POSITIONING AND POWER IN ACADEMIC PUBLISHING: PLAYERS, AGENTS AND AGENDAS

This page intentionally left blank

Positioning and Power in Academic Publishing: Players, Agents and Agendas

Proceedings of the 20th International Conference on Electronic Publishing

Edited by

Fernando Loizides

Emerging Interactive Technologies Lab, University of Wolverhampton, UK

and

Birgit Schmidt

University of Göttingen, State and University Library, Germany

IOS
Press

Amsterdam • Berlin • Washington, DC

ISBN 978-1-61499-648-4 (print)
ISBN 978-1-61499-649-1 (online)
Library of Congress Control Number: 2016940262

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

Cover design: IOS Press
Cover image: Cultural Studies Divisional Library, Göttingen State and University Library; photo by Claudia Hake.

LEGAL NOTICE

The publisher is not responsible for the use which might be made of the following information.

PRINTED IN THE NETHERLANDS

Preface

Running a technology-informed conference such as the *International Conference on Electronic Publishing* (Elpub) for the twentieth time could be taken as a sign of saturation and maturity. However, if we consider technology as only one cultural aspect of our current scholarly communication ecosystem, we have to note that we continue to be in the middle of a digital transformation. Technologies come and go due to social reasons; the positioning of stakeholders and the distribution of economic, technological and discursive power continues to be negotiated. And at times a seemingly given fact of publishing like the transfer of intellectual property rights to third parties gets heavily questioned – as the recent discussion around the shadow library Sci-Hub indicates. To provide room for discussion beyond technology and embed technology in its social and cultural framework, Elpub 2016 will open the floor for emerging alternatives in how scholars and citizens interact with scholarly content and the role dissemination and publishing plays in these interactions. What is the core of publishing today? How does agenda-setting in emerging frameworks like Open Science work and what is the nature of power of the surrounding scholarly discourses? How does this relate to the European and world-wide Open Science and Open Innovation agenda of funders and institutions, and what does this look like in publishing practice? Asking such questions promises to widen our horizons.

Elpub reaches a great milestone with its 20th anniversary this year, to be held in Göttingen, Germany. Since its beginnings twenty years ago, the conference has been a leading forum for electronic publishing topics, attracting people from around the world and facilitating active collaboration and knowledge exchange. Twenty years on, the conference brings together leading stakeholders such as academics, practitioners, policy makers, students and entrepreneurs from a wide variety of fields and countries.

The conference once again has an exciting programme in store for attendees and readers of the proceedings alike. The conference opens with a reflection and celebration of the last twenty years. This year, 17 research papers and 9 posters will be presented. The programme covers a wide variety of topics, including how to maintain the quality of electronic publications, modelling processes, and implementation issues regarding open access. These subjects, and especially the latter, become even more prevalent with reforms such as Britain's Research Excellence Framework rule which allows only open access articles to be eligible for submission – and on this basis, deposit in repositories (as a core element of institutional research information systems) becomes the norm. At Elpub, there will be several new publishing systems and repositories presented and tested, as well as datasets for the delegates to examine.

In addition, four workshops will offer delegates the opportunity to explore "Open Peer Review: Models, Benefits and Limitations" (co-organized by OpenAIRE), "Opening up the collection – reuse and publishing" (LIBER), "Entering the publishing system – Junior Scientist Day (FOSTER)" and "OJS 3.0 and OMP 1.2: The latest in open source software for academia-controlled publishing (PKP)".

We are delighted to have three captivating keynote talks this year: Jean-Claude Guédon from the University of Montréal, Canada, will speak on the topic of "Whither Open Access? Four scenarios and four choices", and our second keynote speaker, Tara

Andrews from the University of Bern in Switzerland talks on the topic of “After the Spring: digital forms of scholarship and the publication ecosystem”. The third speaker Prateek Mahalwar shares with us his views on opportunities and challenges for early career researchers in the context of Open Science. Finally, a panel of experts will investigate the conference topics in an open forum with short stakeholder perspectives and the opportunity for the audience to engage with the discussion.

We would like to express our sincere gratitude to all members of the Elpub Executive Committee who, together with the Programme Committee, helped us to bring together such a diverse and exciting programme. We would also like our sponsors – Altmetric, MDPI and Copernicus Publications (at the time of writing) – for their support as well as their openness to discussion and cooperation in bringing forward the Open Science agenda.

We wish everyone an inspiring conference and a happy 20th Anniversary with many more to come. We look forward to continuing the discussion and seeing you again at the 21st edition of the conference in Cyprus!

Fernando Loizides and Birgit Schmidt

1 June 2016

Organisation

Conference Host

University of Göttingen

General Chair

Birgit Schmidt, University of Göttingen, Germany

Programme Chair

Fernando Loizides, University of Wolverhampton, UK

Programme Committee Members

Ana Alice Baptista	University of Minho	Portugal
Margo Bargheer	University of Göttingen	Germany
George Buchanan	City University London	UK
Leslie Chan	University of Toronto	Canada
Jan Engelen	Katholieke Universiteit Leuven	Belgium
Paola Gargiulo	CINECA	Italy
Stamatios Giannoulakis	Cyprus University of Technology	Cyprus
Herbert Gruttemeier	CNRS	France
Arunas Gudinavicius	Vilnius University	Lithuania
Puneet Kishor	Independent	USA
Alexia Kounoudes	Cyprus University of Technology	Cyprus
Peter Linde	Blekinge Institute of Technology	Sweden
Natalia Manola	University of Athens	Greece
Eva Méndez Rodríguez	Universidad Carlos III de Madrid	Spain
Pierre Mounier	Open Edition	France
Panayiota Polydoratou	Alexander Technological Education Institute of Thessaloniki	Greece
Andreas Rauber	TU Wien	Austria
Laurent Romary	DARIAH	Germany
Anthony Ross-Hellauer	University of Göttingen	Germany
Andrea Scharnhorst	DANS-KNAW	Netherlands

Executive Committee

Sponsors

Altmetric

MDPI

Copernicus Publications
The Innovative Open Access Publisher

nature
COMMUNICATIONS

This page intentionally left blank

Contents

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-1

Time to Adopt: Librarians' New Skills and Competency Profiles

Birgit SCHMIDT[a,1] , Pascal CALARCO[b], Iryna KUCHMA[c], and Kathleen SHEARER[d]

[a] *University of Göttingen, State and University Library*
[b] *University of Waterloo*
[c] *Electronic Information for Libraries (EIFL)*
[d] *Confederation of Open Access Repositories (COAR)*

Abstract. On the one hand, libraries are at the forefront of the digital transformation and digital information infrastructures, on the other, they manage and curate cultural heritage collections. This brings about new ways of engagement with information and knowledge and the need to rethink skills and competency profiles – which enable librarians to support e-research all along the research cycle. This paper presents findings of the joint Task Force on Librarians' Competencies in Support of E-Research and Scholarly Communication.

Keywords. e-research, competencies, job profiles, libraries.

1. Introduction

Rapid changes in technology and associated shifts in research and scholarly communications are profoundly changing the role of libraries in the 21st century. The emergence of e-research, for example, is bringing about new ways of doing science across the globe, compelling libraries to adopt new services, such as assisting with the development of research data management plans, hosting collaborative virtual research environments, managing institutional repositories, and disseminating research outputs through open access mechanisms. These novel services require a range of new skills and expertise within the library community as well as a shift in organizational models for libraries.

In August 2013, the Association of Research Libraries (ARL), the Canadian Association of Research Libraries (CARL), the Association of European Research Libraries (LIBER), and the Confederation of Open Access Repositories (COAR) launched the joint Task Force on Librarians' Competencies in Support of E-Research and Scholarly Communication.[2]

Since then, the Task Force has been working on identifying emerging specialty roles, through performing literature reviews and collaboratively preparing a series of service areas and competencies documents for research data management, scholarly communication and Open Access, digital curation and preservation and support for digital scholarship.

[1] Corresponding Author. E-mail: bschmidt@sub.uni-goettingen.de.
[2] https://www.coar-repositories.org/activities/support-and-training/task-force-competencies.

2. Competencies and Skills under Review

The growing abundance of digital information and data affects the whole research workflow, including methods and tools as well as enabling infrastructures. Accordingly, there is an emerging need for a new type of workforce, and existing and emerging staff competencies and skills are under scrutiny. In particular, a lot of attention has been attracted by "data science", in particular in relation to "big data"; with Hal Varian, chief economist at Google declaring "the sexy job of the next 10 years will be statisticians" (Lohr, 2009). With a view on customers Affelt (2015) pointed out how librarians and information professionals – which have always been well-versed in working with data – can leverage their skills and training for big data applications that resonate with stakeholders.[3][3]

Librarians manage different types of published information and data, and also curate a wealth of information that awaits further exploration and exploitation based on digital methods and tools. To step up skills and competencies of librarians and to some degree research staff, several initiatives have looked into specific areas, e.g. open science, research data management, digital curation, digital humanities, eResearch, data science, etc.[4][4] Some of these initiatives focus on professional training, others target the development of higher education curricula and explore how librarians can contribute.

2.1. Defining Competencies and Skills

According to the European e-Competence Framework (e-CF) competence is the "demonstrated ability to apply knowledge, skills and attitudes to achieve observable results". Hence, a competence is not a skill; on the contrary, a competence *embeds* skills. Whilst competencies are holistic concepts, skills are precise and definite abilities, either hard technical, e.g. make a cost / benefit analysis, develop user interfaces; or soft, e.g. deploy empathy to customer needs, negotiate contract terms and conditions (e-CF, 2014). Job profiles typically combine several competencies, and one single competence may be assigned to a number of different job profiles. A core idea in this context is that competencies can be grouped by areas (plan, build, run, enable, manage) and can be categorized by proficiency levels, ranging from the ability to apply knowledge and skills to solve straightforward problems to the overall accountability and responsibility, and to solve critical problems in an innovative way (the e-CF's levels e-1 to e-5 are related to the European Qualifications Framework's levels 3 to 8) (e-CF, 2014).

In North America, ARL members have been engaged in identifying the issues around evolving competencies needed for the work of research librarians, and how these roles intersect with new functional specialists, documented in a series of

[3] See also: Florida Library Webinars, 31 August 2015, http://floridalibrarywebinars.org/the-accidental-data-scientist-a-new-role-for-librarians-and-info-pros-ondemand

[4] E.g. Facilitate Open Science Training for European Research (FOSTER), https://www.fosteropenscience.eu/; MANTRA Research Data Management Training, http://datalib.edina.ac.uk/mantra/; Essentials for Data Support, http://datasupport.researchdata.nl/en/; Curriculum Framework for Digital Curation, Digital Curator Vocational Education Europe (DigCurV), 2013, http://www.digcur-education.org; EDISON, http://www.edison-project.eu.

publications entitled "New Roles for New Times" (Janguszewski and Williams, 2013; Covert-Vail and Collard, 2012) and by Rockenbach *et al.* (2015).

3. Mapping E-Research and Service Areas

A range of descriptions of service~~s~~ areas and related competencies have been developed and shared with the community for comments by the joint Task Force on Librarians' Competencies in Support of E-Research and Scholarly Communication. Consolidated versions will be published in spring 2016.

3.1. Managing Research Data

Research data management (RDM) involves services and infrastructures in order to support the handling of research data across the data lifecycle (i.e. creating/collecting, processing, analyzing, publishing, archiving/preserving, re-using data). The various aspects of RDM are often distributed across different support services (research office, IT services, library) and academic departments. Interviews with researchers demonstrate that, while researchers need support in numerous areas across the entire research lifecycle: planning, organizing, security, documenting and sharing, preparing datasets for deposit and long-term preservation, as well as issues related to copyright, licensing and intellectual property more generally (e.g. Wilson, 2014; Parsons *et al.*, 2013).

Research data management encompasses a large group of activities that may differ significantly across the research data lifecycle. Generally it requires a high level of interaction with researchers and also working with other support services including technical services and research officers.

There are various strategies for service development and operation, some concentrate on discipline-specific services developed in the context of projects, others highlight multidisciplinary perspectives (e.g. Molloy and Snow, 2012; Carlson, 2015).

For RDM training close work with disciplinary experts is recommended to ensure that the terminology used is accurate and clear; discipline-specific examples and good practices are also highly valuable for engaging the audience and for putting basic principles in context (Molloy and Snow, 2012).

Based on funder requirements, the need to support researchers in creating and implementing data management plans has substantially grown over the last years. Several libraries have set up a service to support such needs, often in collaboration with other service units (e.g. research office, IT services, legal advisor, ethics committee). The development of such a service can even serve as a training ground for librarians and other institutional stakeholders (Davis and Cross, 2015).

Libraries' activities in research data management can be usefully conceptualized as falling into three broad categories: *providing access to data*; *supporting researchers and students in managing their data*; and *managing a data collection*. There are overlaps, but each of these areas has some distinctive roles for librarians. Providing access to data mainly involves consultation and reference services, e.g. to identify datasets, provide advice on discovery and analytic tools as well as how to cite/reference data. Advocacy and support for managing data is a wide area of activities ranging from promoting the institutional data policy, providing support and training, e.g. on how to write a data management plan or how to identify and use data repositories, to develop data curation profiles, and to manage software related to data. The management of collections targets activities such as the preparation of data, its preservation, sharing and publishing. Further details on the core competencies needed to cover these areas can be found in the RDM skills and competency profile (cf. note 1).

3.2. Scholarly Communication and Open Access

Scholarly communication and Open Access (OA) involves changing modes of communication of research made possible in the digital environment. For example, the evolution from the traditional commercial publishing model where the author signs away their copyright to a work of original research scholarship to subscription-based journals to one of several OA models that have been emerging over the past two decades. Scholarly communication can be defined as "the system through which research and other scholarly writings are created, evaluated for quality, disseminated to the scholarly community, and preserved for future use. The system includes both formal means of communication, such as publication in peer-reviewed journals, and informal channels, such as electronic listservs" (ARL). Other informal scholarly communication channels include posts in social media, e.g. blogs, tweets, etc.

One of the most widely used definitions of OA is that from the Budapest Open Access Initiative, from a conference of that name held in 2001: "By 'open access' to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited" (BOAI, 2002).

Some examples of how libraries have been involved in this process are:

- Providing consultation and training that encourages and enables faculty to manage their own copyright and improve the economics of, and access to, published research.
- Developing and contributing to scientific information infrastructures for the dissemination and linking of research outputs, e.g. digital repositories and their networks – on institutional, national and international levels (e.g. SHARE, OpenAIRE, LA Referencia, COAR).[5]
- Offering OA journal and/or book publishing services and other value-added services to scholars on their campus (work with the university press if there is any).
- Working with the acquisition department, library consortia and research funders to develop and maintain a publication fund, OA memberships and OA clauses in licenses.
- Providing access to services and resources that help measure quality and impact of scholarship, from traditional bibliometrics to emerging altmetrics.

[5] SHARE, http://www.share-research.org/; OpenAIRE, https://www.openaire.eu/; LA Referencia, http://lareferencia.redclara.net/; COAR, https://www.coar-repositories.org.

Libraries' activities in scholarly communication and open access typically fall into one of these categories: *scholarly publishing services; copyright and open access advocacy and outreach; scholarly resource assessment.* Some level of subject knowledge is required in most of these roles. In particular, librarians will need to have a broad perspective and understanding of the traditional (commercial, society) and open access models of publishing, intellectual property issues, and economics of scholarly publishing. For example, librarians in this role may support graduate students and faculty members who wish to start new Open Access journals that the library may host, coordinate education and advocacy events such as Open Access Week, and serve on campus copyright committees to assist with campus policy development and interpretation.

3.3. Digital Humanities

Digital Humanities (DH) describes a multidisciplinary research community involving the application of computational methods to humanistic topics of inquiry, or, more broadly, the intersection between the arts and humanities and information technology and/or digital expression (for a discussion of research trends and views on DH see e.g (Burdick *et al.*, 2012; Holm *et al.*, 2015). Digital humanists utilize computational methods and/or digital tools to advance research and pedagogy. Methods and tools include, but are not limited to: 3-D representation, digital mapping, electronic textual analysis, digitization of materials, data visualization techniques, and interactive digital media including game-based systems.

Over the last decade, several universities have stepped up their support of humanities research by creating research centres and workspaces,[6][6] many of them in close collaboration and physically located with their respective university libraries.

Engagement with Digital Humanities is an evolving specialization in librarianship, one that requires a combination of a strong academic background in the arts and humanities (domain expertise), technical grounding in technologies and tools to support computational models of research and teaching in Humanities, and project management. The role has other important components such as advisor, advocate, and partner for special collections curators. The digital humanities librarian's role is also directly related to scholarly communication and data management.

The library's organizational structure will affect the services offered and roles played by digital humanities librarians, but across various models, the digital humanities librarian will likely work in a team environment, making collaboration and the ability to perform in a changing and dynamic role core competencies for this specialization.

When it comes to services and responsibilities digital humanities librarians engage in *scholarly communication and publishing, technical services* (in particular related to interaction with digital resources and collections), *partner with faculty and student for*

[6] Cf. centerNet, an international network of digital humanities centres, http://www.dhcenternet.org/, and European Association for Digital Humanities, Digital Humanities Centres, http://eadh.org/education/digital-humanities-centres. From a much broader perspective an increasing number of universities and colleges are also establishing digital scholarship centres to support researchers in their work with digital tools and large datasets, such as data visualization in the environmental sciences, data mining of large corpora of texts in the humanities, and developing GIS or other geolocation data representations in the social sciences (Lippincott, 2014).

digital humanities research and consulting, provide teaching and training activities, and develop and manage spaces (labs, collaboratories) for digital humanities work.

Accordingly, digital humanities librarians bring together a wide range of compentencies and skills. A base layer is typically advanced academic subject expertise and professional training in library and information science, particularly in scholarly communication and data management. When it comes to technical skills, many job descriptions note the emerging and evolving state of technology by requiring general competencies such as "demonstrated ability and interest in exploring and evaluating emerging technologies in support of digital humanities," and a "willingness to remain current with changing technology and its applications" (cited from job descriptions, cf. note 1). Technical skills and competencies include e.g. data visualization, text mining, metadata standards and schema, text markup and encoding, semantic web technologies. Essential for direct involvement and/or consultation with research activities are also project and program development and collaboration skills, e.g. grant writing and the development of technology-rich work spaces.

4. Conclusions

Not surprisingly, a number of other new areas could benefit from librarians' support – and this again comes with a need for developing/expanding skills to fulfill these new roles. One area which is evolving very fast is text and data mining (TDM), and libraries might already have a range of subscriptions and collections which come with appropriate licenses but have not yet stepped up to provide practical support for researchers to exploit these riches (Okerson, 2013). As already mentioned above support for TDM plays a key role in digital humanities research, e.g. allowing new views on texts, but other research areas such as economical and social sciences will benefit as well (Liber, 2014; ASIS&T, 2015).

In our discussion of competency profiles we have only briefly touched how librarians acquire the skills and competencies needed for these evolving and sometimes already well-established service areas. Strategies will vary depending on institutional and personnel resources and range from attending workshops or conferences and/or joining working groups, the development of institutional training programs for individuals and/or groups, participation in online learning course (e.g. MOOCs), etc. Most beneficial might be to combine newly hired experts and long-term staff in new teams which dedicate their efforts to developing and delivering new types of services. Such teams will often combine staff with different backgrounds, and bring new skills sets to the institution, e.g. from the publishing industry to build up / enhance a publishing unit, or from research disciplines or technology experts to develop and promote specialized data infrastructures and digital work environments. Not surprisingly, involvement in collaborative projects, national and internationally, are a good instrument to contribute own expertise and learn from others to build up prototypical services. However, additional effort will be needed to assess the results of these efforts and for sustaining both personnel and infrastructures.

It should be noted that job descriptions can be excessively demanding in terms of experiences and skills, as if the search is for the "Unicorn Librarian – that magical creature who can be all things to all people" (Johnson, 2014). Therefore, individuals and employers should consult the task force's competency profiles with some caution.

Typically it will be a group of individuals that bring together these competencies and skills, a collaborative work force which strengthens the library's capacities and which may also be an element of new organizational structures.

Acknowledgement. The authors would like to thank Rob Grim, Alicia López Medina, Susan Reilly, Judy Ruttenberg, and Dominic Tate for their contributions to the work of the Task Force and their comments to this paper.

References

Affelt A. (2015) The Accidental Data Scientist: Big Data Applications and Opportunities for Librarians and Information Professionals. Medford, New Jersey: Information Today, Inc.
ARL – Association of Research Libraries, Focus areas: Scholarly communication. Available at: http://www.arl.org/focus-areas/scholarly-communication (Accessed: 11 March 2016).

ASIS&T (2015) 'Text Mining in Libraries, How Librarians Develop Skills Required to Support this Evolving Form of Research', ASIS&T webinar, blog post, 2 November 2015. Available at: https://www.asist.org/events/webinars/text-mining-in-libraries (Accessed: 11 March 2016).

BOAI – Budapest Open Access Initiative (2002): Declaration, Budapest, Hungary, 14 February 2002. Available at: http://www.budapestopenaccessinitiative.org/read (Accessed: 11 March 2016).

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T. and Schnapp, J. (2012) 'Digital Humanities', Cambridge, MA: MIT Press. Available at: https://mitpress.mit.edu/sites/default/files/titles/content/9780262018470_Open_Access_Edition.pdf (Accessed: 11 March 2016).

Carlson, J., Sapp Nelson, M. Johnston, L. R. and Koshoffer, A. (2015) 'Developing Data Literacy Programs: Working with Faculty, Graduate Students and Undergraduates', Bulletin of the American Society for Information Science and Technology 41(6), pp 14–17. doi:10.1002/bult.2015.1720410608.

Covert-Vail, L. and Collard, S. (2012) 'New Roles for New Times: Research Library Services for Graduate Students', Association of Research Libraries (ARL), http://www.arl.org/storage/documents/publications/nrnt-grad-roles-20dec12.pdf (Accessed: 11 March 2016).

Davis, H. M. and Cross, W. M. (2015) 'Using a Data Management Plan Review Service as a Training Ground for Librarians', Journal of Librarianship and Scholarly Communication 3(2), eP1243, doi: 10.7710/2162-3309.1243.

eCF – European e-Competence Framework 3.0 (2014) A common European Framework for ICT Professionals in all industry sectors, CWA 16234:2014, Part 1. Available at: http://www.ecompetences.eu/e-cf-3-0-download (Accessed: 11 March 2016).

Holm, P., Jarrick, A. and Scott, D. (2015) 'The Digital Humanities', Humanities World Report 2015, pp 64–83. Palgrave Macmillan UK. doi: 10.1007/978-1-137-50028-1.

Janguszewski J. M. and Williams, K. (2013) 'New Roles for New Times: Transforming Liaison Roles in Research Libraries', report prepared for the Association of Research Libraries (ARL), August 2013. Available at: http://www.arl.org/storage/documents/publications/nrnt-liaison-roles-revised.pdf (Accessed: 11 March 2016).

Johnson A. (pseudonym) (2014) 'Hiring Data Librarians: Notes on hiring and being hired as a data librarian'. Available at: http://www.scribd.com/doc/265015825/Hiring-Data-Librarians (Accessed: 11 March 2016).

LIBER – Association of European Research Libraries (2014), 'Text and Data Mining: The need for change in Europe'. Available at: http://libereurope.eu/text-data-mining (Accessed: 11 March 2016).

Lippincott, J., Hemmasi, H. and Lewis, V. M. (2014) 'Trends in Digital Scholarship Centers', Educause Review, 16 June 2014, http://er.educause.edu/articles/2014/6/trends-in-digital-scholarship-centers (Accessed: 11 March 2016).

Lohr S. (2009) 'For Today's Graduate, Just One Word: Statistics', New York Times, 5 August 2009. Available at: http://www.nytimes.com/2009/08/06/technology/06stats.html (Accessed: 11 March 2016). [16] Molloy, L. and K. Snow, K. (2012) 'The Data Management Skills Support Initiative: Synthesising Postgraduate Training in Research Data Management', The International Journal of Digital Curation 7(2), pp 101–109. doi: 10.2218/ijdc.v7i2.233.

Okerson, A. (2013) 'Text & Data Mining – A Librarian Overview', paper presented at IFLA World Library and Information Congress, 17-23 August 2013, Singapore. Available at: http://library.ifla.org/252/1/165-okerson-en.pdf (Accessed: 11 March 2016).

Parsons, T., Grimshaw, S. and Williamson, L. (2013) 'Research Data Management Survey', University of Nottingham, February 2013. Available at: http://admire.jiscinvolve.org/wp/files/2013/02/ADMIRe-Survey-Results-and-Analysis-2013.pdf (Accessed: 11 March 2016).

Rockenbach, B., Ruttenberg, J., Tancheva, K. and Vine, R. (2015) 'Pilot Library Liaison Institute', Association of Research Libraries/Columbia University/Cornell University/University of Toronto, Final Report, June 2015. Available at: http://www.arl.org/storage/documents/publications/library-liaison-institute-final-report-dec2015.pdf (Accessed: 11 March 2016).

Wilson, J. A. J. (2014) 'University of Oxford Research Data Management Infrastructure', LIBER Research Data Management Case Study, June 2014. Available at: http://libereurope.eu/wp-content/uploads/2014/06/LIBER-Case-Study-UOX.pdf (Accessed: 11 March 2016).

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-9

SCOAP^3/SCOAP^3-DH – Gold Open Access in High Energy Physics

Angelika KUTZ LL.M.[1]
TIB Hannover
(German National Library of Science and Technology)
Germany

Abstract. SCOAP^3 is a global partnership which converts high-quality subscription journals in the field of High-Energy Physics to Open Access through redirection of existing subscription funds. Since January 1, 2014 the SCOAP^3 Gold Open Access Repository is providing free access to scientific articles in high quality journals in the field of High Energy Physics. This article describes this international pilot which flips the current subscription-based financing of scientific publication to an output-based financing model (fair share). This includes a description of the unique mechanisms of SCOAP^3 as well as its governance structure and short view on the national German contributing partners, especially the German universities (SCOAP^3-DH).

Keywords. SCOAP^3, SCOAP^3-DH, Gold Open Access, High Energy Physics, Flipping Model, CC-BY, No Costs for Authors, Text- and Datamining, CERN, TIB Hannover

1. Introduction

SCOAP^3 stands for Sponsoring Consortium for Open Access Publishing in Particle Physics. The goal of this worldwide open access project initiated by CERN (*European Organization for Nuclear Research*) is to convert scientific articles in the high energy physics field published in high quality journals into Gold Open Access by redirecting subscription fees for the flipping of the financing system.

Its first phase from 2014-2016 was made possible due to the common effort of about 3000 libraries and consortia worldwide, research organisations, cooperative publishers and funding agencies in many countries.

SCOAP^3-DH stands for Sponsoring Consortium for Open Access Publishing in Particle Physics – German Universities. Organised by TIB (*German National Library of Science and Technology*) in Hannover in Germany this national project SCOAP^3-DH is taking care of the participation of the German universities in the worldwide project led by CERN.

[1] Corresponding Author.: angelika.kutz@tib.uni-hannover.de

2. SCOAP^3: a worldwide Gold Open Access Flipping Model

This worldwide pilot originally initiated by CERN includes over three thousand libraries and consortia and is supported by funding agencies in various countries.

The thereby implemented Gold Open Access articles can be read and re-used by anyone as long as the author is named as all articles are published under a CC-BY licence (https://creativecommons.org/licenses/by/4.0/).

During its first phase SCOAP^3 redirects the subscription payments into the SCOAP^3 fund out of which the publishers are centrally paid by CERN for their services.

These concrete services were defined during the tendering process and laid down in a technical specification which is available online under: http://scoap3.org/files/Technical_Specification.pdf.

In order to be able to provide funds for the SCOAP^3 fund the participating publishers reduced the proportional amount of the converted journals in the bills sent to the libraries thereby enabling them to redirect these released amounts to the SCOAP^3 fund.

All final versions of peer-reviewed articles published in the SCOAP^3 journals are immediately available for everyone on the internet and free of charge on both the publishers' websites and the SCOAP^3 Repository (http://repo.scoap3.org).

2.1. Start on January 1, 2014

Parameters of SCOAP^3:

- Gold Open Access (worldwide free accessibility over the internet).
- Key Journals in High Energy Physics have – fully or partially – been converted into Gold Open Access.
- No costs for authors.
- The copyright remains with the author.
- No administrative burden for the author.
- No administrative burden for participating institutions due to National Contact Points dealing with SCOAP^3/SCOAP^3-DH centrally.
- CC-BY Licences (the copyright remains with the author).
- CC0 for metadata.
- Text- and datamining allowance.
- Constantly growing number of articles available upon publication in the SCOAP^3 Repositorium (end of 2015: more than 9.000 Gold Open Access articles) and the publishers' websites

These parameters have been implemented due to an international tendering process led by CERN for the first phase of SCOAP^3.

2.2. Participating Countries

Currently there are over 30 countries worldwide participating, more are expected to join soon. The current international SCOAP^3 Partners are listed on the CERN website: http://scoap3.org/participating-countries.

2.3. The capping mechanism of SCOAP^{3}leading to reasonable average APCs

The SCOAP^3 tendering process included a maximum amount paid to the individual publisher for a certain ("capped") amount of articles of the specific journal ("capping mechanism"). Any article above this "cap" will be published under the same conditions and services (see parameters described in *2.1*) but the publisher will not receive any additional amounts of money for any article exceeding this cap. This capping mechanism thereby steadily leads to sinking average article processing charges (APC) as the publisher receives only a maximum amount for each journal contract.

Due to this capping mechanism the average APC for SCOAP^3 Articles (articles published in SCOAP^3 Journals) is currently about 1.100€.

2.4. Value added services of SCOAP^3

SCOAP^3 reached a very high compliance corresponding with the prerequisites (like CC-BY, XML and CC0 for metadata) implemented by the tendering process due to the central control function of a single well trained team at CERN. SCOAP^3 can offer a near to 100% compliance rate (99.98%) compared to other similar initiatives (e.g. Welcome Trust; 61%).

2.5. Advantages of SCOAP^3

SCOAP^3/SCOAP^3-DH offers the following advantages for SCOAP^3-Partners:

Advantages for the libraries:

No administration effort as after one single payment to SCOAP^3 everything regarding the publication is taken care of by SCOAP^3/SCOAP^3-DH.

There is no need for a single library to deal with individual APC-payments for each single article published and which relieves them from any administrative burden.

At the same time the necessary involvement of time, personnel and costs are eased.

Advantages for Authors:

Authors can publish their articles in high quality journals Gold Open Access upon publication and at no cost. They do not have to deal with any of the administrative burden connected with APC-payment in other publication models.

2.6. First Phase of SCOAP^3 2014-2016

During the first phase of SCOAP^3 a model of redirection of subscription cost was implemented. This included that publishers reduced their bills according to the SCOAP^3 contribution of each institution. The bills of the respective institution were reduced according to their participation portion. By using the money already contained in the system in form of subscription costs SCOAP^3 enabled a real flip from closed access to Gold Open Access in one single step.

The Technical Specification during the international tendering process led by CERN set inter alia the following preconditions:

- CC-BY
- Capping mechanism
- API
- OAI-PMH

There is the important note for articles to be published in journals which have been partially converted: authors have to upload their articles to arXiv before submitting it to the respective publisher together with the arXiv number received by uploading it to arXiv.

2.7. Second Phase of SCOAP^3 (2017-2019)

The second phase of SCOAP^3 is stepping even further. The next phase will change the current model to a mere output-orientated model in which institutions pay a lump sum (flat-rate) according to the share of publications made by their institutions (fair share) in order to enable their authors to publish freely and without any organizational or administrative burden with regards to any Article Processing Charge (APC) handling. By doing this SCOAP^3 enables a kind of an “all-you-can-publish-model” for authors of High Energy Physics Gold Open Access scientific articles in high quality journals.

3. Publishers, Journals and APCs (Article Processing Charges)

The following publishers and journals are participating in SCOAP^3 during its first phase of from 2014-2016. The table reflects the results of the international tendering process held by CERN during 2012 and 2013 in order to reach competitive quality parameters as well as reasonable APCs (Article Processing Charges).

Table 3. Publishers. Journals. Article Processing Charges (APCs)

Publisher	Journal	APCs
Elsevier	Nuclear Physics B	2.000 USD
Elsevier	Physics Letters B	1.800 USD
Hindawi	Advances in High Energy Physics	1.000 USD
Institute of Physics Publishing/ Chinese Academy of Science	Chinese Physics C	1.000 GBP
Institute of Physics Publishing/ Deutsche Physikalische Gesellschaft	New Journal of Physics	1.200 GBP
Institute of Physics Publishing/ SISSA	Journal of Cosmology and Astroparticle Physics	1.400 GPB
Jagiellonian University	Acta Physica Polonica B	500 EUR
Oxford University Press/	Progress of Theoretical and	

Physical Society of Japan	experimental Physics	
Springer/SISSA	Journal of High Energy Physics	1.200 EUR
Springer/Società Italiana di Fisica	European Physical Journal C	1.500 EUR

Due to the SCOAP^3 inherent capping mechanism which pays only for a certain amount of articles determined upfront with the publishers the current number of over 9.000 articles published via SCOAP^3 lead to an average APC of ca. 1.100 EUR.

4. SCOAP^3 Repository

The technical advantages of the SCOAP^3 Repository (http://repo.scoap3.org) are:

- OAI-PMH
- RSS feeds
- API

5. SCOAP^3 Panels

The SCOAP^3 Governance comprises three panels taking care of the decisions concerning future steps of SCOAP^3 as well as the day-to-day administration.

- SCOAP^3 Executive Committee
- SCOAP^3 Governing Council
- SCOAP^3 Forum

The *SCOAP^3 Executive Committee* comprises five to seven members and oversees the SCOAP^3 operations. The *SCOAP^3 Governing Council* has up to forty-five representatives from contributing countries. They take their decisions about the direction of SCOAP^3 during regular meetings at CERN in Geneva at least twice a year. The *SCOAP^3 Forum* is a panel open to anyone interested in the latest developments in SCOAP^3.

6. Why arXiv and SCOAP^3?

Both arXiv and SCOAP^3 are necessary publication tools for scientists. Both are Open Access tools enhancing the visibility and quick dissemination of scientific information which is advantageous for all scientists.

There are good reasons to maintain these two important instruments in order to support the worldwide swift towards more and more Open Access in scientific publication.

6.1. Good reasons for arXiv

arXiv constitutes the daily tool for scientist. Preprints uploaded by the scientists in arXiv are visible very quickly that is why their content is known sometimes long before being published in a journal. This influences the velocity of spreading scientific information in a positive way.

6.2. Good reasons for SCOAP^3

Quality Journals as they are participating in the Open Access project SCOAP^3 are highly necessary and important as the quality of the citation index depends on the quality of the journal an article is published in.

Quality Journals provide for the necessary evaluation by the reputation of the journal itself which means quality which is important for both the scientists' careers and their ability to get funding for their research.

7. National Project SCOAP^3–DH supporting SCOAP^3

7.1. German Partners for SCOAP^3

On a national level there are three German partners supporting SCOAP^3.

- SCOAP^3-DH – participation of the German universities organised by TIB (*German National Library of Science and Technology)*
- HGF (*Helmholtz Association*)/DESY (*German Electron Synchrotron)*
- MPG (*Max Planck Society*)/MPDL (*Max Planck Digital Library*)

All three organisations are National Contact Points (NCPs) and participate in the SCOAP^3 panels for the institutions they represent.

7.2. Advantages of SCOAP^3 for the German universities

Compared to an individual publication fund at each single university SCOAP^3-DH offers the advantage of no administrative burden after having paid the lump-sum and "flat-rate" for as much publications in the SCOAP^3 Journals as ever wanted.

Further advantages in short are:

- SCOAP^3-DH is an all-inclusive solution for the HEP-Community (lump-sum).
- All-you-can-publish-flat-rate for HEP-Publications in SCOAP^3-Journals.
- Reasonably-priced administration – due to centralization at CERN/TIB.
- No administrative burden regarding individual APC-handling for each university through publication funds with their financial restrictions.

All in all SCOAP^3 as well as SCOAP^3-DH are quite ahead of their time.

7.3. German Universities supporting SCOAP^3

During this current first phase of SCOAP^3 thirty university libraries plus one consortium are SCOAP^3 partners contributing financially to SCOAP^3-DH by redirecting their former subscription costs to the SCOAP^3 fund. These current participants are listed on the national website under the following link: http://www.scoap3.de/scoap3-partner/nationale-scoap3-partner.

Once the change to the output-based financing mechanism will have taken place the structure of SCOAP^3-DH for the German universities might change slightly due to the fact that each publishing community of a university will have to support SCOAP^3 financially and politically.

8. The future of SCOAP^3/SCOAP^3-DH

Currently all international partners are preparing a second phase of SCOAP^3 which is planned to consist of prolonged contracts with the current publishers and journals in order to reach Gold Open Access for the respective high energy articles and journals for another three-year-period (2017-2019).

The decision whether APS (American Physical Society) which comprises further important High Energy Physics publications will join SCOAP^3 for its second phase is still pending.

With or without APS the continuation of SCOAP^3 is highly depending on the commitment of the scientific community to keep up both Open Access publication tools, arXiv as well as SCOAP^3, and to enable its further financing in order to support and uphold quick as well as reputation-driven quality articles and their publication.

9. Repository, Websites and Information

Direct link to the SCOAP^3 Repository:
http://repo.scoap3.org

Further information about the international project SCOAP^3 can be found under:
www.scoap3.org

Information about the national project SCOAP^3-DH is provided under:
www.scoap3.de

SCOAP^3 Newsletters provided by CERN can be found under:
http://scoap3.org/news/scoap3-newsletter

Further Articles about SCOAP^3:

- http://cds.cern.ch/record/1735210/files/SCOAP3-APC.pdf
- http://www.scoap3.de/fileadmin/dateien/Dokumente/Prof._Heuer_PJ02_2012_31_PDF.pdf

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-16

COAR Case Study Controlled Vocabularies and PHAIDRA International

Imma SUBIRAT[a,1], Iryna SOLODOVNIK[a], Paolo BUDRONI[b], Raman GANGULY[c], Rastislav HUDAK[c]

[a] Food and Agriculture Organization of the United Nations, Italy; [b] Department PHAIDRA, Vienna University Library and Archive Services, Austria;[c] Vienna University Computer Center, Austria

Abstract. Following the COAR-SPARC conference in Porto, the COAR Controlled Vocabularies Interest Group met on the 16th of April 2015 and had a detailed discussion about the new set of COAR controlled vocabularies, while proposing detailed actions to solve the remaining issues (mainly in the Resource Type vocabulary). Echoing aspects discussed about sustainability and organization (long-term technical support), as well as dissemination and implementation of COAR controlled vocabularies, the immediate projection of these issues has seemed to be quite feasible on PHAIDRA International long-term ecosystem. Nowadays this OAI-PMH environment consists of fourteen Open Access repositories disseminating their contents to EUROPEANA, OpenAIRE, OAPENLibrary, e-infrastructures Austria, and national CRIS etc. PHAIDRA International is taking all necessary steps to be constantly aligned with Trusted Digital Repositories criteria and harmonized (technically and semantically) in line with COAR Roadmap: Future Directions for Repository Interoperability. The COAR Roadmap stresses that still many challenges remain with improving interoperability. These involve standardization of controlled vocabularies in use, as well as metadata and indicators, in connection with state-of-the-art interoperability approaches supporting Linked Open Data. There is an urgent need for PHAIDRA International to harmonize the encoding description (metadata properties) according to LOD-enabling strategies, and to adhere to multilingual COAR controlled vocabularies" shared registry (Knowledge Base) services. COAR Controlled Vocabularies can be easily implemented in PHAIDRA in fixed formats and mapped to related persistent RDF/SKOS versions.

Keywords. COAR, Controlled Vocabularies, Linked Open Data, LODE-BD, PHAIDRA

1. PHAIDRA International

PHAIDRA (*Permanent Hosting, Archiving and Indexing of Digital Resources and Assets*) International[2] consists of fourteen Open Archives Initiative (OAI) persistent

[1] Corresponding Author; E-mail: imma.subirats@fao.org

[2] Available at: http://phaidra.org/; http://bibliothek.univie.ac.at/english/phaidra.html (Accessed: 19 February

digital repositories in use internationally at universities in Austria, Serbia, Montenegro and Italy. PHAIDRA is involved in content aggregation, reliable storing/archiving, linking for creation and archiving of multi-resource content types with a view to long-term data preservation (LIBER, 2014).

Documentation of digital contents - uploaded and aggregated in PHAIDRA repositories is accomplished by means of metadata assignment on several levels: *single file (individual item), collection, container (multiple content datastreams)* and *paper (single file with relations to other objects)*. The metadata structure of PHAIDRA permits clear assignment of data covering contextual and provenance information (both analog and digital). The descriptive metadata structure of PHAIDRA consists of three widely-endorsed descriptive metadata standard schemes such as Metadata Object Description Schema (MODS) crosswalked to Dublin Core (DC), and Learning Object Metadata (LOM). PHAIDRA metadata includes also locally developed metadata (e.g., UWmetadata, crosswalked to DC and some other metadata solutions) implemented as minimum information needed to fulfill requirements of different designated communities (Digital Humanities Centers, OAPEN Foundation, OpenAIRE, e- Infrastructures Austria, EUROPEANA Libraries, local CRIS, institutional repositories). The overall interoperability between PHAIDRA and designated communities is in accordance with policies (based on technical, content, organization agreements) about the perpetual use of data, Open Access (OA) and specific community recommendations such as, for example, COAR Roadmap: Future Directions for Repository Interoperability (COAR, 2015) providing common meaning to the requested and exchanged services and data in order to lower possible (technical, syntactic, functional, semantic) conflicts.

2. Open Repositories towards Linked Open Data

As stated by Zeng & Chan (2015, p.5), "within the spectrum of different perspectives on interoperability, *semantic interoperability* lies at the heart of all matters", and goes "beyond those implementing the canonical search paradigm for seeking relevant information"[3]. Metadata and Knowledge Organization Systems (KOS; names and subject heading authorities, ontologies, classifications, thesauri and other controlled vocabularies) are two related areas of most interoperability efforts. Without semantic interoperability, the meaning of the language, terminology, and metadata values in use cannot be negotiated or correctly understood.

