "'likes', tweets or other expressions of interest on social networks" are "very" or "rather important" only for a very low percentage of the respondents and "not important at all" for the greatest majority.

When requested what they expect from their professional social network profile, the great majority of answers reports "better visibility of my research activity", followed by "better availability of my research papers" (37,09%; average among the different age tiers' results). The prevalent goal among respondents seems thus to consist in seizing the opportunity to be more easily reached and read throughout the global scholarly community, rather than being actively engaged in professional social networking. This is in line with the previous studies according to which the traditional scholarly communication model is not undergoing a final stage of its crisis (e.g.: [5, 2]).

At the same time, traditional peer-review with reviewer's identity unknown is perceived as "ideal" by a very restricted percentage of respondents, whereas different models as for reviewer and reviewed author's mutual awareness result to be considered desirable options by most researchers (which is in line with [6] and others).

The data seem to reveal that fissures as for satisfaction with the traditional evaluation and dissemination models might be emerging among astrophysicists, but at the moment it doesn't seem that the use of the social networks is the key for exploring new models. Astrophysicists seem anyway to be rather keen on experimenting new internet tools and the situation might evolve unpredictably in the future.

References

[1] http://www.ced.inaf.it/anagrafica/zz-eta.php, concerning open-ended contracts.
[2] CIBER & Emerald Group Publishing, *Social media and research workflow*, 2010.
[3] S. Haustein et al., Astrophysicists on Twitter: an in-depth analysis of tweeting and scientific publication behavior, *Aslib Journal of Information Management* **66** (2014), 279–296.
[4] D. Ponte, J. Simon, Scholarly Communication 2.0: Exploring Researchers' Opinions on Web 2.0 for Scientific Knowledge Creation, Evaluation and Dissemination, *Serials review* **37** (2011), 149–156.
[5] D. Harley et al., Astrophysics case study, in id., *Assessing the future landscape of scholarly communication. An exploration of faculty values and needs in seven disciplines,* Berkeley, Center for Studies in Higher Education, 2010.
[6] L. Calvi, M. Cassella, Scholarship 2.0: analyzing scholars' use of Web 2.0 tools in research and teaching activity, *Liber Quarterly* **23** (2013), 110–133.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-181

The Roadmap to Finnish Open Science and Research

Pekka OLSBO[a,1]
[a] *Jyväskylä University Library*

Abstract. Finland published its open science roadmap at the end of November 2014. This roadmap is based on the work of the Open Science and Research Initiative (ATT), a cross-administrative initiative established by the Ministry of Education and Culture. The goal of this initiative is to promote open science and availability of information. Exploration of recent developments of open access in the EU shows that Finland is not among the leading countries in the EU. This paper focuses on the practical action plan of this roadmap and describes how the weakest part of Finnish open science, green open access is to be lifted at international top level.

Keywords. Finland, funder policies, green open access, mandates, open science, university libraries

1. Introduction

Open science and research seek to promote science through openness and increase the societal impact of science by improving the management and utilization of information generated by research. Though openness has always been and will be a fundamental principle of science and research, new open operating models will make science more democratic than ever before [1].

For researchers and research groups, openness conserves resources, improves the quality of research, and potentially offers increased credits and opportunities for cooperation. The future economy of Finland will rely on research, innovation and expertise. Open science and research play a decisive role in all of these [2].

Many international organizations campaign for open science. National openness policies have been and are being made in many European countries. In Finland only few universities and research funders have created their own open access policies and the first steps are still under preparation [3]. More detailed description of open access mandates and situation in Finland can be found in Schmidt & Kuchma [4] and Olsbo et al [5] (in Finnish).

2. The Roadmap to Finnish Open Science and Research

In order to answer the growing demands of open science, Finland has established an Open Science and Research Initiative. The objectives of the Open Science and

[1] Corresponding Author. E-mail: pekka.olsbo@jyu.fi

Research Initiative (ATT) are to make Finland a leading country for openness in science and research by 2017, and for the opportunities afforded by open science and research to be extensively harnessed in society. Dialogue in science and research will be promoted on many levels, both nationally and internationally. This will be achieved through four sub-objectives: reinforcing the intrinsic nature of science and research, increasing openness-related expertise, ensuring a stable foundation for the research process, and increasing the societal impact of research [6].