Different KOS have been already: *shared through different registries* (e.g., VEST, BARTOC.org, CKAN DataHub, European Union Open Data Portal, CIARD RING; COAR controlled vocabularies); *aligned with each other* (e.g., according to ISO 25964); *linked on the frontline of the Semantic Web (SW, Web of Data),* with the help of web-based tools (VocBench, SILK, PoolParty, Amalgame, Protégé) and other concept and (meta)data harmonizing approaches; *published as Linked Open Data*. Linked Data (LD) offers great potential to connect open repository [4] users to a

[3] 2nd Linked Open Data-enabled Recommender Systems Challenge (2015). Available at: http://sisinflab.poliba.it/events/lod-recsys-challenge-2015 (Accessed: 19 February 2016).
[4] OR 2015 10th International Conference on Open Repositories (2015). Retrieved from: http://www.or2015.net/ (Accessed: 19 February 2016).

vast array of interconnected heterogeneous sources found inside and outside of repository virtual walls. With LD every repository can provide access to linked specific datasets, without converting full metadata records. Metadata statements - rather than whole records - can be aligned and mashed up from a variety of datasets.

Where the repository is based on highly specialised formats for the exchange of its metadata and authority files, open web standards (LD included) and programming can be used to process them (LOD-LAM Project, 2015).

The COAR Roadmap stresses that Linked Open Data (LOD) and their provision have the most important interoperability function in the SW environment and, consequently, to be considered the best candidates for interoperable web services for repositories. By engaging both LOD-enabled metadata and KOS data values published as LOD trustworthy (i.e. maintained in long term by authoritative organizations) sources, repositories can augment access to the best (most relevant and most reliable) sources of the required information.

The "supporting Linked (Open) Data" aspect in the COAR Roadmap is cross-linked with other (cross-independent) aspects, such as: improving metadata quality; extending/replacing metadata exposition protocols; supporting long-term preservation and archiving; extending usage of visualization tools (along with implementing best practices for search engine optimization); handling of complex/compound/nested repository objects (e.g., Enhanced Publications); monitoring OA mandate compliance; exposing versioning information; extending the bi-directional connectivity of repositories and their services (technical issues and strategic benefit).

Information, data and knowledge modeled and processed according to LD techniques "practice at different levels of semantic precision surpass the current syntax-based possibilities in a qualitative fashion and thus can be processed, exchanged, referred and linked to different statement levels" (Isaac, Baker, 2015, p.35).

To experience how LOD-ready data (metadata properties and controlled vocabulary value data pairs) represent and support navigation of (external) related contents, one can access AGRIS (International Information System for the Agricultural Sciences and Technology) platform[5] and search for some contents. AGRIS leverages the power of descriptive LOD-enabled metadata (encoded according to LODE-BD Recommendations (Subirats, Zeng, 2015)) as well as AGROVOC LOD Thesaurus mapped (through skos:closeMatch, skos:exactMatch RDF/SKOS syntaxes) to other sixteen LOD-ready controlled vocabularies on its backbone (Subirats, Zeng, 2014), and openly shared via CKAN Data Hub[6].

2.1. PHAIDRA towards Linked Open Data

SKOS, a modern well established SW standard, can potentially support: formal alignments and hierarchical grouping of concepts using different SKOS relations (e.g. skos:exactMatch, skos:closeMatch, skos:narrower, skos:broader, skos:related); translation of concept labels; URI-based mapping to similar concepts in other KOS. (Open Metadata Handbook, 2012).

[5] Available at: http://agris.fao.org/agris-search/index.do (Accessed: 19 February 2016).

[6] Available at: http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus; http://datahub.io/dataset/agrovoc-skos (Accessed: 19 February 2016).

Controlled vocabularies implemented in PHAIDRA International can be processed and aligned with existing trustworthy (based on persistent URIs) controlled vocabularies – such as EUROVOC Thesaurus and Dewey classification encoded in RDF using SKOS, Association for Computing Machinery (ACM) classification and Art & Architecture Thesaurus in LOD version – with help of a web-based, multilingual, editing and workflow tool VocBench v.2.3 (2015). All concepts controlled vocabularies in use should be rendered/identified with persistent URIs and stored in triple store Terminology server, in order to be reused.

2.2. Linked Open Data-enabled bibliographic data

Without syntactic interoperability - underpinned primary by metadata - data and information cannot be handled properly with regard to formats, encodings, properties, values, and data types; and therefore, they can neither be merged nor exchanged. Despite the importance of syntactic interoperability, still different repository data can be difficult to find on the Web. This means, that a number of repositories are losing public influence and impact rather than providing much needed leadership to the information age. To address this dilemma, the following questions may arise: (i) is there any consistent well-structured metadata modeling and encoding methodology (in spite of concurrent proliferation of metadata standards) that can ensure maximum interoperability among digital objects on the frontline of LOD?; (ii) are there significant metadata properties related to the reuse of digital data/information/knowledge and how can these be identified and expressed?; (iii) could repository management and the user experience really benefit from bibliographic data published as LOD?

The already mentioned LODE-BD Recommendations were born to provide clear practical solutions for these cutting edge issues. LODE-BD is a reference tool providing practical support on significant metadata modeling decisions (in both depth and detail), metadata encoding (enabling data re-use) and implementation (satisfying local and/or specific needs), while insuring sharing of meaningful (with clear purport) and comprehensive (both to humans and web engines) bibliographic data. LODE-BD also clarifies what kind of KOS values should be used to produce high-quality LOD-enabled bibliographic data, easily exchangeable through open repositories and sharable on the SW.

The present use-case proposes to align the encoding of metadata properties in all PHAIDRA RDF-aware repositories according to LOD-BD strategies encouraging the reuse of different LD-enabled KOS data values in *property–value* statement pairs.

3. COAR Vocabularies. Some alignment and implementation issues

Focusing on open repository interoperability issues, COAR vocabularies (2015) - published in traditional/fixed formats and anchored to URI-RDF/SKOS with the help of the VocBench tool - represent a high recommendable set of standardized multilingual controlled vocabularies. Concept labels of these latter can be easily used to encode metadata properties, thus providing the cornerstone of effective representation of bibliographical records for different communities of users.

The COAR Resource type controlled vocabulary has largely been compared to other vocabularies (used in repositories), to address a number of alignment issues, as follows:

- *resource type:* comparison performed with DataCite metadata kernel, Elsevier CrossMark, CASRAI Dictionary for publications, CERIF Semantic Vocabulary, dcmi_terms, e-LIS, Gateway to ResearchData (GtR), PubMed, PURE, RIOXX, SWAP schemes;
- *access rights* and *versions*: comparison performed with RIOXX, NISO-ALPS-JAV schemes, EPrints access rights vocabulary;
- *grant agreement identifiers:* comparison performed with RIOXX and FundRef schemes;
- *resource identifier schemes*: comparison performed with LOC identifiers vocabulary;
- *date types and value:* comparison performed with DataCite, dcmi_terms, SWAP, Bibo-Ontology;
- *classification concepts:* LOC classification scheme vocabulary.

COAR vocabularies formats are replicating good practices of other open communities (e.g. CLARI[77]) publishing controlled vocabularies. COAR vocabularies adhere to the following principles: open vocabularies and open standards (under open licenses) are preferred over proprietary standards; formats and protocols are well-documented, verifiable, proven (being used in practice). COAR vocabularies can be implemented in repositories through *plug in* to import an RDF into the systems and integration through XML, with the name space for the vocabularies <http://purl.org/coar/> and redirection of <info:eu> URIs (Figure 1).

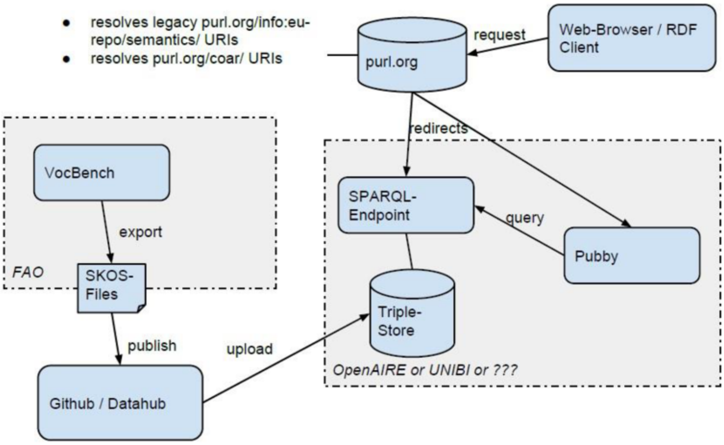


Figure 1. Data flow components for the COAR Controlled Vocabularies.

[7] Available at: https://www.clarin.eu/content/standards-and-formats; https://openskos.meertens.knaw.nl/clavas/web-report/documentation/clavas-overview.html (Accessed: 19 February 2016).

COAR vocabularies will be exported in PHAIDRA in RDF and HTML formats. Afterwards a list of all required (from COAR community) export functionalities will be investigated and a survey to identify the needs in terms of data (vocabulary) ingestion and data publication will be prepared.

Conclusions

In respect to the use-case PHAIDRA International - on which COAR controlled vocabulary community, is working on - there are some final considerations/proposals that can become actions/tasks to be undertaken: (1) all concepts of KOS implemented in PHAIDRA fixed formats should be turned to (persistent) URI-RDF/SKOS notations. Existing URI-RDF/SKOS versions of KOS should be re-used; (2) all concepts of KOS turned into URI-RDF/SKOS should be mapped/switched-off/crosswalked/cross- referenced on COAR vocabulary backbone based on LOD-ready concepts.

The implementation of COAR controlled vocabularies in PHAIDRA repositories will not only expand semantic expressivity of metadata in use, but also provide normalization and validation of metadata properties, according to standard COAR community requirements included as default for research contents, firstly at the European level. Normalization and validation of PHAIDRA (meta)data properties through COAR vocabularies will certificate PHAIDRA as a trustworthy member (Solodovnik, Budroni, 2015) of European research community, highly and reciprocally tuned with OpenAIRE/Zenodo, DANS, Dataverse Network and other national and international research infrastructures.

Last but not least, COAR vocabularies in PHAIDRA International will serve as input for language tools to manipulate metadata records at international level, as well as to improve usability of PHAIDRA multilingual user interfaces according to hierarchical levels of the COAR vocabularies.

By focusing on PHAIDRA International use case, COAR controlled vocabulary group (in cooperation with other COAR working groups) is planning to document serialization of specific formats, and fetching strategies to retrieve COAR vocabularies from web services and to import them into repository platforms.

References

COAR Vocabularies. Available at: https://www.coar-repositories.org/activities/repository-interoperability/ig-controlled-vocabularies-for-repository-assets/coar-vocabularies/. (Accessed: 19 February 2016).

COAR (2015) COAR Roadmap: Future Directions for Repository Interoperability. Available at: https://www.coar-repositories.org/activities/repository-interoperability/ (Accessed: 19 February 2016). Isaac, A., Baker, T. (2015). Linked data practice at different levels of semantic precision: The perspective of libraries, archives and museums, Bulletin of the Association for Information Science and Technology, 41(4), 34-39. Retrieved from: https://www.asist.org/publications/bulletin/apr-15/ (Accessed: 19 February 2016).

Kent State University (updated 2015). LOD-LAM Project. Retrieved from http://lod-lam.slis.kent.edu/about.html (Accessed: 19 February 2016).

LIBER (2014). LIBER Case Study: Factors for Enabling Sharing and Reuse of Research Data – Library and Archive Services at the University of Vienna. Retrieved from: http://libereurope.eu/wp- content/uploads/2014/06/LIBER-Case-Study-UNVIE1.pdf (Accessed: 19 February 2016).

Open Metadata Handbook (2012), in Wikibooks. Retrieved from: http://en.wikibooks.org/wiki/Open_Metadata_Handbook/Recommendations#Switching-accross (Accessed: 19 February 2016).

Solodovnik I., Budroni P. (2015) "Preserving digital heritage: At the crossroads of Trust and Linked Open Data ', IFLA Journal, 41 (3). Available at: http://www.ifla.org/files/assets/hq/publications/ifla-journal/ifla-journal-41-3_2015.pdf (Accessed: 19 February 2016).

Subirats, I., Zeng, M. L. (2015) "LODE-BD Recommendations 2.0: How to select appropriate encoding strategies for producing Linked Open Data (LOD)-enabled bibliographic data". Food and Agriculture Organization of United Nations. Retrieved from: http://aims.fao.org/lode/bd; http://eprints.rclis.org/22454/1/LODE-BD-2.pdf; http://aims.fao.org/lode/bd-2/step-forward#why (Accessed: 19 February 2016).

Subirats, I., Zeng, M.L. (2014). „Using KOS as the Connectors of Linked Datasets". The 77th Annual Meeting of the Association for Information Science and Technology, Seattle, WA, USA. Retrieved from http://www.asis.org/asist2014/ (Accessed: 19 February 2016).

VocBench 2.3 (2015). Available at: http://vocbench.uniroma2.it/downloads/; http://vocbench.uniroma2.it/; http://vocbench.uniroma2.it/support/; http://lists.w3.org/Archives/Public/public-esw-thes/2015Apr/0004.html; VocBench User Manual: http://vocbench.uniroma2.it/documentation/VocBench_v2.1_user_manual.pdf (Accessed: 19 February 2016).

Zeng, M. L., Chan, L.M. (2015) 'Semantic Interoperability", in Encyclopedia of Library and Information Sciences 4th ed

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-23

"If there are documents you really care about: Print them out!" (Vint Cerf, 2015)

Bernd KULAWIK[1]
Stiftung Bibliothek Werner Oechslin, Einsiedeln, Switzerland

Abstract. With theses words (only "documents" substituted for the original "photos") Vint Cerf, one of the 'fathers of the internet' and now Google Vice President, warned in 2015 that all our photos – and obviously, documents and research data, too – might disappear soon and that our century may become the "Digital Dark Age". To avoid this, Cerf is working on a solution named "digital vellum": It shall provide a platform that can preserve any documents, the software used to create and work with them, the operating system needed for this software and even an emulation of the appropriate hardware. But it may take quite some time before this platform will be available. In the meantime, the good old paper is the only medium that surely can and will survive more than 50 years – the maximum now expected for simple formats like .txt and .pdf files. Even Microfilms (also not usable without technical means) may not survive more than 200 years. But how do we print out digital documents created for and by research: short miscellanea, articles and papers, collections of them and monographs, and in recent years even facebook postings or twitter messages? We write these documents in a dedicated (text) program, sent them to the publisher, who may forward them after several transmissions forth and back with the author(s) to a layouter, again followed by some corrections requiring exchange of the file(s) … and finally they may appear in print and / or online repositories. Taking into account that all participants in the process today are (or should be) well familiar with web-based Content Management Systems and – hopefully – the concept of markup languages, it is simply astonishing that there is no system yet combining the advantages of both. Such a combination could not only serve to shorten the publishing process but also provide the ecosystem for online repositories and web-based collaboration while the results – printable documents – could be updated regularly and made available via book-on-demand and as ePublications. There may be some solutions providing such a system used by publishers "in-house", but if so, they are not available for free. The paper will propose such a system based on Free and Open Source Software with a simple *proof-of-concept*.

Keywords. LaTeX, Content Management Systems, Free Software,

Electronic publication in the humanities and other fields is still a process following the centuries-old model developed for paper:

1. Authors write papers about research results, including images and tables, using a 'word processor' producing a digital document in a proprietary format that hardly can be opened with other software without formatting or information loss.
2. The digital document is sent to the publisher.

[1] Corresponding Author: bernd.kulawik@bibliothek-oechslin.ch

3. The publisher (a person working at the publishing house and responsible for this publication) sends the electronic file to someone checking it for errors and guidelines.
4. Before the paper is regarded as 'finished', it is sent at least once back to the author for some sort of 'polishing'.
5. Steps 2 – 4 usually are repeated several times …
6. Finally, the 'final' version of the paper is sent to the layouter who transforms the file from the word processor's format into another usually also commercial and proprietary format, and reworks the entire text according to the layout guidelines.
7. The result may be sent back to the author again for final corrections in a format that he can not change.
8. After the last final 'final' reworkings, the file is transformed again into a format suitable for printing and sent to the printer.
9. The printing machine produces the (e.g.) book with binding etc. in the number of the first edition.
10. The printed books are sent to the bookstores or kept in storage until they are ordered.
11. For a few years now, this process is split up after step 7 into two: the second way is the publication in an e-book format.

Except for the usage of digital files sent around several times in several versions via e-mail or some cloud storage, all of this is still identical with the 'good ol' paper process' – and often authors and publishers repeat some of these steps by using prints on paper. So, could this really be called "electronic publishing"? My answer would be: *No!*

In 1991 Tim Berners Lee developed the World Wide Web mostly based on already existing techniques and HTML, to shorten this process. Since the year 2000 many steps of this process could be shortened and united with the help of web-based, Content Management Systems. But 25 and 15 years later, respectively, these systems with databases and versioning are still not used for (printable, well-formatted) publication(s).

The title of my paper is a slight derivation of a quotation from Vint Cerf. Cerf is the (co-) developer of the basic protocol used for the internet, the TCP/IP, and, therefore, one of the 'fathers of the internet'. Since the early 1970s he took part in crucial developments. Today he is one of Google's Vice Presidents. Therefore, we should take his warnings regarding the looming 'Digital Dark Age' of our data as well as the lack of any solution for the long-time preservation of digital at least seriously. He said these words in February 2015 at the annual meeting of the AAAS. The suggested solution he is working on is the "digital vellum": It shall provide a soft- and hardware environment able to preserve not only digital documents but also the software used for their creation and the operating systems as well as the (emulation of) the hardware needed.

Let's put aside the (crucial) questions of (commercial) licences which usually would not allow to run the software in everchanging virtual environments. And let's also put aside other questions about those digital data that are not documents in a broad sense: Most of the databases today may still be able to print out the entries in one set of data, but its form surely will not be sufficient to be regard as a (printed) scientific publication. And, of course, the pure amount of data (sets) will make it impossible to print them out after every change. Let's in addition put aside questions regarding online platforms where data, layout styles and their definitions as well as software for specific functions etc. are distributed over servers all over the world and combined *ad-hoc*. (This should cause a fundamental problem for the "digital vellum".)

So, after letting aside all of these *fundamental* questions and problems: should we follow Vint Cerf's advice anyway and print out our photographs and other documents as long as his "digital vellum" is not available yet? — Of course! Because no-one can guarantee the preservation (let alone: usability) of *any* digital data format for more than 20, let alone 50 years. Archives *plan* to make sure that simple and open formats like TXT and PDF and image formats like TIFF or JPEG will be available for *up to* 50 years, but even this is not sure. File formats like Microsoft Word's .doc are even definitely excluded by archives. Based on experiences with digital data formats from the past 30 years, I surely would doubt any possibility for such a long timespan of preservation: When I started programming in the early 1980s, the answer to the question "How should we store your data for a longest time possible?" would have been: "I put the paper punched tape into a dry and cold armoire." — So, at the moment there is no way to make sure that the digital data we create and work with will be usable even during our own lifetime! Should the creation of such data not be regarded as waste, even willful destruction of life and working time as well as resources? So: Let's print them out, at least those worth surviving the next 50 years!

But with this decision, another problem arises: The layout of our digital documents as they appear on the screen is usually not very satisfactory, often not even acceptable for scientific publishing. This is even more astonishing when we take into account that the basic tools for scientific publication, mostly word processors and type setting programs, have by now been available for more than 25 years. In addition, the tools for web-based content management able to replicate the publishing processes described above also have been available as Content Management Systems by now for at least 15 years. And both, publishing tools as well as web-based CMS are using Markup Languages.

So why is there no unification of both, one widely spread and used as a standard tool and based on Free Software? As far as I know, at the moment there is no such software system available. Some publishing houses use web-based editing tools that their authors may use, but — according to their warnings — even those tools do not generate a document that looks exactly like the one that finally will be printed. And in some of the cases versioning or commenting seems to be difficult. But these tools are 'private', closed source, 'in-house' applications, not available to others. The by far larger group of professional publishing houses does not even offer such tools: They require their authors to follow their specific guidelines, written in MS Word or PDF documents with more than one hundred pages that authors are expected to read first and fully keep in mind! Some publishers offer MS Word templates that can be filled by the authors with their own texts. Only a small number of publishers also offers LaTeX templates. But these are mostly directed to the natural sciences; to humanists and historians, LaTeX and its free interface tools usually are unknown. So, especially these authors are bound (or bind themselves) to non-free word processing software and operating systems and regularly encounter problems should they, for instance, try to reuse or re-work their own documents written some 10–15 years ago with earlier versions of this very same office software.

The solution, from my point of view, would be a combination of a free and open source Content Management System like *Plone* (based on the free Web Application Server ZOPE and its object-oriented database ZODB, both written in Python) and

LaTeX: Both, Plone and LaTeX, can be run on any operating system, and a working *proof of concept* solution already exists: It is called *ftw.book*, provided as a Plone module by the Swiss software company *4teamwork*. Because it is free, it could be extended by anyone with some experience in Python and LaTeX into a more general tool, e.g. offering different LaTeX-based designs and document classes or new layouts preferred by the publisher or institution.

What are the advantages of such an extended version of ftw.book?

- Because of the customizable user rights and role management, the entire process described above can be applied to the CMS and adapted to almost any special need of institutions or publishers.
- The CMS versioning allows to go forth and back in the editing process and keep control over the versions at any time.
- No document versions have to be sent multiple times between the participants of the publishing process, because they remain in one place, available to any authorised person.
- The available formatting functions offered in the web interface can be limited to avoid authors breaking them — a big problem in all word processing programs, causing lots of additional work.
- The final print layout is always available for controls.
- This process is protected by the CMS against unauthorised access.
- The final document can be made (un) available over the internet with a few clicks, even if a print version is not yet on or intended for the market. But any authorised reader may print a PDF copy.
- Changes can be easily done while the original is still available.
- Different editions are available at any time, so that links set to an old edition will not break because a new one has been published.
- This would allow, e.g., to update a book on an almost daily basis: When a change has been made to its content, it could not only immediately be online, but also appear in the next printed copy.

While these and other advantages regard the publishing process of scientific documents, there are even more important advantages:

- Not only publishing houses could use such a system, but any institution, group or private person. The software could be used to build up scientific repositories, e. g. for Open Access strategies.
- Because all components of the software scale very well from laptops to (groups of) servers, it would be possible to have a copy of the system run on the personal computers of members of an institution or students, e.g.: They could work on their texts even when they are offline, syncing all preserved versions later while observing the layout required by their institution or publisher.
- A simple syncing tool available for ZOPE guarantees the identity and integrity of the documents on the 'official' servers with those on the local computers or laptops.
- With the rapidly growing technical possibilities of handhelds, these should be able to run the entire software system very soon and serve the data to the

internet or synchronise them with the server(s). (Any Ubuntu-based tablet already could do this today.)

- The freedom of all components allows for constant development and adaptation of the entire system: from the underlying operating systems to the hardware. Of course, it would be useful to have large research institutions provide central repositories of the freely available parts. Those institutions could even provide hosting services for projects and, vice versa, require these projects to make their results available online via their repositories in any Open Access strategy suitable.
- Of course: not every paper, article, book etc. would have to be printed: But every one *could* be printed and, therefore, according to Vint Cerf's suggestion, be preserved even for a distant future.

So, everything seems to be wonderful with this suggested solution — or are there disadvantages? Of course, there are some: For instance, if the usability and standard conformity of the system should be preserved, this would radically restrict the many 'bells & whistles' often used in the research projects: Everything that does not fit on a (large) page would have to be excluded. Well, not completely: It would be possible, e.g., to have large images with very high resolutions, annotations, links etc. connected to the reduced images or the data in the printable version. But, of course, such high resolution images, documents or data sets surely will not survive as long as the printed counterpart or 'mother document'.

Another problem could arise from projects where data and information are intrinsically very closely linked to each other. This would make it almost impossible to represent them in a printable form. In these cases a solution could lie in the generation of reduced 'abstracts' or reduced data sets that would be imported automatically from the original database(s) into the suggested system and then be formatted for printing. Again, one would lose some data — but, depending on the decisions made regarding the exported data sets, at least part of the work and resources put into these projects could be preserved for 'eternity'.

The proposed system would not only establish a *real* environment for electronic publishing for the first time, but also provide a solution for the looming dangers of the 'Digital Dark Age' that Vint Cerf and others are warning about and archivists and librarians are or should be aware of. One could even imagine that such system could develop into a general standard for publishing *and* digital preservation. Commercial software then would have to offer plugins to allow its users to publish their texts without having to leave their 'familiar' word processor. For scientific database projects it could offer a solution in the form of repositories that would help to avoid masses of data compiled over years being lost after a short time — just because, e.g., the financial support has been turned off. In cases where server systems spread all over the world are used, e.g. som 'facebook' of science, there should at least be plug-ins to the suggested solution to create printable documents at any time. For such already or soon also very common cases, I do not even see a future in Vint Cerf's "digital vellum".

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-28

Interactive Public Peer Review^{TM}: an innovative approach to scientific quality assurance

Xenia VAN EDIG [a,1]
[a]*Copernicus Publications, Bahnhofsallee 1e, 37075 Göttingen, Germany*
xenia.van.edig@copernicus.org

Abstract. Besides providing open access to the article, Copernicus Publications provides open access to the peer review via its Interactive Public Peer Review^{TM}. In this process, a public discussion among the author, two independent referees, and interested members of the scientific community builds the core of the peer-review process.

Keywords. Peer review, open access, transparency

1. Introduction

The discussions surrounding peer review are ongoing. Several authors are claiming a crisis of peer review with regard to its length (Nguyen et al. 2015; Powell 2016) and effectiveness (Lee et al. 2013; Walker R. and Rocha da Silva, 2015), and researchers are calling for more openness in the process (Aleksic et al. 2015).

Copernicus Publications already developed a new form of peer review in 2001 (Pöschl 2012). Since then, the process has been implemented in different scientific disciplines and enhanced continuously. Today, 18 open-access journals published by Copernicus Publications apply this form of peer review. In addition, an economy journal also applies this kind of peer review.

In the following, the initial idea and the development of the process of Interactive Public Peer Review^{TM} are described.

2. Interactive Public Peer Review^{TM}

When the concept of interactive open-access publishing and Interactive Public Peer Review^{TM} was developed by Ulrich Pöschl and his fellow scientists in 2000, they faced the problem that the traditional journal publication and peer-review process were not sufficient for thorough quality assurance, constructive discussion, and integration of scientific knowledge: the majority of studies did not build on related earlier publications, and some studies were not even self-consistent even though they had been published in reputable journals with high impact factors. After long discussion, Pöschl and his colleagues were convinced that public review on the Internet would provide the

[1] Corresponding author: xenia.van.edig@copernicus.org

opportunity to resolve or at least improve many of these issues. With the Nobel Prize winner Paul Crutzen, the new concept found a prominent supporter (Pöschl 2011). Through the rapid publication after a swift access review, scientists receive a fast record of their research as a discussion paper. The process enhances transparency as referee comments, author comments, and the comments of the scientific community are published in the interactive public discussion (online and open access). However, the process meets the criteria of traditional quality insurance as papers undergo revisions and are only published as final revised papers in the journal after final acceptance by the editor. In summary, the process is designed to

- foster scientific discussion;
- maximize the effectiveness and transparency of scientific quality assurance;
- enable rapid publication of new scientific results;
- make scientific publications freely accessible.

Thus, the new process was intended to provide both rapid scientific exchange and thorough quality assurance (Pöschl 2012).

In contrast to post-publication peer review, the process of scientific quality assurance takes place prior to the formal journal publication. The discussion paper is just the manuscript submitted by the authors and therefore the starting point of the peer-review process. In addition, reviewers can disclose their names, but they do not have to do so as in open peer review.

In 2001, the first journal to apply this new peer-review process, *Atmospheric Chemistry and Physics* (*ACP*), was launched by Copernicus Publications with the support of the European Geophysical Society (EGS), which has been part of the European Geosciences Union (EGU) since 2002 (Pöschl 2011). Since 2001, 17 other journals (14 sister journals of *ACP* and 3 journals not affiliated to EGU) have adopted this innovative review process. In addition to the journals published by Copernicus Publications, the *Economic E-Journal* has also adopted this form of peer review. But how does it work?

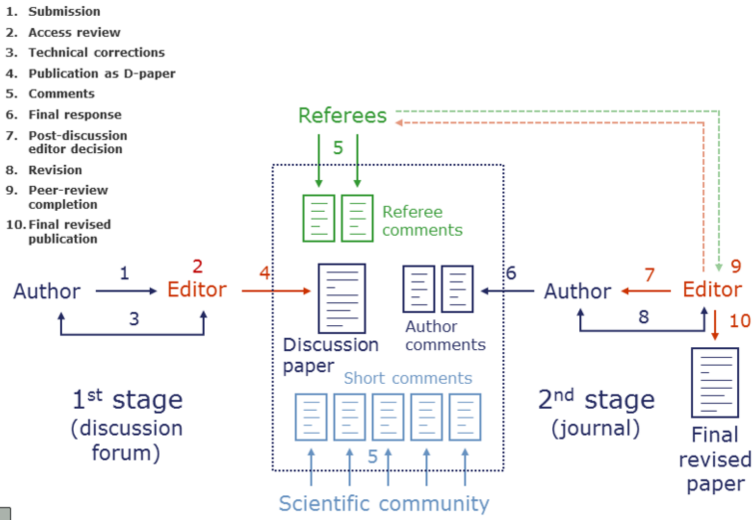


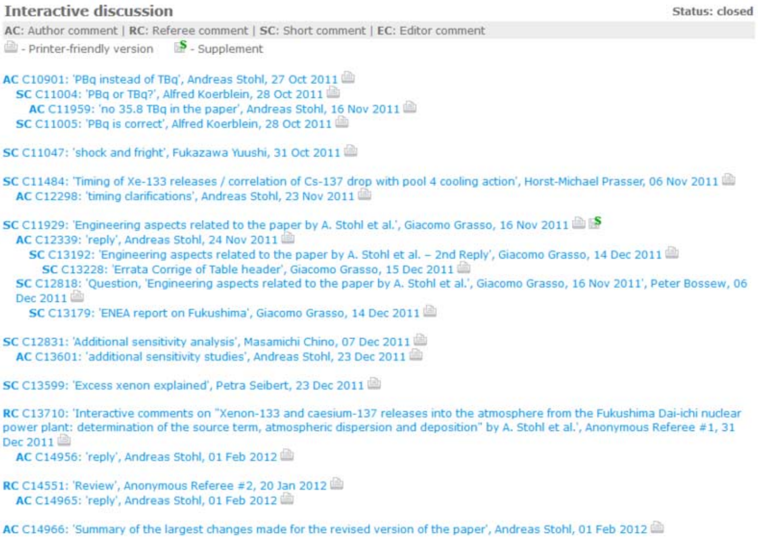
Figure 1. Example workflow of Interactive Public Peer Review^{TM}

2.1 Access review

After submission, the manuscript is swiftly reviewed by the topical editor who agreed to handle the review process. In this first assessment, the topical editor decides whether to start the discussion or not. Reasons for not starting the discussion might be a lack of basic scientific or language quality or the manuscript is not within the journal's scope. Some journals provide the possibility to request the feedback of independent referees already at this stage. However, experience shows that consulting referees at this point unnecessarily prolongs the process. In addition, referees who are not used to the process sometimes already provide full referee reports, which are not needed prior to the discussion. During this stage, only technical corrections or minor revisions can be requested.

2.2 Interactive public discussion

After a positive outcome of the access review, the author's manuscript is published as a discussion paper. At least two independent referees – who are nominated by the topical editor – review the manuscript and post their referee reports as referee comments (RCs) on an online discussion forum. This forum is openly accessible on the Internet. While the reports are open access, the referees can decide whether they want to disclose their names during the discussion or not. Research shows that about 80% of the referees decide to stay anonymous during the Interactive Public Peer Review^{TM}, while ca. 20% of them decide to disclose their name. In addition to the referees, the scientific community is invited to participate in the discussion and to post short comments (SCs). The authors of short comments have to register, and their names and contact details are shown in the discussion (Pöschl 2012). Usually, the interactive public discussion lasts 6–8 weeks depending on the journal. Before a discussion can be closed, at least the two RCs have to be published alongside the discussion paper. The author can decide to answer each comment individually or to address all comments collectively.



Interactive discussion **Status: closed**

AC: Author comment | **RC**: Referee comment | **SC**: Short comment | **EC**: Editor comment

🖶 - Printer-friendly version S - Supplement

AC C10901: 'PBq instead of TBq', Andreas Stohl, 27 Oct 2011
SC C11004: 'PBq or TBq?', Alfred Koerblein, 28 Oct 2011
AC C11959: 'no 35.8 TBq in the paper', Andreas Stohl, 16 Nov 2011
SC C11005: 'PBq is correct', Alfred Koerblein, 28 Oct 2011

SC C11047: 'shock and fright', Fukazawa Yuushi, 31 Oct 2011

SC C11484: 'Timing of Xe-133 releases / correlation of Cs-137 drop with pool 4 cooling action', Horst-Michael Prasser, 06 Nov 2011
AC C12298: 'timing clarifications', Andreas Stohl, 23 Nov 2011

SC C11929: 'Engineering aspects related to the paper by A. Stohl et al.', Giacomo Grasso, 16 Nov 2011
AC C12339: 'reply', Andreas Stohl, 24 Nov 2011
SC C13192: 'Engineering aspects related to the paper by A. Stohl et al. – 2nd Reply', Giacomo Grasso, 14 Dec 2011
SC C13228: 'Errata Corrige of Table header', Giacomo Grasso, 15 Dec 2011
SC C12818: 'Question, 'Engineering aspects related to the paper by A. Stohl et al.', Giacomo Grasso, 16 Nov 2011', Peter Bossew, 06 Dec 2011
SC C13179: 'ENEA report on Fukushima', Giacomo Grasso, 14 Dec 2011

SC C12831: 'Additional sensitivity analysis', Masamichi Chino, 07 Dec 2011
AC C13601: 'additional sensitivity studies', Andreas Stohl, 23 Dec 2011

SC C13599: 'Excess xenon explained', Petra Seibert, 23 Dec 2011

RC C13710: 'Interactive comments on "Xenon-133 and caesium-137 releases into the atmosphere from the Fukushima Dai-ichi nuclear power plant: determination of the source term, atmospheric dispersion and deposition" by A. Stohl et al.', Anonymous Referee #1, 31 Dec 2011
AC C14956: 'reply', Andreas Stohl, 01 Feb 2012

RC C14551: 'Review', Anonymous Referee #2, 20 Jan 2012
AC C14965: 'reply', Andreas Stohl, 01 Feb 2012

AC C14966: 'Summary of the largest changes made for the revised version of the paper', Andreas Stohl, 01 Feb 2012

Figure 2. Example of an interactive public discussion

To guarantee the author's publication precedence and to provide a lasting record of the review process, every discussion paper and its comments remain online and are individually citable (Pöschl 2012). This occurs regardless of whether or not a manuscript is accepted for publication as a final revised paper in the journal.

2.3 Final response and peer-review completion

After the discussion has ended, the author should address all comments in a final response, if he or she did not do so during the open discussion. During this stage also the editor has the opportunity to post comments and suggestions (Pöschl 2012). Formal editorial recommendations and decisions shall be made only after the authors have had an opportunity to respond to all comments, or if they request editorial advice before responding.

Depending on the journal, the next step is one of the following:

- The authors submit their revised manuscript. In this case, the topical editor – in view of the access peer review and interactive public discussion – either directly accepts/rejects the revised manuscript for publication in the journal or consults referees in the same way as during the completion of the traditional peer-review process. If necessary, additional revisions may be requested during peer-review completion until a final decision about acceptance/rejection for the journal is reached (Atmospheric Chemistry and Physics, website, 2016).
- The topical editor makes a post-discussion decision in which he or she, based on the responses, either invites the authors to submit a revised manuscript or directly rejects the manuscript. If necessary, he or she may also consult referees in the same way as during the completion of the traditional peer- review process (Biogeosciences, website, 2016).

2.4 Publication of final revised paper

In the case of acceptance, the final revised paper is typeset and proofread. Then it is published on the journal's website, and the preceding discussion paper and the interactive discussion are displayed in a "peer-review tab" alongside the article. In addition, many journals display all referee and associate editor reports, the authors' response, as well as the different manuscript versions of the peer-review completion. All publications (original paper, interactive comments, and final revised paper) are permanently archived and remain accessible to the public via the Internet, and final revised papers are also available as print copies. The articles are also distributed via various abstracting and indexing services as well as other databases worldwide.

2.5 Interactive Public Peer Review^{TM} in various disciplines

This model is mainly utilized in the geosciences. However, it is also applied to other disciplines such as drinking water engineering and wind energy science.

In the table below all journals published by Copernicus Publications that apply the Interactive Public Peer Review^{TM} are listed. One can see that it is applied in various subdisciplines within the geosciences ranging from geophysics to atmospheric sciences:

Table 1. Journals applying the Interactive Public Peer Review^{TM} published by Copernicus Publications

Title	Access review with referee quick reports	Post-discussion editor decision
Atmospheric Chemistry and Physics (ACP)	yes	no
Atmospheric Measurement Techniques (AMT)	yes	no
Biogeosciences (BG)	yes	yes
Climate of the Past (CP)	no	yes
Drinking Water Engineering and Science (DWES)	no	yes
Earth Surface Dynamics (ESurf)	no	no
Earth System Dynamics (ESD)	no	no
Earth System Science Data (ESSD)	no	no
Geoscientific Instrumentation, Methods and Data Systems (GI)	yes	no
Geoscientific Model Development (GMD)	no	no
Hydrology and Earth System Sciences (HESS)	no	yes
Natural Hazards and Earth System Sciences (NHESS)	no	yes
Nonlinear Processes in Geophysics (NPG)	no	no
Ocean Science (OS)	yes	no
SOIL (SOIL)	no	yes
Solid Earth (SE)	no	no
The Cryosphere (TC)	no	no
Wind Energy Science (WES)	no	yes

In the past years, a range of journals have switched from the traditional peer-review model to public peer review.

In many cases, the interactive public discussion only consists of two referee comments and the author’s reply. However, providing the scientific community withthe opportunity to contribute to the discussion is a crucial aspect. Other papers on “hot topics” such as climate change or radioactivity sometimes receive 30–100 comments. There is an overview of these “most commented papers” in each journal's online library.

Prior to the discussion (i.e. during the access the review) rejection rates vary among journals and are found to be 8–37%. After the discussion, which aims to improve the quality of the manuscripts, the rejection rate is only 8% on average.

3. Recent developments

After the implementation of the accelerated access review (i.e. the access review without the possibility of consulting referees), the launch of the post-discussion editor decision (more guidance for authors after the discussion), and the adoption of the post-discussion report publication by most journals (i.e. disclosure of all reports from peer-review completion after final acceptance), a major adjustment to the concept was the decision to no longer typeset discussion papers from 2016 onwards and to merge the libraries of discussion papers and final revised papers.

Thus, discussion papers now look less like a publication and more like pre-print papers. The discussion paper is the PDF uploaded by the author, with an added header indicating the journal to which the manuscript was submitted for review. The manuscript is still citable, but the citation will indicate that the paper is under review (Copernicus Publications 2015). The discussion papers do not receive a subsequent pagination anymore, but a DOI is still registered for them.

In order to emphasize that the discussion paper is only the first step to the final paper, discussion papers are no longer archived in volumes and issues in a separate online library. With the new concept, a final revised paper and its corresponding discussion paper are archived together. There is a main page that includes all the information relating to the paper in separate tabs, such as metrics, related articles, and the list of peer-review comments and the discussion paper (Copernicus Publications, 2015)

These actions should address two main obstacles which occurred in the past: on the one hand, it should prevent authors from citing the discussion paper instead of the final paper; on the other hand, it should help authors whose discussion papers were rejected by indicating more clearly that discussion papers are not to be regarded as formal publications and thus can be submitted to other journals.

With the new concept for our interactive journals, we also introduced a new payment concept. Before 2016, authors were obliged to pay solely for the publication of their discussion paper if the respective journal had APCs. This concept is now obsolete since we no longer provide formatting services for discussion papers. Furthermore, funders welcome the new payment concept since now they are paying for the final revised paper and hence the version of record.

References

Aleksic J, Alexa A, Attwood TK et al. (2015) An Open Science Peer Review Oath [version 2; referees: 4 approved, 1 approved with reservations] *F1000Research*, 3:271, doi: 10.12688/f1000research.5686.2

Atmospheric Chemistry and Physics (2016) *http://www.atmospheric-chemistry-and-physics.net/peer_review/interactive_review_process.html* (accessed 07.01.2016)

Biogeosciences (2016) *http://www.biogeosciences.net/peer_review/interactive_review_process.html* (accessed 07.01.2016)

Copernicus Publications (2015) New concept for interactive journals, news item of 14 October 2015, http://www.copernicus.org/news_and_press/2015-10-14_new-concept-for-interactive-journals.html (accessed 07.01.2016)

Lee, C. J., Sugimoto, C. R., Zhang, G. and Cronin, B. (2013) Bias in peer review. *J. Am. Soc. Inf. Sci.*, 64, pp. 2–17, doi: 10.1002/asi.22784

Nguyen V.M., Haddaway N.R., Gutowsky L.F.G., Wilson A.D.M., Gallagher A.J., Donaldson M.R., et al. (2015) How Long Is Too Long in Contemporary Peer Review? Perspectives from Authors Publishing in Conservation Biology Journals. *PLoS ONE* 10(8), e0132557, doi:10.1371/journal.pone.0132557

Powell, K (2016) The Waiting Game, *Nature,* 530, pp. 148–151, doi:10.1038/530148a

Pöschl U. (2012) Multi-stage open peer review: scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation, *Front. Comput. Neurosci*, 6:33, doi: 10.3389/fncom.2012

Pöschl U. (2011) On the origin and development of Interactive Open Access Publishing, *A short History of Interactive Open Access Publishing*, Copernicus Publications, Göttingen

Walker R. and Rocha da Silva P. (2015) Emerging trends in peer review—a survey. *Front. Neurosci.* **9**:169, doi: 10.3389/fnins.2015.00169,

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-34

Renegotiating Open-Access-Licences for Scientific Films

Elke BREHM[1]
Technische Informationsbibliothek (TIB) // German National Library of Science and Technology

Abstract. Scientific publishing is not limited to text any more, but more and more extends also to digital audio-visual media. Thus services for publishing these media in portals designed for scientific content, oriented towards the demands of scientists and which comply with the requirements of Open Access Licenses must be provided. Among others, it is the goal of the Competence Centre for Non-textual-materials of TIB to collect, archive and provide access to scientific audio-visual media in the TIB AV-Portal under the best possible (open) conditions. This applies to older films, as for example the film collection of the former IWF Knowledge and Media gGmbH i. L. (IWF) and to new films. However, even if the acquisition of the necessary rights for audio-visual media is complex, the renegotiation of Open-Access-Licenses for older films is very successful. This paper focuses on the role of Open Access in the licensing strategy of TIB regarding scientific films, the respective experience of TIB and the presentation in the AV-Portal, but also touches upon prerequisites and procedures for the use of Orphan Works.