In this paper, the focus is on the first sub-objective, *reinforcing the intrinsic nature of science and research*. Openness is both a prerequisite and a means of promoting science and research. However, the research system needs structures and working methods for openness to be extensively utilized. Open science and research are a long-term continuum that consists of openness in both the research process and working culture [7].

One of the key targets is that the availability of research results is self-evident, and no separate solutions for openness are required. This is especially true in the case of green open access. Centralized services have to be created so that a minimum effort by the researcher is needed.

Organizing centralized services for researchers and creating fluent and stable processes for self-archiving is only the first step. The goal of the developmental work has to be a model which connects both the reporting of research activities and the depositing of research results into an institutional repository to a single comprehensive cost-effective process. Cost-effectiveness is based on knowledge accumulation and centralization, the interoperability of information systems and thus minimizing the working time used in the process. The cost analysis of the UK research institutions made by the Research Consulting shows that even 50 % of green open access related costs can be cut down by making the deposit process as quick and easy for authors as possible and by working to achieve greater clarity in publisher policies [8]. All this can be done only if centralized services are created.

3. From Plans into Action

If we look at statistics and services that measure the openness of science in European countries[2], we can see that Finland is clearly behind the leading countries in Europe. This is especially true if we measure the activity of green open access. There are three universities in Finland that have organized self-archiving of articles in larger scale based on a mandate by the university. Only the University of Helsinki[3] and the University of Jyväskylä[4] require parallel depositing of research articles. The low level of green open access activity is mostly due to lack of standardized processes and centralized services regarding the green open access within the research institutions in Finland.

The University of Helsinki was the first to establish an institutional mandate on open access in 2010. Soon after Helsinki the University of Tampere[5] and the University of Jyväskylä followed and developed their own open access policies. The development

[2] E.g. Ranking web of Universities, Ranking web of Repositories, OpenAIRE, BASE search engine, ROARMAP and DOAJ.

[3] Open Access in the University of Helsinki: http://www.helsinki.fi/kirjasto/en/get-help/open-access/

[4] Open Access in the University of Jyväskylä: http://openaccess.jyu.fi/en

[5] Open Access in the University of Tampere: http://www.uta.fi/english/research/OA/index.html

of green open access in Helsinki and Tampere has stopped, because no other measures beside the mandate have been taken. This can be seen in table 1, where the development of green Open Access at these three universities is shown.

Table 1. The number of deposited articles in university institutional repositories 2010-2014.[6]

University	2010	2011	2012	2013	2014
Helsinki	309	240	265	302	286
Jyväskylä	77	141	147	223	447
Tampere	93	169	189	130	105

At the University of Jyväskylä the development of open access and especially the promotion of green open access have been on the agenda since 2009 and first open access policy was made in 2011. In April 2014, the reporting of research activities of the University was moved to the University Library. This was done in order to improve the quality of the metadata of reporting and to increase the activity of depositing the articles in university repository, JYX[7]. It was possible to perceive a clear rise in the number of the deposited articles by the end of the year 2014. Results can be seen in table 2 below. In December 2014, the Rector of the University refined the open access policy of the University. The parallel depositing of research articles is now required [9].

Table 2. The development of open access at the University of Jyväskylä.[8]

Year	Number of research publications[1]	Total open access %	Green open access %[2]	Share of Green open access out of total OA
2012	2822	32,90	16,90	59,31
2013	2968	33,30	20,80	73,57
2014	2887	42,30	27,40	74,02

[1] Including all research articles, dissertations and articles for larger public.
[2] Deposited articles at institutional repository JYX and other repositories such as arXiv.

The beginning of the year 2015 seems to show that the share of the green open access out of total open access is now over 85% and the total open access percentage will grow to over 50%. These figures show that the share of the green open access and the usage of the institutional repository as the location for depositing articles is much above the average compared with studies gathered in Björk et al. 2014 [10].

The development of green open access at the University of Jyväskylä has shown that by keeping the focus on centralized services and fluent processes, rapid changes in open access activity can be achieved.

The objective of the Finnish Open Science and Research Initiative is to make Finland a leading country for openness in science and research by 2017. This means that we need quick amendment in the open access atmosphere in Finland. In order to enable this "green turn" in Finland, the Ministry of Education and Culture now funds an 18-month joint project by the University of Jyväskylä and the University of Eastern

[6] Situation in 2nd of May 2015. Numbers downloaded from repositories HELDA (https://helda.helsinki.fi/), TamPub (http://tampub.uta.fi/) and JYX (https://jyx.jyu.fi).
[7] https://jyx.jyu.fi
[8] Numbers are collected from the University´s research database TUTKA (tutka.jyu.fi) and library catalogue JYKDOK (https://jyu.finna.fi/).

Finland. The goal of this project is to create a model for implementing the green open access activity as a day to day routine in all Finnish universities. The model will utilize the experiences and developmental work done at the University of Jyväskylä. It will be refined and piloted for the needs of the University of Eastern Finland and then duplicated to other universities in Finland. During the project the green open access activity in Finland is foreseen to be doubled.

References

[1] Ministry of Education and Culture, Open science and research leads to surprising discoveries and creative insights. Open science and research roadmap 2014-2017. Reports of the Ministry of Education and Culture, Finland 2014:21, 7.
[2] ibid, 10.
[3] OpenAIRE. OA in Finland. Available at: https://www.openaire.eu/finland/noads/oa-finland [cited 1.6.2015].
[4] B. Schmidt & I. Kuchma. *Implementing Open Access Mandates in Europe - OpenAIRE Study on the Development of Open Access*. Universitätsverlag Göttingen, Göttingen, 2012.
[5] P.Olsbo, J. Ilva and workgroup. Taustaselvitys EU:n, Pohjoismaiden ja Suomen avoimen julkaisemisen tilanteesta. [in Finnish] (Final report of the ATT Open Access publications working group). Ministry of Education and Culture, Open Science and Research Initiative, 2015. Available at: http://avointiede.fi/documents/10864/12232/Julkaisujen+avoimen+saatavuuden+edist%C3%A4minen+2015+loppuraportti/f606ec8d-716c-4768-a394-c9ff7b866a75
[6] Ministry of Education and Culture, Open science and research leads to surprising discoveries and creative insights. Open science and research roadmap 2014-2017. Reports of the Ministry of Education and Culture, Finland 2014:21, 14-18.
[7] ibid, 15.
[8] Research Consulting, Counting the Costs of Open Access, November 2014, 18. Available at: http://www.researchconsulting.co.uk/wp-content/uploads/2014/11/Research-Consulting-Counting-the-Costs-of-OA-Final.pdf
[9] M. Manninen, Rinnakkaistallentaminen ja tutkimusjulkaisujen avoin saatavuus Jyväskylän yliopistossa. [in Finnish] (Parallel publishing and the openness of research publications in the University of Jyväskylä). Decision by the Rector of the University, 12.12.2014. Available at: https://www.jyu.fi/hallinto/rehtori/intra/rinnakkaistallentaminen_ja_tutkimusjulkaisujen_avoin_saatavuus [cited 22.5.2015].
[10] B.-C. Björk, M. Laakso, M., P. Welling, and P. Paetau. Anatomy of green open access. *Journal of the Association for Information Science and Technology* **65** (2014), 237–250. doi: 10.1002/asi.22963.