Keywords. Open Access, Scientific Films, IWF Knowledge and Media, Licensing

1. Introduction

Scientific publishing is not limited to text anymore but more and more comprises other objects such as research data, 3D objects and digital audiovisual material among others. Thus opportunities are needed to publish different types of scientific content in an environment specifically designed for the object type and adapted to the needs of scientists. Among others, TIB collects, archives and provides access to scientific audiovisual media in its TIB AV-Portal[2]. TIB puts a special emphasis on Open Access and strives to offer its content under free and open conditions as close to the definition of Open Access contained in the Berlin Declaration of 2003[3] as possible. This is valid for older films such as the film collection of the former IWF Knowledge and Media as well as for the new scientific films which TIB acquires continuously.

When libraries acquire analogue film copies, the legal situation is relatively clear because the possibilities of utilisation are already defined in the German Copyright Act and copying constitutes a practical hurdle for analogue film formats. The legal situation for digital films and films posted on the internet, however, is more complex: The permits for analogue material regulated in the German Copyright Act cannot simply be

[1] Corresponding Author: Elke.Brehm@tib.eu .
[2] TIB AV-Portal. Available at https://av.tib.eu (Accessed: 4 March 2016).
[3] Berlin Declaration on Open Access to Knowledge in Sciences and Humanities.

transferred to digital material. In order to be able to make a film available online, all of the permits required must be agreed upon in license agreements. In contrast, creating copies of digital copyrighted material is child's play. There is a great danger that films will be passed on as digital copies (against the creator's will) because the process is such a simple one. In particular with regard to the publication of films on the internet, not only copyright has to be observed, but also publicity and personality rights of the filmed or vocally recorded individuals. This factor can usually be neglected in the case of texts.

At the end of 2012, TIB was entrusted with taking on the scientific film collection of the former IWF Knowledge and Media and, ideally, making it available to the public via its online portals. The collection comprises around 11,500 analogue and digital scientific films related to various subjects. The collection revolves mainly around technical and scientific subjects, as well as biology and ethnology. Although most of the publications were created between the 1950s and 1980s, the collection also contains a number of earlier cinematographic works. Unfortunately not all films are available in digital form. In addition to a lack of funding for digitizing such materials, TIB faces the problem of having to clarify how the films may be used. The rights in the IWF Knowledge and Media Film collection are very heterogeneous. After all, particularly when films are rather old, the agreements with the creators involved in the film production were also concluded at a time before digital use options on the internet became known and customary. In addition, many necessary changes were made in copyright law to adapt it to modern needs in the digital age. These options are therefore not included in the agreements concluded at the time with the creators. Moreover, usually many different people are involved in many different ways in the creation of a film – people who may potentially have rights to the film. Scriptwriter, director, cameraman, cutter, producer, performing artists, narrator … Often it is not easy, or no longer possible in retrospect, to establish who was actually involved in the production of the film. When attempts are made to renegotiate rights, many film authors cannot be traced.

Thus, before being able to make the films available to the public via the TIB AV-Portal or through other services of TIB, it is obligatory that the legal situation regarding each film is examined and each author or rightsholder is contacted individually. As far as necessary and feasible, TIB renegotiates the necessary rights to be able to offer an up-to-date service to the customers. The renegotiation of rights among other things focuses on the rights needed to offer access to the films via the TIB AV-Portal, to create derivates in alternative film formats and perform the video analysis in the TIB AV-Portal. It has become customary to make bibliographic metadata freely available under the Creative-Commons-License CC0 1.0 Universal Public Domain to spread information and knowledge about the collection. When contacting each author individually, TIB offers the authors and rightsholders the possibility to release the films under an open access-license of the Non-Profit-Organisation Creative Commons[4] in the TIB AV-Portal and allow for a greater and facilitated distribution of the films in the public.

[4] Creative Commons. Available at: http://creativecommons.org/ (Accessed: 4 March 2016).

Challenges for the renegotiation of the rights are manifold:

- The effort undertaken needs to be in balance with the results that can be achieved.
- In many cases it is impossible to trace the original film authors.
- Generally, the Creative Commons Licenses are unknown to this particular group of scientists.
- Many films contain material, to which the rights are not held by the original film authors (copyright of third parties, personality and publicity rights, etc.).
- Film authors expect a financial compensation for the use of the films.
- A lot of films are only available in analogue formats.

In spite of the challenges, the renegotiation of Creative Commons Licenses for many of the films is very successful: 59 % of all film authors whose films can be made available online, decide to release the films under the Creative Commons License. In the process, films for which the copyright protection has expired and orphan works are identified and made available online and through other services of TIB. In the poster, more detailed information is given about the background, process and success rates.

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-37

ROAD: the Directory of Open Access Scholarly Resources to Promote Open Access Worldwide

Nathalie CORNIC[a,1]

[a] *Data, Network & Standards Department, ISSN International Centre*

Abstract. ROAD, the Directory of Open Access scholarly Resources, is a new service implemented by the ISSN International Centre. ROAD provides a free access to a selection of worldwide, multidisciplinary scholarly resources in open access that have been identified by the ISSN Network. This paper will present ROAD background, its innovative concept and how it is positioned in the open access ecosystem.

Keywords. open access, scholarly resources, serial publications, ISSN

1. Introduction

ROAD [2], the Directory of Open Access scholarly Resources, gathers all serial publications that can be identified by an ISSN such as journals, conference proceedings, monographic series, academic repositories and blogs. Launched in December 2013 by the ISSN International Centre[3] and supported by the Communication and Information Sector of UNESCO[4], ROAD provides a free access to a subset of the ISSN Register[5]. This subset gathers nearly 14,000 bibliographic records describing and pointing to worldwide, multidisciplinary scholarly resources in open access that have been identified by the ISSN Network[6]. ROAD innovative concept is that ISSN bibliographic records are enhanced with external information aggregated from data sources like indexing and abstracting services, metrics and registries. These OA resources are thus selected for their compliance with the core components of the open access spectrum guide[7].

[1] Corresponding Author, Data, Network & Standards Department, ISSN International Centre, 45 rue de Turbigo, 75003 Paris, France; E-mail: nathalie.cornic@issn.org

[2] http://road.issn.org/en

[3] http://www.issn.org/

[4] http://www.unesco.org/new/en/communication-and-information/

[5] The ISSN international Register gathers 1,9 million bibliographic records available on subscription: http://www.issn.org/understanding-the-issn/the-issn-international-register/

[6] The ISSN Network comprises 89 national centres worldwide: http://www.issn.org/the-centre-and- the-network/our-organization/le-reseau-issn-en/

[7] In terms of readership, reuse, copyright, author and automatic posting, and machine readability, according to the guide How Open Is It? https://www.plos.org/open-access/howopenisit/

2. Project background

A significant growth of open access scholarly resources have been observed over the previous years.

2.1 The needs of the scholarly community

Scholars and librarians have increasing difficulties to select reliable, open access resources for their research in this ocean of publications. The ISSN standard (ISO 3297 : 2007) mainly used to identify serial resources in an unambiguous way and like ISBN[8], it does not guarantee any scientific and editorial quality. However, a key role remains to be played for the ISSN Network. Facing the emergence of questionable publishing practices, ISSN national centres receive recurring questions about the reliability of OA journals, and there is some confusion around ISSN being mistakenly associated to a quality label by some young researchers (Pelegrin 2014). There is also a need for scholars:

- to find out how many journals are indexed or ranked in their country,
- which OA scholarly resources are available in their discipline in a specific language.

And of course, scholars need to be guided to select the trusted journals they should submit their research to.

2.2 The gaps to be filled

In 2013, despite the existence of many complementary (Stenson 2012) directories like the DOAJ[9], the DOAB[10], OpenDOAR[11], there was no global system referencing both, on the one hand, the resources published through the green and gold open access models, and on the other hand, the digitized versions of dead print scholarly journals made available online for free by institutions. As a matter of fact, the scope of these existing directories is focused on a certain type of resources, and consequently, of disciplines, STM being predominant in journal publishing, whereas humanities and social sciences are monograph-centred. Some other directories like AJOL[12] and Latindex[13] have a specific specific geographical coverage.

Referring to scholarly communication evaluation, in 2013, there was no comprehensive, global and multidisciplinary directory enabling to know in which indexes or databases a given publication appears in. Conventional indicators of reputation are traditionally very narrowly defined, and built mainly around one scholar activity. And the numerous emerging reputation platforms like ResearchGate or Kudos are still in their infancy, and none cover the whole gamut of activities (Nickolas 2015).

[8] International Standard Book Number: https://www.isbn-international.org/
[9] The Directory of Open Access Journals: https://doaj.org/
[10] The Directory of Open Access Books: http://www.doabooks.org/
[11] The Directory of Open Access Repositories: http://www.opendoar.org/
[12] African Journals Online: http://www.ajol.info/
[13] Latindex: http://www.latindex.org/

Existing tools (directories, abstracting and indexing services, metrics) may lack accuracy in the bibliographic description, relying on the metadata provided by the publishers.

Finally, UNESCO needed some global, worldwide statistics about open access scholarly communication.

These are the reasons why ROAD was conceived.

3. ROAD's main innovative features

3.1 Identifying scholarly resources is a constant priority for the ISSN system

The ISSN Network, along its 40 years of expertise in identifying serials and maintaining the authoritative database for serials[14], has gathered 1,9 million records, and among them, 160,000 online publications. Identifying scholarly resources has been a constant priority, as the 89 ISSN centres are hosted by National libraries and National Centres for Scientific Information and Technology. Thus, a strong relationship exists between the ISSN Network and the scientific community, through a widespread use of the ISSN. As seen on **Table 1** and **Figure 1** below, the cooperation of the ISSN national centres is essential to make ROAD a unique service based on a continuous identification of 5 types of multidisciplinary resources. Not only content criteria for inclusion in ROAD are strict[15], but also, in a second phase of acceptance, ISSN IC performs some quality control on the metadata in terms of bibliographic accuracy, updates and relevancy to meet the inclusion criteria.

Table 1. Top 10 participating countries in ROAD

Countries	No of resources
India	1400
Brazil	1126
United Kingdom	817
United States	793
France Poland	551
Germany	478
Russian Federation	466
Iran	448
Spain	448
	423

Source: ISSN Register

[14] The ISSN International Register: http://www.issn.org/understanding-the-issn/the-issn-international-register/

[15] Criteria applied: open access to the whole content of the resource, no moving wall, the resources comprise research papers addressed to the scholarly community.

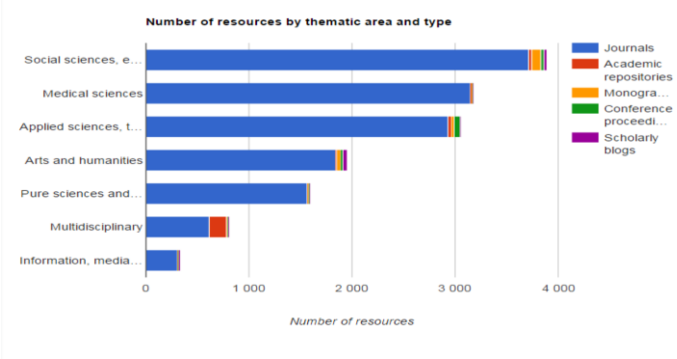


Figure 1. Number of resources in ROAD by thematic area and type as of December 2015.

3.2 *Partnering databases*

Another strength of ROAD is that ISSN bibliographic records are enriched by external data sources[16] through long-standing ISSN partnerships like with Edina[17] and Latindex, or through new collaborations with DOAJ and Scopus for instance. Indexing and abstracting services, evaluation services and registries complete the bibliographic metadata provided by ISSN centres, adding publishing information not traditionally collected by libraries, like business models, article processing charges, re-use licences and metrics about the visibility and the quality. All these external databases aggregate peer reviewed content harvested by the ROAD system. Partnering databases are chosen for their specialized scopes, the learned societies they are supported by, and for their positive, white list approach.

3.3 *Statistics*

On UNESCO's request, ROAD provides statistics computed through automated searches against the database. They cover all types of OA resources, showing the evolution in the number of publications per type, geographic and thematic area, and countries, as well as the coverage of the resources by indexing and abstracting services.

[16] data sources so far have been associated to ROAD: see http://road.issn.org/ Data Sources

[17] Edina maintains The Keepers Registry: http://thekeepers.org/registry.asp

[18] The Global Open Access Portal (GOAP), http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/

4. Current status and future steps

The ISSN International Centre is well positioned to create a global, comprehensive, multidisciplinary portal of open access continuing resources. The support of UNESCO was crucial, ROAD being complementary with, and accessible through the Global Open Access Portal.[18]

ROAD development strategy is oriented on diversifying the thematic content and balancing the non-journals resources through ISSN assignment campaigns among the network and through implementing a data quality plan, to strive to become the authoritative, global database about open access resources.

References

Pelegrin, F.X., *ROAD: A New Free Service for Identifying and Selecting OA Scholarly Resources, 2014, Charleston Conference*

Stenson, L., (2012). Why all these directories? An introduction to DOAJ and DOAB. *Insights*. 25(3), pp.251–256. DOI: http://doi.org/10.1629/2048-7754.25.3.251

Nicholas, D., Herman, E., Jamali, H., Rodríguez-bravo, b., Boukacem-Zeghmouri, c., dobrowolski, t. And pouchot, s. (2015), New ways of building, showcasing, and measuring scholarly reputation. Learned Publishing, 28: 327. http://ciber- research.eu/Dave_Nicholas_Publications.html

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-42

Sustaining the growth of library scholarly publishing

Graham STONE[a,1]
[a] *University of Huddersfield*

Abstract. In 2012, the University of Huddersfield Press presented a paper at the 16th International Conference on Electronic Publishing on its new open access journals platform. At the time, the Press was one of the only New University Presses (NUP) in the UK and one of the first to publish open access journals, open access monographs and sound recordings. This paper will develop Hahn's programme and publication level business plan and relate this to the sustainability of the Press. It will demonstrate how the Press has been able to show value to the University in order to secure funding. The paper will conclude with a discussion around the need for collaboration between library led NUPs.

Keywords. Library, publishing, university press, open access, business models

1. Introduction

The University of Huddersfield Press was re-launched in 2010 as a library led publishing initiative with decisions taken by an academic led Editorial Board. In 2012, the Press presented a paper at the 16th International Conference on Electronic Publishing on its new open access journals platform (Stone, 2011). At the time, the Press was one of the only New University Presses (NUP) in the UK and one of the first to publish open access journals, open access monographs and sound recordings. Since then it has published seven journals, ten scholarly monographs and nine music recordings. The library as publisher or library scholarly publishing is now a growing worldwide movement (Simser, Stockham & Turtle, 2015) and Huddersfield has followed the lead from NUPs in the United States and Australia (Lynch, 2010).

This paper develops Hahn's (2008) programme and publication level business plan and relates this to the sustainability of the Press. It demonstrates how the Press has been able to show value to the University in order to secure future funding. It concludes with a discussion about the need for collaboration between library led NUPs.

2. Business models

Business model development for NUPs is an area that needs significant work (Hahn, 2008; Withey et al., 2011). Issues with open access business models have also been discussed an refined for much of the last decade (Thatcher, 2007). However, they are still based on the principles of rigorous peer review and close engagement with faculty

[1] Corresponding Author: G.Stone@hud.ac.uk

and strategic leadership through an advisory board with representatives from all faculties (Missingham & Kanellopoulos, 2014). Like many NUPs, the University of Huddersfield Press developed without a clear business model in its early years. This has issues for sustainability and scalability.

3. Sustainability

In 2012, a report to SPARC found that only 15% of libraries surveyed had a documented sustainability plan (Mullins et al., 2012). Hahn (2008) found that very few library publishers were able to '…support even 10 journal titles or more than a handful of monographic works' (p.25). Thus, library presses hesitate in more aggressive marketing due to fears that this could generate more demand than could be satisfied. This leads to the question of scalability. If a library publisher wishes to expand, it has to identify the resources needed and this is a long-term commitment. This could result in resources being diverted from other areas (Xia, 2009). A more successful press will create a need to reallocate greater staffing resources unless new resources are identified.

4. Programme level planning

Regarding the development of the business model, Hahn (2008) suggests two levels of business plans for library publishers:

- Programme level planning
- Publication level planning

A NUP operating without a business model at the programme level is effectively operating at a publication level. Moving from one publication to the next without a clear plan. Staffing and funding challenges need to be resolved at a programme level for the library as publisher to be sustainable. In addition, planning is needed at both programme level and publication level in order for the initiative to become a success.

4.1. Scalability of library publishing services

NUPs offer a truncated list of services when compared to traditional publishers (Hahn, 2008). However, this represents a leaner version of traditional 'legacy' publishers. Once presses begin to grow there is a question of scalability and sustainability and this is what programme level planning provides. This is the case for the University of Huddersfield Press, and is echoed by comments made by other library presses (Mullins et al., 2012). There is a fear that greater demand could lead to the press becoming a victim of its own success.

4.2. Staffing

The issue of staffing and the resulting effect of increased success verses a limited staff base have been the focus of discussion for many successful presses as over time this inhibits growth (Perry et al., 2011). The SPARC study found the number of staff allocated to publishing activities ranged between 0.9-2.4 FTE, with staff dedicated to library publishing programmes described as relatively rare (Mullins et al., 2012).

4.3. Business models and funding

As part of the UK Crossick report (London Economics, 2015), theoretical tests established that each open access business model has its own strengths, weaknesses, opportunities and threats. This was further developed as part of the OAPEN-UK project (Beech & Milloy, 2015). The predominant business model for the University of Huddersfield Press is the institutional subsidy model where the Press receives subsidies from the University, either centrally, from faculty or the library, or from a funder.

5. Publication level planning

In order for publication level planning to work, programme level planning needs to be in place. For example, planning at the programme level leads towards a business plan. This plan can outline the case for growth of the press. The plan at Huddersfield suggests a more robust funding allocation and modest increase in staffing. This in turn supports a greater number of publications and improved publication level planning. An annual plan, which includes a budget, key dates and an evaluation process, could then be produced. An example of this at Huddersfield is *Fields: journal of Huddersfield student research* (Stone, Jensen & Beech, 2016). The Press worked with the University's Teaching and Learning Institute to ring-fence funding for the publication. Publication level planning helps to address issues that have arisen in the process. The journal is now entering its third year of publication and lessons learned from volume 1 have led to a revision in the notes for contributors, a writing retreat for authors, conference attendance for student authors and marketing around campus.

6. Cash flow and profit and loss forecast

At Huddersfield, a paper on staffing was taken to the Press Board in 2015. As a result annual staffing costs for the Press of around £40K have been absorbed by Computing and Library Services (CLS) as part of the staffing budget. Institutional repository costs (the publication platform for the Press) are also covered by CLS. As part of the Press business plan, the following costs were identified in order to grow the Press at a sustainable level.

- DOI costs for seven existing journals, with a growth rate of an extra two journals per year
- Set-up costs for the additional journals
- Two monographs to be published in 2016, three in 2017, four in 2018 and five in 2019
- Recurrent costs including appropriate memberships and marketing

Sales forecasts were also included for print copies of monographs, although these are not guaranteed. Income from print sales would enable the Press to publish additional titles to those highlighted above. This model also allows the Press to run a fee waiver model for peer reviewed monographs and journals from Huddersfield authors as this would be underwritten by programme level funding. This model is also being adopted by other NUPs in the UK such as UCL (2015) and the recently launched White Rose University Press (2016).

7. Value and impact

NUPs are not for-profit enterprises, they are an exercise in scholarly communication. In order to attract programme level funding and to justify a local subsidy the press must demonstrate its value to the university rather than monetarize the work of the press. An example of how to do this is to show that financial returns, which do not come back to the Press directly, have the potential to earn research income for the university. At Huddersfield this has been done by showing how the Press can contribute to 'quality-related research funding' (QR funding) from the 2014 Research Excellence Framework (REF) (HEFCE, 2015).

As part of the 2014 REF, the University submitted 100 research outputs from its staff to the music Unit of Assessment (REF, 2014). University Press publications were included in eleven of these outputs (some of these were as part of portfolio outputs). While REF scores cannot be associated with individual outputs, 85% of music research was judged to be internationally excellent (3* and 4*), which attracts QR funding. The assumption here is that at least some of the Press output was ranked in these categories. In addition, Press output also contributed to the wider impact and environment statements, which were also ranked highly.

If all 100 outputs were treated equally, then six outputs from the Press (three books and three CDs) have contributed to 11% of the University's QR funding for music in 2016. This is a not inconsiderable sum, indeed far more than the overheads of the Press for all publications forecast in the Press's four year plan.

In February 2016 a discussion paper was tabled at the University of Huddersfield Press Editorial Board. It invited the Board to discuss the four-year plan, which outlined the funding required at programme level in order for the Press to become sustainable and scalable. This included a detailed cash flow and profit and loss forecast and evidence of the value and impact of the Press on QR funding in the University. It was suggested that there was potential for this to have impact on other disciplines such as history, politics and English, which rely on monograph publishing as the gold standard. The Board approved the proposal for a programme level funding model in principle. As a result the Press has now had funding confirmed for the 2016/17 and 2017/18 academic years. This is in addition to staffing costs and will allow the Press to finance additional monographs and journals described above as part of a programme level plan. The Press will also be able to offer a fee waiver to researchers at Huddersfield who submit proposals for new titles subject to satisfying the Press's peer review process.

8. Collaboration

In addition to programme and publication level planning, NUPs also need to collaborate to achieve scalability. The 2012 report to SPARC recommends that collaborations should be used to, "…leverage resources within campuses, across institutions, and between university presses, scholarly societies, and other partners"

(Mullins et al., 2008, p.19). This paper suggests that the follow areas of collaboration are required.

- **Landscape survey**. In the UK there is uncertainty as to how many library led open access university presses are operating. Huddersfield, White Rose and UCL presses have all been mentioned in this paper. However, there are others emerging in both the UK and the rest of Europe. A data gathering exercise is required in order to assess the current state of play regarding NUPs and library publishing ventures in Europe

- **A Library Publishing Coalition for Europe**. This paper suggests that NUPs in Europe establish a European Library Publishing Coalition (LPC). This would be based upon the LPC in the United States (Educopia Institute, 2013) and could become a hub for best practice and innovative approaches

- **Best practice/efficiencies in the workflows**. The Landscape study will give intelligence on where the new and proposed library presses are, a LPC would help to establish a community. It is hoped that this will lead to further collaboration and therefore sustainability for NUPs. It is suggested that a series of best practice guidelines could be developed providing useful tools for NUPs. For example, licences, workflows, business models and recommendations for appropriate membership, e.g. COPE, OASPA, DOAJ and DOAB. Best practice around establishing value and impact would also allow these NUPs to flourish in the future.

9. Conclusion

This paper has shown how the University of Huddersfield Press has used evidence of value and impact based on REF output to secure funding for the next two years. An understanding of Hahn’s programme and publication level business plan has allowed the Press to achieve sustainability going forward. This will allow the scaling up of publications with a view to the next REF in the UK. The next steps for the Press are to produce a plan for the next two years in order to secure further funding going forward. In addition, the Press needs to work alongside other NUPs in order to establish best practice for library-led open access publishing.

References

Beech, D. and Milloy, C. (2015) Business models for open access monographs: a summary document. Available at http://oapen-uk.jiscebooks.org/files/2015/04/Annex-A-OAPEN-UK-Business-Models-SWOT-Workshop-preparatory-document.docx (Accessed: 7 March 2016).

Educopia Institute (2013) Library Publishing Coalition. Available at http://librarypublishing.org/ (Accessed: 7 March 2016).

Hahn, K.L. (2008) Research library publishing services: new options for university publishing, Washington DC: Association of Research Libraries. Available at http://www.arl.org/storage/documents/publications/research-library-publishing- services-mar08.pdf (Accessed: 7 March 2016).

HEFCE (2015) How we fund research. Available at http://www.hefce.ac.uk/rsrch/funding/mainstream/ (Accessed: 7 March 2016). London Economics (2015) Economic analysis of business models for open access monographs: Annex 4 to the Report of the HEFCE Monographs and Open Access Project. London: HEFCE. Availableat http://www.hefce.ac.uk/media/hefce/content/pubs/indirreports/2015/Monographs,and,o pen,access/2014_monographs4.pdf (Accessed: 7 March 2016).

Lynch, C. (2010) 'Imagining a university press system to support scholarship in the digital age', Journal of electronic publishing, 13(2). http://dx.doi.org/10.3998/3336451.0013.207

Missingham, R. and Kanellopoulos, L. (2014) 'University presses in libraries: Potential for successful marriages', OCLC Systems & Services: International Digital Library Perspectives, 30(3), 158-166. http://dx.doi.org/10.1108/OCLC-01-2014-0001

Mullins, J.L. et al. (2012) Library publishing services: Strategies for success. Final research report. Washington, DC: SPARC. Available at http://docs.lib.purdue.edu/purduepress_ebooks/24/ (Accessed: 7 March 2016).

Perry, A.M., Borchert, C.A., Deliyannides, T.S., Kosavic, A., Kennison, R. and Dyas- Correia, S. (2011) 'Libraries as journal publishers', Serials Review, 37(3), 196-204. http://dx.doi.org/10.1016/j.serrev.2011.06.006

Research Excellence Framework (2014) UoA 35 Music, Drama, Dance and Performing Arts, University of Huddersfield. Available at http://results.ref.ac.uk/Results/BySubmission/2521 (Accessed: 7 March 2016).

Simser, C.N., Stockham, M.G. and Turtle, E. (2015) 'Libraries as publishers: a winning combination', OCLC Systems & Services: International Digital Library Perspectives, 31(2), 69-75. http://dx.doi.org/10.1108/OCLC-01-2014-0006

Stone, G. (2011) 'Huddersfield Open Access Publishing', Information Services and Use, 31(3/4), 215-223. http://dx.doi.org/10.3233/ISU-2012-0651 Stone, G., Jensen, K. and Beech, M. (2016) 'Publishing undergraduate research: linking teaching and research through a dedicated peer reviewed open access journal', Journal of scholarly publishing, 47(2), 147-170. http://dx.doi.org/10.3138/jsp.47.2.147

Thatcher, S.G. (2007) 'The challenge of open access for university presses', Learned Publishing, 20(3), 165-172. http://dx.doi.org/10.1087/095315107X205084 UCL (2015) UCL launches UK's first fully Open Access university press. Available at https://www.ucl.ac.uk/news/news-articles/0515/270515-ucl-press (Accessed: 7 March 2016).

White Rose University Press (2016) White Rose University Press, 2016. Available at http://universitypress.whiterose.ac.uk/ (Accessed: 7 March 2016).

Withey, L. et al. (2011) 'Sustaining scholarly publishing: New business models for university presses: a report of the AAUP task force on economic models for scholarly publishing', Journal of Scholarly Publishing, 42(4), 397-441. http://doi.org/10.1353/scp.2011.0035

Xia, J. (2009) 'Library publishing as a new model of scholarly communication', Journal of Scholarly Publishing, 40(4), 370-383. http://dx.doi.org/10.3138/jsp.40.4.370

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-48

Genome sharing projects around the world: how you find data for your research

Fiona NIELSEN[a,b,1] and Nadezda KOVALEVSKAYA[a,b]

[a] *Repositive Ltd, Future Business Centre, Kings Hedges Road, Cambridge CB4 2HY, United Kingdom*

[b] *DNAdigest, Future Business Centre, Kings Hedges Road, Cambridge CB4 2HY, United Kingdom*

Abstract. Access to raw experimental research data and data reuse is a common hurdle in scientific research. Despite the mounting requirements from funding agencies that the raw data is deposited as soon as (or even before) the paper is published, multiple factors often prevent data from being accessed and reused by other researchers. The situation with the human genomic data is even more dramatic, since on the one hand human genomic data is probably the most important data to share - it lies at the heart of efforts to combat major health issues such as cancer, genetic diseases, and genetic predispositions for complex diseases like heart disease and diabetes. On the other hand, since it is sensitive and personal information, it is often exempt from data sharing requirements. DNAdigest investigates the barriers for ethical and efficient genomic data sharing and engages with all stakeholder groups, including researchers, librarians, data managers, software developers, policy makers, and the general public interested in genomics. Repositive offers services and tools that reduce the barriers for data access and reuse for the research community in academia, industry, and clinics. To address the most pressing problem for public genomic data: that of data discoverability, Repositive has built an online platform (repositive.io) providing a single point of entry to find and access available genomic research data.

Keywords. Genomic data, data access, data sharing, genomic data repositories, tools

1. Introduction: data access and reuse is a common hurdle in scientific research

Research organisations, both public and private, are producing ever increasing volumes of data. Irrespective of whether the research is funded publicly or privately, there is increasing pressure to provide evidence that the maximum benefit is obtained from generated data. Recent years have seen a concerted effort by providers of public funds for research to require that the results of that research be publicly available (Collection of UK funders' policies 2015).

While the benefits of data sharing are becoming more widely accepted (Toronto International Data Release Workshop Authors 2009), human genomic data (i.e.,

[1] Corresponding Author: fiona@repositive.io

information about the composition of our DNA and RNA) is often exempt from data sharing requirements from major funders that all experimental data must be placed in publicly accessible repositories. This is because of concerns that making human genomic data public exposes potentially sensitive personal information to the world (Richards 2015).

2. The special case of human genomic data: it is there but it is mainly inaccessible

It is estimated that, in 2015, the world human genome sequencing capacity will exceed 80 petabyte of sequence a year. However, as of 2014, the largest public repository for human genomics data (the NIH database of genotypes and phenotypes dbGaP) holds only about 0.5 petabytes of clinical genomics data.

This gap between the availability of genomic information and the production of it can be at least partially attributed to the absence of tangible benefits for the individuals who make data available and, at the same time, to the existence of sanctions for improper handling of personal information. However, when data donors give consent for their data to be used for research, they set their expectation that the data will actually be used for this purpose. To not utilise their data in the best possible way within the consent given goes against the data donor's interests and expectations. Ironically, human genomic data is probably the most important data to share, since it lies at the heart of efforts to combat major health issues such as cancer, genetic diseases, and genetic predispositions for complex diseases like heart disease and diabetes. In particular, the promise of personalised medicine (where treatment is tailored to the individual) is unlikely to be realised without widespread access to large amounts of genomic data.

Existing data sharing initiatives generally take the form of some kind of repository for storing data or some kind of service to help find collaborators or data. Examples of repositories include publicly funded repositories (e.g. SRA, ENA, dbGaP, EGA, ArrayExpress etc), biobanks, and data repositories set up by individual institutions or projects (e.g. LOVD). Examples of services to help find data include, for example, GenomeConnect, PhenomeCentral and the Beacon project.

Public data repositories do an excellent job of storing data, a crucial task to enable data availability. The mentioned services do great jobs at servicing specific needs, e.g. connecting clinicians who have found similar phenotypes for their patients with genetic diseases. Currently no public initiatives are have successfully addressed the problem of discovering the existence of datasets (data discoverability) for specific diseases and specific data types across locations and repositories. In addition, all of the mentioned initiatives face challenges of funding and sustainability of their initiatives, due to their reliance on research grants.

3. A solution to increase access to human genomic data: the community platform

To address the most pressing problem for public genomic data: that of data discoverability (van Schaik 2014), Repositive has built an online platform (repositive.io) providing a single point entry to search and access public genomic data sources. The Repositive platform enables users to search through all its indexed data sources in a single click via an easy-to-use interface free of charge. To address the problem of varying quality and type of metadata associated with data across data sources and public repositories, the Repositive platform allows users to comment on the content and quality of datasets and add descriptions to the listed metadata. If a research scientist has data that he/she would like to share but cannot for any reason, he/she can announce the existence of the data on the Repositive platform. In this case, other scientists that have similar or complementary data can contact the author to start a collaboration or to discuss for instance the conditions under which they can exchange their data. Similarly, a user can post a request for data and another user, who has the data stored but not used, can respond and find an application for their otherwise unused data.

By listening to our users and concentrating on a specific use case for the genetics researcher – the problem of finding and accessing human genomic research data - and supporting best practices for data annotation, accessibility and reuse, we offer the Repositive platform and services as a contribution to ease the workflow for research in human genomics for health and disease.

The Repositive business model is built around the Repositive freemium features of the online platform for data discovery. The online platform is open for all to sign up and search for free (see above), but at the same time Repositive offers premium products and services to both data providers and data consumer organisations. Our premium services include: customised data scouting; data access applications; automating data access workflows; setting up and maintaining public data catalogues; setting up data collaborations between different organisations, e.g. across industry and academia, etc. With this business model, Repositive can deliver a free service on the online platform which supports researchers across academia and industry, while our revenue comes from related professional products and custom services.

References

Collection of UK funders' policies. In: Research Data Management Blog. (2015) Available at: http://www.data.cam.ac.uk/funders. (Accessed: 15 February 2016).

Richards, M., Anderson, R., Hinde, S., Kaye, J., Lucassen, A. et al. (2015) 'The collection, linking and use of data in biomedical research and health care: ethical issues'. Nuffield Council on Bioethics. Report. Available at: http://nuffieldbioethics.org/wp- content/uploads/Biological_and_health_data_web.pdf. (Accessed: 15 February 2016).

van Schaik, T.A., Kovalevskaya, N.V., Protopapas, E., Wahid, H. and Nielsen, FGG (2014) 'The need to redefine genomic data sharing: A focus on data accessibility'. Applied&Translational Genomics, 3(4), pp.100-104 (doi:10.1016/j.atg.2014.09.013)

Toronto International Data Release Workshop Authors (2009) 'Prepublication data sharing'. Nature, 461, pp. 168-170. (doi:10.1038/461168a)

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-51

Stakeholders in academic publishing: text and data mining perspective and potential

Maria ESKEVICH[1]

Radboud University, Nijmegen, The Netherlands

Abstract. In this paper we discuss the concept of open access in academic publishing with the focus on the right to mine the data once the right to read is granted. Thus we envisage the roles and types of the stakeholders in academic publishing from the perspective of the potential text and data mining (TDM) applications. Further on, we briefly introduce FutureTDM project that aims to improve TDM uptake in Europe.

Keywords. digital libraries, text and data mining, FutureTDM project

1. Introduction

The main incentive for academic publishing is to share the knowledge acquired through experimental and/or empirical observations with an overall aim to promote further scientific development and knowledge distribution. Hence, initially the publishing empowered more researchers and thinkers to improve their expertise and to expand the overall knowledge ontology. The dialogue was set up between the content creators through the medium of written and printed text providers. However, the societal and technological development in combination with population growth over the recent centuries lead to a more complicated framework of agents in the field of scientific knowledge sharing, as the same agents can play different roles.

Th The publication process, while having a target to broaden the access to the knowledge across communities, is a service that is being provided, and thus over the years it converted into a business model which raised a pay wall between the ultimate content consumers, i.e. researchers, general audience, and the content itself. As a vast amount of research is being carried out on public funding, the new frameworks for efficient scientific knowledge transfer are discussed and promoted, with the Open Access (OA) strategy being the main focus (De Grandis, Lomazzi, Rettberg). Following the OA principles defined in Budapest and Berlin Declarations, the European Commission (EC) defines Open Access as 'the practice of providing online access to scientific information that is free of charge to the end user and that is reusable', where scientific information can refer to (i) peer-reviewed scientific research articles (published in scholarly journals) or (ii) research data (data

[1] Corresponding Author: Maria Eskevich, The Netherlands; E-mail: m.eskevich@let.ru.nl

underlying publications, curated data and/or raw data). These definitions describe ’access’ in the context of open access as including not only basic elements such as the right to read, download and print, but also the right to copy, distribute, search, link, crawl, and mine2. Mining of both the publication text itself and of the corresponding data sets can be carried out with the help of diverse text and data mining (TDM) tools. Based on the last decades development in this domain, it is evident that TDM mechanisms are present throughout scientific and cultural environments, but not in a systematic or infrastructural way. TDM could help in solving scientific problems, which is why we see it in the heart of the future of Open Science. It is often used in domains that are rather advanced in their open and interoperable practices, e.g. bioinformatics, signifying a change in the modus operandi of performing research already showcasing a shift in approach to organizing Science. However, as reported in the Royal Society 2012 report ”Science as an Open Enterprise”, new text-mining technologies and developments in multidisciplinary research would be empowered if TDM barriers were lowered, and there are global policy and political signals that this is not only scientifically desirable, but ultimately inevitable.

In this paper we outline the field of text and data mining that in our view should be incorporated into the publishing practice framework in order to profit from the state-of-the-art TDM research technologies which can be helpful across all fields of science. Thus we regard the structure of the academic publishing stakeholders from the angle of TDM technologies involvement. The remainder of this paper is structured as follows: Section 2 introduces the open access publication agenda (2.1) and introduces the concept of usefulness of the text and data mining as its next step (2.2); Section 3 describes in details what kind of different potential roles (3.1) the diverse players in publishing, research and overall society (3.2) can take in order to promote further beneficent interaction of TDM technologies and the digital publishing; Section 4 reports the state-of-the-art research that can represent potential implementation use cases; Section 5 introduces the FutureTDM project that analyses current TDM uptake across different fields and outlines its potential; and finally Section 6 gives conclusions and outlines directions for future work.2.

2. Academic publishing: challenging background

The amount of academic publications is steadily growing across different fields of research with thousands of papers being produced each year, e.g. Figure 1 illustrates the case of Computer Science publications over the past 20 years. The sheer volume of publications pool and the growing trend impede research community, as well as generally the society members, to track all the trends within one field, and it becomes even more challenging to target multidisciplinary domains or to promote cross-domains methods applications. Having the access to this content, TDM can help researchers to cope with the tripling rate of growth of scientific output (Laren 2010).

2.1 Open Access Publication Agenda

The European Commission has made open access a general principle of Horizon 2020 in order to boost innovation capacity1. 'Open access' publications make scholarly literature freely available on the Internet, so that it can be read, downloaded, copied, distributed, printed, searched, text minded, or used for any other lawful purpose, without financial, legal or technical barriers, subject to proper attribution of authorship. Open access improves the pace, efficiency and efficacy of research. It heightens the visibility of authors and the potential impact of their work.

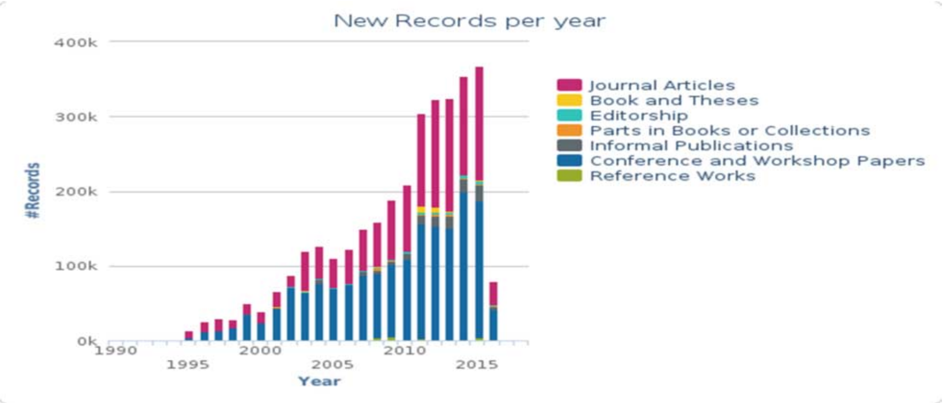


Figure 1: Total number of publications of the different publication types according to DBLP Computer Science Bibliography. [Accessed on 03.2016]

It removes geographical and structural barriers that hinder the free circulation of knowledge. Thereby contributing to increased collaboration, and ultimately strengthening scientific excellence and societal progress. It would seem therefore that open access is a major factor in increasing the uptake of TDM. Yet, it seems that the potential of open access as a means to facilitate data-driven innovation may be undermined by lack of interoperability between licenses and the proliferation of licenses which prohibit the creation of derivatives. This transition requires cooperation of all the stakeholders in the field.

2.2 Is the right to read becoming a right to mine?

The right to mine that is stated as the principle of the open access implies the availability and development of the TDM techniques which in reality requires a framework of data storage, access and processing that may be built with the collaboration between different stakeholders in the field. The TDM research is rapidly growing, but its incorporation into the publishing agenda is still affected by several factors (Hargreaves): economic issues of the market practices changes that it should bring on, the legal issues of copyright (Handke 2015), the lack of awareness among key potential stakeholders, the need for additional training of librarians, researchers, etc.

3. Stakeholders in the field and their different yet overlapping agendas

Stakeholders in the field may be actively engaged in publishing and/or text and data mining directly in their day to day activities, as service providers or developers; or they may have an indirect interest in knowledge discovery, analyze and/or make use of the information gleaned through content mining.

3.1 Stakeholder Roles

We assume a number of general roles that can be taken by different/same stakeholders in the field. Each role is associated with a different step in the circle of knowledge sharing:

Direct work with the TDM process, legal and financial support of this work. Building of a sustainable infrastructure for TDM requires the main stakeholders to undertake the following roles:

- Data Provider: in the framework of academic publishing it implies both papers writing, editing for publication, indexing in the database of publications and associated resources;
- Processing Techniques Developer: the core TDM research is to be implemented based on the state-of-the-art scientific accomplishments in the field;
- Service Providers: once the TDM techniques are developed into a software, the results of the automatic analysis in terms of trends analysis, building solutions based on TDM trends can be released as a service;
- User of TDM techniques and results: the new insights into problems and extended knowledge based on the TDM extracted data and trends that can and should be accessible to use for the research community and general audience.

3.2 Stakeholder Types

The community that can benefit from the new academic publishing framework expanded with the TDM perspective is broader than simply researchers, publishers, and librarians. We foresee the involvement of both public/non-profit and industry sectors. Figure 2 exemplifies how the different stakeholders roles listed in Section 3.1 can be associated with different agents in the community.