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-185

Infrastructures for Policies: How OpenAIRE Supports the EC's Open Access Requirements

Najla RETTBERG[1], Birgit SCHMIDT and Anthony ROSS
University of Göttingen, State and University Library

Abstract. Recently launched, OpenAIRE2020 is an Open Access (OA) Infrastructure for Research which supports open scholarly communication and access to the research output of European funded projects. This brief paper outlines how such an infrastructure can support an OA policy, the efforts required to successfully implement the mandate and the overall benefit that an infrastructure can bring.

Keywords. Open Access, Open Science

1. Introduction

Guided by the principles that the outputs of publicly-funded research should most fully benefit the public and that science works best when it is most efficient, transparent and accessible, the movement towards open access and open science gathers ever more momentum. The European Commission (EC) defines open access (OA) as "the practice of providing online access to reusable scientific information that is free of charge to the end user" [1]. The EC, a major research funder, has adopted OA as a core strategy in its efforts to improve the circulation of knowledge and hence innovation. The EC is clear: OA ensures better, more efficient science, and greater innovation in the public and private sectors. In 2008, the EC launched the Open Access Pilot [1], requiring selected beneficiaries of its then current funding programme, the Seventh Framework Programme (FP7), to strive to ensure OA to peer-reviewed articles resulting from their FP7-funded research. The EC's new funding programme, Horizon 2020, goes much further. From 2014 to 2020, Horizon 2020 will invest nearly 80 billion Euros in competitive research [2] and all research publications stemming from this investment must be made openly accessible [3]. Moreover, Horizon 2020 also includes an Open Research Data Pilot, which aims to improve and maximise access to the raw research data generated by EU-funded projects in order to increase efficiency, quality and transparency, and an FP7 post-grant Open Access Pilot to cover the open access publishing fees for publications arising from finished FP7 projects.

[1] Corresponding Author. E-mail: najla.rettberg@sub.uni-goettingen.de.

2. OpenAIRE as a Technological Infrastructure

Open Access Infrastructure for Research in Europe (OpenAIRE) exists to support and foster the EC's OA policies and their take-up and implementation [4]. It does so in various ways. Firstly, OpenAIRE tracks publications and associated outputs resulting from EC-funded research by harvesting metadata information from a network of Open Access repositories, aggregators and OA journals. It connects this data with data regarding projects and research data to give a rounded picture of the impact of the research outcomes of EC-funded projects. By harnessing the contents of "compatible" publication and data repositories (both institutional and disciplinary), it encourages the exposure of metadata, including funding information, in a uniform way. OpenAIRE also collects information from OA journals and other services that collect and display disparate collections. OpenAIRE currently operates an interoperable and validated network of more than 590 repositories and OA journals, integrating more than 11 million OA publications and 7,000 datasets, with 50,000 organizations and 30,000 projects from the EC and other funders. It has identified over 100,000 FP7 publications from about half the 26,000 FP7 projects, and offers literature-data integration services.

OpenAIRE provides guidelines for datasource managers (incl. data and literature repositories, ejournals, aggregators and Current Research Information Systems or CRIS's) to assist them in exposing their metadata in a way that is compatible with the OpenAIRE infrastructure. It also offers a CrossRef service to establish the links between publications and FP7 projects and provides a generic data repository, Zenodo.org. Zenodo is a simple and innovative service that enables researchers, scientists, EU projects and institutions to share and showcase multidisciplinary research results (data and publications). It complements the existing institutional or subject-based repositories of the research communities. Zenodo enables researchers, scientists, EU projects and institutions to:

- Easily share the long tail of small research results in a wide variety of formats including text, spreadsheets, audio, video, and images across all fields of science.
- Display their research results and get credited by making the research results usable and integrate them into existing reporting lines to funding agencies like the European Commission.
- Easily access and reuse shared research results.

Zenodo currently encompasses 10,600+ research related artefacts (6,172 publications, 611 datasets, 3,085 software items, 480 presentations, 208 posters, 26 videos, 95 images) in 300 communities.

3. OpenAIRE as a Human Infrastructure

Equally as important as this technical infrastructure, however, is OpenAIRE's extensive human support network. Europe has a diverse repository landscape, with the maturity and implementation of OA policies and mandates varying hugely across the continent. Thus OpenAIRE operates a network of 33 pan-European advocacy nodes, known as National Open Access Desks (NOADs), which reach out to institutions

locally to inform and advise regarding the EC's OA mandate and to align local infrastructures with a common European platform.

OpenAIRE's community support network works to publicise and clarify the EC's OA policies and to advance open access initiatives at national levels through the following channels:

- **Helpdesk:** As part of its localized outreach, a ticketing system distributes requests (often policy or repository related) to NOADs. This proves to be an efficient way of allocating resources across the network.
- **Online resources:** The OpenAIRE portal gives access to many FAQs, guides, and factsheets. Common questions such as Intellectual Property Rights or copyright also have to be answered. The portal also features EC projects that successfully implement OA [5]. A strong social media presence helps attract new users and foster a sense of community. A blog and monthly newsletter also raise awareness of scholarly communications issues [6, 7].
- **Webinars / Training:** The OpenAIRE networking team is adept at holding webinars and training on different aspects of OA and over the last few year has held a series of European workshops on scholarly communication topics for example legal issues in data sharing, and effective data management practices and policies [8].

The scholarly communication landscape is changing rapidly and OpenAIRE exploits synergies with key players to align its strategic goals. It works closely with a number of initiatives that promote open science, training and common standards, for example the EC-funded projects *Facilitate Open Science Training for European Research* (FOSTER), *Open Access Policy Alignment Strategies for European Union Research* (PASTEUR4OA) and the *Confederation of Open Access Repositories* (COAR) [9, 10, 11]. Outside Europe, OpenAIRE collaborates with many international organisations such as SHARE, US and the Australian National Data Service and the Research Data Alliance, as well as data infrastructures such as Data Dryad and DataCite.

4. OpenAIRE2020

OpenAIRE is now in its third phase of project funding: OpenAIRE2020. In this phase, OpenAIRE is maintaining and developing its established services while going further in researching and developing the new scientific commons. New areas of investigation include:

Open Research Data Pilot: Valuable information produced by researchers in many EU-funded projects will be shared freely as a result of a pilot on open research data in Horizon 2020. Researchers in projects participating in the pilot are asked to make the underlying data needed to validate the results presented in scientific publications and other scientific information available for use by other researchers, innovative industries and citizens. This will lead to better and more efficient science and improved transparency for citizens and society. It will also contribute to economic growth through open innovation.

FP7 post-grant OA Pilot: The EC has launched a pilot to fund peer-reviewed OA publications arising from finalized FP7 projects through the OpenAIRE project [12]. The FP7 post-grant Open Access Pilot provides an additional instrument to

improve access to research results from FP7 projects, but does not affect authors' choice on how their project publications are made OA. According to its policy guidelines [12], the Pilot will cover OA Article Processing Charges (APCs) for FP7 projects up to two years after they end, funding a maximum of three peer-reviewed research articles or monographs per FP7 project. The Pilot will be subject to an early 2016 review based on the data gathered to ensure that the guidelines continue to support the project objectives. An intensive dissemination effort is currently taking place towards NOADs, institutions and publishers in order to ensure that eligible researchers will become aware of this funding opportunity.

Literature-Data Brokerage: Domain-specific organisations with a track record of excellence in dataset-literature management (EBI, PANGEA, DANS) are working to provide mechanisms for data citation. The team is working with publishers and repositories to build community consensus regarding data citation standards. Advancing the work of the RDA/WDS Data Publishing Services working group, this work aims to develop a data citation interlinking service that collects and resolves DOI-to-DOI cross-references between datasets and publications from publishers, data repositories and other infrastructures, while ensuring the principles of openness and reciprocity.

Linked Open Data (LOD): This work extends technical interoperability to provide scholarly communication content as LOD objects. It will map to and engage with related open content initiatives, such as Open Educational Resources, Public Sector Information (often used as datasets in Social Sciences and Humanities), and DBpedia.