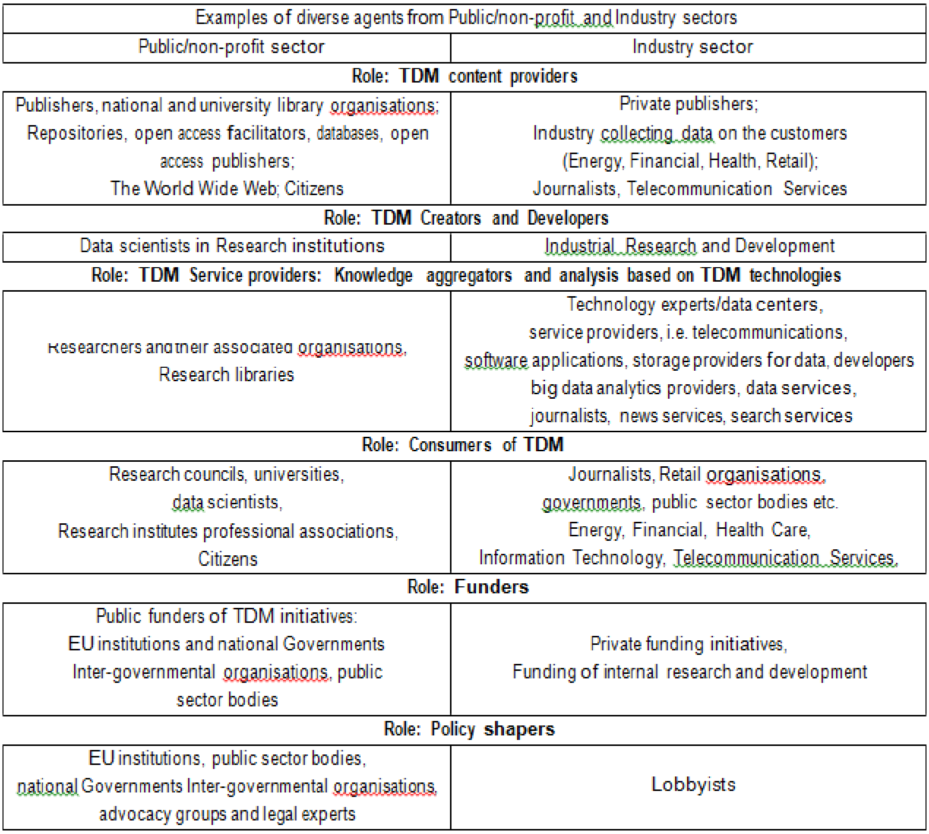
Within this framework, content providing means both the provision of the content of the publications, as well as general data that the paper's discussion and experiment sections can be built on. TDM research is being carried out within the research institutions or departments in both public and industrial sectors, and its outcome is available both through the relevant publications or through the services to the general audience.

The algorithms behind TDM research are created and thoroughly investigated within natural language processing (NLP) communities. However, due to the legal restrictions of access, these studies are more often carried out on limited corpora when directly applied to the scientific publications, or otherwise the NLP researchers test their theories on the other datasets with an assumption of potential further technology transfer across datasets.

4. TDM techniques use cases

In this section we outline current trends in the applied computational linguistics research that are already directly applied to academic publications in order to extract the knowledge or demonstrate potential within the outlined framework. Overall, there are three directions for these applications: information extraction of the content across a set of publications; summarization of the information across a set of publications; use the TDM techniques to reinvent the impact measurements of the scientific publications:

- Information extraction: Each section of a paper can be treated separately when different information is to be extracted for further analysis. It varies from simple detection of the papers published across different venues within the same project and funding scheme to more complicated cases of citations sentiment detection which allows better comprehension of the relevance of the current paper to the work in the field (Hong 2015).
- Information summarization Once the separate facts are extracted from the papers, this information can be automatically summarized for further analysis using TDM and NLP techniques.



Examples of diverse agents from Public/non-profit and Industry sectors	
Public/non-profit sector	Industry sector
Role: TDM content providers	
Publishers, national and university library organisations; Repositories, open access facilitators, databases, open access publishers; The World Wide Web; Citizens	Private publishers; Industry collecting data on the customers (Energy, Financial, Health, Retail); Journalists, Telecommunication Services
Role: TDM Creators and Developers	
Data scientists in Research institutions	Industrial Research and Development
Role: TDM Service providers: Knowledge aggregators and analysis based on TDM technologies	
Researchers and their associated organisations, Research libraries	Technology experts/data centers, service providers, i.e. telecommunications, software applications, storage providers for data, developers big data analytics providers, data services, journalists, news services, search services
Role: Consumers of TDM	
Research councils, universities, data scientists, Research institutes professional associations, Citizens	Journalists, Retail organisations, governments, public sector bodies etc. Energy, Financial, Health Care, Information Technology, Telecommunication Services.
Role: Funders	
Public funders of TDM initiatives: EU institutions and national Governments Inter-governmental organisations, public sector bodies	Private funding initiatives, Funding of internal research and development
Role: Policy shapers	
EU institutions, public sector bodies, national Governments Inter-governmental organisations, advocacy groups and legal experts	Lobbyists

Figure 2: Examples of diverse agents from Public/non-profit and Industry sectors

- Impact measurement Evaluation of the impact of the content, raising the profile of the publications using novel approaches to bibliometrics (Mayr 2015, Athar 2012), as not only the sheer number of citations might be representative of the quality of a certain research paper, but such details as the context of citations can bring better insight into understanding of papers overall value and mutual relevance.

5. Future TDM Project

FutureTDM2 project supports the uptake of TDM across all sectors of economy, considering publishing sector being of high importance, as so much of scientific information is confined within the deluge of publications which can be profitable for both commercial and non-profit use. In this paper, we discuss the scientific publications context, while within this project in general we aim to contact various types of stakeholders from different sectors in order to identify how their progress in the field can be supported when embracing TDM on the large scale, and how the structure of the roles is to be readjusted for each specific sector accordingly.

6. Conclusions

In this paper we have discussed the importance of TDM technologies for the future development of the academic publishing, and introduced the structure of the roles and types of the stakeholders in the field accordingly. This discussion should raise the awareness of the TDM potential and bring better understanding of the interaction structure.

7. Acknowledgments

This work is supported by the European Commission's HORIZON 2020 Programme (H2020) under GARRI-3-2014 – 665940 (FutureTDM).

References

Hargreaves et al. Expert Group Report on standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining, http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf [Accessed on 12.2015]

Berlin Declaration of 2003 http://openaccess.mpg.de/67605/berlin_declaration_engl.pdf [Accessed on 12.2015]

De Grandis G., Neuman Y. Measuring Openness and Evaluating Digital Academic Publishing Models: Not Quite the Same Business. Journal of Electronic Publishing, Volume 17, Issue 3: Metrics for Measuring Publishing Value: Alternative and Otherwise, Summer 2014. DOI: http://dx.doi.org/10.3998/3336451.0017.302

Lomazzi L., Chartron G. The implementation of the European Commission recommendation on open access to scientific information: Comparison of national policies. Inf. Services and Use, vol. 34 (3-4), pp. 233–240, 2014.

Rettberg N., Schmidt B., RossA. Infrastructures for Policies: How OpenAIRE Supports the EC's Open Access Requirements. New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust, ELPUB 2015, pp. 185 - 189, IOS Press.

C. Handke, L. Guibault, J.-J.Vallbé. Is Europe falling behind in data mining? Copyright's impact on data mining in academic research. New Avenues for Electronic Publishin in th eAge of Infinite Collections and Citizen Science. ELPUB 2015, pp. 185 - 189, IOS Press.

P.O. Laren, M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84(3): 575603. 2010.

P. Mayr, P.Schaer, A. Scharnhorst, P. Mutschke. Editorial for the Bibliometric-Enhanced Information Retrieval Workshop at ECIR 2014. CoRR abs/1404.7099 (2014). Amsterdam, the Netherlands.

P. Mayr, I. Frommholz P. Mutschke. Editorial for the 2nd Bibliometric-Enhanced Information Retrieval Workshop at ECIR 2015. CEUR Workshop Proceedings, Vol. 1344. Vienna, Austria, March 29th, 2015.

A. Athar, S. Teufel. Detection of Implicit Citations for Sentiment Detection. Proceedings of the Workshop on Detecting Structure in Scholarly Discourse at ACL '12, Jeju, Republic of Korea, pp. 18–26, 2012.

K. Hong, M. Marcus, and A. Nenkova. System Combination for Multi-document Summarization. EMNLP, page 107-117. The Association for Computational Linguistics, (2015).

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-58

Referencing of complex software environments as representations of research data

Sven BINGERT [a,1], Stefan BUDDENBOHM [b] and Daniel KURZAWE [c]
[a] *Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen*
[b] *Max-Planck-Institut zur Erforschung multiethnischer und multireligioser Gesellschaften*
[c] *Niedersächsische Staats- und Universitätsbibliothek Göttingen*

Abstract. Complex software environments, like virtual research environments or visualisation frameworks, are increasingly used to conduct research and present its results. While there is a growing amount for solutions facilitating the (granular) citation of publications and research data, the citation of complex software environments remains a challenge. This abstract outlines the challenges and introduces an approach for referencing software environments developed in the Humanities Data Centre project: the application preservation.

Keywords. digital humanities, complex software environment, application preservation

1. Introduction

Progress and transparency in science largely depends on the capability of researchers to cite and reference the various aggregations of research data, instruments, and publications. Regardless what the subject of referencing may be, it is inevitable to sustain access to stable object representations that have to be documented in a transparent and proper way. For this end standards and infrastructure are necessary to serve the discipline- specific procedures of citing and referencing. Whereas these standards and infrastructure are quite established and harmonised for publications of research results - a traditional duty of libraries - the field is only developing with regard to research data. Beyond this an comprehensive overview and discussion of the evolving landscape of research data types and some implications can be found in [Sahle & Kronenwett 2013]. The challenges related to referencing these new types of data will be described in detail in this paper, focussing on complex software environments as representations of research data. The proposed solution revolves around an adapted persistent identification (PID) [Kalman 2015] approach applying fragment identifier and template handles. The insights base to a large part on the design phase of the Humanities Data Centre[2] , a research data centre for the Humanities currently under construction.

[1] Corresponding author. Gesellschaft fuer wissenschaftliche Datenverarbeitung Goettingen
[2] http://www.humanities-data-centre.org; last visited March 2016

2. What do researchers reference?

The short answer is: Everything. There is no hard limitation regarding the objects of research. Every distinguishable object might be addressed and referenced or cited [Kalman 2015]. The range of objects is almost unlimited: files in a file system, database entries, web sites, books, places, people or journal articles. For many of those object classes solutions in form of services or tools to create and resolve references are available. The most prominent examples are ISBN[3] for books or DOIs [Paskin 2010] for digital publications, but also unique stable references to people realised by ORCID[4] are prevailing. However the unsolved problem remains, that a large share of research data does not fit in these categories. Despite this the heterogeneity with regard to size, format or structure doesn't make it easier to handle. So far for static objects as representations of research data. But what lies beyond these conventional, static object classes? It is obvious that researchers want to reference a broad range of research data types that do not fit into the static definition and that are only evolving. The development of new content or data types is closely aligned to the development of the working environment, methods and instruments of the researchers and for this reason quite difficult to be foreseen for an infrastructure provider. Also the citation of research data fulfils various purposes ranging from impact and reputation to transparency and reproducibility of research results so the scope of the archives can be widened. The latter point for example makes the archiving of different aggregations of one same data set interesting as it allows reproducing the certain processes. So in terms of content types not only conventional formats of data should be taken into account but also software environments, virtual research environments[5], complex databases[6], visualisation frameworks[7], collections, or processed data. As a common thread appears the complex character of these kinds of data, meaning that they can have various consecutive layers, aggregations or components. To some extent software is depending on its environment, but also on its way of usage. Both is evolving over time, also ontologies, terms or references in data bases. The state of software environments, like an interactive visualisation tools is also highly fluid but is, as we argue, also a research object by itself. In which way can these classes of objects be referenced? A reference will likely point to a specific fixed state of the object such as a query term or search string and not to the virtual research environment or the database as such. A workaround could be a reference pointing to a jump page of a research data centre or a repository providing the query term or search string. This can work out as long as the database remains stable. Out of the question is the inconvenience of this procedure as it requires additional steps in the absorption process on side of the reader. Therefore a more convenient solution is needed. This solution must provide a citation which includes descriptors of such a quality that the reference can point to clearly one specific data set, ideally integrated in one PID that just has to be resolved by the reader.

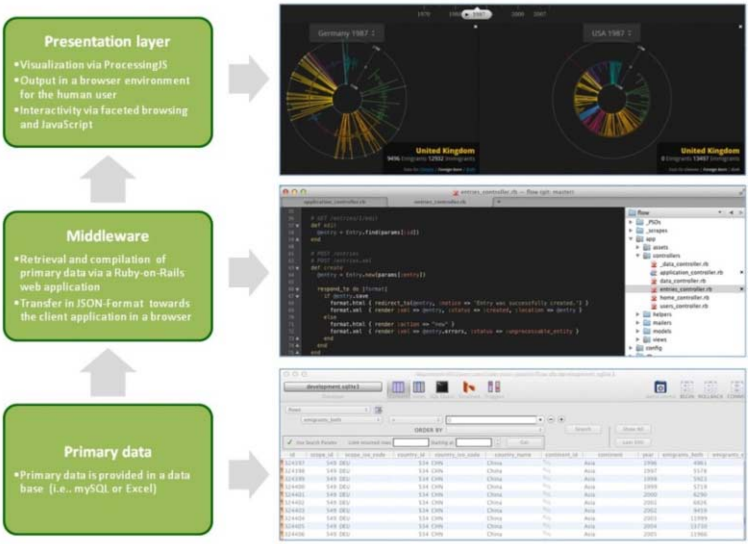
[3] DIN ISO 2108

[4] http://www.orcid.org; last visited March 2015

[5] e.g. http://textgrid.de; last visited March 2016

[6] e.g. http://www.berliner-klassik.de; last visited March 2016

[7] .g. http://media.mpi.mmg.de, last visited March 2016


Presentation layer
•Visualization via ProcessingJS
•Output in a browser environment for the human user
•Interactivity via faceted browsing and JavaScript
Middleware
•Retrieval and compilation of primary data via a Ruby-on-Rails web application
•Transfer in JSON-Format towards the client application in a browser
Primary data
•Primary data is provided in a data base (i.e.. mySQL or Excel)
Germany 1987
USA 1987
United Kingdom
entries_controller.rb — flow (git: master)
development.sqlite3
ORDER BY

```
# GET /entries/1/edit
def edit
  @entry = Entry.find(params[:id])
end

# POST /entries
# POST /entries.xml
def create
  @entry = Entry.new(params[:entry])

  respond_to do |format|
    if @entry.save
      format.html { redirect_to(@entry, :notice => 'Entry was successfully created.') }
      format.xml  { render :xml => @entry, :status => :created, :location => @entry }
    else
      format.html { render :action => "new" }
      format.xml  { render :xml => @entry.errors, :status => :unprocessable_entity }
    end
  end
end
```

Figure 1. The multi layer character of a complex software environment using the example of the visualisation of global migration flows.

2.1. Excursus: visualisation framework

As an illustration for the above mentioned new object classes we describe a visualisation framework because it not only demonstrates the fluid character of the data but also what additional layers and dependencies have be taken into account when archiving and referencing this kind of data. For example there can be questions of a technical nature (granularity of the reference) or legal questions (licence status of content components) that influence the proposed infrastructural solution. The term visualisation framework may be seen in this specific context as synonym for a complex software environment. Other embodiments of complex software environments may be digital editions or virtual research environments. Our visualisation framework in this context [Aschenbrenner et.al. 2015] can be characterised as an attractive presentation of research data with interactive components aimed at the human user. Basically it is a database-service visualising the result of a search string. The illustration in Figure 1 depicts the Global Migration Flows[8] , allowing the user to create individual data sets visualising migration movements between selected countries and over selected periods of time. The visualisation framework is accessible via a common browser and is based on data from the United Nations Population Division ranging from 1970 to 2011. The data visualisation is closely bound to its presentation environment, therefore an archival solution for this kind of data has to address this multiple layer character to sustain its added value. Usually this kind of data visualisation, based on a browser as access interface, consists at least of three layers (c.f. Figure 1):

[8] http://flow.mmg.mpg.de; last visited March 2016

1. The normalised and enriched data provided through a database (data layer or primary data), which can be as simple as an Excel chart or ranging to more complex forms of databases.

2. A processing layer **(middleware)** that transfers the normalised and enriched data to the client application of the end user, the web browser. The necessary resources for the development and coding of the processing layer and the following presentation layer cannot be provided in each research project because of spare competences or resources. This advocates the refuse of these kind of application and data for other research projects.

3. On this rests the user interface, usually a common web browser. This presentation layer is normally out of the scope of action of the researcher. The visualisation framework has no influence on the browsers used by the end used and only can try to cover a range of most common standards.

The above standing remarks allow us to identify two overarching aspects that have to be addressed by any technological solution: the sustainability of the system (c.f. Sec. 3) and the citation of actual states of the system (c.f. Sec. 4).

3. Challenges related to the sustainability of complex software environments

Current software versions are often transient. With the end of a research project or project financing occurs the risk that developed software components and systems are no longer maintained - meaning they will be outdated and inaccessible soon. This is either due to (a) security reasons, (b) incompatibility with new technologies (changing web standards, new operating systems standards, hardware incompatibility) or (c) dependencies to external systems (e.g. changing APIs). Whereas these issues - or the proposed solutions - are more technology-related, there are other areas of challenges, e.g. social, legal, or financial. The above described example of a visualisation framework provides an illustrative example of this context. The discontinuation of software environments may not pose a problem as long as publications of the research results are available. But as explained above this is not in the general interest of researchers, research funding and research institutions for various reasons such as transparency, reproducibility or refuse of research.

3.1. System security

Common problems are security issues in the software stack. These might be vulnerabilities in the operating system or software modules that are either third party or developed within the project. Some may be fixed with simple updates, which can be handled by the hosting institution, like updates of the operating system. But it is only a matter of time before the software is at the end of its support cycle or major upgrades might then threaten its stability or functionality. Separation is a relatively simple step for handling security issues. With access control and strict firewall rules on the network level, the system can be separated from other environments. The strongest way of separation is complete isolation from other systems. Direct access is not possible in this case. Just a single gateway service can access the system and communicate between the user and the system. But this limitation might cause problems: External dependencies, like the access to other databases, are limited. Another security layer of archived system states can be provided with the usage of templates to achieve safe states: A selected state of the research system will be transferred into a template. With each session, a new instance will be cloned from the template. This has the advantage, that every user will get a clean configuration in a defined state. However, in this scenario the state is fixed to a defined state: The session management and storage of data will become complex. System security poses a general challenge to the infrastructure provider as it is a fluid concept, advancing over time.

3.2. Compability

Software systems continuously have to adapt to new standards and technologies. Even if some technology stacks are relatively stable, their usage might change over time because of possible semantic drifts. There are basically two concepts to tackle hardware compatibility issues: either virtualisation or emulation (see also [van der Hoeven et.al. 2005]) of components. While emulation allows for running of operating systems and software on hardware they were not developed for, it costs extra computing power to emulate the needed hardware environment. A change of the already replaced hardware would cause additional work to reimplement the emulation. Virtualization has the clear benefit that the hardware appear as physical device to the system and the software modules. It comes with the disadvantage that the virtualised environment must be already capable directly on that hardware. But for most of the use cases considered here only standard hardware is used. Therefore the advantage of running more than one virtualised system and the relatively simple management of these system makes the virtualisation the state of the art approach.

3.3. Dependencies

With the increasing bandwidth and speed of data networks (including the internet) it became common to outsource data and software modules that are only (down-)loaded upon request. This approach of course enables compatibility and simplifies maintenance of widely used information. On the other hand it causes dependencies of a rapidly changing environment where the data provider has no influence on. Additional steps need to be taken in order to reduce these number as far as possible whereas the limitations are e.g. legal issues or data volume. On the one hand dependencies form a challenge for the infrastructure provider, on the hand the benefits for the research are obvious: it is possible to integrate very different types of data and content and in this way to enhance research. As a relativisation one has to take into account that the integration of external libraries, modules or applications is only possible with a certain level of standardisation. In this standardisation may also lay the key to the technological solution for archiving and referencing these kinds of data.

4. The citation of unambiguous system states via Fragment PIDs

The usage of PIDs is common to ensure stable references to publications and is increasingly accepted for file?based data to. The uniqueness and stability of PIDs has to be guaranteed and realised by the PID service provider. These service providers use a specific PID system that may be distinguished by its functionality. E.g. the ePIC PID service offers the creation of Fragment PIDs. These are identifiers that not only can be resolved to a given location, but also allow to forward parameters when the PID is resolved. With this it becomes possible to make a stable reference to defined system conditions - but it also requires some effort on the side of the referenced system. The Fragment PID can be used to present the user predefined configurations of a system. This could for instance be a visualisation with specific parameters or a simulation of a specific state. As mentioned in the introduction PIDs are a common means to cite a very broad range of various objects. In the humanities PIDs are used to identify collections, content or objects. PIDs are not only able to reference to definite objects but may also reference object fragments with the usage of a Fragment PID. This may be passages of text or illustrations or links to certain sections in digital media by following examples:

- http://www.domain.org/book1@page=10
- http://www.domain.org/video1@begin=10&end=20

where in this example *book1* and *video1* represent the PID. The naming schema for such PIDs differs between the existing PID systems and is not subject of discussion in this paper. But important to note is that with those identifiers an unlimited number of fragments in an entity can be referenced and provide the level of granularity that is necessary for scientific citations.

5. The Humanities Data Centre as use case for referencing

The infrastructure for long-term storage and provision of research data in the humanities is only beginning to emerge, but is currently not as developed as in other domains such as astrophysics or climate research. Therefore, the Humanities Data Centre (HDC), aims to establish a data centre for research data from the humanities. The project will enter the construction phase to a later point. Therefore the described approach has a prototypical character. Above all looms the question of sustaining a research data centre, meaning that the proposed solutions not only have to meet the demand of the researchers but also have to address the conditions and resources available to the infrastructure provider (in terms of costs). As a consequence we wanted to identify suitable and already available solutions and components to compile the HDC service portfolio (c.f. Figure 2). For a detailed description of the HDC's initial service portfolio consult the project website. With regard to the question of referencing complex software environments (as representations of complex research data) and addressing the above mentioned general challenges of sustainability and referencing we introduce the application preservation as service component.

5.1. Application preservation

The initial service portfolio of the HDC can on the one hand be described as modular to address complex use cases, and contains on the other hand innovative components such as the application preservation. The application preservation can be used for complex forms of research data that fit into the above described pattern. It shall guarantee access to a stable version of a component (sustainability) - such as a virtual research environment, a complex database, or a visualisation framework - and enables the user to reference not only the component as a whole but also individual fixed states (referencing), e.g. to generate a certain query term or a specific visualisation set.

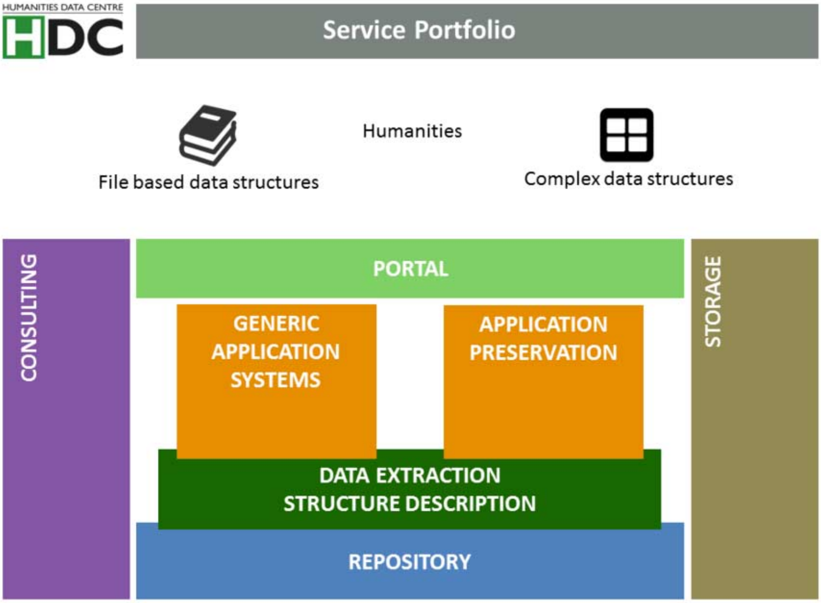


Figure 2. The HDC service portfolio that is currently under construction.

Data structures are preserved in their functional handover status and are only technically changed or maintained by the research data centre to extend its accessibility. Subject of the handover is ideally the whole data structure (e.g. client server structure, dependent libraries and applications). The service focusses on the presentation and reproducibility of research results and methods, not implicitly on the re-usability of the data. It is obvious that preserved applications can only be provided for a limited time by the research data centre for reasons of security gaps or outdating components. An ordinary archive case will utilise a combination of the above depicted services. Nevertheless the application preservation seems as an attractive service as it allows an ample presentation of research compared to archiving of raw data or documentations and for that reason demanding a solution for referencing.

5.2. Architecture of the application preservation

The architecture of the application preservation prototype as described below is the result of a process of raising requirements by the researchers and of evaluating already available components. Beside its nature as prototype it has also to be seen as a compromise which will improve and be further developed only in practical use. The application preservation allows for three ways to access the research result according to the security level. Directly after the project the software is up to date and safe and may be accessed directly (option 1 in Figure 3). When the software module is vulnerable the access will be restricted via an archived browser only (options 2 and 3 in Figure 3). Basically the application preservation consists of these layers:

1. A cloud infrastructure providing the storage and computing capability for the preserved application (its snapshots) and the application itself.

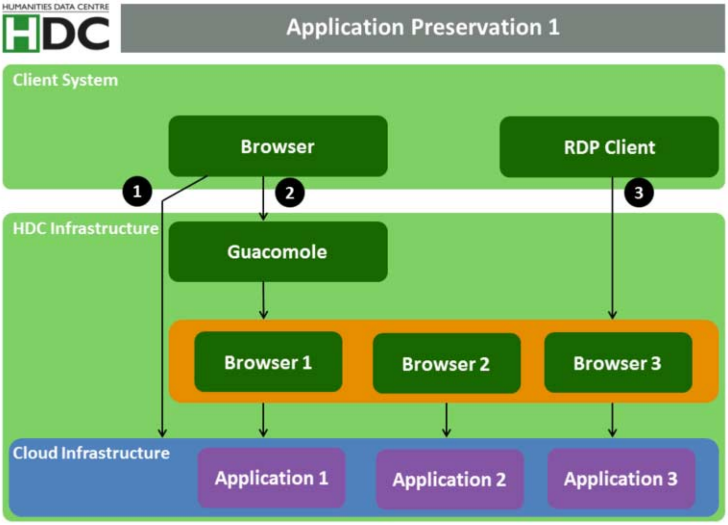


Figure 3. HDC infrastructure for application preservation. The numbered options depict different ways to access the preserved application.

2. A set of archived **browsers** of different brands and versions needed to visualise the application in the optimal way.

3. **Guacamole** [9, 10] serves as bridge between the user’s environment - typically his browser - and the HDC infrastructure. It is able to handle RFP/VNC and RDP protocols to enable a remote access from the user to the preserved application without the inconvenience of installing additional software on the user site. A simple browser is sufficient to access the research data.

[9] http://www.pidconsortium.eu; last visited March 2016
[10] http://guac-dev.org

4. **PIDs** are used to reference the application or specific system state, e.g. a query term resulting in a specific data visualisation or database result.

The usage of PIDs in case of the application preservation can be divided in the following use cases:

1. A PID pointing directly to the URL of the visualisation (option 1 in Figure 3). This is the simplest way of referencing but the PID will become unresolvable quickly due to the reason of the application becomes outdated (c.f Sec. 3).

2. PID pointing at a specific connection configured in the guacamole client (option 2 in Figure 3). This allows a long term access to the application stored in a secure environment. The downside is an enhanced PID management effort to assure the reference to be stable.

3. In case of the third access method (option 3 in Figure 3) a PID usage is not possible due the a configuration of the RDP client on the user system. It still can be used to create a snapshot of an system state that later can be restored and referenced by a PID.

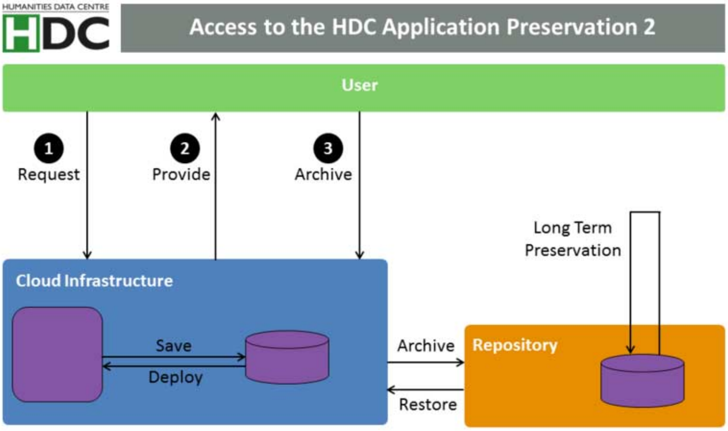


Figure 4. HDC cloud infrastructure for application preservation.

Also one need to distinguish between PIDs generated to reference the software model (the research result) or the HDC infrastructure. In the first case the software offers the service of citable states directly, e.g. by a URL search string. But if this functionality is missing the problem becomes more complex. The user queries or browser interactions have to be monitored and used to recover the desired state. Together with these informations the full state of the virtual environment has to be archived (c.f. Figure 4) which will then be referenced by PIDs that may also activate the system via the fragment mechanism. This functionality is only possible for option 2 in Figure 3.

6. Conclusion

This paper introduced a technological solution for the problem of referencing complex software environments. Complex software environments in this context serve as exemplary representation for the emerging new forms of complex research data, not only in the humanities. A data visualisation framework has been introduced as an individual example for complex software environments. There are other classes of new research data but complex software environments are particularly suitable to demonstrate the challenges related to referencing these new forms of research data. There is a growing demand by researchers for referencing these forms of research data as they allow an ample view of a research project. The referencing of complex software environments could become prospectively a substantial element of the scientific impact and reputation of a researcher, therefore research data centres have to develop solutions of it. We continued with a description of the main characteristics of research data in the humanities, focussing on the more and more complex characteristics of data, which poses new requirements for referencing compared to the long-established referencing of conventional types of content, e.g. text, and formats (monographs, journal articles). Following this, the main challenges in referencing complex forms of research data have been outlined: first ensuring the sustainability of actual software environments and second enabling to reference specific system states, such as a specific search string or query term in a database to ensure a certain granularity.

At this point it became clear that these new forms of research data pose crucial challenges for referencing and infrastructure providers in other, non-technological fields, e.g. legal issues. For instance the dependencies of complex software environments regarding external libraries or the licence status of raw data require coverage on side of the research data centre. A prototypical solution - the HDC application preservation - was introduced as a technological approach for referencing complex software environments. The approach mainly consists of a cloud infrastructure allowing remote access of users to preserved applications via an easy to use additional access layer. The problem of pointing to specific system states - such as a database query term or a specific visualisation data set - is solved via Fragment PIDs. Depending on the application these Fragment PIDs are used to forward URL extensions, resolve into connections to specific applications, or contain pointers that activate the complex software environment in exactly the state created as a snapshot by the editor of the reference. Although the paper remained on a technological level, there are challenges to be addressed reaching beyond the technological level such as legal, financial, interoperability or organisational questions. These challenges also have to be addressed by a research data centre but are not be solved by one research data centre alone. Beyond such research data centres stand libraries and data centres, in the case of the HDC the Gottingen State and University Library (SUB) and the Gesellschaft für wissenschaftliche Datenverarbeitung (GWDG) as important providers of information infrastructure.

References

Sahle & Kronenwett 2013 P. Sahle & S. Kronenwett, Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner "Data Center for the Humanities"; Libreas 23 (2013); urn:nbn:de:kobv:11-100212726.

Kalman 2015 Tibor Kalman, ”Fragment Identifiers, Template Handles”. Presentation at the RDA-D Conference 2015.

Paskin 2010 Norman Paskin, ”Digital Object Identifier (DOI) System” , Encyclopedia of Library and Information Sciences (3rd ed.), Taylor and Francis, pp. 1586-1592.

Aschenbrenner et.al. 2015 Andreas Aschenbrenner, Stefan Buddenbohm, Claudia Engelhardt, Ulrike Wuttke: Humanities Data Centre - Angebote und Abläufe für ein geisteswissenschaftliches Forschungsdatenzentrum; HDC-Projektbericht Nr. 1. 2015; http://humanities-data-centre.org/?page_id=1036

van der Hoeven et.al. 2005 Jeffrey van der Hoeven, Hilde van Wijngaarden: Modular emulation as a long-term preservation strategy for digital objects. In: Lecture Notes in Computer Science. Volume 3652 2005. Research and Advanced Technology for Digital Libraries. 9th European Conference, ECDL 2005, Vienna, Austria, September 18-23, 2005. Proceedings. http://link.springer.com/chapter/10.1007%2F11551362_47

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-69

Towards New Metrics for Bioresource Use

Laurence MABILE[a1], Paola DE CASTRO[b], Elena BRAVO[b], Barbara PARODI[c], Mogens THOMSEN[a], Samuel MOORE[d] and Anne CAMBON-THOMSEN[a,e]

[a]UMR1027 INSERM-Université Paul Sabatier Toulouse III, Toulouse, France
[b]Istituto Superiore di Sanita, Rome, Italy
[c]IRCCS AOU San Martino - IST, Genova, Italy
[d]Ubiquity Press, London, United Kingdom
[e]BBMRI-ERIC, Graz, Austria

Abstract. The BRIF is an ongoing initiative that encompasses reflections and actions from various stakeholders (researchers, funders, industrials, editors) towards i/ standardised identification schemes and reporting for better visibility and tracing of bioresources on the web; ii/ incentive policies from hosting institutions; iii/ creation of tools allowing follow up of their use. Tracing the use of bioresource is the first step in this process and for this purpose we have published the CoBRA (Citation of BioResources in journal Articles) guideline, launched the Open Journal of Bioresources and started developing new metrics. The CoBRA guideline aims to standardise the citation of bioresources in scientific articles in order to trace their use on the web. The Open Journal of Bioresources (OJB) was created in close collaboration with the open access publisher Ubiquity Press allowing both the resources and the OJB papers to be cited, and also providing authors with tools to get metrics on reuse and impact. New better adapted metrics are being worked out in a dedicated BRIF working subgroup. A first list of relevant parameters to take into account in the impact measure of bioresources has been provided. The tools proposed here foster easier access to samples and associated data as well as their optimised use, sharing and recognition for data producers. Input from the scientific editorial community would be highly appreciated at this stage.

Keywords. Bioresources citation; guideline; open access journal; metrics; impact

1. Introduction

For several years, the BRIF (Bioresource Research Impact factor)[2] (Cambon-Thomsen et al 2011) initiative has focused on specifying the framework to facilitate sharing of bioresources [3] through incentives and tools. The basis of the BRIF concept is that making feasible to trace the use of a bioresource and to calculate

[1] Corresponding Author: laurence.mabile@univ-tlse3.fr

[2] http://gen2phen.org/groups/brif-bio-resource-impact-factor

[3] Bioresources are defined as any collection of biological samples with associated data, biological related databases independent of physical samples or other collections of biomolecular and bioinformatics research tools.

a corresponding impact factor should encourage institutions, researchers, bioresource managers and other actors involved in bioresource work, to share them. Sharing would then be seen as a gain rather than a loss of control or than an additional non recognised work, as often felt, so far. These issues are a concern in many biology and biomedical communities. Although the concept could be used in many areas (for example for primary resources in humanities and for ecological collections) we focus on human biological and biomedical resources because their very existence is depending directly on the willingness of patients and participants to give their samples and to allow the use of their data and there is an ethical imperative of making their contribution useful and recognised.

BRIF is an ongoing initiative that encompasses reflections and actions from various stakeholders (researchers, funders, industrials, editors) within dedicated working groups towards i/ standardised identification schemes and reporting for better visibility and tracing of bioresources on the web; ii/ incentive policies from hosting institutions; iii/ creation of tools allowing follow up of their use. Tracing the use of bioresource is the first step in this process and new tools have been or are being developed to make it feasible: the CoBRA guideline (Citation of BioResources in journal Articles), the Open Journal of Bioresources (OJB) and the BRIF metrics.

2. Citing bioresources: the CoBRA guideline

At present, bioresources are either cited in a confusing, heterogeneous way or they are not cited at all. The use of a bioresource in a research article is not retrievable systematically via PubMed or other bibliographic databases (Mabile et al 2013). Traceability and visibility of bioresources in scientific literature or in other (online) sources would highlight their use. By being properly cited, bioresource use would be valued and their sharing thus encouraged. The CoBRA guideline (Mabile et al 2013) was hence developed to standardise citation of bioresources in scientific articles in order to trace their use on the web. This was achieved through close collaboration between the BRIF journal editors' subgroup with scientific journal editors, the EQUATOR[4] (Enhancing the QUAlity and Transparency Of health Research) network and the research community managing and/or using bioresources. It recommends mainly that each individual bioresource used to perform a research work should be mentioned in the Method section and should be cited as an individual "reference [BIORESOURCE]" according to a delineated format, using a unique identifier when possible. The detailed recommendation is given by the CoBRA checklist reported on the EQUATOR's website[5].

CoBRA needs now to be implemented and points to the necessity of integrating scientific editorial policies in the loop using several strategies. One way to enforce CoBRA use in articles is to include it in instructions to reviewers as part of the checklist used to process manuscripts. A second way is to add CoBRA in the list of reporting guidelines that is usually part of the instructions to authors. We also aim to obtain recommendation by the International Committee of Medical Journal Editors (ICMJE). In any such case though, compliance to the guideline is not guaranteed unless it is strictly verified by either reviewers or editorial staff (or made mandatory).

[4] http://www.equator-network.org/

[5] http://www.equator-network.org/wp-content/uploads/2015/03/Cobra-check-list.pdf

Associations of editors such as the European Association of Science Editors (EASE[6]) are of great help in reaching and empowering journal editors and authors of scientific publications. EASE Guidelines for authors and translators of articles to be published in English already include the necessity to mention in the methods section the origin and identity of experimental materials used and refer to the CoBRA guideline. The more key associations or committees of scientific journals editors will be aware of CoBRA, the more it will be applied. There is a need to go beyond the European dimension. Worldwide asociations such as WAME (World Association of Medical Editors), AMERBAC, Canadian Editors Association and CSE (Council of Science Editors) must be informed of the existence of CoBRA and should promote it.

Other stakeholders are also key players in developing good practices and could contribute to the implementation of CoBRA. Institutions hosting bioresources as well as funding agencies can guide researchers in good reporting of bioresource use. In France, the National Institute of Science and Techniques Information (INIST[7] - CNRS) has been a great support in disseminating and promoting the guideline. The European Research Infrastructure of biobanking and biomolecular resources (BBMRI-ERIC[8]) has actively supported the BRIF initiative and included it in its 2015-2016 workplan to facilitate notably the implementation of CoBRA among its members. It will be added to the MTA/DTA and specified in publication policies. Other infrastructures could be interested in helping implementing CoBRA as one of the tools of their own strategy. As a matter of fact, "Research infrastructures in the biological and medical thematic area of the European Strategy Forum on Research Infrastructures (ESFRI[9]) roadmap are committed to provide access to the most advanced, unique, and large-scale biological resources, instruments and expertise in Europe to support research and development in all life sciences." On a global scale, consortia or scientific societies such as the Public Population Project in genomics and society (P3G[10]), the International Society for Biological and Environmental Repositories (ISBER[11]) and the European, Middle Eastern & African Society for Biopreservation & Biobanking (ESBB[12]) would help in extending these actions. Patient's associations could have a role in this too. Contributors to bioresources give importance to the fact that they are used and not sleeping resources. Thus accessing data on the use of such resources would be valuable for them too.

Over the last years, other initiatives throughout the world have flourished within the open access and sharing move to better identify and trace different types of resources (OpenAire, DataCite, CODATA, Force 11, ORCID and others). Among them, Research Data Alliance Working Group on Dynamic Data Citation has provided recommendations about making subsets of data citable. Connecting to these groups would certainly facilitate CoBRA implementation and foster a better granularity by using suitable identifiers. Such identifiers of subsets or combination of subsets of bioresources must first be worked out with the idea of keeping traceable their "genealogy" (origin of parental resources). In general, coordination of all these actions has become an urge if one wishes to develop standard citation tools and improve good reporting practices.

[6] http://www.ease.org.uk

[7] http://www.inist.fr/

[8] http://bbmri-eric.eu/

[9] http://eu-openscreen.eu/index.php?id=130

[10] http://www.p3g.org/

[11] http://www.isber.org/

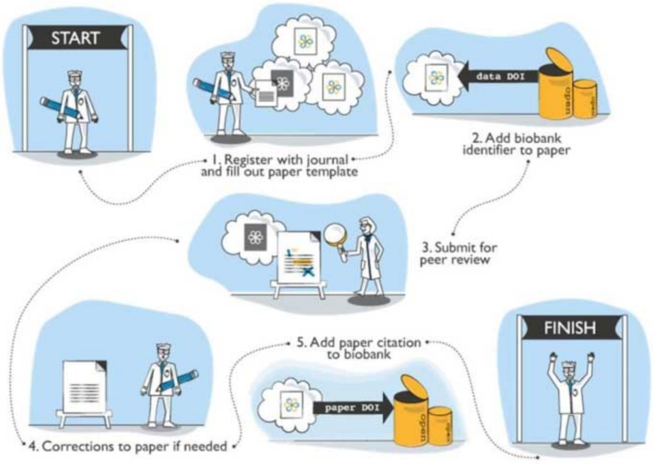
[12] http://www.esbb.org/

3. Publishing a bioresource: a new type of journal

The *Open Journal of Bioresources* is one journal in a suite of so-called ‘metajournals’ published by Ubiquity Press. These journals are dedicated to opening up and aiding the discoverability of all research elements involved in the research lifecycle, such as data, software, bioresources and hardware (forthcoming). The idea behind the metajournals is that researchers need to be able to discover and cite these research elements, but they also want credit for sharing them and the ability to track their impact. Given this, the metajournals offer credit – in the form of citation and altmetric data – for researchers making their resources permanently available and discoverable in accordance with community norms.

In the case of bioresources, the idea behind this journal is to provide a permanent marker paper so that users can definitively cite a bioresource they have accessed or referred to. The best way to do this is by integrating the bioresource into the traditional process for obtaining scholarly credit: the peer-reviewed journal article. In this way, users simply cite the bioresource as they would do any other journal article – and this is facilitated by the application of a digital object identifier (DOI) to all articles. This means that each article acts as a permanent marker for a bioresource and conforms to the standard processes for citing research.

OJB publishes bioresource papers, which are structured summaries of bioresources that are peer-reviewed to ensure they are accurately described. Papers are published in accordance with a structured template that describes the bioresource, outlines how it is preserved, the methods used in its creation, and how it can be accessed in the biobank.


START
data DOI
open
1. Register with journal and fill out paper template
2. Add biobank identifier to paper
3. Submit for peer review
FINISH
5. Add paper citation to biobank
paper DOI
open
4. Corrections to paper if needed

These papers are not lengthy descriptions of bioresources but more akin to a short online form. Contents are therefore structured not by paragraphs, but by individual sentences and one-word answers. The result is a highly structured, objective description of a bioresource.

Because the bioresource paper is an objective description, so too is the peer review process. Importantly, OJB papers are not peer reviewed for their significance but rather that the information is accurately filled out and presented in accordance with the standards set by the CoBRA guidelines (see above). Because of this, the peer review process is relatively quick and articles can be published within a matter of weeks from submission. Articles are published open-access under the CC BY licence, ensuring anyone can access the final contents. For this, the journal charges a small APC of £100 – which is completely waivable if an author does not have access to funding for publication fees.

The published article then becomes a permanent marker paper for the described bioresource. Users cite the paper directly when they have accessed, used or simply referenced a bioresource. Citations are tracked and displayed on the article page alongside numbers of article views, tweets and Facebook likes. In this way, the bioresource paper allows authors to understand the true impact of their bioresource, which would not have been possible previously.

Articles are also sent to various scholarly indexes to aid discoverability, ensuring they become part of the permanent scholarly record. We have also been in discussion with PubMed about indexing articles there – which we're confident will happen in the future.

4. Towards a new metrics: the BRIFs

Once the bioresource is fully traceable and indexed, the impact of its use could be measured using the metrics tools offered on the net. Those tools are mainly based on citation indexes and assume that citation reflects the 'success' of the enterprise. But in the case of bioresources this is not sufficient. They do not reflect the full range of utility of a bioresource. For example, a clinical and biological collection of rare diseases will be used by a restricted community, whereas the resource has a high value, requiring a worldwide coordination effort and the contribution of different stakeholders. Other metrics are needed that take relevant parameters into consideration.

As part of the BRIF initiative, a dedicated working subgroup worked out this issue and provided a first list of relevant parameters to take into account in the impact measure. An online survey was sent to selected biobanks in order to assess those parameters in the evaluation of the impact of a bioresource. The answers from 28 biobanks (mainly from Italy and France) were used to classify parameters of scientific impact for bioresources. Several groups of parameters were defined according to their availability and to the feasibility of their retrieving for calculating the impact using one or several specifically designed algorithm(s). The main parameters relate to indicators of research productivity and sustainability; indicators of sample/data value; indicators of workflow and efficiency and indicators of collaboration and visibility. An extended study on various types of bioresources in more countries will allow refining the list and characteristics of such parameters.

On the basis of the selected parameters an algorithm will be proposed in close collaboration with BBMRI-ERIC IT service for measuring the use and impact of bioresources. It will be tested in the wider context of European biobanks covered by the National Nodes of this European research infrastructure. A major step in this process is the proper identification of bioresources, including the physical resources; this point is presently being discussed between BBMRI-ERIC and DataCite.

The tools proposed here foster easier access to samples and associated data, their optimized use and sharing as well as the recognition of data producers. Input from the scientific editorial community would be highly relevant at this stage. This work could benefit from initiatives in other domains, in particular the long standing work performed in astronomy to provide mechanisms for quoting astronomical databases [5] and could serve as a reference for other communities, beyond human biological and medical bioresources.

Ackowledgements

We wish to thank Federica Napolitani, Anna Maria Rossi, Alessia Calzolari from Istituto Superiore di Sanità, Rome Italy, as well as Petr Holub, Jan Eric Litton, Outi Törnwall and Michaela Mayrhofer from BBMRI-ERIC, Graz, Austria for very fruitful discussions and constructive interactions.

The BRIF initiative was supported by funds from BBMRI-LPC project, grant agreement number 313010, theme FP7-–INFRA-2012-1.

References

Cambon-Thomsen, A., Carpenter, J., Dagher, G., Dalgleish, R., Deschênes, M., di Donato, J.H., Filocamo, M., Goldberg, M., Hewitt, R., Hofman, P., Kauffmann, F., Leitsalu, L., Lomba, I., Mabile, L., Melegh, B., Metspalu, A., Miranda, L., Napolitani, F., Oestergaard, M.Z., Parodi, B., Pasterk, M., Reiche, A., Rial-Sebbag, E., Rivalle, G., Rochaix, P., Susbielle, G., Tarasova, L., Thomsen, M., Thorisson, G.A., Zawati, M.H., Zins, M.; BRIF workshop group (2011) 'The role of a Bioresource Research Impact Factor as an incentive to share human bioresources', Nature Genetics, 43 (6), p. 503-504.

Mabile, L., Dalgleish, R., Thorisson, G.A., Deschênes, M., Hewitt, R., Carpenter, J., Bravo, E., Filocamo, M., Gourraud, P.A., Harris, J.R., Hofman, P., Kauffmann, F., Muñoz-Fernàndez, M.A., Pasterk, M., Cambon-Thomsen, A.; BRIF working group (2013) 'Quantifying the use of bioresources for promoting their sharing in scientific research', GigaScience, 2 (7), p. 1-8. Available: http://www.gigasciencejournal.com/content/2/1/7 [7 Mar 2016]

Bravo, E., Calzolari, A., De Castro, P., Mabile, L., Napolitani, F., Rossi, A.M. and Cambon-Thomsen, A. (2015) 'Developing a guideline to standardize the citation of bioresources in journal articles (CoBRA)', BMC Medicine, 13 (33), p. 1-12. doi: 10.1186/s12916-015-0266-y. Available: http://www.biomedcentral.com/1741-7015/13/33 [7 Mar 2016] [4] http://openbioresources.metajnl.com/

Uhlir, Paul F. (2012) For Attribution - Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop, Washington, D.C: The National Academies Press.

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-75

Open Access, Open Science, Open Society

Thomas MARGONI[a,1], Roberto CASO[b], Rossana DUCATO[b], Paolo GUARDA[b], and Valentina MOSCON[b]
[a] *School of Law – University of Stirling*
[b] *Faculty of Law – LawTech Group - University of Trento*

Abstract. Open Access' main goal is not the subversion of publishers' role as driving actors in an oligopolistic market characterized by reduced competition and higher prices. OA's main function is to be found somewhere else, namely in the ability to subvert the power to control science's governance and its future directions (Open Science), a power that is more often found within the academic institutions rather than outside. By decentralizing and opening-up not just the way in which scholarship is published but also the way in which it is assessed, OA removes the barriers that helped turn science into an intellectual oligopoly even before an economic one. The goal of this paper is to demonstrate that Open Access is a key enabler of Open Science, which in turn will lead to a more Open Society. Furthermore, the paper argues that while legislative interventions play an important role in the top-down regulation of Open Access, legislators currently lack an informed and systematic vision on the role of Open Access in science and society. In this historical phase, other complementary forms of intervention (bottom-up) appear much more "informed" and effective. This paper, which intends to set the stage for future research, identifies a few pieces of the puzzle: the relationship between formal and informal norms in the field of Open Science and how this impact on intellectual property rights, the protection of personal data, the assessment of science and the technology employed for the communication of science.

Keywords. Open Science – Open Access – Intellectual Property – Copyright – Privacy and Data Protection – Law and Technology – Comparative Law

1. Open Access, Science and Society

Open Access (OA) is a term that in recent years has acquired popularity and widespread recognition (Willinsky, 2006; Suber, 2012; Frosio 2014). International definitions and scholarly analysis converge on OA main characteristics: free availability on the public internet, permission to any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited (BOAI, 2002; Bethesda Statement on OA, 2003; Berlin Declaration on OA, 2003). Suber defines OA as digital, online, free of charge, and free of most copyright and licensing restrictions. (Suber, 2012). However, while OA main features appear more or less

[1] Corresponding Author: thomas.margoni@stir.ac.uk

known to many, its real function is often overlooked (Guédon, 2001). Open Access' main goal is not the subversion of publishers' role as driving actors in an oligopolistic market characterised by reduced competition and higher prices. Of course, an open and competitive market should certainly be seen with favour by economists and also by the community of scholars and the society at large, as this is usually synonym of faster innovation and better conditions for consumers --a larger consumer surplus, economists would say (Shavell, 2010; Ramello, 2011). Nevertheless, OA's main function is to be found somewhere else, namely in the ability to subvert the power to control science's governance and its future directions, a power that is more often found within the academic institutions rather than outside. By decentralising and opening-up not just the way in which scholarship is published but also the way in which it is assessed, OA removes the barriers that helped turn science into an intellectual oligopoly even before an economic one.

What is more, science is not only a key component of many social organisations, but can be seen as a form of social organisation in its own right (Merton, 1942; Polanyi, 1962). Therefore, changing those mechanisms that have explicitly or implicitly governed science and scientific institutions over the last few decades towards a more transparent and accountable model, will contribute to advance science in a more open, collaborative, democratic, and transparent system. This will in turn contribute to reach a more open, collaborative and transparent society (Bucchi, 2004). Consequently, the main argument presented in this paper, which sets the stage for future work, is that OA is not just an academic or scientific phenomenon, but is one that affects science in general and therefore society. Stronger Open Access will empower a thriving Open Science, which will enable a wealthier Open Society (Fecher and Friesike, 2013).

This thesis is followed by a logic corollary. Precisely within the rules and dynamics of a more open paradigm for science and society can be found the normative guidance that can help to reform the tools that regulate academic and scientific outputs: intellectual property, privacy and data protection, rules on scientific assessment and the role of technology.

The scope of this paper is limited to only some of the pieces of this intricate puzzle and accordingly attention is paid only to some of the legal aspects of Open Science policy: legislation on Open Access, Text and Data Mining and data protection.

The structure of the paper is as follows. After this general introduction (1), the main function of OA will be discussed under the assumption that OA is not only about scientific publications. On the contrary, the promotion of a model based on the wide availability of knowledge and on a decentralised system of scientific assessment will directly impact the way we imagine not only science but society in general (2). This theoretical analysis is followed by a survey of the main legislative and policy initiatives and by a brief discussion of how these initiatives have contributed (or not) to the achievement of OA/OS goals (3). New areas of scientific analysis where OA principles are in high demand such as in the field of data and databases will be presented in relation to Text and Data Mining (4), as well as in relation to the creation and use of databases and the dissemination of results containing personal data (5). In the last chapter brief conclusions and future work are identified (6).

2. Open Science: the unfinished revolution

Open Science – i.e. the abstraction and general implementation of basic OA principles such as sharing, cooperation, democracy and transparency to the entire scientific field – is an unfinished revolution. Despite a large consent on the benefits of Open Science (OS) in terms of progress of knowledge, innovation, pluralism, transparency and preservation, the most part of scientific results and publications is under the "control" of traditional closed access publishers who base their business models on vast commercial databases protected by intellectual property (IP), contracts and technological protection measures (Björk, 2013).

Moreover, the oligopolistic power of commercial publishers is much stronger than before the digital age (Larivière et al., 2015). The most important reason for the marginal impact in quantitative terms of OS is likely linked to the phenomenon of commodification of scientific and academic research, which has characterised the last forty years (Radder, 2010). However, shaping scientific and academic research on pure market logics has many side effects. Amongst the most relevant is conceiving competition as a value in itself. For example the "publish or perish" logic, strengthened by bibliometrics, imposes on scientists a mentality shift that often privileges quantity and impact factor over quality and impact on society. According to this logic, publications are no longer expressions of critical thinking but commercial "products" (Pievatolo, 2015). Naturally, this form of hyper-competitive science reflects a system of power: referees, members of editorial boards, learned societies, commercial publishers and bibliometrics databases (e.g., ISI Web of Science and Scopus), universities, national agencies for quality assurance in higher education; all act under the control or at least the influence of the same market logic that sees science as a product. To illustrate this point with one example ex pluris, we can look at the fact that often the scientific achievements of a department are assessed also in the light of the number of patents that the department was able to secure. This is done on the assumption that more patents are always synonym of more or better innovation. While in many cases this is certainly true, a large amount of literature is emerging which demonstrates that there are extreme variations in the correctness of this assumption depending not only on the scientific field but also on the nature and structure of the patentee (Lemley, 2008). The main problem here is that the equation "more-patents-more-innovation" was applied to the academic field in total absence of any sound analysis of the economic and funding structure of these institutions, nor was it supported by any serious empirical data. This is a direct effect of assuming – i.e. not proving – that a pure market system of incentives would work smoothly in the field of scientific and academic research, which is only partially moved by market incentives. As a result, many university patents are not effectively used, representing a cost for the institution and a barrier for other researchers.

As a matter of fact, science is not only a competitive game, it is also, sometime mostly, a collaborative one, where standard market incentives are only partially valued. In particular, OS is essentially based on collaborative action. In an OS model, the Mertonian norms of "communalism", "disinterestedness" and "organised scepticisms" are not only present "by design" but also enhanced by digital technologies. Illustratively, institutional and disciplinary OA repositories based on a common interoperable standard (Open Access Initiative-Public Metadata Harvesting) feature a great example of the interaction between the Mertonian scientific norms and the use of technology.

Until recently, OS has been driven by a bottom-up approach based on technological infrastructures and solemn declarations such as Budapest, Bethesda and Berlin declarations; but in the last years we are facing a new top-down approach based on legislative tools (de Roman Perez, 2012; Caso, 2013; Moscon, 2015b; Guibault, 2015a, 2015b; Visser 2015; Todolí Signes, 2015; see paragraph 3). This mix of bottom-up and top-down initiatives can be particularly effective. Nevertheless, especially in the case of top-down initiatives, legislators have often showed a lack of systematic view, which caused their interventions to lack real effectiveness. If we want to make science really open we have to study with more attention the interaction between social norms (and ethics), legal rules and technology. Without a new scientific thrust centred on cooperation, OS will remain an unfinished revolution. From this perspective we have to deeply rethink IP and copyright (Reichman, Okediji, 2012), the assessment and the technological infrastructures of science. Furthermore, we also need to rethink the education of scientists and lawyers putting at the centre of undergraduate and PhD programs a critical perspective on IP, assessment of science and technology. Mertonian CUDOS can be seen as a set of normative elements – already clearly present in OA – where to start from.

3. Open Access legislative and policy interventions

Recent empirical studies have shown that the implementation of OA policies varies by country and discipline (Migheli and Ramello, 2014; Eger et al., 2013). While one of the difficulties in unfolding the full potential of OA can be found in the hostility found in traditional publishers towards the OA paradigm, obstacles to OA publishing are present within the scientific community itself. This is largely due to the aforementioned commodification phenomenon (Radder, 2010).

A bottom-up approach based on ethical rules and social norms is likely the key element in guaranteeing success and future viability to OA (Lametti 2010; Geiger, 2013). However, a top-down complementary intervention may play an equally important role in addressing cultural and social change towards a broad dissemination of, and access to, research outputs (Reichman and Okediji, 2012; Priest, 2012). Within top-down approaches we can distinguish between institutional policies and legislative interventions. Institutional policies are adopted by research and funding bodies in accordance with organisational and regulatory choices and are crucial in promoting OA. Various options have emerged and prima facie institutional policies can be grouped into two main categories: voluntary and mandatory (Suber, 2012). The first category provides recommendations encouraging university departments to publish or re-publish in OA according to the gold or green road (Harnad et al., 2004)

Mandatory policies require the publication in OA following the green or gold road. In particular, the gold road may be more problematic as it is usually costly, requiring the payment of Article Processing Charge (APC), at least when Gold OA is combined with an author's pays business model. A distortion of this model is emerging as hybrid OA publishing, that is to say, traditional journals that offer the author of a given article the possibility to "buy back" the right to OA (Adams, 2007; Bjork, 2012).

In legal systems that encourage publication in gold OA such as the UK the institutional policies provide for specific funding mechanisms for OA publications. Gold OA funding was recently discussed at the Berlin Conference on the reorganisation of funding models for scholarly journals[2]. A process was initiated there to transform subscription journals into Open Access. The key element in this discussion is strictly connected with the scientific institutions and their sponsors' policies: public resources that are currently spent on journal subscriptions could be converted into open-access publishing funds with clear savings for Universities libraries.

Yet, mandatory green OA institutional policies are subordinated to the author's ownership of copyright. Given the weakness of the author in the contractual bargaining with publishers (especially when the author has to publish in specific high impact journals for assessment purposes) often authors will have transferred the right to (OA) publish. An example of an extra EU policy that found a solution to this problem can be seen in the model adopted by Harvard University. Harvard's OA policy introduced a legal mechanism through which, at the start of the publishing process, the university is automatically considered the non-exclusive licensee of the right to archive and publicly distribute all faculty-produced scholarly articles (Priest, 2012).

Moving the analysis to legislative interventions, some European governments have taken steps towards proper recognition of OA principles through the approval of specific Acts (i.e. Spain, Artículo 37 "Difusión en acceso abierto", Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación; Italy, § 4, Law October 7 2013, no. 112; Germany, Law October 1 2013 (BGBl. I S. 3714) amending Article 38 Copyright Act; Netherlands, Law June 30, no. 257 amending Article 25fa Copyright Act). Since 2006, the European Commission favours OA to publications and scientific data. The EU Commission requires that research funded by at least 50% with its money (i.e. FP7 and Horizon 2020 framework programs) be published in OA and has recently developed a pilot that covers also data. The EU also encourages Member States to take measures aimed at promoting Open Access as witnessed by the EU Communication "Towards better access to scientific information: Boosting the benefits of public investments in research" COM (2012) 401, and by the Commission Recommendation on "Access to and preservation of scientific information" (2012/417/EU) of 17 July 2012. The European approach promotes a multilayer system involving lawmakers, national legislatures, funding bodies and research entities that manage public funds.

Interesting national implementations can be seen in Spain where the legislature implemented Art. 37 (Difusión en acceso abierto) of Law 14/2011, of 1st of June "de la Ciencia, la Tecnología y la Innovación" (de Roman Perez, 2012; Todolí Signes, 2015). The scope of the rule is limited to serial or periodical publications and requires research that is more than 50% state-funded to be published as soon as possible – no later than 12 months after the first publication – in the form of the final version accepted for publication in an open-access disciplinary or institutional repository (Green Road). It is worth mentioning that the version of the publication which is republished in open- access repositories is available for consideration in the evaluation procedures of public administration. The main limit of this provision is that it – explicitly – does not override agreements that transfer to third parties the rights on the publication. A similar approach was adopted by the Italian legislature in Law of 7 October 2013, n. 112, G.U. n. 236, 8.10.2013). The Act seeks to bring Italian law in line with the aforementioned EU Recommendation, by requiring that all the subjects involved "implement the necessary measures for the promotion of Open Access" with regard to works publicly funded (at least 50%) and published in periodical collections

[2] http://openaccess.mpg.de/Berlin-Conferences

(at least biannually). The new Act requires research institutions to adopt policies that promote OA by following both the gold road and the green road. Similarly to the Spanish example, the new Italian law does not address the issue of IP rights. Consequently, authors may find themselves in the need of assigning their copyright thereby losing the power to determine how their research will be published (Caso, 2013; Moscon 2015b).

A completely different approach can be seen in the "German model" which was source of inspiration also to the Dutch Legislator (Guibault, 2015a; 2015b; Visser, 2015). The Law of 1 October 2013, amending Section 38 of the German Copyright Act (Urheberrechtsgesetz—UrhG) aims to remove one of the main obstacles to OA, i.e. the loss of the right to republish the work as a consequence of assigning the copyright to the publisher. The new law allows the author of a scientific work, published in a periodical collection (at least biannually) and created in the context of a research activity that "was at least 50% publicly funded", to make his work publicly available for non-commercial purposes 12 months after the publication. The provision is mandatory and cannot be limited by contract. Whether rules on conflict of laws, i.e. to say whether a publishing agreement between a publisher and an author which contains a choice of law provision excluding the applicability of the national OA provision, can constitute a quick and viable circumvention of said provision is not certain; But this hardly could have been the intention of the legislator (Guibault, 2015b).

4. Open Access, Text and Data Mining and the benefits for science and society

Text and Data Mining (TDM) is the process of extracting (new) information from newly created or already existing knowledge. The process of information extraction is performed using automated statistical analysis tools. In particular, TDM is emerging as a powerful tool "for harnessing the power in data by analysing datasets and content at multiple levels" in order to discover concepts and entities in the world, patterns they may follow and relations they engage and on this basis annotate, index, classify and visualise such content (OpenMinTeD, 2015). From a legal standpoint, it is important to note that these datasets and content (e.g. data, alphabetic or numerical entries, texts, articles, papers, collections of words such as vocabularies and corpora, databases) can receive different types of protection. Firstly, there is copyright, usually protecting the single elements of the database when these are original works of authorship (e.g. scientific papers, drawings, images). Secondly, the sui generis database right (SGDR) on databases that were made thanks to a "substantial investment" (Bently and Sherman, 2014; Derclaye, 2008; Wiebe and Guibault, 2013). As a matter of fact, copyright could also protect the database as such, but this is only possible when the database structure (the selection or arrangement of contents) is original in the sense of the author's own intellectual creation. This latter situation is not common for many databases in the scientific field and more importantly the scope of protection only extends to the structure of the database and not to its content. Therefore, for the purpose of most, if not all, TDM activities this form of protection is not relevant. What can represent a real barrier to TDM are the two other forms of protection: copyright on the elements of the database and the SGDR on the database itself (Guibault and Margoni, 2015).

Copyright on the elements of a database (DB): copyright protects works of authorship such as scientific, literary or artistic works. Therefore, when a DB is composed by journal articles, original photographs, musical compositions, etc. these will most likely be protected by copyright. Other items such as sound recordings, non original photographs (only in some cases), broadcasts, performances, fixations of films (e.g. the audiovisual recordings of birds hatching in their natural environment) can constitute protected subject matter even though technically speaking these do not constitute works protected by copyright, but "other subject matter" protected by rights related to copyright, also known as neighbouring rights. Copyright prevents acts such as making copies (total or partial, permanent of temporal) and redistribution of those copies in verbatim or modified form in absence of authorisation. Neighbouring rights offer similar, though not identical, protection.

The SGDR is a peculiar EU form of protection for databases which are protected regardless of any originality. What is protected is the "substantial investment" in quantitative or qualitative terms that the maker of the database puts in it. This substantial investment can take the form of time, money, labour or any other resources spent in the making of a DB. Importantly, when talking about "making" the database, the substantial investment has to be in the obtaining, verification and presentation of the data and not in their creation (Hugenholtz and Davison, 2005). The extent to which scientific databases can be said to be constituted by created or obtained data is not clearly settled in case law. In particular, the dichotomy between creating and obtaining data is not necessarily solved at the epistemological level.

TDM often, if not always, requires the making of a usually temporal copy of the datasets or works to be mined. The EU legal framework sketched above has been drafted in an era when methods such as TDM were unknown. However, said framework is based on the assumption that authors deserve a high level of protection (InfoSoc Directive, Recital 9) which has led to the formulation of very broad definitions of protected rights (e.g. the right of reproduction regulated in Art. 2 InfoSoc) and to the creation of special rights such as the SGDR. On the contrary, the set of rules intended to balance this exclusivity has been drafted in very loose terms and accordingly exception and limitations to copyright and to the SGDR are exhaustively listed in the InfoSoc and Database directive, but are not made mandatory (except for Art. 5.1 InfoSoc). The resulting situation, which has been referred to as an "accident" (Copyright Society Opinion 2014), is one were, at least in the EU, TDM is an act that most likely infringes copyright and/or the SGDR, absent a specific nationally implemented exception (to date only the UK has created a TDM exception limited to non-commercial purposes). Contrast this situation to countries such as the US, where TDM and web-mining have been held to be a transformative use covered by fair use (Authors Guild, Inc. v. Google Inc., 954 F. Supp. 2D 282 (S.D.N.Y. 2013), Aff'd 2015 2d Circuit). Other countries such as Japan have likewise clarified the legitimacy of this technology (Guibault and Margoni 2015). Unfortunately, the EU, despite general declarations, seems to be falling behind in this strategical field of science and technology.

Consequently, given the likely – but not certain – presence of the aforementioned forms of protection, content and databases to be TDM have to be licensed under licenses capable of addressing the identified rights. In fact, when those rights are present, the default situation is that of "all rights reserved" and even if the database is publicly available on the Internet acts such as reproduction and distribution are not permitted, unless of course specific exceptions and limitations to copyright apply. Currently, most exceptions to copyright and to the SGDR under EU law are not fit to fully cover the needs of TDM. Furthermore, as it is known, of the 21 exceptions listed in Art. 5 InfoSoc only 1 is mandatory, while the remaining 20 are implemented at the discretion of each of the 28 European Member States. This situation is clearly unsatisfactory in terms of legal certainty and even though some countries (such as the UK) have shown foresight by creating a dedicated TDM exception the presence of a non-commercial limitation still represents a competitive barrier if compared to other more dynamic legal systems (e.g. the US).

Licences such as the Creative Commons Public License (CCPL) version 4 are a technically viable alternative to the lack of proper legislative intervention in this field. CCPLv4 addresses both copyright and SGDR in the licensed work. In particular, by applying a CCPL 4.0 to a DB such as a website or a repository of journal articles the licensor (the person who applies the licence and who needs to be the right holder or be authorized by the right holder to do so) is giving permission to reuse: a) the SGDR in the database; b) copyright in the DB in the limited cases in which copyright applies to the DB structure; and c) copyright and/or related rights in the elements (works such as journal articles and original photographs) composing the DB.

While other open content licenses may also achieve the same results, the convergence towards one, or a few, licenses that can be seen as a de facto standard is not only desirable but also essential in order to lower the transitive costs associated with license compatibility and therefore to facilitate use and reuse of resources for goals such as TDM.

5. Open Science and Data Protection: specific v. any purpose?

To facilitate the appropriate understanding and study of OS, it is crucial to take into account the rules stated by data protection regulations: a research study, a scientific paper or any product of scientific investigation (i.e., databases, slides, blog, etc.) may contain personal (i.e. any information relating to an identified or identifiable natural person) or even sensitive (i.e. data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life) information.

In the field of data protection, the European reference framework is that of Article 8 of the Chart of Fundamental Right of the EU (recognizing the protection of personal data as an autonomous fundamental right) Article 16 of the TFEU (Treaty on the Functioning of the European Union), and Directive 95/46/EC (Data Protection Directive, hereinafter: DP Directive) (Bygrave, 2014)[3]. As known, a General Data Protection Regulation ("GDPR") has been recently approved and will replace the DP Directive, updating the European privacy rules to the digital era and overcoming the existing fragmentation in the application of data protection law across the EU member states (De Hert and Papakonstantinou, 2012)[4].

For the purpose of this paper, we will take into account two phases in the data processing cycle. Firstly, the phase of collection and use of personal data. At this stage,

[3] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Official Journal L 281, 23/11/1995 p. 31 - 50). The European data protection framework is complemented by Directive 2002/58/EC on privacy and electronic communication.

[4] Pending the drafting of this paper, the European legislative process has arrived to its final stage. The agreement on the final text of the Regulation has been reached on December 2015, therefore any reference to the European Regulation in this paper shall be construed as referring to the consolidated text available at: http://static.ow.ly/docs/Regulation_consolidated_text_EN_47uW.pdf.

the fundamental legal tool is formed by the combination of two concepts: consent (Article 7.a, DP Directive) and the information to be given to the data subject (Articles 10-11, DP Directive). In particular, the latter (in addition to the elements set out in Article 10) must indicate the purposes of the processing for which the data are intended, in conformity with the principle of the "specific purpose", within the meaning of Article 6.1.b, according to which data must be: " collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes". Thus, at the time of recording personal data, the controller must obtain a specific and informed consent for the purposes for which the processing is intended.

However, the DP Directive states a very important principle in our context by making a general presumption of compatibility of the research purposes with any previous processing: "Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards" (Article 6.1.b, DP Directive). This means that in case of "secondary uses" for research purposes, the processing is presumed to comply with the principles enshrined in the European legal framework. In this context national legislators shall furnish suitable safeguards. This setting is also confirmed in the GDPR (Recital 40).

The second important phase of the processing is represented by the dissemination of research results containing personal data. In this case as well, the detailed operational rules and procedures applicable are determined by the Member States, as established by DP Directive (Article 13.2). For example, according to Italian law, which explicitly recalls the recommendations of the Council of Europe[5], research results shall be published or otherwise disseminated only as aggregate data or in ways that the data subject cannot be identified. Furthermore, sensitive data processed for research purposes has to be anonymised. The GDPR underlies the crucial role of research results, especially in the medical and life sciences field (see Recital 125aa). However, the provision regarding the processing of personal data for scientific, historical and statistical purposes has been radically changed during the trilogue's meetings. In the proposal made by the EU Commission in 2012, Article 83 contained a specific regulation on the publication of personal data for research purposes, while the consolidated text now entrusts the adoption of specific safeguards to Member States and Union law. Therefore, in this sensitive sector the unifying purpose of the Regulation is likely to have missed an important opportunity.

In the light of this investigation, the most interesting legal issue concerns the possible clash between the different purposes of the processing, on the one hand, and the circulation of content governed by an OA license, on the other hand. While in the privacy context the focus on the "specific purpose" principle of the processing forms the hub of the whole system of protection, the Open Access expressly stresses the ability to reuse data "for any purpose".

[5] Ex pluris, Council of Europe, Committee of Ministers, Recommendation No. R (83)10 on theprotection of personal data used for scientific research and statistics (Sept. 23, 1983); Recommendation No. R (92) 3 on genetic testing and screening for health care pursposes (Feb. 10, 1992); Recommendation No. R (97) 5 on the protection of medical data (Feb. 13, 1997); Recommendation No. R (97) 18 concerning the protection of personal data collected and processed for statistical purposes (Sept. 30, 1997).

6. Conclusions and future work

The goal of this paper is to demonstrate the fundamental relation between Open Access, science and society. Not only OA can influence scientific and social institutions towards a more open and transparent model, but a more open paradigm in science and society can offer the normative guidance needed to adjust some of the basic rules that regulate the Information Society: intellectual property, the protection of personal data, the assessment of scientific and academic outputs and the role of technology. Furthermore, it emerged that while legislative interventions play an important role in the top-down regulation of Open Access, legislators currently lack a general and systematic vision of the role of Open Access in science and society. In this historical phase, other complementary forms of intervention (bottom-up) appear much more "informed" and effective. Legislative interventions mandating the green road or conferring an unalienable right of publication to the author are useful instruments but only partially effective. These top-down interventions must be combined with bottom- up solutions such as institutional policies that mandate green road archiving. A particularly well drafted example of this latter policy can be found in the French INRIA institutional policy that requires to deposit in the French OA archive HAL the results of research, establishing that only the deposited articles will be considered for assessment[6].

Future work will investigate in more depth other pieces of the puzzle that this study has started to analyse. In particular, it is important to analyse the relationship that exists between formal rules and informal norms in the field of Open Science, intellectual property rules, personal data protection, the assessment and the communication of scientific and academic research.

References

Adams, A. (2007). Copyright and research: an archivangelist's perspective. SCRIPTED, 4(3), pp. 285-290.

Bently, L. and Sherman, B. (2014). Intellectual property law. Oxford: Oxford University Press.

Björk, B. (2013). Open Access—Are the Barriers to Change Receding?. Publications, 1(1).

Björk, B. (2012). The hybrid model for open access publication of scholarly articles: A failed experiment?. J Am Soc Inf Sci Tec, 63(8), pp.1496-1504.

Bucchi, M. (2004). Science in society. London: Routledge.

Budapestopenaccessinitiative.org, (2016). Budapest Open Access Initiative | Read the Budapest Open Access Initiative. [online] Available at: http:/www.budapestopenaccessinitiative.org/read [Accessed 3 Mar. 2016].

Bygrave, L. (2014). Data Privacy Law: An International Perspective. Oxford: OxfordUniversity Press.

Caso, R. (2013). La legge italiana sull'accesso aperto agli articoli scientifici: prime note comparatistiche. Il diritto dell'informazione e dell'informatica, 4, pp.681-702.

Derclaye, E. (2008). The legal protection of databases. Cheltenham, UK: Edward Elgar.

[6] https://iww.inria.fr/hal/aide/spip.php?article327&lang=fr

De Hert, P. and Papakonstantinou, V. (2012). The proposed data protection Regulation replacing Directive 95/46/EC: A sound system for the protection of individuals. Computer Law & Security Review, 28(2), pp.130-142.

De Román Pérez, R. (2012). Acceso abierto a los resultados de investigación del profesorado universitario en la Ley de la Ciencia. Diario La Ley, (7986).

Eger, T., Scheufen, M. and Meierrieks, D. (2013). The Determinants of Open Access Publishing: Survey Evidence from Germany.

SSRN Electronic Journal.European Copyright Society, (2014). Answer to the EC Consultation on the review of the EU copyright rules. [online] Available at: https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs_answer_to_ec_consultation_on_copyright_review.pdf [Accessed 3 Mar. 2016].

Fecher, B. and Friesike, S. (2013). Open Science: One Term, Five Schools of Thought. [online] Dx.doi.org. Available at: http://dx.doi.org/10.2139/ssrn.2272036 [Accessed 3 Mar. 2016].

Frosio, G. (2014). Open Access Publishing: A Literature Review. SSRN Electronic Journal.

Geiger, C. (2013). The social function of intellectual property rights, or how ethics can influence the shape and use of IP law. In: G. Dinwoodie, ed., Intellectual Property Law: Methods and Perspectives, 1st ed. Cheltenham: Edward Elgar, p.153.

Guédon, J. (2001). In oldenburg's long shadow. [Montreal]: Assoc Of Research.

Guibault, L. (2015a). Almost there! In Support of the Green Road to Dutch Science!. [Blog] Kluwer Copyright Blog. Available at: http://kluwercopyrightblog.com/2015/04/09/almost-there-in-support-of-the-green-road-to-dutch-science/ [Accessed 3 Mar. 2016].

Guibault, L. (2015b). Back on the Green Road: How Imperative are Imperative Rules?. [Blog] Kluwer Copyright Blog. Available at: http://kluwercopyrightblog.com/2015/04/19/back-on-the-green-road-how-imperative- are-imperative-rules/ [Accessed 3 Mar. 2016].

Guibault, L. and Margoni, T. (2015). Legal Aspects of Open Access to Publicly Funded Research. In: OECD, ed., Enquiries into Intellectual Property's Economic Impact, 1st ed. [online] Available at: http://www.oecd.org/sti/ieconomy/KBC2- IP.Final.pdf [Accessed 3 Mar. 2016].

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohanns, H. and Hilf, E. (2004). The Access/Impact Problem and the Green and Gold Roads to Open Access. Serials Review, 30(4), pp.310-314.

Hugenholtz, P. and Davison, M. (2005). Football fixtures, horseraces and spinoffs: the ECJ domesticates the database right. EIPR, 3, pp.113–118.

Lametti, D. (2010). How virtue ethics might help erase C-32's conceptual incoherence. In: M. Geist, ed., From "Radical Extremism" to "Balanced Copyright": Canadian Copyright and the Digital Agenda, 1st ed. Toronto: Irwin Law, p.309.

Larivière, V., Haustein, S. and Mongeon, P. (2015). The Oligopoly of Academic Publishers in the Digital Era. PLOS ONE, 10(6), p.e0127502.

Legacy.earlham.edu, (2016). Bethesda Statement on Open Access Publishing. [online] Available at: http://legacy.earlham.edu/~peters/fos/bethesda.htm [Accessed 3 Mar.2016].

Merton, R. (1942). Science and Technology in a Democratic Order. Journal of Legal and Polotical Sociology, 1, p.115.

Migheli, M. and Ramello, G. (2014). Open Access Journals & Academicss Behaviour. SSRN Electronic Journal.

Moscon, V. (2015a). University Knowledge Transfer: From Fundamental Rights to Open Access within International Law. In: G. Bellantuono and T. De Rezende Lara, ed., Law, Development and Innovation, 1st ed. Berlin: Springer, pp.147 – 189.

Moscon, V. (2015b). Academic Freedom, Copyright, and Access to Scholarly Works: A Comparative Perspective. In: R. Caso and F. Giovanella, ed., Balancing Copyright Law in the Digital Age - Comparative Perspectives, 1st ed. Berlin: Springer, pp.99-135.

Openaccess.mpg.de, (2016). Berlin Declaration. [online] Available at: http://openaccess.mpg.de/Berlin-Declaration [Accessed 3 Mar. 2016].

OpenMinTeD, (2016). Home - OpenMinTeD. [online] Available at: http://openminted.eu/ [Accessed 3 Mar. 2016].

Pievatolo, M. (2015). Publishing without perishing. Are there such things as "research products"?. In: Aisa 1st annual conference Nostra res agitur: open science as a social question.

Polanyi, M. (1962). The Republic of science. Minerva, 1(1), pp.54-73.

Priest, E. (2012). Copyright and the Harvard Open Access Mandate. Northwestern Journal of Technology and Intellectual Property, 10, p.377.

Radder, H. (2010). The commodification of academic research. Pittsburgh Pa.: University of Pittsburgh Press

Ramello, G. (2010). Copyright & Endogenous Market Structure: A Glimpse from the Journal Publishing Market. Review of Economic Research on Copyright Issues, 7(1), pp.7 29

Reichman, H. and Okediji, R. (2012). When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale. Minnesota Law Review, 96(4).

Shavell, S. (2010). Should Copyright of Academic Works be Abolished?. Journal of Legal Analysis, 2(1), pp.301-358.

Suber, P. (2012). Open access. Cambridge, Massachusetts: MIT Press.

Todolí Signes A. (2015). El open access en la regulación española, in J.A. Altès Tárrega, Investigación, docencia universitaria y derechos de propriedad intelectual, Tirant lo blanch, Valencia, 119.

Visser, D. (2015). The Open Access provision in Dutch copyright contract law. Journal of Intellectual Property Law & Practice, 10(11), pp.872-878.

Wiebe, A. and Guibault, L. (2013). Safe to be open - Study on the protection of research data and recommendations for access and usage. Gottingen: Universitatsverlag.

Willinsky, J. (2006). The access principle. Cambridge, Mass.: MIT Press.

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-87

Jupyter Notebooks—a publishing format for reproducible computational workflows

Thomas KLUYVER[a,1], Benjamin RAGAN-KELLEY[b,1], Fernando PÉREZ[c], Brian GRANGER[d], Matthias BUSSONNIER[c], Jonathan FREDERIC[d], Kyle KELLEY[e], Jessica HAMRICK[c], Jason GROUT[f], Sylvain CORLAY[f], Paul IVANOV[g], Damián AVILA[h], Safia ABDALLA[i], Carol WILLING[d] and Jupyter Development Team[j]

[a] *University of Southampton, UK*
[b] *Simula Research Lab, Norway*
[c] *University of California, Berkeley, USA*
[d] *California Polytechnic State University, San Luis Obispo, USA*
[e] *Rackspace*
[f] *Bloomberg LP*
[g] *Disqus*
[h] *Continuum Analytics*
[i] *Project Jupyter*
[j] *Worldwide*

Abstract. It is increasingly necessary for researchers in all fields to write computer code, and in order to reproduce research results, it is important that this code is published. We present Jupyter notebooks, a document format for publishing code, results and explanations in a form that is both readable and executable. We discuss various tools and use cases for notebook documents.

Keywords. Notebook, reproducibility, research code

1. Introduction

Researchers today across all academic disciplines often need to write computer code in order to collect and process data, carry out statistical tests, run simulations or draw figures. The widely applicable libraries and tools for this are often developed as open source projects (such as NumPy, Julia, or FEniCS), but the specific code researchers write for a particular piece of work is often left unpublished, hindering reproducibility. Some authors may describe computational methods in prose, as part of a general description of research methods. But human language lacks the precision of code, and reproducing such methods is not as quick or as reliable as it should be. Others provide code separately as supplementary material, but it may be difficult for readers to cross reference between code and prose, and there is a risk that the two become inconsistent as the author works on them.

Notebooks—documents integrating prose, code and results—offer a way to publish a computational method which can be readily read and replicated.

[1] Corresponding Author.

2. Notebooks

Notebooks are designed to support the workflow of scientific computing, from interactive exploration to publishing a detailed record of computation. The code in a notebook is organised into cells, chunks which can be individually modified and run. The output from each cell appears directly below it and is stored as part of the document. This is an evolution of the interactive shell or REPL (read-evaluate-print loop) which has long been the basis of interactive programming (Iverson, 1962; Spence, 1975). However, whereas the direct output in most shells can only be text, notebooks can include rich output such as plots, formatted mathematical equations, and even interactive controls and graphics. Prose text can be interleaved with the code and output in a notebook to explain and highlight specific parts, forming a rich computational narrative.

The notebook interface first became popular among mathematicians. The proprietary computer algebra systems Mathematica and Maple both feature notebook interfaces, as does the open source SageMath.

Jupyter aims to bring notebooks to a broader audience. Jupyter is an open source project, which can work with code in many different programming languages. Different language backends, called kernels, communicate with Jupyter using a common, documented protocol; over 50 such backends have already been written, for languages ranging from C++ to Bash. Jupyter grew out of the IPython project (Pérez & Granger, 2007), which initially provided this interface only for the Python language. IPython continues to provide the canonical Python kernel for Jupyter.

The Jupyter Notebook is accessed through a modern web browser. This makes it practical to use the same interface running locally like a desktop application, or running on a remote server. In the latter case, the only software the user needs locally is a web browser; so, for instance, a teacher can set up the software on a server and easily give students access. The notebook files it creates are a simple, documented JSON format, with the extension '.ipynb'. It is simple to write other software tools which access and manipulate these files.

3. Sharing and reproducibility

Notebooks record a computation in order to explain it in detail to others, and a variety of tools help users to conveniently share notebooks. The Jupyter project includes *nbconvert*, which converts notebook files into a variety of file formats, including HTML, LaTeX and PDF, so that they are accessible without needing any Jupyter software installed. Nbconvert uses a powerful templating engine (Jinja), so the conversion process can be completely customised to produce different kinds of output.