Usage statistics: OpenAIRE2020 is working to develop a sophisticated set of indicators to measure the impact of EC funded research across subjects. To facilitate comparison, the exchange and aggregation of Altmetrics in general and usage statistics in particular must be improved on an international level. This includes the development of best practices for gathering data corpora and the dissemination of alternative research measures. OpenAIRE2020 is aiding this effort by investigating and describing multidimensional indicators to assess scholarly performance, cooperating with data providers (repositories, CRIS) to collect, aggregate and analyse usage data and participating in the RDA/WDS group.

Open peer review: OpenAIRE2020 is also conducting research into new peer review approaches. This work involves a landscape scan and evaluation of new forms of (open) peer review, gathering information via literature research and a stakeholder survey. As a prototype for the Humanities and Social Sciences, hypotheses.org, a platform hosting academic blogs, will carry out experiments to model the workflow from blog articles to peer reviewed publications. Two further small-scale open peer review prototypes will be commissioned via tender.

Global interoperability: OpenAIRE2020 is bringing international repository initiative representatives into dialogue under an independent and global organization (COAR), so as to form a common strategic vision and develop shared strategies. In practical terms this work aims at accelerating current activities in the area of interoperability by promoting alignment, and facilitating the exchange of good practices and the adoption of shared indicators, services and technologies across regional networks.

5. Summary

Although some believe that the argument in favour of OA has already been won (the current authors amongst them), it is clear that the road to successful implementation still stretches far ahead. At the core of OpenAIRE are technical activities and community coordination in support of the EC's commitments to open scholarship and open science. The outreach effort needed to translate these commitments, including the EC's OA mandate, into researcher/project coordinator awareness, and hence foster OA compliance and best practice, must not be underestimated. At the same time, the EC's need to monitor this compliance presents a technological challenge for distributed metadata aggregation which requires bleeding-edge research into data management and citation, usage statistics and other metrics, and infrastructure interoperability. OpenAIRE is achieving both aims through the incremental development of trusted services, aligning and implementing policies that promote the free flow of research outcomes, and so supporting and strengthening research collaboration and excellence, and serving as a platform for innovation.

References

[1] European Commission, *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020*, 2013. Available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (accessed 12 June 2015).
[2] European Commission, *Press Release: Better Access to Scientific Articles on EU Research*, 2008). Available at: http://ec.europa.eu/research/swafs/index.cfm?pg=policy&lib=pilot (accessed 12 June 2015).
[3] European Commission, *What is Horizon 2020?* Available at: http://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020 (accessed 12 June 2015).[1]
[4] European Commission, *Press Release: Scientific Data: Open Access to Research Results Will Boost Europe's Innovation Capacity,* 2012. Available at: http://europa.eu/rapid/press-release_IP-12-790_en.htm (accessed 12 June 2015).
[5] OpenAIRE. Available at: www.openaire.eu (accessed 12 June 2015).
[6] OpenAIRE, Featured Projects. Available at: https://www.openaire.eu/featured-projects/ (accessed 12 June 2015).
[7] OpenAIRE Blog. Available at: https://blogs.openaire.eu/ (accessed 12 June 2015).
[8] OpenAIRE Newsletter. Available at: https://www.openaire.eu/newsletter/view (accessed 12 June 2015).
[9] OpenAIRE Workshops. Available at: https://www.openaire.eu/workshops-tutorials/workshops-*tutorials*/workshops/ (accessed 12 June 2015).
[10] FOSTER. Available at: https://www.fosteropenscience.eu/ (accessed 12 June 2015).
[11] PASTEUR4OA. Available at: http://www.pasteur4oa.eu/ (accessed 12 June 2015).
[12] COAR. https://www.coar-repositories.org/ (accessed 12 June 2015).
[13] FP7 post-grant Open Access publishing funds pilot. Available at: https://www.openaire.eu/goldoa/fp7-post-grant/pilot (accessed 12 June 2015).

This page intentionally left blank

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

Subject Index

New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

Author Index

This page intentionally left blank