Another Jupyter project, *nbviewer*, is a hosted web service built around nbconvert. nbviewer provides an HTML view of notebook files published anywhere on the web. The primary instance runs at https://nbviewer.jupyter.org/, but as it is open source, anyone can run their own instance—for example on an internal network, to view notebooks which should not be made public. These HTML views have a major advantage over publishing converted HTML directly: they link back to the notebook file, so interested readers can download it, run it and modify it themselves.

While nbconvert and nbviewer facilitate sharing statically rendered notebooks, a new project called *Binder* (http://mybinder.org/) enables sharing of live notebooks, including a computational environment in which users can execute the code. Authors can publish notebooks on GitHub along with an environment specification in one of a few common formats. By pointing the Binder web service at the repository, a temporary environment is automatically created with the notebooks and any libraries and data required to run them. This allows authors to publish their code in an interactive and immediately verifiable form.

Together, these tools allow the preservation and reuse of scientific code, the computational environment to run that code, and data within the size constraints of a git repository. Third party tools such as *noWorkflow* can integrate with this to track provenance: how inputs, code and generated files relate to one another. noWorkflow captures the execution of a marked notebook cell, or a script run through its command line tool, as a 'trial', recording in a database the code that was used, the environment in which it ran, the versions of modules that were used, and the files read and written.

4. Notebooks in academic publishing

Several papers have been published with supporting notebooks to reproduce the analysis, or the creation of key plots. The detection of gravitational waves by the LIGO experiment (LIGO Scientific Collaboration and Virgo Collaboration et al., 2016), announced earlier this year, is one such: the researchers posted a notebook on their website illustrating in detail how to filter and process the data to reveal the signature of a distant black hole merger (LIGO collaboration). Others quickly made this available through Binder, as described above (https://github.com/minrk/ligo-binder), allowing anyone to replicate the analysis even without downloading or installing anything. Other papers published in fields from geology to genetics to computer science have used notebooks as supporting material (e.g. Sylvester et al., 2013; Olson & Roberts, 2015; Brown et al., 2012).

Authors have also written books as a collection of IPython notebooks. Some of these have been published in hard copy (e.g. Unpingco, 2014; Davidson-Pilon, 2015; Rossant, 2014), but with the internet blurring traditional categorisations, similar collections of notebooks are being published purely online. Of these, course materials are a notable group, both to accompany teaching and for learners to work through independently (e.g. Caporaso; Barba; Johansson).

It is not yet very practical to write academic papers themselves as notebooks, but we are working towards this. One tricky point is inserting academic citations, which require structured data about sources to be formatted in a very precise way which may depend on the journal. One of us (TK) has an experimental plugin cite2c (https://github.com/takluyver/cite2c), which allows the author to search their reference library stored in the Zotero service, and insert citations into a Markdown cell. The citations and bibliography are rendered by the citeproc-js package (Bennett), using the common Citation Style Language format (http://citationstyles.org/).

Notebooks also fit well into novel publishing paradigms, such as post publication review. Digital objects such as GitHub repositories, which may contain notebooks, and blog posts, which may be made from notebooks, can now be archived and given permanent DOI references (GitHub; Yarkoni, 2015), making it practical to cite them in other publications. The Jupyter Project is part of the coalition around Hypothes.is, an open source tool to annotate documents on the web (Perkel, 2015; Hypothes.is, 2015). Finally, work is under way to support real-time collaboration in notebooks. This will let multiple authors work on a notebook together, with the changes instantly visible to all, reducing the chance of two people trying to change the same thing in different ways.

References

Barba, L.A. CFD Python: 12 Steps to Navier-Stokes, Available from: <http://lorenabarba.com/blog/cfd-python-12-steps-to-navier-stokes/> [Accessed: 4 March 2016]

Bennett, F. Citeproc-Js, Available from: <https://bitbucket.org/fbennett/citeproc-js> [Accessed: 4 March 2016]

Brown, C.T., Howe, A., Zhang, Q., Pyrkosz, A.B. & Brom, T.H. (2012) A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data, arXiv:1203.4802 [q-bio] Available from: <http://arxiv.org/abs/1203.4802> [Accessed: 4 March 2016]

Caporaso, G. An Introduction to Applied Bioinformatics, Available from: <http://readiab.org/> [Accessed: 4 March 2016]

Davidson-Pilon, C. (2015) Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference, New York: Addison-Wesley Professional GitHub Making Your Code Citable, Available from: <https://guides.github.com/activities/citable-code/> [Accessed: 4 March 2016]

Hypothes.is (2015) Annotating All Knowledge, Available from: <https://hypothes.is/annotating-all-knowledge/> [Accessed: 4 March 2016]

Iverson, K.E. (1962) A Programming Language, New York, NY, USA: John Wiley & Sons, Inc.

Johansson, R. QuTiP Lectures as IPython Notebooks, Available from: <https://github.com/jrjohansson/qutip-lectures> [Accessed: 4 March 2016]

LIGO collaboration Signal Processing with GW150914 Open Data, Available from: <https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.html> [Accessed: 4 March 2016]

LIGO Scientific Collaboration and Virgo Collaboration, Abbott, B.P., Abbott, R., Abbott, T.D., Abernathy, M.R., Acernese, F., et al. (2016) Observation of Gravitational Waves from a Binary Black Hole Merger, Physical Review Letters 116 (6): 061102

Olson, C.E. & Roberts, S.B. (2015) Indication of Family-Specific DNA Methylation Patterns in Developing Oysters, bioRxiv: 012831

Pérez, F. & Granger, B.E. (2007) IPython: A System for Interactive Scientific Computing, Computing in Science Engineering 9 (3): 21–29

Perkel, J.M. (2015) Annotating the Scholarly Web, Nature 528 (7580): 153

Rossant, C. (2014) IPython Interactive Computing and Visualization Cookbook, Packt Publishing Spence, R. (1975) APL Demonstration, Imperial College London Available from: <https://www.youtube.com/watch?v=_DTpQ4Kk2wA> [Accessed: 4 March 2016]

Sylvester, Z., Pirmez, C., Cantelli, A. & Jobe, Z.R. (2013) Global (latitudinal) Variation in Submarine Channel Sinuosity: COMMENT, Geology 41 (5): e287–e287

Unpingco, J. (2014) Python for Signal Processing, Springer

Yarkoni, T. (2015) Now I Am Become DOI, Destroyer of Gatekeeping Worlds, The Winnower Available from: <https://thewinnower.com/papers/282-now-i-am-become-doi-destroyer-of-gatekeeping-worlds> [Accessed: 4 March 2016]

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-91

Peer Review on the Move from Closed to Open

Birgit SCHMIDT[a,1], Arvid DEPPE[a], Julien BORDIER[b]
and Tony ROSS-HELLAUER[a]
[a] *University of Göttingen, State and University Library*
[b] *Open Edition*

Abstract. Openness in peer review is no longer a terra incognita. However, there remains a need for further experimentation and careful evaluation of its advantages and disadvantages in practice. OpenAIRE, the European digital infrastructure for Open Scholarship, offers a unique environment for such experiments. This paper describes the design and early results of three such experiments, which are currently under development in close collaboration with selected publishing and repository communities.

Keywords. Open peer review, eInfrastructures, open evaluation, peer commentary

1. Introduction

Open peer review (henceforth OPR) is no longer a terra incognita, with the first implementations and trials to explicitly categorize themselves as such emerging in the late 20th Century (van Rooyen *et al*, 1999). Indeed, some variation of OPR is now the established mode of peer review for many journals and publishers (Amsen, 2014). OPR is best defined in contradistinction to traditional or classical peer review. Traditional peer review is generally (1) *anonymous*, with either the reviewer unknown to the author (single-blind review) or both author and reviewer unknown to each other (double-blind review); (2) *selective*, with reviewers selected by editors; and (3) *opaque*, with neither the review process nor the reviews themselves made public. OPR, although often narrowly defined as peer review where author/reviewer identities are disclosed to one another (see e.g., Ford, 2015), is best understood as an umbrella term for a variety of innovative review methods that remove one or more of these conditions and thus add *transparency* to the peer review process. Hence, in our definition, 'openness' can refer to the absence of anonymity (*open identity*), self-selecting reviewers (*open participation*), public processes and reviews (*open access*), or some mixture of the three.

These elements are often complementary, and can be combined in various ways to produce a broad continuum of 'openness' in OPR. For example, some journals publish the entire multi-staged review process: the manuscript under review, the review reports and the authors' responses, and the revised manuscript(s), while inferring links between the earlier released version(s) and the final version of record (Pöschl, 2004; Pöschl, 2012; Sandewall, 2012; Ford, 2013; Walker and da Silva, 2015). Sometimes reviewers

[1] Corresponding Author. E-mail: bschmidt@sub.uni-goettingen.de.

themselves may decide how much information they would like to disclose during the review process (for a discussion of a wide range of examples see e.g. Walker and da Silva (2015)). Additionally, some journals open up the process to readers, allowing them to join the discussion of the paper through open peer commentary. *Table 1* gives an indicative (not necessarily exhaustive) overview of this continuum as it applies to various aspects of peer review variations of openness as currently implemented. (Note the table only takes into account the roles of author, reviewer, readers/commenters – journal editors typically moderate the review process and will continue to play an important role, e.g. in providing practical and ethical advice on open review processes).

Table 1. Options for openness in peer review processes.

Category	Fully open	Gradually open	Closed
Submitted manuscript	Published online as discussion paper	Available to reviewers with author names disclosed	Available to reviewers, author names blinded
Reviewer names	Publicly available on time of publication of reviews	Reviewer names are disclosed if they opt in	Reviewer names not disclosed
Access to review reports	Available to the public	Available to the author(s), reviewers may opt in to disclose reports to the public (blinded or non-blinded)	Only available to the author(s)
Release of review reports	Immediately available to the public (incl. the author(s))	Published after the review process is closed	Not published
Accepted vs. rejected papers	All review reports made available	Only for accepted papers	Not published

2. On Benefits, Biases and Limitations

Several research studies and reports from publishers setting up OPR processes have explored its benefits, possible biases and limitations. When authors and reviewers are asked about their preferences regarding peer review they continue to prefer the classical double-blind model (Taylor & Francis). However, such assessments may not be representative and some questions may be biased (Davis, 2015). Among the benefits reported by journal publishers who implemented OPR include more civil language, more thorough dialogue between authors and reviewers, better understanding of why the research was conducted and the decisions taken, and the use of review reports as educational tools and as case studies to provide guidance for reviewers (PeerJ, 2014). In addition, authors in transparent (open access) review "have a much higher incentive to maximize the quality of their manuscript prior to submission" and it also "prevents authors from abusing the peer-review process by delegating some of their own tasks and responsibilities to the referees during review and revision behind the scenes", where reviewers often make substantial contributions to the quality of the paper (Pöschl, 2004).

One issue often raised about OPR is accountability: Disclosing reviews and identities forces reviewers to stand openly by what they believe. According to Kowalczuk, this also makes reviews more constructive (Kowalczuk, 2015). Further, OPR is said to prevent abuse and reduce biases (e.g. reputation of author/institution, degree of conservatism / conformity, language, sex, age, against 'negative results', etc.) by its transparency and the wider engagement with the scientific community (e.g. Godlee, 2002; Smith, 2006; Perakakis, 2011).

Sometimes a higher quality[2] of review reports is expected (Prug, 2010; Boldt, 2011) but this does not seem generally result from openness (Vinther *et al. 2012*; Van Rooyen *et a.l*, 2010, Kowalczuk, 2015). Epistemologically, OPR and its traceability can strengthen the professional discourse and the scientific community as a whole and in particular the exchange between authors and reviewers (Ford, 2013) (see also the concept of 'extelligence' (Friedman *et al*., 2010)). Pragmatically, open review can prevent unnecessary duplication of effort in the sense that rejected papers' reviews can be reused if the paper is resubmitted to other journals (Hames, 2014).[3]

OPR, and in particular publishing review reports, also aims at raising the recognition and reward of the work of peer reviewers. Adding review activities to the reviewer's professional record is common practice; author identification systems currently also add mechanisms to host such information (e.g. via ORCID) (Hansen, 2016).

However, some of the benefits of open peer review may also be closely linked to possible pitfalls. Nobarany and Booth's findings indicate that politeness in reviewer - author communication can affect the clarity and effectiveness of criticism, and can turn out to make the process more time-consuming. They suggest that a careful approach should be taken based on respective community norms, in terms of politeness level but also through structured reports (which ask for pros and cons for the primary aspects of the submission) and a technical system that allows interactive discussion (Nobarany and Booth, 2015).

While OPR can reduce several biases, openness may present an obstacle for some reviewers – especially junior researchers – who might be reticent to publicly criticize more senior researchers in the field. This effect might be avoided by not disclosing reviewers' names if a paper is rejected (Pöschl, 2004). In the context of reviewing a special track of a computer science conference, Nobarany and Booth found "that less experienced researchers tended to express unmitigated criticism more often than did experienced researchers"; the authors could find no evidence that less experienced researchers avoided reviewing more experienced ones (Nobarany and Booth, 2015). Moreover, "reviewers tended to use more positive politeness strategies (e.g., compliments) towards less experienced authors" (Nobarany and Booth, 2015).

Furthermore, Blanes i Vidal and Leaver found that in settings where reviewers and reviewee share the same rank (in the studied case: the English Superior Courts), reviewers were reluctant to reverse the judgements of reviewees, in particular when a reviewer knows that he or she will soon work with the reviewee (Blanes i Vidal and Leaver, 2015). The authors conclude that to some degree this could be prevented through a change in the system of assignments. However, in very specialized disciplines where the community is small and interaction between reviewer and reviewee is likely, OPR might not be appropriate.

[2] The quality of peer review can be rated based on Van Rooyen *et al.*'s established Review Quality Instrument (RQI) (van Rooyen, 1999), applied and reproduced with permission in (Kowalczuk, 2015).

[3] It must be noted that this is also an option and actually implemented in traditional or mixed settings, e.g. some publishers offer authors of rejected papers the choice to resubmit their manuscript together with the referees' report to a different journal of the same publisher or within a disciplinary peer review consortium. Compare e.g. the Neuroscience Peer Review Consortium which brings together subscription-based and open access journals, http://nprc.incf.org/nprc-overview.

Open reviews can be considered as a new kind of publication. This allows reviewers’ contributions to be fully acknowledged in the final published paper (Godlee, 2002) (for an example see Ford (2015) who reviews four open peer review implementations at STM journals and cites two review reports). However, this incentive might not yet be particularly strong: Van Rooyen *et al.* found that “the rate of refusal of reviewers to participate in the study was high at 55%”. This reluctance might be due to anxieties related to public exposure and an expectation of an additional workload. Indeed, the study reported an “increase in the amount of time taken to write a review“, which was not the case for papers which were accepted directly but statistically significantly higher for papers which were eventually accepted (reviews of rejected papers were not published) (van Rooyen *et al*., 2010). Overall, authors seem to be less reluctant to participate in OPR than reviewers (80% vs. 40% for the journal PeerJ (2014), although this difference was found to be less pronounced by Taylor & Francis (2015)).

Table 2. Open peer review’s benefits and limitations

Category	Benefits	Limitations
Language used in the review report	More civil language	Less direct criticism, may result in lack of clarity
Efficiency of the review towards reviewees	- Polite language can help to maintain authors’ willingness to accept criticism. - Potential reuse of review reports in resubmissions to other journals.	- More time-intensive for reviewers and authors - Follow-up reviews might perpetuate existing (negative) judgements.
Education about peer review	Good and bad practice can be highlighted, case studies serve as advice	Exposure as bad example can cause embarrassment
Quality of submitted manuscripts	- Authors submit more mature manuscripts - Less abuse of the review process by delegating tasks or responsibilities to referees - Reviewers contributions to quality are acknowledged and made transparent	
Quality of review	- Potentially higher quality vis-á-vis a larger and public audience - quality can be directly assessed, e.g. based on the Review Quality Instrument (RQI)	- More politely phrased but in substance generally the same quality - In some cases a higher quality could be shown
Early career researchers	Visible engagement with community members	Undesirable exposure of communication of criticism
Senior career researchers	Sharing of experience through providing access to high-quality reviews	Undesirable exposure,
Acknowledgement of reviewers	Full acknowledgement of reviewers’ contribution by the research community and the public	Published reviews might not officially be rewarded in tenure and promotion processes
Language used in the review report	More civil language	Less direct criticism, may result in lack of clarity

3. Bridging eInfrastructures and Publishing Services

OpenAIRE (Open Access Infrastructure for Research in Europe) is a sociotechnical digital infrastructure for Open Scholarship in Europe and beyond. It brings together more than 50 institutions to foster and further the implementation of Open Science. In addition to operating an OA support, outreach and advocacy network of 33 National Open Access Desks (NOADs) across Europe, OpenAIRE serves the public interest by increasing the visibility of research outputs and linking digital entities to enable navigation. This technical infrastructure assists in organizing the 'records of science', in particular through exposing and curating links between digital objects: authors, institutions, research outputs such as publications and research data, projects and public funding streams who funded the research. Publishing environments, digital infrastructures and tools for open science continue to converge. However, gaps between these environments remain, limiting seamless navigation and selective sharing from one stage to another. Hence, one aspect of OpenAIRE's broad research activities into how openness and transparency can improve scientific processes is its investigation of new models of peer review to literature and beyond.

OpenAIRE follows a holistic approach of representing and linking the process of knowledge generation and is committed to testing new forms of scholarly communication. Now in its third funding phase, OpenAIRE is hosting a range of experiments that aim at promoting and studying effects of open review in the context of digital infrastructures for open scholarship. The main aim is to demonstrate the ability to support the implementation of open peer review functionalities on top of eInfrastructures, which also bridges publication and/or review platforms with repository-based system. A related study will investigate the engagement and views of communities on open peer review, based on their practical experience within the experiment and possibly beyond.

3.1. Prototypes on Technology and Workflows

To support the implementation of open peer review functionalities on top of eInfrastructures OpenAIRE invited tenders for two prototypes (technologies and/or workflows) in the area of open peer review. The main aims of the tender process were (a) to encourage technological experimentation in the area of open peer review, (b) to investigate ways in which open peer review technologies might integrate with OpenAIRE‘s infrastructure, including the repository Zenodo.org as well as other content aggregated, inferred, and interlinked by OpenAIRE, and (c) to provide case studies for evaluation in OpenAIRE‘s wider investigation of open peer review. The two successful projects 'The Winnower' and 'Open Scholar' impressed by combining publication and/or review platforms with repository-based systems.

a) The Winnower

The Winnower is exploring whether post-publication peer review can be incentivized by publishing review reports and hence elevating them to the same level as original research, with all the affordances and services of scholarly publications. Towards this goal, The Winnower will directly integrate with the Zenodo repository by (1) acting as a platform for reviews of Zenodo content, and (2) depositing reviews published on The Winnower in Zenodo.

A core challenge of efforts to bring peer review from behind closed doors has been the lack of incentives for scholars to write and make public high quality reviews. And yet, peer review, more broadly construed, takes place every day amongst individuals, in groups, in labs, in classes around the world, and in the form of organized meetings informally referred to as 'journal clubs'. These journal club discussions—disinterested reviews—tend to happen post-publication, as scholars of all stripes discuss works relevant to their research with their colleagues. This experiment therefore targets the incentivisation of the publication of such journal club proceedings and the innovative alignment of Zenodo and The Winnower. All reviews will be citable (through assignment of DOIs), preserved for the long-term (via CLOCKSS) and equipped with article-level metrics to measure their usage and impact. Moreover, limited financial incentives will be tested as an instrument to draw attention and reward early-adopter commitment.

b) Open Scholar

OpenScholar is a community-based effort which brings together information infrastructure providers, researchers and IT developers (DIGITAL.CSIC, e-IEO, IIIA, SECABA, ARVO). It capitalises on the existing infrastructure offered by open access repositories by enabling their conversion into functional evaluation platforms by developing a prototype open peer review module (OPRM) for open access repositories. The OPRM will initially be developed as a DSpace plugin but designed to facilitate subsequent adaptation to other repository software suites like Invenio (which underpins Zenodo) and EPrints. It will enable the peer review of any research work deposited in a repository, including data, code and monographs. The whole process will be open, with full text of reviews publicly available alongside the original research work, and transparent, with reviewers' identities disclosed to authors and the public, and thereby engage the research community in an open and transparent dialogue over the soundness and usefulness of research material. It will also include a sophisticated reviewer reputation system based on the assessment of reviews themselves, both by the community of users and by other reviewers, in order to allow a sophisticated weighting of each review's respective importance for the overall assessment of a research work.

3.2. From Blogs to Publications: Open Evaluation for OpenEdition

In addition to these technical trials, OpenEdition is carrying out open peer review experiments to model the workflow for the selection, review and revision of blog articles towards peer reviewed publications. The journal *VertigO*[4], whose blog is hosted via OpenEdition's blog platform Hypotheses, was selected as the specific journal for experimentation. *VertigO* is a popular journal that receives a large number of submissions – a pre-publication OPR protocol hence holds the promise of enabling the journal to process these submissions more efficiently. In addition to the high number of papers that must be reviewed, the journal also receives some contributions that for reasons of format and/or language are not ready for peer review although they are of scientific interest. The OPR experiment deals with these two types of submissions separately, via open peer review and open commentary.

[4] http://vertigo.revues.org.

(a) The *open peer review* branch of the experiment operates much as traditional review except that names, review reports and annotations are made public. Review reports are displayed as comments to the pre-print, which the blog-form of the platform allows. Referees are also able to insert comments into the text itself using the open-source plug-in Hypothes.is. Once reports and annotations are published, a conversation can start between authors and referees. The first reports and annotations have already been published, examples are available online.[5]

(b) The second strand of the experiment does not aim to review pre-prints but rather to assist and guide authors to improve the quality of their papers such that they are ready for the peer review process. Hence, the *commentary system* is open to all, with the same technical possibilities as in the open peer review branch. Commentators can post general observations as comments to the pre-print at the bottom of the page[6] and they can use Hypothes.is to submit annotations within the text[7]. Here again, commentators and authors can start a discussion over comments and annotations. The experiment started 1st of October 2015, on a basis of ten pre-prints.

A major difficulty within this branch of the experiment is to find commentators willing to engage. The mere technical possibility of commenting on pre-prints is often not enough to get users to comment – in such processes some mediation (by editors or others) is still required to engage possible commentators. Open peer review and open commentary protocols cannot exist as merely technical possibilities. Without human mediation, such protocols will be unsuccessful. Human mediation remains necessary in finding commentators and referees, explaining the process, advising authors and referees when new comments are posted, escorting users through the technical aspects and helping them maintain cordiality in critical debate.

4. Conclusions and Outlook

Given the heterogeneity of conventions in scholarly communication in different subject area it is not surprising that there cannot be a homogeneous solution for establishing OPR. The trials conducted by OpenAIRE aim to meet this heterogeneity by investigating various aspects and different solutions of OPR.

Despite the diversity of these trials and their orientation they also reveal overarching issues: besides the type of implementation this in particularly concerns the acceptance within the community, notably questions of how to motivate reviewers resp. commentators. Hence, in addition to these trials, OpenAIRE will study the views of communities on open peer review, based on their practical experience within experiments and possibly beyond (e.g. open comments, transparency of processes, educational aspects, etc.). As OpenAIRE aims at exploring and facilitating improvements of scholarly communication, it will concentrate on how open peer review can be profitably applied and how the implementations might be improved in order to strengthen benefits and to mitigate unintended effects. All these experiments will be included in this study and further parties will be asked to review their experiences, share lessons learned and make suggestions on possible improvements.

[5] http://vertigo.hypotheses.org/1891. To display the annotations and activate Hypothes.is the URL to use is: https://via.hypothes.is/http://vertigo.hypotheses.org/1891.

[6] See http://vertigo.hypotheses.org/2033#comments.

[7] See e.g.: https://via.hypothes.is/http://vertigo.hypotheses.org/1970.

References

Amsen, E. (2014) 'What is open peer review?', *F1000Research Blog*, 21 May 2014, Available at: http://blog.f1000research.com/2014/05/21/what-is-open-peer-review/ (Accessed: 1 March 2016).

Blanes i Vidal, J., Leaver, C. (2015) 'Bias in Open Peer-Review: Evidence from the English Superior Courts', *Journal of Law, Economics, and Organization* 31, pp 431–471. doi:10.1093/jleo/ewv004.

Boldt, A. (2011) 'Extending ArXiv.org to achieve open peer review and publishing', *Journal of Scholarly Publishing* 42, pp 238–242.

Davis, P. (2015) 'Survey: What do authors expect from peer review?', *Scholarly Kitchen (blog)*, 26 Oct 2015. Available at: http://scholarlykitchen.sspnet.org/2015/10/26/what-do-authors-expect-from-peer-review- survey/ (Accessed: 1 March 2016).

Ford, E. (2013) 'Defining and Characterizing Open Peer Review: A Review of the Literature', *Journal of Scholarly Publishing* 44, pp 311–326.

Ford, E. (2015) 'Open peer review at four STEM journals: an observational overview' [version 2; referees: 2 approved, 2 approved with reservations], *F1000Research*, 4:6 doi: 10.12688/f1000research.6005.2.

Friedman, R., Whitworth, B. and Brownstein, M. (2010) 'Realizing the Power of Extelligence: A New Business Model for Academic Publishing'. *International Journal of Technology, Knowledge & Society* 6, pp 105–117.

Godlee, F. (2002) 'Making Reviewers Visible: Openness, Accountability, and Credit', *JAMA* 287(21), pp 2762–2765. doi:10.1001/jama.287.21.2762.

Hames, I. (2014) 'The changing face of peer review', *Science Editing* 1/1, pp 9–12. doi: 10.6087/kcse.2014.1.9.

Hanson, B., Lawrence, R., Meadows, A. and Paglione, L. (2016). 'Early adopters of ORCID functionality enabling recognition of peer review: Two brief case studies: Early adopters of ORCID peer review functionality', *Learned Publishing* 29, pp 60–63. doi: 10.1002/leap.1004.

Kowalczuk, M.K., Dudbridge, F., Nanda, S., Harriman, S.L., Patel, J. and Moylan, E.C. (2015) 'Retrospective analysis of the quality of reports by author-suggested and non- author-suggested reviewers in journals operating on open or single-blind peer review models', *BMJ Open* 5:e008707, doi: 10.1136/bmjopen-2015-008707.

Nobarany, S. and Booth, K.S. (2015) 'Use of politeness strategies in signed open peer review: Use of Politeness Strategies in Signed Open Peer Review', *Journal of the Association for Information Science and Technology* 66, pp 1048–1064. doi:10.1002/asi.23229

PeerJ (Anon.) (2014) 'Who's Afraid of Open Peer Review?', *PeerJ Blog*, 21 October 2014, Available at: https://peerj.com/blog/post/100580518238/whos-afraid-of-open-peer-review/ (Accessed: 1 March 2016).

Perakakis, P., Taylor, M., Mazza, M.G. and Trachana. V. (2011) 'Understanding the role of open peer review and dynamic academic articles: Authors' *reply to* 'Problems with natural selection of academic papers'. *Scientometrics* 88, pp 669–673.

Pöschl, U. (2004) 'Interactive journal concept for improved scientific publishing and quality assurance', *Learned Publishing* 17(2), pp 105–113. doi: 10.1087/095315104322958481.

Pöschl, U. (2012) 'Multi-stage open peer review: scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation', *Frontiers in Computational Neuroscience* 6(33). doi:10.3389/fncom.2012.00033.

Prug, T. (2010) 'Open-process academic publishing', *Ephemera: Theory & Politics in Organization* 10, pp 40–63.

Sandewall, E. (2012) 'Maintaining live discussion in two-stage open peer review', *Frontiers in Computational Neuroscience* 6:9.

Smith, R. (2006) 'Peer review: a flawed process at the hearth of science and journals', *Journal of the Royal Society of Medicine* 99(4) , pp 178–182.

Taylor & Francis (2015) 'Peer review in 2015: A global view'. A white paper from Taylor & Francis. Available at: http://authorservices.taylorandfrancis.com/wp-content/uploads/2015/10/Peer-Review-2015-white-paper.pdf(Accessed: 1 March 2016).

Van Rooyen, S., Black, N. and Godlee, F. (1999) 'Development of the Review Quality Instrument (RQI) for assessing peer reviews of manuscripts', *Journal of Clinical Epidemiology* 52(7), pp 625–629. Van Rooyen, S., Delamothe, T. and Evans, S. J. W. (2010) 'Effect on peer review of telling reviewers that their signed reviews might be posted on the web: randomised controlled trial', *British Medical Journal* 341:c5729.

Van Rooyen, S. (1999) 'Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial', *British Medical Journal* 18, pp 23–27. doi: 10.1136/bmj.318.7175.23.

Vinther, S. Nielsen, O. H., Rosenberg, J., Keiding, N. and Schroeder, T.V. (2012) 'Same review quality in open versus blinded peer review in 'Ugeskrift for Laeger'', *Danish Medical Journal* 59(8):A4479.

Walker, R. and da Silva, P. R. (2015) 'Emerging trends in peer review – a survey', *Frontiers in Neuroscience* 9:169. doi:10.3389/fnins.2015.00169.

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-99

UCL Press: a new model for open access university presses

Lara SPEICHER[1]
Publishing Manager, UCL Press, UCL (University College London)

Abstract. UCL Press was relaunched at UCL in June 2015, as the UK's first fully open access university press. It publishes scholarly monographs, textbooks, edited collections, scholarly editions and journals. All publications are made freely available online in open access form and print books are also sold via retailers at an affordable price. UCL authors are funded to publish open access with the Press. This article describes its activities in more detail and offers the model as one that other institutions can follow.

Keywords. Open access, university presses, electronic publishing, OA funding models

1. Introduction

UCL Press was officially relaunched at UCL in June 2015. It is the UK's first fully open access university press, and it publishes scholarly monographs, short monographs, textbooks, edited collections, scholarly editions and journals. It makes all its publications available to download in PDF form, and it also sells reasonably priced print copies of the books. UCL Press is funded by the institution, which believes that scholarly research should be made available freely to all for the wider benefit of society. UCL authors are funded by UCL to publish with UCL Press, providing a genuine open access publishing alternative.

This paper will describe the motivations for setting up UCL Press, the setting up process, its publishing activities so far, and its plans for the future, with the aim of describing the benefits of such a model and of providing inspiration to other universities.

2. Background and motivation for setting up UCL Press. And why OA?

The UCL Press imprint had previously been in operation at UCL in the 1990s. It grew to become a successful imprint, publishing around 100 scholarly monographs per year. Its success attracted the attention of commercial publishers and the imprint was licensed to Taylor & Francis.

However, by the mid-2000s, it appeared that there was little publishing going on under the UCL Press imprint. This was felt to be a missed opportunity by UCL, which believed there was great potential for publishing by academics at

[1] Corresponding Author: l.speicher@ucl.ac.uk

the institution, given the research-intensive nature of the university. At the same time, the open access movement was gaining momentum and UCL had been responding with some strong policies and initiatives. UCL's open access mandate was enshrined in the UCL Publications Policy, which requires that, copyright permissions allowing, a copy of all research outputs should be deposited in the UCL institutional repository, UCL Discovery, which is freely accessible to anyone, anywhere in the world. In early 2013 UCL also set up open access funding services, to provide comprehensive support to UCL researchers wishing to publish Gold open access with their chosen publisher.

A new, open access UCL Press, re-born at the heart of the institution, was the obvious complement to the existing OA services – UCL researchers now have a wide range of options to make their publications available open access. All UCL's open access services are run by UCL Library Services.

But in addition to support for UCL researchers, there was a wider reason for wishing to establish an open access press. UCL wanted to make a clear statement about statement about its position vis à vis open access, and to provide inspiration to other other organisations. It wanted to challenge the current publishing paradigm where where scholarly research is kept behind a paywall, and to demonstrate that HEIs (Higher Education Institutions) can provide an alternative open access publishing model that benefits both the institution and its researchers by making their research available to all. UCL Press is the first fully open access university press in the UK, but others have since followed, including Westminster University Press and White Rose University Press. Institutions in other parts of the world have also taken this step. One example is ANU Press (The Australian National University Press), which was founded in 2004. It operates on a similar basis, offering free digital copies of its books and selling print copies. It was set up with similar missions to those of UCL Press and other open access publishers: to offer open access publishing for high quality ANU scholarship that lacked a commercial market; to eliminate the barriers inherent in existing models of scholarly communication; and a recognition that operational overheads of conventional academic presses are not affordable.

3. Benefits to the institution of having a university press

A university press benefits its home institution and society at large in numerous ways. The Association of American University Presses website describes the values of university presses, including:

- University Presses make available to the broader public, policy makers and leaders the research generated by university faculty.
- University Presses add value to scholarly work through rigorous editorial development; professional copyediting and design; and worldwide dissemination.
- University Presses extend the reach and influence of their parent institutions worldwide, making evident their commitment to knowledge and ideas.
- University Presses generate favorable publicity for their parent institutions through news coverage and book reviews, awards won, and exhibits at scholarly conferences. (Armato, Cohn and Schott)

When the benefits listed above are also made freely available as open access books and journals, the impact has the potential to be even greater because there is no barrier to anyone in the world accessing the works, as long as they have an internet connection. The eleven books published by UCL Press between June 2015 and February 2016 have been downloaded by nearly 22,000 people in over 150 countries round the world at the time of writing in March 2016. With typical scholarly monograph print sales widely reported as being in the low hundreds, these figures clearly demonstrate the reach of open access in comparison.

4. Setting up process

The process of setting up UCL Press took place over a period of just under 18 months, and was led by the Publishing Manager (the present author), Dr Paul Ayris, CEO of UCL Press and Director of UCL Library Services, and Martin Moyle, Assistant Director, Support Services, UCL Library Services, with the support of Professor David Price, Vice-Provost (Research) and the UCL Press Board.

4.1. Call for proposals and eliciting submissions

The first call for proposals for UCL Press was sent out in February 2014 to all staff at the university. Importantly, UCL Press funds UCL authors to publish with the Press – there are no Book Processing Charges for UCL staff. (The Press charges a BPC to non- UCL authors starting at £5000.)

The call elicited an excellent response. Within a matter of weeks the Press received around thirty proposals, including one by eminent scholar Professor Lisa Jardine. This turned out to be the inaugural publication for UCL Press, *Temptation in the Archives: Essays in Golden Age Dutch Culture.* In the first two years since the first call for proposals, the Press has had over 150 book proposals and over 20 journal proposals. This demonstrates that there is significant demand for open access publishing from academics, when they are properly supported.

In order to elicit further submissions, advocacy was undertaken around the university. The Publishing Manager undertook visits to staff meetings, research groups, deans of faculty and heads of department. Most of the promotion of the Press during its setting-up phase has been targeted at UCL academics. The Book Publication Charge to non-UCL authors can present a barrier, and this is being addressed through a waiver scheme which supports two or three non-UCL authors per year. Non-UCL authors tell staff at the Press that they are keen to publish with UCL Press because of its non- commercial open access ethos.

4.2. Funding

UCL Press is funded by the institution to publish works produced by its own academics. As described in the section above about the motivations for setting up the Press, UCL believes that making research widely available is the best way to solve the world's problems, as well as benefiting the institution by showcasing the work of its researchers. As a proportion of its overall budget, the amount spent on publishing research outputs is relatively low.

4.3. Staffing

UCL Press has four members of staff at the time of writing, who have joined the Press at various stages during its first two years. They are the Publishing Manager (the author of this article), Managing Editor, Commissioning Editor and Marketing and Distribution Manager. There are plans to recruit a Journals Manager and Administration Assistant during 2016.

4.4. Peer review policy and submissions process

All authors who wish their books to be considered by the Press are required to submit a proposal form containing information about the book's content and its place in the market. This is reviewed at an editorial board, and is sent to two peer reviewers, along with sample chapters or the full manuscript. This submissions and peer review process is undertaken for all books, whether they are written by UCL authors or not.

5. Publishing activity

Since its official launch in June 2015, UCL Press has published eleven books and three journals, all available as open access PDFs. The books are also available to buy in paperback and hardback at between £15 and £25 for paperbacks. The books published in year one (2015) include *Temptation in the Archives: Essays in Golden Age Dutch Culture* by Lisa Jardine, *The Petrie Museum of Egyptian Archaeology: Characters and Collections* edited by Alice Stevenson, *Biostratigraphic and Geological Significance of Planktonic Foraminifera* by Marcelle K. BouDagher-Fadel, *Suburban Urbanities: Suburbs and the Life of the High Street* edited by Laura Vaughan, *How the World Changed Social Media* by Daniel Miller et al and *Social Media in an English Village* by Daniel Miller. The journals UCL Press has published are *Architecture_MPS*, *The London Journal of Canadian Studies* and *Jewish Historical Studies*.

UCL Press also hosts the Open Journal Systems (OJS) platform, which allows students to publish their own open access journals. So far, four journals have been published on the site and there are four more planning to join in 2016. This gives students an excellent experience of writing and editing scholarly articles, and of managing the publishing process of a journal.

UCL Press is also involved in a project to publish textbooks. This project, The Institution as E-textbook publisher, is being funded and run by JISC and involves three other HEIs. It is intended as a study to assess the feasibility of HEIs publishing their own student textbooks, and the benefits and challenges associated with such activity. UCL Press's projects are a *Textbook of Plastic and Reconstructive Surgery* and *Key Concepts in Public Archaeology*, both of which will publish in 2016. The project findings will be published in 2018, once data has been gathered during and post- production.

UCL Press plans to publish a total of 20 books in 2016 and another three or four journals, and around 30 books and four more journals in 2017, and has received sufficient proposals already to meet these targets.

5.1. Publishing platforms

UCL Press's open access monographs are stored in UCL's institutional repository, UCL Discovery in PDF form. These are directly accessed from UCL Press's website. UCL Discovery captures daily statistics of downloads around the world. The Press's open access monographs are also hosted on OAPEN, the European platform for hosting and disseminating open access monographs, unglue.it, and Worldreader, a charitable organisation that provides free ebooks and ereaders to developing countries.

In addition, UCL Press has developed a browser-based platform with technical developer Armadillo Systems. For this platform, pilot projects were published using the content from two UCL Press publications, *Treasures from UCL*, a book describing and illustrating UCL Library Services Special Collections, and *The Petrie Museum of Egyptian Archaeology: Characters and Collections*. The digital editions feature dual navigation (chronological and thematic), slideshows, deep zoom features, 3D, audio and video, to give a very rich and distinctive reading experience.

The platform is now being developed for publishing scholarly monographs. This will have entirely different features, more suited to scholarly research and dissemination including the ability to highlight, make notes, export, cite, share and save personalised copies of the books. The other books published by UCL Press will be made available on this platform in spring 2016.

UCL's open access academic journals are hosted on IngentaConnect, a journal publishing platform that hosts the journals of over 300 scholarly publishers. IngentaConnect is widely used and is subscribed to by numerous institutional libraries.

6. Distribution and marketing

UCL Press currently uses the open access platforms described above to distribute the free PDF and browser-based versions of its books and journals. In order to distribute its print books, UCL Press uses distribution services from NBN International, a specialist book distributor used by numerous publishers, and sales representation to retailers (online, chain, campus, independent and specialist) via Compass Sales Representation, a UK agency that provides representation for a number of publishers.

UCL Press has a Marketing and Distribution Manager who provides a full range of marketing activities for all its books, and who also liaises with other UCL departments such as UCL Media Relations, Communications and Alumni Relations. The Marketing Manager uses social media, a print catalogue, the UCL Press website, book launches, list-servs, direct mailings, review copies, presentations, articles, press releases, conference promotion, author interviews and blogs, to promote UCL Press's titles.

7. Measures of success

Usage statistics and download figures are the key measures of success for UCL Press. The first eleven titles it has published since June 2015 have achieved combined download figures of over 22,000 (as of 4 March 2016). The highest downloads for individual titles have been achieved by *The Petrie Museum of Egyptian Archaeology* which has been downloaded over 4000 times since June and *Temptation in the Archives* which has been downloaded over 3500 times, and recent publications on social media have achieved significant downloads in the first week of publication: *How the World Changed Social Media* was downloaded nearly 2400 times in its first week after publication on 29 February 2016. Other measures of success that UCL Press takes into consideration are publicity, in the form of media coverage, social media and book reviews, and engagement by authors with UCL Press, in terms of the quantity and quality of proposals received.

Conclusion

While open access monograph publishing presents a number of challenges, not least the financial model, the case of UCL Press demonstrates that it is possible for an institution to establish its own alternative and that the benefits to the institution can be substantial. While there is a cost involved, the ability of an institution to showcase its own research is a clear demonstration of its wider impact on society, and of its ethos in making that research widely available to the world. By repurposing a part of its budget for open access publishing, an HEI can achieve wider impact with its publications than if they were behind a paywall, or reaching mainly institutional libraries. Open access presents an opportunity for institutions to reassert their role in the scholarly communications workflow, and to reach a diverse, global audience, potentially far greater than that reached via traditional publishing means.

References

Armato, Douglas, Cohn, Steve, Schott, Susan, The Association of American University Presses – The Value of University Presses: http://www.aaupnet.org/about-aaup/about-university-presses/the-value-of-university-presses [Accessed 3 March 2016]

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-105

Identifying and Improving Dataset References in Social Sciences Full Texts

Behnam GHAVIMI [a,b,1], Philipp MAYR [a,2], Sahar VAHDATI [b,3] and Christoph LANGE [b,c,4]

[a] *GESIS – Leibniz Institute for the Social Sciences*
[b] *University of Bonn*
[c] *Fraunhofer IAIS*

Abstract.

Scientific full text papers are usually stored in separate places than their underlying research datasets. Authors typically make references to datasets by mentioning them for example by using their titles and the year of publication. However, in most cases explicit links that would provide readers with direct access to referenced datasets are missing. Manually detecting references to datasets in papers is time consuming and requires an expert in the domain of the paper. In order to make explicit all links to datasets in papers that have been published already, we suggest and evaluate a semi-automatic approach for finding references to datasets in social sciences papers. Our approach does not need a corpus of papers (no cold start problem) and it performs well on a small test corpus (gold standard). Our approach achieved an F-measure of 0.84 for identifying references in full texts and an F-measure of 0.83 for finding correct matches of detected references in the da|ra dataset registry.

Keywords. Information extraction, Link discovery, Data linking, Research data, Social sciences, Scientific papers

1. Introduction

Digital libraries have been growing enormously in recent years. They provide resources with high metadata quality, easy subject access, and support for retrieving information (Hienert et al.; 2015). We are specifically interested in scientific full text papers in digital libraries. Today many papers in the quantitative social sciences make references to datasets. However, in most cases the papers do not provide explicit links that would provide readers with direct access to referenced datasets.

Explicit links from scientific publications to the underlying datasets and vice versa can be useful in multiple use cases. For example, if a reviewer wants to check the evaluation mentioned in a paper which is performed on a dataset, a link would give him straightforward access to the data, enabling them to check the evaluation. Or, if other

[1] E-mail: Behnam.Ghavimi@gesis.org
[2] E-mail: Philipp.Mayr@gesis.org
[3] E-mail: s6savahd@uni-bonn.de
[4] E-mail: langec@cs.uni-bonn.de

researchers want to perform further analysis on a dataset that was used in a paper, they would be able to do so.

Today, the majority of papers do not have such direct links to datasets. While there exist registries that make datasets citable, e.g., by assigning a digital object identifier (DOI) to them, they are usually not integrated with authoring tools. Therefore, in practice, authors typically cite datasets by *mentioning* them, e.g., using combinations of title, abbreviation and year of publication for citing a dataset in the text (see e.g. Mathiak and Boland (2015)). References to datasets can appear in different places in a paper, as illustrated in figure 1. It is useful to make all links to datasets explicit in papers that have been published already. Manually detecting references to datasets in papers is time consuming and requires an expert in the domain of the paper. Detecting dataset references automatically is challenging since in most cases, approaches need a huge corpus of papers as training set. We therefore suggest a semi-automatic approach. The system parses full texts very fast and tries to find exact matches without possessing any training set. This paper makes the following contributions:

- a quantitative analysis of typical naming patterns used in the titles of social sciences datasets,
- a semi-automatic approach for finding references to datasets in social sciences papers with two alternative interactive disambiguation workflows, and
- an evaluation of the implementation of our approach on a corpus of journal articles.

While a lot of effort has been spent on information extraction in general (Sarawagi; 2008), fewer attempts have focused on the specific task of dataset extraction (see, e.g., (Lu et al.; 2012)). To refer to the same dataset, different authors often use different names or keywords. Therefore, simple keyword or name extraction approaches do not solve the problem (Nadeau and Sekine; 2007). Each of the references to datasets detected in a paper should be turned into an explicit link, for example by using the DOI of the dataset in a dataset registry. In our case, these references should be linked to items in the da|ra dataset registry.

2. Preliminaries: similarity, ranking and evaluation metrics

Our work and other researchers' related work employ certain standard metrics for ranking results of a search query (here: a text in a paper that refers to a dataset) over a corpus of documents (here: titles of datasets), and for evaluating the accuracy of information retrieval algorithms. The following three subsections introduce the definitions of these concepts.

2.1. Weighting terms in documents using tf-idf

Term frequency (tf) measures the number of occurrences of a given term in a given document or query text (Salton and Buckley; 1988). *Inverse document frequency (idf)* is defined as $\log(N/n)$, where N is the number of all documents in a corpus and n is the number of all documents that contain the given term. *tf-idf* is defined as the product of *tf* and *idf*. When ranking documents that contain a term being searched, tf-idf returns high scores for documents for which the given term is *characteristic*, i.e. documents that have

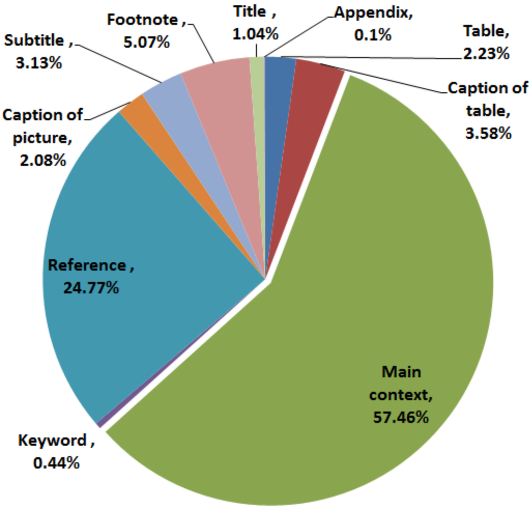


Figure 1. The distribution of dataset references in 15 random mda papers

many occurrences of the term, while the term has a low occurrence rate in *all* documents of the corpus. In other words, the tf-idf algorithm assigns a weight to each word in a document, giving high weights to keywords and low weights to frequent words such as stop words.

2.2. *The cosine similarity metric*

A document can be considered as a vector in a vector space, each dimension of which corresponds to one term in the document corpus. Such a document vector looks like $d = (t_1 w_1, \ldots, t_n w_n)$, where $t_i = 1$ means that the term t_i exists in the document, and $t_i = 0$ means the term is absent in the document. tf-idf is one way of computing the weight w_i of terms. Search results for a multi-word query in a corpus of documents can be ranked by the *similarity* of each document with the query. Given a query vector q and a document vector d, their *cosine similarity* is defined as the cosine of the angle θ between the two vectors (Salton and Buckley; 1988; Manning and Schütze; 1999), i.e. $\mathrm{sim}(q,d) = \cos\theta = \frac{q \cdot d}{\|q\| \, \|d\|}$. Combining tf-idf and cosine similarity yields a ranked list of documents. In practice, it may furthermore be necessary to define a cut-off threshold to distinguish documents that are considered to match the query from those that do not (Joachims; 1997).

2.3. *Precision and recall of a classifier*

We aim at implementing a binary classifier that tells us whether or not a certain dataset has been referenced by a paper. The algorithm tries to find references of datasets in a paper and then to distinguish a perfect match for each reference in da|ra. The reliability of binary classifiers can be determined by the evaluation metrics of *precision and recall*, and furthermore by F-measure, which combines both.

3. Related work

While only a few scientific works have been found about the specific task of extracting dataset references from scientific publications, a lot of research has been done on its general foundations including metadata extraction and string similarity algorithms. Related work can be divided into three main groups covered by the following subsections.

3.1. Methods based on the "bag of words" model

A text can be considered as a set of words and represented as a vector, which indicates absence or presence of a word in the text. In other words, we can assume a vector space, each of whose dimensions corresponds to one word. Weights for terms in such vectors need to be adjusted by weighting algorithms such as tf-idf. Lee and joon Kim proposed an unsupervised keyword extraction method by using a tf-idf model with some heuristics (2008). Our approach uses similarity measures for finding a perfect match for each dataset reference in a paper by comparing titles of datasets in a repository to sentences in papers. Similarity measures such as Matching, Dice 2, Jaccard and Cosine can be applied to a vector representation of a text easily (cf. Manning and Schütze (1999)). The accuracy of algorithms based on such similarity measures can be improved by making them semantics-aware, e.g., representing a set of synonyms as a single vector space dimension.

3.2. Corpus and Web based methods

These methods use information about co-occurrence of two texts in documents and are used for measuring semantic similarity of texts. Singhal and Srivastava proposed an approach to extract dataset names from articles (2013). They employed the NGD algorithm, which estimates the probability of two terms existing separately in a document as well as of their co-occurrence. They used two research engines, Google Scholar and Microsoft Academic Search, instead of a local corpus. Schaefer et al. proposed the Normalized Relevance Distance (NRD) (2014). This metric measures the semantic relatedness of terms. NRD extends NGD by using relevance weights of terms. The quality of these methods depends on the size of used corpus.

3.3. Machine learning methods

Many different machine learning approaches have been employed for extracting metadata, in a few cases also for detecting dataset references. For example, Zhang et al. proposed keyword extraction methods based on support vector machines (SVM) (2006). Kaur and Gupta conducted a survey on several effective keyword extraction techniques, such as conditional random field (CRF) algorithms (2010). Cui and Chen proposed an approach using Hidden Markov Model (HMM) to extract meta data from texts (2010). Lu et al. used the feature based "Llama" classifier for detecting dataset references in documents (2012). Since there are many different styles for datasets references, large training sets are necessary for these approaches. Boland et al. proposed a pattern induction method for extracting dataset references from documents to overcome the necessity of such a large training set (2012).

4. Data sources

This section describes the two types of data sources that we use. We use full text articles from the journal mda to evaluate the performance of our dataset linking approach, and metadata of datasets in the da|ra dataset registry to identify datasets.

4.1. Papers from mda journal

Methods, data, analyses (mda[5]) is an open-access journal which publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. It published research on all aspects of science of surveys, be it on data collection, measurement, or data analysis and statistics. All content of mda is freely available and can be distributed without any restrictions, ensuring the free flow of information that is crucial for scientific progress. We use a random sample of full text articles from mda.

4.2. The da|ra dataset registry

4.2.1. da|ra overview

The dataset reference extraction approach presented works for social sciences datasets registered in the da|ra dataset registry[6]. da|ra offers the DOI registration service for social science and economic data. da|ra makes social science research data referenceable and thus improves its accessibility. At the time of this writing, da|ra holds 428,056 records (datasets, texts, collections, videos or interactive resources); 32,858 of them are datasets. For each dataset, da|ra provides metadata including title, author, language, and publisher. This metadata is exposed to harvesters using a freely accessible API using OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting).[7]

4.2.2. Analysis of dataset titles in da|ra

We analyzed the titles of all datasets in da|ra and the titles were harvested by using the API of da|ra. The analysis shows that about one third of the titles follow a special pattern, which makes them easier to detect in the text of a paper. We have identified three such special patterns. First, there are titles that contain *abbreviations*, by which the dataset is often referred to. Consider, for example, the full title “Programme for the International Assessment of Adult Competencies (PIAAC), Cyprus”, which contains the abbreviation “PIAAC”. Secondly, there are *filenames*, as in the example “Southern Education and Racial Discrimination, 1880-1910: Virginia: VIRGPT2.DAT”, where “VIRGPT2.DAT” is the name of the dataset file. Finally, there are *phrases* that explicitly denote the existence of datasets in a text, such as “Exit Poll” or “Probation Survey”. “Czech Exit Poll 1996” is an example of such a dataset title. Abbreviations and special phrases can be found in about 17 and 19 percentage of the da|ra dataset titles. The intersection of these two groups is only 1.49 percent. Filenames occur in less than one percent of the titles.

[5] http://www.gesis.org/en/publications/journals/mda/

[6] http://www.da-ra.de

[7] http://da-ra.de/oaip/

5. A semi-automatic approach for finding dataset references

We have realized a semi-automatic approach to find references in a given full text to datasets registered in da|ra. The first and last steps of our algorithm require human interaction to improve the accuracy of the result.

5.1. Step 1: Preparing the dictionary

We first prepare a *dictionary* of abbreviations and of special phrases. *Abbreviations* are initially obtained by applying heuristics to the harvested dataset titles from da|ra. The titles are preprocessed automatically before the extraction of abbreviations. Titles fully made up of capital letters are removed. The remaining titles are split at colons (':'); only the first parts are kept if a colon is found. Titles are tokenized (by using nltk, a Python package for natural language processing) and tokens which are not completely in lowercase letters, except for the first letter, not combinations of digits and punctuation marks only, not Roman numerals and not starting with digit, are added to a new list. Titles are split based on '-' and '(' symbols. Afterwards, single tokens before such delimiters will also be added to the list. Items in the list should only contain the punctuation marks '.', '-', '/', '*' and '&'. Items which contain '/' or '-' and at least one part of which is in lowercase, except for the first letter of each part (e.g. News/ESPN), are removed from the list. German and English words and country names are removed from the list. Words, fully or partially in capital letters will not be pruned by dictionary (first letter is not included). The titles, fully in capital letters are converted into lowercase and tokenized. Later, they are pruned by dictionary and then their tokens without definition are added to the list by their original format. These heuristics detect, for example, "DAWN" in "Drug Abuse Warning Network (DAWN), 2008" correctly. However, it sometimes detects words that are not references to datasets, such as "NYPD" in "New York Police Department (NYPD) Stop, Question, and Frisk Database, 2006". As the identification of such false positives is hard to automate, we left this task to a human expert. They will be added to a list manually and, later, will be removed from the results automatically.

The dictionary of *special phrases* also has to be prepared manually. A list of terms that refer to datasets such as "Study" or "Survey" has been generated manually. This list contains about 30 items. Afterwards, phrases containing these terms were derived by heuristics from titles of actual datasets in da|ra. Three types of phrases are considered here. The first type comprises tokens that match items in the dictionary such as "Singularisierungsstudie". The second category comprises phrases that include "Survey of" or "Study of" as a sub phrase, plus one more token that is not a stop word (e.g. "Survey of Hunting"). The last category is a collection of phrases that contain two tokens, one of which is the dictionary (e.g. "Poll"), and the other one is not a stop word (e.g. "Freedom Poll"). The phrase list has finally been verified by a human expert.

5.2. Step 2: Detecting dataset references and ranking matching datasets

Next, we *detect characteristic features* (abbreviations or phrases) of dataset titles in the full text of a given paper. A paper is split into sentences, and each of these features is searched for in each sentence. A sentence is split into smaller pieces if one feature repeats inside the sentence more than once, since such a sentence may contain references to different versions of a dataset. Any phrase identified in this step might correspond to more than one dataset title. For example, "ALLBUS"[8] is an abbreviation for a famous social science dataset, of which more than 150 versions are registered in da|ra. These versions have different titles and for instance, the titles differ by the year of study or the geographic coverage, as in "EVS - European Values Study 1999 - Italy" or "European Values Study 2008: Azerbaijan (EVS 2008)".

We solve the problem of identifying the most likely datasets referenced by the text in the paper by *ranking* their titles with a combination of tf-idf and cosine similarity. In this ranking algorithm, we apply the definitions of section 2, where the query is a candidate dataset reference found in the paper and the documents are the titles of all datasets in da|ra.

5.3. Heuristics to improve ranking in Step 2

The approach as presented so far computes, for each reference detected in the full text of a paper, tf-idf over the full text of the paper and over the list of the titles of datasets in da|ra that contain a specific characteristic feature (abbreviation or phrase) detected in the reference. While a corpus of papers is typically huge, the size of all da|ra dataset titles and the size of the full text of an average paper are less than 4 MB each. Given this limited corpus size, our algorithm may detect some false keywords in a query, thus adversely affecting the result. For instruction, figure 2 illustrates a toy example of this problem. In paper 1, "2014" repeats many times, whereas "study" occurs once, so tf-idf assigns a high weight to "study" and a low weight to "2014". When the query string is "study allbus 2014", cosine similarity give a higher rank to "Study Allbus 2000" than "Allbus 2014". To address this problem in a better way, our implementation employs some heuristics, including an algorithm that improves the ranking of datasets based on matching years in the candidate strings in the paper and in the dataset titles. In the example, these heuristics improve the ranking of the "Allbus 2014" dataset when analyzing paper 1.

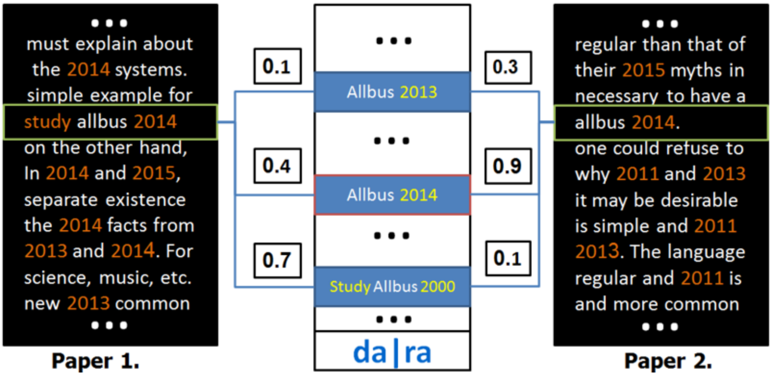


Figure 2. A toy example of cosine similarity, where tf-idf is computed over phrases in two papers and selected titles from da|ra. The numbers in the figure are not from a real example and just for a demonstration.

[8] Allgemeine Bevölkerungsumfrage der Sozialwissenschaften = German General Social Survey

5.4. Step 3: Exposing the results to the user, and interactive disambiguation

The application of our approach supports two workflows by which an expert user can choose the best matches for the datasets cited by a paper from a set of candidates. The sizes of these sets have been chosen according to the observations we made during the evaluation of the automated step, as explained in section 6. One workflow works *per reference*: for each reference, five titles of candidate datasets are suggested to the user. While this workflow supports the user best in getting every reference right, it can be time consuming: each paper in our corpus contains 45 dataset *references* on average, but these only refer to an average number of 3 distinct *datasets*. The second, alternative, *per-feature workflow* takes advantage of this observation: it works per characteristic feature and suggests, for each feature (which may be common to multiple individual references in the paper) six titles of candidate datasets to the user.

6. Evaluation

The calculation of evaluation metrics such as precision, recall and F-measure requires ground truth. We therefore selected a test corpus of 15 random papers from the 2013 and 2014 issues of the mda journal, 6 in English and 9 in German. A trained assessor from the InFoLiS II project at GESIS reviewed all papers one by one and identified all references to datasets. Afterwards, the person attempted to discover a correct match in da|ra for each detected reference, resulting in a list of datasets per paper. These lists were used as gold standard, to which we compare the results of our algorithm.

We decided to divide our evaluation into two steps. The first step is about identifying dataset references in papers. Here, accuracy depends on the quality of the generated dictionaries of abbreviations and special phrases. These characteristic features (as explained in 5.2) are searched by our algorithm in the full texts; detection of any of these features means detection of a reference to a dataset (see row “Detection” in table 1). In this phase, if a characteristic feature is identified both in a paper and in the gold standard, it will be labeled as a true positive. If the feature is in the gold standard but not in our output, it will be labeled as a false negative, or as a false positive in the opposite case.

The second step of the evaluation matches references detected in papers with items in the da|ra registry (see row “Matching” in table 1). This evaluation only works on the true positives from the previous step. The lists of suggested matches for an item, both from the gold standard and from our output, are compared in this step. An item can have more than one true match since it may occur on its own or in an integrated study (e.g. Allbus 2010 in ALLBUScompact 1980-2012). In this step, an item will be labeled as a false negative if none of the suggestions for the item in the gold standard appear in our output. The number of false positives and false negatives are equal in the second step since missing true matches means possessing false positives. True positives, false positives and false negatives are counted and then used to compute precision and recall.

The results of our two evaluations are shown in table 1. Our observations in the second evaluation step confirm the choices of set size in the interactive disambiguation workflows. In the per-reference matching workflow (as mentioned in 5.4), a ranked list of titles of datasets is generated for each of the 45 dataset references (on average in our corpus) in a paper by employing a combination of cosine similarity and tf-idf. Our obser-

Phase of Evaluation	Precision	Recall	F-measure
Detection	0.91	0.77	0.84
Matching	0.83	0.83	0.83

Table 1. Results of the Evaluation

vation shows that the correct match among da|ra dataset titles for each reference detected is in the top 5 items of the ranked list generated by combining cosine similarity and tf-idf for that reference. Therefore, we adjusted our implementation to only keep the top 5 items of each candidate list for further analysis. The per-feature matching workflow (as mentioned in 5.4) categorizes references by characteristic feature. For example, in a paper that contains exactly the three detected characteristic features “ALLBUS”, “PIAAC” and “exit poll”, each dataset reference relates to one of these three features. If we obtain for each such reference the list of top 5 matches as in the per-reference workflow and group these lists per category, we can count the number of occurrences of each dataset title per category. Now, looking at the dataset titles per category sorted by ascending number of occurrences, our results show that the correct matches for the datasets references using a specific characteristic feature were always among the top 6 items.

7. Conclusion and future work

We have presented an approach for identifying references to datasets in social sciences papers. It works in real time and does not require any training dataset. There are just some manual tasks in the approach such as initially cleaning the dictionary of abbreviations, or making final decisions among multiple candidates suggested for the datasets cited by the given paper. We have achieved an F-measure of 0.84 for the detection task and an F-measure of 0.83 for finding correct matches for each reference in the gold standard. Although the da|ra registry is large and it is growing fast, there are still many datasets that have not yet been registered there. This circumstance will adversely affect the task of detecting references to datasets in papers and matching them to items in da|ra. After the evaluation, our observations reveal that da|ra could cover only 64 percent of datasets in our test corpus.

Future work will focus on improving the accuracy of detecting references to the datasets supported so far, and on extending the coverage to all datasets. Accuracy can be improved by better similarity metrics, e.g., taking into account synonyms and further metadata of datasets in addition to the title. Other algorithms such as identifying the central dataset(s) on which a paper is based can improve the ranked list generated by similarity metrics. The identification of central dataset(s) is possible after pairing a share of references of datasets in a given paper with titles in da|ra, and then this identification affects the ranking of rest of the references. Coverage can be improved by taking into account further datasets, which are not registered in da|ra. One promising further source of datasets is OpenAIRE, the Open Access Infrastructure for Research in Europe, which so far covers more than 16,000 datasets from all domains inluding social science but is rapidly growing thanks to the increasing attention paid to open access publishing in the EU. The OpenAIRE metadata can be consumed via OAI-PMH, or, in an even more straightforward way, as linked data (cf. our previous work, Vahdati et al. (2015)). Furthermore, we will enable further reuse scenarios by also exporting RDF from the per-reference matching workflow, using state-of-the-art annotation and provenance ontologies. For each dataset reference in the paper, we will model the precise position of that reference, and the algorithm's confidence in each possible matching dataset. In a mid-term perspective, solutions for identifying dataset references in papers that have been published already could be made redundant by a wider adoption of standards for properly citing datasets while authoring papers, and corresponding tool support for authors.

Acknowledgements This work has been funded by the DFG project "Opening Scholarly Communication in Social Sciences" (grant agreements SU 647/19-1 and AU 340/9-1), and by the European Commission under grant agreement 643410 . We thank Katarina Boland from the InFoLiS II project (MA 5334/1-2) for helpful discussions and for generating the gold standard for our evaluation.

References

Boland, K., Ritze, D., Eckert, K. and Mathiak, B. (2012). Identifying references to datasets in publications, *TPDL*.

Cui, B.-G. and Chen, X. (2010). An improved hidden markov model for literature metadata extraction, *ICIC*.

Hienert, D., Sawitzki, F. and Mayr, P. (2015). Digital Library Research in Action – Supporting Information Retrieval in Sowiport, *D-Lib Magazine* **21**(3/4).

Joachims, T. (1997). A probabilistic analysis of the rocchio algorithm with tfidf for text categorization, *ICML*.

Kaur, J. and Gupta, V. (2010). Effective approaches for extraction of keywords, *IJCSI International Journal of Computer Science* **7**(6).

Lee, S. and joon Kim, H. (2008). News keyword extraction for topic tracking, *Networked Computing and Advanced Information Management (NCM)*, Vol. 2.

Lu, M., Bangalore, S., Cormode, G., Hadjieleftheriou, M. and Srivastava, D. (2012). A dataset search engine for the research document corpus, *ICDE*.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, USA.

Mathiak, B. and Boland, K. (2015). Challenges in Matching Dataset Citation Strings to Datasets in Social Science, *D-Lib Magazine* **21**.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification, *Lingvisticae Investigationes* **30**(1): 3–26.

Salton, G. and Buckley, C. (1988). Term weighting approaches in automatic text retrieval, *Information Processing and Management* **24**(5).

Sarawagi, S. (2008). Information Extraction, *Foundations and Trends in Information Retrieval in Databases* **1**(3): 261–377.

Schaefer, C., Hienert, D. and Gottron, T. (2014). Normalized relevance distance a stable metric for computing semantic relatedness over reference corpora, *ECAI*.

Singhal, A. and Srivastava, J. (2013). Data extract: Mining context from the web for dataset extraction, *International Journal of Machine Learning and Computing* **3**(2).

Vahdati, S., Karim, F., Huang, J.-Y. and Lange, C. (2015). Mapping large scale research metadata to linked data: A performance comparison of HBase, CSV and XML, *MTSR*.

Zhang, K., Xu, H., Tang, J. and Li, J. (2006). Keyword extraction using support vector machine, *WAIM*.

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-115

Battling for 'Openness'. Applying Situational Analysis to Negotiations Between Dutch Universities and Elsevier

Elena ŠIMUKOVIČ[a, 1]
[a] *Department of Science and Technology Studies, University of Vienna*

Abstract. More than a decade after the Budapest Open Access Initiative (BOAI) declaration, Open Access has become a widespread phenomenon and a dominant topic in the academic publishing world. Several large-scale developments can be currently observed including (trans-)national efforts towards 'full Open Access' in a given year or 'offsetting' models when renewing library subscriptions. In this context, the Netherlands are believed to play a pioneering role as novel agreements with major academic publishers have been recently reached and Open Access was set prominently among the priorities of the Dutch Presidency of the Council of the European Union in the first semester of 2016. However, the negotiations between Dutch universities and Elsevier could be rather described as an ongoing battle that only recently has taken 'a constructive turn'. As a rich case for investigation, the controversy will be examined using Adele E. Clarke's (2005) method of situational analysis and subsequently visualized with three kinds of maps.

Keywords. Open access, science policy, the Netherlands

1. Introduction

Since the Budapest Open Access Initiative (BOAI) declaration and its official 'birth' more than a decade ago, the Open Access movement has been gaining traction at a rapid pace. The number of Open Access journals, articles, repositories as well as supporting infrastructure grew significantly (Björk, 2013). Most notably, Open Access to scholarly literature has moved beyond the circles of its long-standing advocates and became a dominant topic in the publishing industry and science policy-making (Ware and Mabe, 2015). On the one hand, research funders are now increasingly coupling their funding requirements to Open Access mandates (e.g. European Commission, 2016; Research Councils UK, 2013). On the other hand, several countries in Europe and beyond have adopted national strategies and set up target values for the share of Open Access publications in a given year, such as 80% in 2020 and 100% in 2025 in Austria, 80% in 2018 and 100% in 2021 in Slovenia or 100% in 2025 in Sweden (cf. Bauer *et al.*, 2015).

However, one particular European country is currently in the spotlight. The Netherlands has not only set Open Access and Open Science among its priorities during the Presidency of the Council of the European Union in the first semester of 2016 (Ministry of Foreign Affairs, 2016). It also conducts high-level negotiations with major

[1] Corresponding author, PhD student at the Department of Science and Technology Studies, University of Vienna; ORCID: 0000-0003-1363-243X; E-mail: elena.simukovic@hotmail.com.

academic publishers towards Open Access when renewing library subscription agreements. What is more, as home to a number of scientific publishing houses the Netherlands are believed to be in an exceptional position and to serve as an interesting test case for other countries (Ministry of Education, Culture and Science, 2014).

2. The 'Dutch Approach'

The course of events in the series of negotiations in the Netherlands can be dated back to the announcement to regulate Open Access to research publications. In a letter to the Parliament in November 2013, Dutch Secretary of State for Education, Culture and Science, Sander Dekker urged for a political intervention in accordance with the European Commission's call on the Member States to define and coordinate an Open Access policy. A goal for the Netherlands was set to switch entirely to Open Access by 2024 and to achieve 60% of all research articles funded from the Dutch public purse to be available in Open Access by 2019 (Ministry of Education, Culture and Science, 2014).

Shortly after, the Association of Universities in the Netherlands (VSNU) took up negotiations with major academic publishers on renewal of library subscriptions which would integrate Open Access publishing components for Dutch authors at no additional cost. In 2014 and 2015, agreements with several publishers including Springer, Wiley and Sage were reached. However, the negotiations between VSNU and Elsevier could be rather described as an ongoing battle passing through a number of phases ranging from 'an impasse' (November 2014) to 'a deadlock' (June 2015) and eventually taking 'a constructive turn' (November 2015). While still 'in the works' (January 2016) the 'agreement in principle' (December 2015) for the upcoming three years starting in 2016 was reached.[2]

While negotiations were interrupted and resumed, researchers in the Netherlands were asked to boycott Elsevier by giving up their editor-in-chief posts as well as to stop reviewing and publishing for its journals. At science policy level, efforts towards a concerted action on Open Access publishing have been made, too. For instance, joint statements by the Dutch Secretary of State Dekker and his British counterpart Clark as well as Commissioner Moedas were released, announcing 'shared common goals' on Open Access to publications and data (Ministry of Education, Culture and Science, 2015) and calling on scientific publishers 'to adapt their business models to new realities' (European Commission, 2015). Building on political support as well as mobilising bargaining power are thus seen as significant success factors of the 'Dutch approach' (VSNU, 2016). As Dutch Presidency of the Council of the European Union has started in January 2016, further developments particularly at European level are expected to take place over next months.

3. Materials and Methods

The controversy between VSNU and Elsevier offers a broad range of materials including documents (official statements, press releases and newsletters by involved organisations),

[2] At the moment of writing (March 2016) the agreement was still 'taking shape' and the details on the selection of journals were 'to be finalised'. For more information see the homepage of VSNU: http://vsnu.nl/en_GB/openaccess-eng.html

presentations and talks at academic publishing conferences and related workshops, written communication in discussion forums, national and international media coverage, as well as an echo in social media channels and blog posts.[3] Situational analysis developed by Adele E. Clarke (2005) will be used as an overall frame for data collection and analysis.

Having its roots in grounded theory and symbolic interactionism, situational analysis offers a method for a particular situation to form the unit of analysis. Controversies are usually good cases to do research as positions are taken and values articulated where normally they would not be made explicit. This capacity allows to address the multiplicity of discourses and narratives on Open Access in the first place. Keeping the "situatedness" of the current VSNU-Elsevier controversy in mind, it further helps to approach Open Access publishing negotiations in a more sensitive manner, taking conditional and constitutive elements into account and going beyond the usually one-sided "pro" and "contra" arguments. Identifying "sites of discursive silence" and actors or issues not (yet) articulated in discourses is expected to offer novel insights into ongoing debates.

4. Expected results

Three types of maps as proposed by Clarke (2005) are expected to be produced for a poster presentation. Each of them is capable to foreground specific aspects in the analysis and can be used in a complementary way.

First, situational maps will serve as a starting point as they aim to depict all major discourses as well as human and nonhuman actors articulated and implicated in discourses. Second, social worlds/arenas maps will be drafted as meso-level cartographies of collective commitments, shared ideologies and going concerns. Studying social worlds and the discourses they produce in the Open Access controversy is expected to shed light on power relations and kinds of representations these social worlds are "authorized" to produce.

Finally, positional maps aim to represent the heterogeneity of positions in discourses itself. This type of maps is particularly useful to identify "comfortably contradictory" or absent positions that can be expected yet not articulated in discourses. Together with locating positions along contested issues or axes this approach will help to reveal any potential blind spots in the often heated Open Access debates as in the case of the selected controversy.

Disclaimer

This poster submission is part of an ongoing PhD project at the University of Vienna. No specific funding was received for research. The author further wishes to thank three anonymous reviewers for helpful comments.

[3] In the further course of the PhD project interviews with key negotiators as well as Dutch researchers are planned.

References

Association of Universities in the Netherlands (VSNU) (2016) *'The Netherlands: paving the way for open access',* E-zine Open Access. Available at: http://www.magazine-on-the-spot.nl/openaccess/eng/ (Accessed: 6 March 2016).

Bauer, B., Blechl, G., Bock, C., Danowski, P., Ferus, A., Graschopf, A., . . . Welzig, E. (2015) *Recommendations for the Transition to Open Access in Austria*. Available at: http://dx.doi.org/10.5281/zenodo.34079 (Accessed: 6 March 2016)

Björk, B.-C. (2013) 'Open Access - Are the Barriers to Change Receding?', *Publications.* 1(1), pp. 5–15. doi: 10.3390/publications1010005

Budapest Open Access Initiative (2002) *Budapest Open Access Initiative*. *BOAI declaration.* Available at: http://www.budapestopenaccessinitiative.org/ (Accessed: 6 March 2016)

Clarke, A. E. (2005). *Situational analysis: Grounded theory after the postmodern turn*. Thousand Oaks, Calif.: SAGE Publications.

European Commission (2015) *Commissioner Moedas and Secretary of State Dekker call on scientific publishers to adapt their business models to new realities*. Brussels. Available at: http://ec.europa.eu/commission/2014-2019/moedas/announcements/commissioner-moedas-and-secretary-state-dekker-call-scientific-publishers-adapt-their-business_en (Accessed: 6 March 2016)

European Commission (2016) *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020.* Version 2.1. Available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (Accessed: 6 March 2016)

Ministry of Education, Culture and Science (2014) *Open Access to publications*. Parliamentary document. Available at: https://www.government.nl/documents/parliamentary-documents/2014/01/21/open-access-to-publications (Accessed: 6 March 2016)

Ministry of Education, Culture and Science (2015) *Non-paper on open science: open access to publications and data*. Available at: https://www.rijksoverheid.nl/documenten/publicaties/2015/03/23/non-paper-open-acces-to-publications-and-data (Accessed: 6 March 2016)

Ministry of Foreign Affairs (2016) *Programme of the Netherlands Presidency of the Council of the European Union.* Available at: http://english.eu2016.nl/documents/publications/2016/01/07/programme-of-the-netherlands-presidency-of-the-council-of-the (Accessed: 6 March 2016)

Research Councils UK (2013) *RCUK Policy on Open Access and Supporting Guidance.* Available at: http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/RCUKOpenAccessPolicy.pdf (Accessed: 6 March 2016)

Ware, M. and Mabe, M. (2015) *The STM Report: An overview of scientific and scholarly journal publishing*. Fourth Edition. Available at: http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf (Accessed: 6 March 2016)

F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-119

Insights from Over a Decade of Electronic Publishing Research

Fernando Loizides[1] and Sam A. M. Jones
Emerging Interactive Technologies Lab
University of Wolverhampton, UK

Abstract. The work in this article presents findings from a text mining exercise of over a decade of research into electronic publication. We give readers insights into the past, present and possible future directions in a structured way, and further allowing them access to the extracted data in order to produce their own analysis and conclusion. We also produce our working methodology which can be replicated to produce systematic similar findings over the years as well as comparisons to other data souces.

Keywords. Electronic Publishing, Trends, Data Mining

1. Introduction and Methodology

The advent of the internet has brought about changes in the way that publishing takes place, both in terms of speed and cost. Research on electronic publishing has been reporting on current policies, stakeholder behavior and economic repercussions as well as other state of affairs. In order to understand the past, present, and perhaps even be able to hypothesize some of the near future of trending topics in electronic publishing, we present a text mining (Gupta 2009) exercise performed on over a decade of published material on electronic publications. In order to do this we take a sample from a leading electronic publication conference within Europe. The aim of the work is to identify to the reader areas in which research has steered over the years, report on likely trends and allow each reader to make their own inferences based on the data provided. Within the limitations such as scope, we aim to stimulate discussion and identify likely trends rather than absolute findings (a problem that would be more suitable to a big data analysis approach).

For our study we selected the corpus of the Electronic Publishing Conference from the years 2003 until 2015. A total of 462 full texts and 564 abstracts were extracted and converted into plain text format using a custom built scraper and format translator, built in the programming language python (*https://www.python.org*). The abstracts included in the full text documents but were also extracted as separate files in order for more specific analysis to be performed on these. We then performed a series of 'cleaning' activities on the data. This was required in order for a more effective lexical analysis to occur. The pseudocode for the cleaning process is as follows. (1) Remove common words (e.g. 'the' 'and') (2) Remove Specific Characters (e.g. '-' '.' '*' and punctuation

[1] Corresponding Author: fernando.loizides@wlv.ac.uk

marks) (3) Transform all text to lowercase (e.g. INFORMATION to information) (4) Remove Digits (it was deemed appropriate for our research that digits would not contribute to the findings) (4) Strip excess whitespace (5) Match Spelling (change all American to English or vice-versa) (6) Apply Porter Stemming Algorithm (Porter 1980) Stemming reduces words to their most basic state in order to identify similar words - e.g. 'visualize' and 'visualizing' would become 'visual'). The cleaning process took place using R (*https://www.r-project.org*). Once the cleaning process takes place, the documents were queried (again using R) as to the cumulative term frequency of each word across the documents (full texts and abstracts). This method is similar to the established 'analysis of co-occurring terms' (Buzydlowski et al 2002).

2. Findings and Discussion

The purpose of this section is to give the reader the high level findings from the data. The complete data findings can be found at (http://www.eitlab.net/wp-content/uploads/2016/03/elpub2016_DataMining.pdf). We focus on reporting a summary of the 'popular' areas in which electronic publishing has been reporting on. There are four subsections. The first presents the holistic findings from 2003 until 2015. The second section presents the findings from the last year of publication (2015) to report on the most recent trends but primarily to distinguish between the three year findings. The third section presents the cumulative findings of 2013, 2014, 2015 in order to distinguish a rate of change. The final section gives a more detailed year by year overview off the main findings from the last six years 2010-2015 to show longer term changes and suggest more stable areas.

2.1. 2003-2015

From the findings in the abstracts (See Table 1), we can see that almost no terms appear on more than 50% of the abstracts. Ignoring the words 'paper', 'use', 'publish', 'inform' and 'research' which are naturally occurring in an article, we highlight the words 'access' and 'develop'.

Table 1: Number of abstracts containing the same terms

Range	No of terms	Terms
324-524	0	
274-323	2	paper use
224-273	5	access develop inform publish research
174-223	5	base digit open present provid
124-173	16	also can content describ electron journal librari new project public result system technolog user web will

From the full texts (See Table 2) available since 2003, we are also able to highlight (occurring in over 75% of the documents) the terms 'access' and 'avail' (available, availability), while the documents also present several technology related terms such as 'http', 'technolog' and 'develop'.

Table 2: Number of documents (Full Texts) containing the same terms

Range	No of terms	Terms
412-462	4	can inform publish use
362-411	39	abstract access also avail base can confer data develop differ electron follow http import includ introduc keyword make need new one paper possibl present process provid refer relat research result system technolog time univers user web well will work
312-361	47	allow author case conclus content creat current digit discuss document exampl exist first form format group howev initi institut interest intern june level librari like manag mani may model number open order part proceed project public requir scienc search servic set specif support term text two way
262-311	83	activ addit anoth applic approach articl associ becom chang collect common communic communiti comput consid contain databas defin describ design direct distribut environ even experi field figur final find focus full function futur general high identifi implement increas integr issu journal knowledg languag larg link list made main mean object offer onlin particular point practice problem produc report repres resourc review see sever show sinc softwar sourc standard start still structur studi take technic three tool type version view wide within world year
212-261	119	abl academ accord ad address aim alreadi although among analysi appear appli archiv area aspect better book build call clear combin compar complet concept concern consist context contribut de descript detail effect eg elpub enabl end engin establish etc evalu expect fact featur found generat give given help human improv index indic individu internet involv last learn limit look mail major materi metadata method much must name nation natur necessari network non note now oper organ origin page perform person place potenti print product propos purpos qualiti question reason recent record relev repositori retriev right role scientif second select share signific similar social solut state step subject tabl th therefor thus toward tradit understand us valu various without word

2.2. 2015

The most recently available data come from the 2015 corpus and can be seen in Tables 3 (abstract) and Table 4 (Full Texts). Unsurprisingly we are able to distinguish the words ‘access’ and ‘open’ occurring in more than 50% of the abstracts, emphasizing the open access initiative that is highlighted and ever increasing in perceived importance by all stakeholders.

Table 3: Number of abstracts containing the same terms

Range	No of terms	Terms
20-26	0	
15-19	1	research
10-14	5	access also inform open paper
5-9	37	academ activ articl avail base benefit can current data develop digit experi find implement journal knowledg librari main new number particular practic present project provid public publish requir scienc scientif servic share studi use well within work

From the full texts for 2015, we are able to distinguish terms such as 'model', 'project' and 'manag', beyond the technological plethora of terms which are dominant.

Table 4: Number of documents (Full Text) containing the same terms

Range	No of terms	Terms
20-26	77	abstract access activ addit also associ author avail base can case consid correspond current data describ develop differ first follow form group howev http import includ increas inform institut interest introduc keyword knowledg level librari mail make model need new number one open order paper part particular practic present process project provid public publish refer relat requir research result review scienc servic share support system technolog term time univers use way web well will within work year
15-19	130	achiev aim allow alreadi among analysi anoth approach articl becom challeng chang collect common communic communiti compar conclus confer content context continu creat digit direct discuss distribut document enabl establish european even exampl exist expect experi field figur final format framework full futur general high https identifi impact implement improv initi integr intern issu journal key lead like link list made main major manag mani materi may mean measur method much must nation network now offer onlin organ particip place possibl potenti produc purpos qualiti question recent relev report repositori repres resourc right role scholar scientif search second see set sever sinc small social societi softwar solut sourc specif standard start state still structur studi subject suggest take technic text therefor third three topic toward two type user version wide

2.3. Three Years

Table 5 and 6 show the cumulative number of the terms occurring in abstracts and full texts respectively.

Table 5: Number of abstracts containing the same terms

Range	No of terms	Terms
46-57	0	
36-45	1	research
26-35	3	access open paper
16-25	12	base can data develop inform present provid public publish scienc use work

Following with the pattern of the previous observations in 2015, open access is reported in 50% or over of the documents. No other term (apart from the expected 'research' and 'paper') occur at this frequency.

Table 6: Number of documents (Full Text) containing the same terms

Range	No of terms	Terms
45-56	48	abstract access also author avail base can case current data develop differ follow http import includ inform institut introduc keyword level make mani need new one open paper possibl present process project provid public publish refer relat research result support system time univers use way web will work

35-44	98	activ addit aim allow analysi approach articl associ becom call chang collect communic communiti conclus consid content correspond creat describ digit discuss document even exampl exist experi figur final first focus form format futur group high howev identifi implement improv increas initi integr interest intern issu journal knowledg librari like link list made main major manag may mean method model nation number onlin order part particular practic recent report repositori require resourc review scienc search servic set share sinc social sourc start state structur studi take technic technolog term text therefor tool two type user well within year
25-34	158	abl academ accord achiev address alreadi among anoth appli applic archiv area basic best better challeng clear combin come common compar complet comput concept conduct confer consist contain context continu contribut core creation databas defin descript design detail direct distribut effect effort electron enabl end engin environ establish european expect extend fact factor field find found framework full function fund general give given global help human idea impact indic individu instanc interact investig involv lack languag larg lead learn least less long mail materi measur metadata might much must name natur network non now object offer often oper organ origin other output page perform place platform point polici potenti problem proceed produc propos purpos qualiti question read regard relev repres respons right role scholar scientif second see select sever show similar singl small societi softwar solut specif stage standard statist step still subject success suggest tabl third three thus topic toward valu various version view wide without world

The full texts of the 2013-2015 years provided more diversity in the results over the 50% level, with terms such as ‘social’, ‘knowledge’ and ‘management’ infiltrating and complementing the technological dominance and focus.

2.4. Year by Year (2010-2015)

We systematically studied each individual year between 2010 and 2015 (inclusive) and looked at the frequency of occurrences of words throughout the abstracts and full texts for that year. From the data we can see that there are 9518 terms appearing in either one, two, three or four papers. In 2014 the number of terms in this category is just 3159, an almost threefold decrease. One might be tempted to conjecture that there were more small clusters of similar papers in 2014 than in 2010 but if one considers the infamous birthday problem we are led to believe that there is less evidence for such a social explanation. The birthday problem (Wagner 2002) states that as the size of some collection of objects increases the number of ways of finding pairs within this collection increases much faster than the size of the collection. In 2014 there were 15 papers and in 2010 there were 35 (our dataset contains 32 of these 35 papers) so it is not surprising that there are many more terms in the one to four paper category – there are more ways of picking a pair or triplet or quartet of papers in 2010. In general it does look like there is some correlation between words appearing often in the abstracts and that same word appearing often in the papers (note again that we are looking across all documents here. A word appearing often in one particular abstract does tend to imply that it will occur often in the full text of that paper). It also appears that those words which do occur often in abstracts are words relating to general research, they do not necessarily appear to be particularly closely related to electronic publishing. When looking at the full text of the papers it appears that words which appear often are more likely to have something to do with publishing or electronic publishing in general. 2010 was a year with a large number of documents and so the numbers towards the bottom of the table tend to get very large. Interestingly, the phrase "http" appears in every single document from 2010. At a first glance it may seem that http appears as it is the standard prefix of URLs which often appear in the references section of academic papers. However, after some inspection it is clear that http appears in many of the papers outside of the references section and in some papers as a stand-alone term, although in the vast majority of cases it is still the prefix of a URL. There are seven words appearing in all of the papers from 2011. This number is 21 for 2012. In 2012 there were 30 accepted papers but our dataset contains only 19 of them. In 2011 there were 24 papers but the dataset also contains 19 documents. Whilst there are clearly many papers missing from the 2012 dataset it is interesting that the datasets have equal size and yet there are three times as many terms appearing in every paper in 2012. From the table we can see that the words which do not appear in every paper in 2011 are words such as "system", "technology" and "digital" it might be reasonable to assume that the papers in 2012 were more focused on the systems and technological aspects of electronic publishing.

3. Conclusions and Future Work

In this work we present the findings from a text mining exercise of over 550 documents from over a decade of articles on electronic publication. We present the data in a structured form in order to give the readers, and different stakeholders, insight into the findings. Some high level comments identify interesting areas where terms correlate, without making any claim on statistical or predictive models for future directions for electronic publishing. The fragmented nature of the data set has a larger impact when trying to analyse trends from individual years than when trying to analyse trends over all the years. It is clear therefore that this section could be improved by improving the quality of the dataset. This could be done by obtaining access to more of the papers or by improving the quality of the text extraction from the PDF files. A possible direction for further work would be to extract some sort of contextual data from the files. Much of the analysis in this paper is based around the frequency of which words appear but if one were able to extract some contextual data it may be that more insightful inferences could be made. This is likely to be a difficult problem involving complicated NLP and text mining techniques and big data volume rendering techniques. We aim to continue the work on a much larger scale to achieve two goals. Firstly, verify if our data has further ecological validity and secondly, to identify further insights.

References

Buzydlowski, J.W., White, H.D. and Lin, X., 2002. Term co-occurrence analysis as an interface for digital libraries. In Visual interfaces to digital libraries (pp. 133-144). Springer Berlin Heidelberg.

Gupta, V. and Lehal, G.S., 2009. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), pp.60-76.

Porter, Martin F. "An algorithm for suffix stripping." Program 14, no. 3 (1980): 130-137.

Wagner, D., 2002. A generalized birthday problem. In Advances in cryptology—CRYPTO 2002 (pp. 288-304). Springer Berlin Heidelberg.

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-125

Article Processing Charges: A New Route to Open Access?

Gerald BEASLEY [1]
University of Alberta, Canada

Abstract: Article Processing Charges (APCs) have recently been studied as a means towards a sustainable Open Access (OA) environment for scholarly communications. However, APCs at any level represent a substantial economic barrier to the authors, institutions, funding agencies and governments that many of its advocates most wish to serve through OA initiatives.

Keywords: Article Processing Charges, Open Access

1. Introduction

Article Processing Charges (APCs) used as a means to offset publication expenses are not particularly new features on the scholarly communications landscape. A good summary of their usefulness is to be found on the publications page of the Public Library of Science (PLOS) web site:

"To provide Open Access, PLOS uses a business model that includes Article Processing Charges (APCs) to offset expenses – including those of peer review management, journal production and online hosting and archiving. Authors, institutions or funders are charged this publication fee for each article published." [2]

Furthermore, APCs are increasingly adopted by funding agencies as eligible expenses in policies seeking to promote open access to publicly funded research outputs.[3] They are, in fact, a generally accepted and acceptable way to unlock scholarly journal articles from behind subscription paywalls. It is well documented that recently subscription fees paid by libraries for academic journals have been increasing beyond the rate of inflation in North America and elsewhere (Whitehead 2016). APCs, on the other hand, are generally paid to publishers by the author(s), their parent institutions or granting agencies, with the expectation that subscription fees are either eliminated or at least substantially reduced as a result.

[1] Corresponding Author: gbeasley@ualberta.ca

[2] https://www.plos.org/publications/ (Accessed: March 12, 2016)

[3] See for example Canada's 2015 Tri-Agency Open Access Policy on Publications which includes as an eligible expense "Page charges for articles published, including costs associated with ensuring open access to the findings (e.g., costs of publishing in an open access journal or making a journal article open access)." Available at: http://www.science.gc.ca/default.asp?lang=En&n=F6765465-1 (Accessed: March 12, 2016)

In the wake of the Finch Report (Finch 2012) and its subsequent implementation in the United Kingdom there are several research projects under way which seek to analyze the financial viability of providing open access more globally to academic journal literature. One idea is to replace many or all subscription fees with APCs. The present article – actually more of a thinkpiece than a scholarly contribution – offers a tentative response to this more recent phenomenon[4].

2. Two Research Projects

Two research projects are particularly notable in this regard. The University of California, under the leadership of UC Davis and the California Digital Library and with support from the Andrew W. Mellon Foundation, is undertaking "Pay It Forward: Investigating a Sustainable Model of Open Access Article Processing Charges for Large North American Research Institutions." The project includes partnerships with three major research libraries (Harvard University, Ohio State University and the University of British Columbia) as well as the Ten University of California campus libraries. The rationale for this project is articulated in the grant submission, available online:

"The key question that the proposed project asks is whether a large-scale conversion to open access scholarly journal publishing funded via APCs would be viable and financially sustainable for large North American research-intensive institutions, whose faculty currently author a significant percentage of the world's research."[5]

Similarly, a Max Planck Digital Library Open Access White Paper was published on April 28, 2015. The White Paper provides data to support the authors' conclusion that:

"… a large-scale transformation of the underlying business model of scientific journals is possible at no financial risk. Our own data analysis shows that there is enough money already circulating in the global market – money that is currently spent on scientific journals in the subscription system and that could be redirected and re-invested into open access business models to pay for APCs" (Schimmer 2015).

The scope of these two major contributions to the scholarly communications debate are by no means identical. The quality and value of the data on which their findings are based need not be doubted. However, they are united in their focus on exploring APCs as a means towards a sustainable Open Access environment for scholarly communications. Both studies appear ready to indicate that large research libraries would realize a substantial economic gain on behalf of their parent institutions

[4] Although the conclusions are my own, I would like to thank several individuals who helped me to arrive at them: Prof. Michael McNally; Kathleen Shearer; Janet Williamson; Prof. Heather Zwinger.

[5] University of California Libraries (2014). Pay It Forward: Investigating a Sustainable Model of Open Access Article Processing Charges for Large North American Research Institutions. Available at: http://icis.ucdavis.edu/wp-content/uploads/2014/06/UC-Pay-It-Forward-narrative-2014-FINAL.pdf (Accessed: March 12, 2016)

if APCs replaced subscription costs as the principle underlying business model for scientific journals.[6]

3. APCs and Social Inequality

I believe these two studies – both of which may be regarded as ongoing – present longstanding advocates for OA such as myself with a new challenge. It seems to me that a business model for scholarly communications largely or entirely based on APCs, whether or not it is financially advantageous for research-intensive institutions, needs to be questioned: does it not represent a substantial economic barrier to the authors, institutions, funding agencies and governments we would most wish to serve through OA initiatives?

Such a question leads in many directions, one of which involves a reconsideration of the initial motivations behind the open access movement. Not surprisingly, the Wikipedia entry for Open Access offers an excellent summary of these and provides extensive references to Peter Suber's outstanding 2012 monograph on the topic (Shuber 2012). The entry notes the technological and economic rationale for OA before stating that

"*The OA movement is motivated by the problems of social inequality caused by restricting access to academic research, which favor large and wealthy institutions with the financial means to purchase access to many journals, as well as the economic challenges and perceived unsustainability of academic publishing*."[7]

It is this concept of social inequality that seems not to be sufficiently well addressed by business models for scholarly communications based on APCs. A system based on author payments continues to favor those authors affiliated with organizations that have the means to pay. Both philosophically and literally, no matter how low the charge, there will be authors, institutions, funding agencies and governments unable to afford the cost of APCs.

This is not an original insight: in a seminal article by Peterson, Emmett and Greenberg in the Journal of Librarianship and Scholarly Communication, the authors note that *"Lurking behind the joy of "the reader gets free access" are subtle assumptions and ethical dilemmas that arise on the author side of the equation. Averting new inequities as the OA movement gathers momentum is critical"* (Peterson 2013).

These inequities will be seen and experienced differently by different groups. Red flags have already been raised in relation to scholars working in Humanities disciplines by one of the initial Co-Directors of the Open Library of Humanities (Eve 2014). But in truth the problem is even larger: political scientists may legitimately argue that OA models proposing APCs form part of an unintended neocolonialism – or should that be neoimperialism? – in which first world publicly and privately funded research smothers the work of academic researchers in developing nations as well as the *bona fide* research of other excluded or underprivileged communities, at home and abroad. Such an outcome would perpetuate and even reinforce an already well-

[6] This forseeable economic gain is not based solely on an analysis of current variable costs compared to the new model. Both studies also contain valuable assessments of potential financial implications for research institutions in the future.

[7] https://en.wikipedia.org/wiki/Open_access#CITEREFSuber2012 (Accessed: March 12, 2016)

documented system of discrimination by which important groups are denied the privilege of seeing their research disseminated through generally accepted vehicles of scholarly communication.

APCs do address many of the financial concerns that have been used by OA advocates ever since the movement was founded. Sadly, however, it seems to me that they provide a route to satisfying the terms of foundational documents such as the Berlin Declaration on Open Access[8] and the Budapest Open Access Initiative[9] without solving the more knotty problem posed by the concept of social inequality. Thanks to the internet, I think most of us would agree that there is no longer any technological justification for this. I have heard the financial argument that if APCs were adopted in place of subscription costs then research institutions would save so much that they could collectively repurpose a portion of the savings to address any inequities in the new system. Unfortunately, in addition to being vulnerable to all kinds of ethical pitfalls, there is insufficient evidence of this rebalancing effort in the current subscription-based environment, where it would seem to be equally valid. Stronger arguments may be based on the very real danger that without APCs we are left with an untenable *status quo*. There is no easy answer to this, but one way to develop a response is by briefly considering a theoretical framework for understanding APCs in relation to disruption theory.

4. APCs and Innovation

Disruptive innovation is a much-discussed business term coined by Harvard Business School Professor Clayton M. Christensen:

"*Disruptive innovation ... describes a process by which a product or service takes root initially in simple applications at the bottom of a market and then relentlessly moves up market, eventually displacing established competitors.*"[10]

Since introducing the concept some 20 years ago, Prof. Christensen and his followers have had to spend more and more time distinguishing disruptive innovation from other forms of market change (Christensen 2015). Although postsecondary education has frequently been a target of the theory, it has not to my knowledge been applied to OA. In my opinion, however, OA was originally developed as a proposal that would lead to exactly the kind of disruptive innovation described by Christensen's theory, i.e., allowing new market entrants – specifically OA journal publishers – to displace their established, subscription-based rivals. An OA world based on APCs, by contrast, appears instead to meet Christensen's definition of a *sustaining* innovation:

"Disruption theory differentiates disruptive innovations from what are called 'sustaining innovations.' The latter make good products better in the eyes of an incumbent's existing customers: the fifth blade in a razor, the clearer TV picture, better mobile phone reception. These improvements can be incremental advances or major breakthroughs, but they all enable firms to sell more products to their most profitable customers."[11]

[8] http://openaccess.mpg.de/Berlin-Declaration (Accessed: March 12, 2016)

[9] http://www.budapestopenaccessinitiative.org/read (Accessed: March 12, 2016)

[10] http://www.claytonchristensen.com/key-concepts/ (Accessed: March 12, 2016)

[11] Ibid, 47-8

Setting aside the many obvious differences between academic and commercial business, it does seem possible to assert that an open access world based on APCs would make the journal article look better in the eyes of existing customers by providing more value for money. However, like all sustaining innovations, it might also maintain a kind of exclusive *status quo* for both market incumbents and their customers.

5. Conclusion

If the widespread and largescale adoption of APCs does not provide an optimal route to an OA environment for scholarly communications, because it fails to address the issue of social inequality, then what is the better path? empowering the voices of those who are currently disenfranchised: the research- oriented authors, institutions, funding agencies and governments who already face economic barriers, whether to research access or dissemination, but whom we would like to include in our more equitable – and therefore more productive – future. Given the heterogeneity of these disenfranchised groups it seems obvious there will be no single solution. Forums open to diverse international and multidisciplinary perspectives such as Elpub might well provide good starting points.

It also seems obvious, at least to this author, that such a question cannot be answered without questioning the long term future of the academic journal. After all, last year was declared to be the 350th anniversary of its birth, marked by the publication of the first volumes of the *Philosophical Transactions* (London, 1665) and *Le Journal des scavants* (Paris, 1665)[12]. Such an anniversary surely provided an excellent opportunity to reflect on the value of perpetuating the dominance of the academic journal: a first world product intended for a first world audience. This opportunity seems to have been largely missed. Yet the question remains: designed to use a particular 17th century technology to solve a particular 17th century problem, does history in the modern period justify the academic journal's – and journal publisher's – continued hegemony in the world of scholarly communications?

References

Whitehead, M. and Owen, B. (2016). *Canadian Universities and Sustainable Publishing (CUSP): A White Paper prepared on behalf of the Canadian Association of Research Libraries.* Available at: http://www.carl-abrc.ca/uploads/pdfs/Can_Univ_Sustainable_Publishing_2016.pdf (Accessed: March 12, 2016)

Finch, *Dame* J. (2012). *Accessibility, sustainability, excellence: how to expand accessto published research findings: Report of the Working Group on*

[12] For an excellent overview of the growth and present state of the scholarly journal industry, see Larivière V, Haustein S, Mongeon P. (2015) 'The Oligopoly of Academic Publishers in the Digital Era.' PLOS ONE 10(6): e0127502. doi: 10.1371/journal.pone.0127502

Expanding Access to Published Research Findings. Available at: http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report- FINAL-VERSION.pdf (Accessed: March 12, 2016)

Schimmer, R., Geschuhn, K.K., & Vogler, A. (2015). *Disrupting the subscription journals' business model for the necessary large-scale transformation to open access.* Doi:10.17617/1.3

Suber, P. (2012). *Open Access* (The MIT Essential Knowledge Series ed). Cambridge, Mass.: MIT Press.

Peterson, A.T., Emmett, A., Greenberg, M.L. (2013) 'Open Access and the Author-Pays Problem: Assuring Access for Readers and Authors in a Global Community of Scholars', *Journal of Librarianship and Scholarly Communication,* 1(3):eP1064. http://dx/doi.org/10.7710/2162-3309.1064

Eve. M. (2014). 'All that Glisters: Investigating Collective Funding Mechanisms for Gold Open Access in Humanities Disciplines', *Journal of Librarianship and Scholarly Communication* 2(3):eP1131. http://dx.doi.org/10.7710/2162-3309.1131

Christensen, C.M., Raynor, M., & McDonald, R. (2015) 'Disruptive innovation? Twenty years after the introduction of the theory, we revisit what it does – and does not explain', *Harvard Business Review,* 93(12):44-54.

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-131

A glance on presence of Shahid Beheshti University scholars in Research Gate

Amir Reza ASNAFI[1]
Faculty member of Information Science and Knowledge Department, Shahid Beheshti University

Abstract. Using web 2.0 capabilities in research fields have provided various facilities for scholars. By these capabilities, people can interact together and share their publications with large range of other scholars. Current research aim is study on presence of Presence of Shahid Beheshti University Scholars in Research Gate. Used approach in this paper is Scientometrics with Altmetrics method. For data gathering, page of Shahid Beheshti University in Research Gate was used. Findings indicated that courses of Chemistry, Laser and Plasma and Physics had the most presence in Research Gate. This paper revealed that Humanities courses in Shahid Beheshti University had not any serious activities in Research Gate. Establishing some workshops on using academic social networks can effect on knowledge of scholars, faculty member and students and direct their information seeking way for these people.

Keywords. Shahid Beheshti University, Scientific Collaboration, Academic Social Network, Research Gate

1. Introduction

Ben Shneiderman (2008) based on collaborative approach, suggested a new kind of research. He called it Research 2.0. Innovation, information sharing, fast access to information, creating online brain storming and so on, is some productions of new research approaches. So, using scientific social networks scholars have a scientific channel for information seeking and can present their publications in order to receive more citations. Mohammadi & Thellwall (2014) found that Mendeley readership data could be used to help capture knowledge transfer across scientific disciplines, especially for people that read but do not author articles, as well as giving impact evidence at an earlier stage than is possible with citation counts. Thellwall & Kousha (2014) stated that Research Gate is a social network site for academics to create their own profiles, list their publications and interact with each other. It provides a new way for scholars to disseminate their publications and hence potentially changes the dynamics of informal scholarly communication.

So, current research has two major aims: 1. Identification Shahid Beheshti University action in Research Gate based on RG score and Impact Points indicators. And 2. Identification action of Shahid Beheshti University courses in Research Gate based on indicators like: number of members, documents, visits and downloads.

[1] Corresponding Author: *aasnafi@gmail.com*

2. Research Method

In current research using Scientometrics and Altmetrics approaches, Research Gate as an academic social network has been used. In February 2015 Shahid Beheshti University name was searched in Research Gate. Then active members were extracted. Profile of active scholars of this university in Research Gate was studied. Indicators like: RG Score, Impact Points, Publications, Citations, Downloads and Views were used for studied scholars.

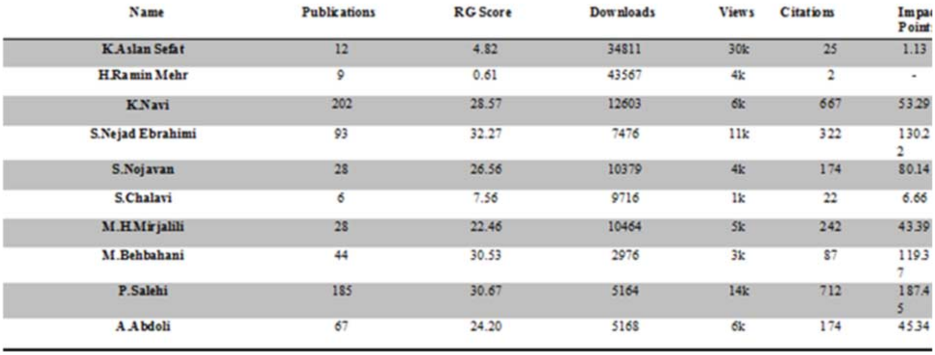
3. Findings

Findings indicate that 1604 people from Shahid Beheshti University are members of Research Gate. 1636 works were uploaded in this scientific social network. RG Score of Shahid Beheshti University Scholars was 3481.57 and their Impact Points was 2364.2. Current paper revealed that scholars from USA, Iran and China have the most download papers of Shahid Beheshti University scholars.

Table 1: Presence of Shahid Beheshti Faculty members in Research Gate

Members	RG Score	Publications	Impact points
1604	3481.57	1636	2364.2

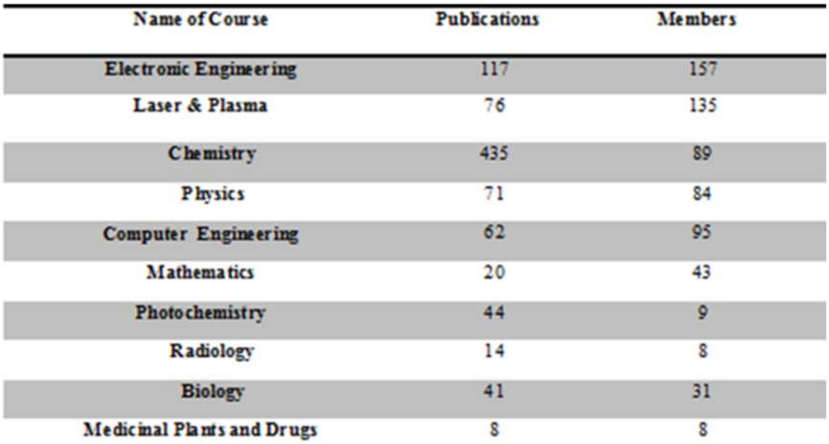
Table 2: Active scholars of Shahid Beheshti University in Research Gate



Name	Publications	RG Score	Downloads	Views	Citations	Impa Point
K.Aslan Sefat	12	4.82	34811	30k	25	1.13
H.Ramin Mehr	9	0.61	43567	4k	2	-
K.Navi	202	28.57	12603	6k	667	53.29
S.Nejad Ebrahimi	93	32.27	7476	11k	322	13022
S.Nojavan	28	26.56	10379	4k	174	80.14
S.Chalavi	6	7.56	9716	1k	22	6.66
M.H.Mirjalili	28	22.46	10464	5k	242	43.39
M.Behbahani	44	30.53	2976	3k	87	11937
P.Salehi	185	30.67	5164	14k	712	1874 5
A.Abdoli	67	24.20	5168	6k	174	45.34

So, it can be found that P.Salehi and S.Nejad brahimi were the most active scholars of Shahid Beheshti University in Research Gate.

Table 3: The most active courses of Shahid Beheshti University in Research Gate



Name of Course	Publications	Members
Electronic Engineering	117	157
Laser & Plasma	76	135
Chemistry	435	89
Physics	71	84
Computer Engineering	62	95
Mathematics	20	43
Photochemistry	44	9
Radiology	14	8
Biology	41	31
Medicinal Plants and Drugs	8	8

Current paper indicated that Humanity Sciences courses in Shahid Beheshti University have not presence in Research Gate.

3. Findings

Current research revealed that scholars from Chemistry, Laser and Plasma and Physics have the most presence in Research Gate, This represents that Shahid Beheshti University researchers examined the use of science and modern communication tools for interaction and cooperation between research, teaching and other researchers. This paper showed that Humanity Sciences scholar's presence was very weak. This could be due to their lack of familiarity with the network or the inability to work and interact with web based social networks. It should be noted that in the new era of research and education, science social networking websites such as: Academia, Research Gate Mendeley have essential role in the development of education and research. Millions of users around the world use these networks and easily can access to their needed information. They are not limited to space of libraries or search engines and their information seeking channels have been changed. So scholars should adapt themselves with new Information context. Interaction of Humanity Sciences scholars with the web setting to do research or to communicate with other researchers can lead to emergence of new ideas. Presence of Shahid Beheshti University researchers in the scientific social network helps to visibility of their publications.

References

Mohammadi, E. and Thelwall, M., 2014. Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. Journal of the Association for Information Science and Technology,65(8), pp.1627-1638.

Schneiderman, B 2008. Science 2.0. Science.319: 1349-1350. Available in: http://www.cs.umd.edu/~ben/papers/Shneiderman2008Science.pdf

Thelwall, M. and Kousha, K., 2015. ResearchGate: Disseminating, communicating, and measuring Scholarship?. Journal of the Association for Information Science and Technology, 66(5), pp.876-889.

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-135

FOSTER's Open Science Training Tools and Best Practices

Astrid ORTH[a,1], Nancy PONTIKA[b] and David BALL[c]
[a] *Georg-August-Universität Göttingen, Niedersa chsische Staats- und Universitätsbibliothek, Göttingen, Germany*
[b] *Open University, UK*
[c] *SPARC-Europe*

Abstract. FOSTER is an EU project aiming at identifying, enriching and providing training content on relevant Open Science topics in support of implementing EC‘s Open Science Agenda in the European Research Area. During the previous two years a wealth of training resources have been collected, which are now presented in a dedicated training portal. The paper describes how to use the FOSTER training platform and the tools available to identify suitable training materials and create modular e-learning courses.

Keywords. Open Science, e-Learning, Training Materials, European Research.

1. Introduction

Recently we have witnessed significant debate and activity surrounding the movement to make research papers, data, and scientific information available, free of cost and with limited rights restrictions, to all readers online. Open Access as well as Open Data policies have been championed across the European Research Area (ERA), and feature prominently in the recommendations of Horizon 2020, the European Commission's research and innovation programme (EC, 2014). Moreover, since the launch of the FOSTER Project, Open Science has seen both grass-roots demand by young researchers (McDowell et al., 2015), as well as becoming a centerpiece of new agenda on Open Innovation (EC, 2015). In this environment, FOSTER[2] , a two-and-a-half year EU project, is designed to facilitate Open Science adoption by early career researchers, established scholars, librarians, library managers, research administrators, funders and other research stakeholders. Particular focus is placed on key skills necessary to adopt Open Science in the daily research routines. Open Science, as defined by the project, is the practice of research in a transparent, sharable and collaborative manner, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproducibility of methods and/or results (FOSTER, 2015).

[1] Corresponding Author: orth@sub.uni-goettingen.de.

[2] FOSTER - Facilitating Open Science Training for European Research - is funded through the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant Agreement No 612425. See http://fosteropenscience.eu.

The project therefore focuses on identifying, enriching and providing training contenttent on all relevant topics in the area of Open Science for the European research community. It started in February 2014 with the following objectives:

- Support different stakeholders, especially young researchers, in complying with the open access policies and rules of participation set out for Horizon 2020;
- Integrate open access principles and practice in the current research workflow by targeting the young researcher training environment;
- Strengthen the institutional training capacity (beyond the FOSTER project);
- Facilitate the adoption, reinforcement and implementation of open access policies from other European funders in partnership with the PASTEUR4OA project.

These objectives were realized through the combination of 3 main activities:

1. Identify existing re-usable training content, repackage and reformat them to be used within FOSTER, and develop/enhance content if/where needed;
2. Create a portal to support e-learning and dissemination of training materials and a help desk;
3. Deliver face-to-face training, especially training multipliers, to carry on further training and dissemination activities within their institutions, countries or disciplinary communities.

Materials and tools developed by project partners and collected with the help and support of relevant communities of researchers, librarians and other stakeholders, are now available on the FOSTER portal and will be introduced in detail below.

2. FOSTER Open Science Training Portal

Our dedicated platform, the FOSTER Open Science Training Portal, serves as a single hub of information for collecting, storing and disseminating training content on Open Science to a variety of stakeholders at different knowledge levels, in various formats and for different usage scenarios. Launched as an early preview version in September 2014, it has now developed into a mature system. Functions that support the project's objectives can be grouped into the two areas introduced below.

2.1. Training Content: Upload, Categorisation and Navigation

Materials suitable for training targeted stakeholders were collected using two methods:

- Project partners identified and reviewed existing but widely scattered materials in a joint exercise. An open Call for Content was launched engaging interested stakeholders in the search for and contribution of re-usable training content.

- Training organisers of all FOSTER-funded training events (cf. activity 3 above) provided their training materials for re-use.

Training content collected by the means described above was analysed to identify a suitable classification system. Based on previous efforts to classify open science, and in particular research data management training outputs, the classification of the DaMSSI-ABC project[3] was chosen as a starting point. DaMSSI-ABC aimed to classify course offerings to ensure that participants are able to select the training that best meets their particular learning objectives. The following is the list of metadata fields adapted for use within the FOSTER project: Title, General description of the resource, Author/creator, Date, URL of the resource, Language, Target audience, Scientific discipline, Level of knowledge, Main topic, Secondary topic, Resource licence, Media type, File type, and Size.

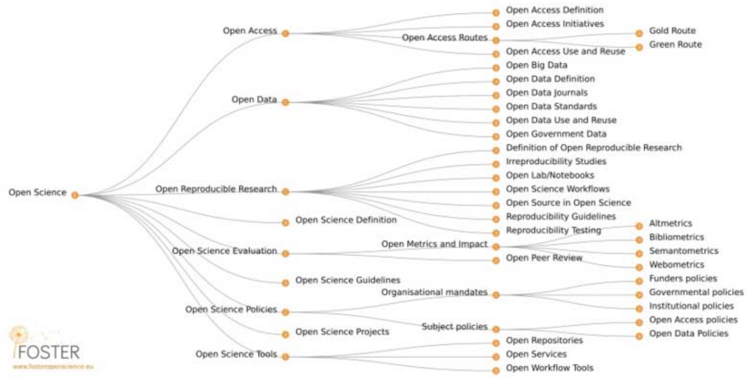


Figure 1. FOSTER Open Science Taxonomy.

For the 'Topic' part of the classification scheme the FOSTER Open Science Taxonomy was developed. As explained by Pontika et al. (2015) this taxonomy was not only created to enable navigation and searching of the portal content. It has also been used to link and recommend related content items; to provide a structure to which users can subscribe to receive content updates; and for content experts to be notified of items that need review or raise questions that should be answered. Furthermore the map of topics provides an overview to inform learners of the existence and relationships of areas that comprise Open Science (cf. Figure 1). Finally it serves as a means to check whether the FOSTER training content covers all relevant areas in the field and whether there are any gaps in the topics explored in the Open Science agenda.

Collecting Open Science training resources in one place under a common classification scheme helps individuals identify their training needs and contribute to satisfying them. Readers may search for topics (e.g. by using the taxonomy), map their own interests by comparing indicated audience and learning levels, and use the provided material for self-study. Any individuals and contributors in the Open Science

[3] DaMSSI: Data Management Skills Support Initiative - Assessment, Benchmarking, Classification. See http://www.dcc.ac.uk/training/damssi-abc.

community are permitted and invited to add content, by creating an account to the portal, to store their materials for greater distribution and re-use.

2.2. e-learning Courses: self-learning and course creation

In this paper so far we have discussed, how the content of the FOSTER training portal is being collected and how learners can identify resources that help them fulfil their learn training activities. In today's distributed and networked environments it might be useful for trainers to think also about e-learning and blended learning formats. The advantages of self-learning (at students' own pace, location-independent and asynchronously with self-assessment) can be combined with benefits of teacher-led training (synchronous interaction with teacher and classmates, higher engagement and motivation).

The FOSTER training portal offers tools for creating online training courses, including definition of learning objectives, multimedia course content, quizzes for self-assessment and discussion forums for supporting learners. Several courses on different areas of Open Science, such as the 'Introduction to Open Science', 'Open Access to publications' and the 'Horizon2020 Open Data Pilot', are already available. Several of these were run as blended learning courses combining self-learning with interactive webinars. Similar to the content on the FOSTER portal, the e-learning courses can be proposed and created by any individuals, again registration to the portal is necessary. This functionality is critical in order to allow various recombinations of the content into modules that best fit local learning needs, and train all stakeholders in the academic ecosystem.

3. FOSTER Training Programme

By means of funding training events throughout Europe a large number of participants was educated on different aspects of Open Science on various knowledge levels. Two open Calls for Training were conducted: for 2014 45 proposals from 19 countries were submitted and 80 proposals from 28 countries for 2015. Consortium partners additionally held training sessions and were invited to give presentations through the Speaker Directory available from the project website[4].

Preliminary results of the FOSTER training programme indicate that in total more than 100 training events with over 4600 participants were co-organised and/or funded by FOSTER. A great diversity of approaches (institutional, national, and discipline-specific), geographies and languages could be observed. Most prominent themes of the training courses were Research Data Management/Open Data as well as Open Access and Open Science. Subjects covered included policies, legal and ethical issues as well as evaluations. Training events aimed at all stakeholder groups (researchers, students, project managers, research administrators, librarians and policy makers; cf. Figure 2).

Materials from training events were added to the FOSTER training portal and are available for individual learning or for the creation of new training activities. Two tools

[4] FOSTER Speaker Directory: see https://www.fosteropenscience.eu/project/index.php?option=com_speaker&view=speakers&Itemid=192

developed by project partners to help in re-using these training materials are described below.

3.1. Training Toolkit

The FOSTER training toolkit[5] explains how an instructor can organize a successful Open Science training course. The purpose of its creation was to maintain a consistent quality of the FOSTER funded training activities.

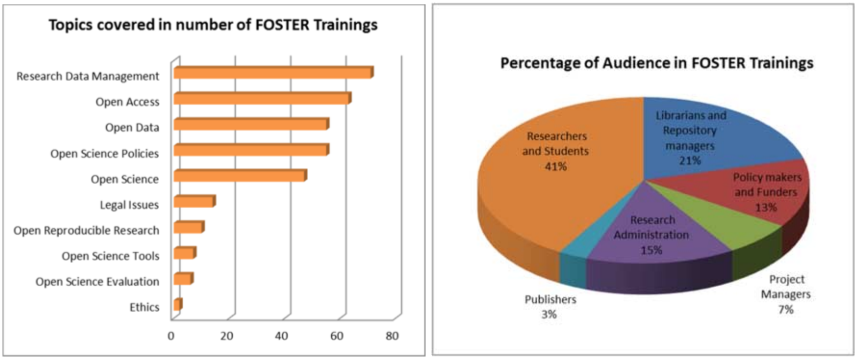


Figure 2. FOSTER Training Topics and Audiences.

Additionally, it ensured that resulting training materials are re-usable, for instance by giving tips and best practices on video recording of training sessions and licensing of contributions. It contains many training examples, which can serve as templates and best practices.

3.2. Learning Objectives

When selecting training materials either for self-learning or inclusion in new training contexts, learning objectives are core elements for mapping the learners' needs to available training resources. As detailed by Grigorov et al., (2015) FOSTER learning objectives are structured by Open Science Topics according to a functional Open Science Taxonomy (Pontika et al., 2015), reflecting the main responsibilities of each stakeholder along the Research Lifecycle. Specific Learning Objectives are structured in increasing levels of competence, frequently ending with successful integration of Open Science best practices in the daily research routine, facilitating self-assessment of personal workflow. The FOSTER learning objectives assist trainers with identifying related core learning elements and resources. These are additionally mapped to audiences and divided into three knowledge levels (introductory, intermediate, and advanced).

[5] FOSTER Training Toolkit: https://www.dropbox.com/s/ccrgkd0d6cizj4u/D4.2%20-%20Toolkit%20for%20Training.pdf?dl=0

4. Experiences and Recommendations

First numbers and feedback on the FOSTER training events demonstrate the successful approach of funding community-driven training courses. By means of engaging the targeted stakeholder groups their different perspectives and necessities could be integrated in design and content of the training sessions. The resulting training programme had a wider reach, clearer focus on the needs of the respective participants, and could also be held in many more languages than the consortium would have been able to provide on its own. Training organisers valued not only the financial support, but also consulting and recommendations on formats, topics and speakers for events. Participants ranked the quality of the trainings between 'good' and 'excellent'. According to the evaluation forms, attendants highly valued the quality of speakers and training materials provided, and appreciated the wide range of topics that were discussed during the majority of trainings. (Schmidt et al., 2016) The training programme is currently being evaluated in more detail. Case studies on training experiences and recommendations will be published.

5. Conclusions

FOSTER aims to identify and enrich training content in Open Science and assist researchers in searching and locating it in one central point for self-learning purposes. In addition, its goal is to provide a collection of resources that will help trainers during the planning of their training activities, which can also be used as primary resources of information in these sessions. For this purpose a wealth of high quality training materials has been collected in the FOSTER Open Science Training Portal. The tools that help with identifying and re-using the training resources described in this paper are the: Open Science Taxonomy, Training Toolkit, Learning Objectives and e-learning courses. Case studies with best training practices will complete the offering.

6. Acknowledgements

The authors gratefully acknowledge the contribution of all project partners to the work which is described in this article. The preliminary data about FOSTER training events were contributed by Matteo Cancellieri (Open University) and Gwen Franck (EIFL) and analysed by Remedios Melero (IATA-CSIC).

References

European Commission. (2014) Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. Available at: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (Accessed: 6 March2016).

European Commission. (2015) Open Innovation, Open Science, Open to the World EC SPEECH/15/5243 22 June 2015. Available at: http://europa.eu/rapid/ press-release_SPEECH-15-5243_en.htm (Accessed 6 March 2016).

McDowell GS, Gunsalus KTW, MacKellar DC et al. (2015) Shaping the Future of Research: a perspective from junior scientists [version 2; referees: 2 approved]. F1000Research, 3:291. Available at: http://dx.doi.org/10.12688/ f1000research.5878.2 (Accessed 6 March 2016).

FOSTER. (2015) Open science definition. Available at = https://www.fosteropenscience.eu/taxonomy/ term/100 (Accessed: 6 March 2016).

Grigorov et al. (2015) FOSTER Open Science Learning Objectives. Zenodo. Available at: http://dx.doi.org/ 10.5281/zenodo.15603 (Accessed: 6 March 2016).

Schmidt et al. (2016) 'Stepping up Open Science Training for European Research', to appear in Publications, Special Issue on Current Operational Issues in Open Access.

Pontika, N., Knoth, P., Cancellieri, M. and Pearce, S. (2015) 'Fostering Open Scienceto Research using a Taxonomy and eLearning Portal', i-Know - 15th International Conference on Knowledge Technologies and Data Driven Business. 21 - 22 October. Graz, Austria.

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-142

Stepping Up Towards Greater Alignment of Repository Networks

Katharina MÜLLER[a,1], Arvid DEPPE[a], Maxie GOTTSCHLING[a], Eloy RODRIGUES[a], and Kathleen SHEARER[a]

[a] *Confederation of Open Access Repositories (COAR e.V.)*

Abstract. Over the past few years, the Confederation of Open Access Repositories (COAR) has established the initiative “Aligning Repository Networks” in order to foster global repository interoperability and align policies and practices internationally. This paper outlines several activities to align major repository networks across the globe strategically, technically and on the service level.

Keywords. Repositories, interoperability, harmonization, standards

1. Introduction

Digital repositories are becoming increasingly important components of the scientific communication system. Their value lies in the potential to interconnect them to create a network of repositories, a network that can provide unified access to research outputs and be (re-) used by machines and researchers. Although there are unique requirements in each jurisdictional context, repository networks must be aligned across the world in order to support the truly global nature of research and scholarly communication.

Aligning repository networks is important for a number of reasons (Shearer 2014):

- To develop a seamless scientific infrastructure that supports the needs of researchers at the global level.
- To provide evidence to national funding bodies and governments that local and regional services are being developed in parallel with international activities.
- To provide uniform information to governments and funding agencies about the impact of open access policies.
- To avoid duplication of work across networks and enable cost synergies in areas of common interest.

[1] Corresponding Author. COAR e. V. Office, c/o State and University Library Goettingen, Platz der Goettinger Sieben 1, 37073 Goettingen, Germany; E-Mail: office@coar-repositories.org

COAR[2] has been working on international alignment of repository networks at three levels: (1) strategic, (2) technical and semantic interoperability and (3) common services.

The alignment work is being undertaken via several mechanisms including bilateral meetings and agreements between the major networks, international working groups, and informal conversations with regions/countries not yet involved.

2. Aligning of Regional Repository Networks

COAR has undertaken several activities over the last year to foster global repository interoperability through the initiative "Aligning Repository Networks"[3]. In this initiative several major international organizations emphasize their support for immediate open access to research articles and harmonized standards and common vocabularies. The joint committee for Aligning Repository Networks acts as a forum for exchange between repository networks and provides a global voice promoting the role of repositories as critical research infrastructure. Therewith, these joint programs strengthen open access managed as a commons within the scholarly community. Participating repository networks are benefiting both strategically and pragmatically from the international collaboration.

A meeting of senior representatives from Africa, Asia, Europe, Latin America, and North America (April 2015/Porto/Portugal) showed that many networks have evolved significantly over the last year and are ready to collaborate with others more extensively. Activities shall include closer cooperation around the development of guidelines and tools, and several bilateral collaborations between different networks. In addition, supporting communication shall raise the visibility of repositories and their networks as key infrastructure components. It will underpin efforts at both the international and local levels to raise awareness about the value of repositories. Support was also expressed for further engaging with policy makers and other stakeholders, especially the Global Research Council (GRC) to ensure adoption of balanced open access policies (Shearer & Müller 2015).

2.1. Cooperation between Large Regional Networks in Europe, North- and South-America

There is already significant repository infrastructure in many countries and regions of the world, which are connected through national and thematic networks. Worldwide, three continents stand out in implementing strong regional ties in setting up e-infrastructure networks based on repositories. Certainly, these networks have evolved based on the unique local requirements and mandates, and are at different stages of

[2] The Confederation of Open Access Repositories (COAR) is an international association with over 100 members and partners from around the world representing libraries, universities, research institutions, government funders and others. COAR brings together the repository community and major repository networks in order build capacity, align policies and practices, and act as a global voice for the repository community.

[3] See https://www.coar-repositories.org/activities/advocacy-leadership/aligning-repository-networks-across-regions/

development. It should be noted that there are also repository systems in other regional and national contexts including Australia, China, Canada, and Japan among others (Shearer 2015).

In *Europe*, OpenAIRE aggregates the research output of EC-funded projects and makes them available through a centralized portal. All EU member states are participating in this project, as well as five associate countries, making a total of 33 countries. OpenAIRE currently aggregates the metadata from over 6,000 repositories and OA journals across Europe.[4] In *Latin-America*, LA Referencia maintains a centralized harvester, promotes common standards across Latin America and works at the strategic level to further promote open access. These services reflect the public policy agreements of the science and technology authorities in nine countries. LA Referencia currently harvests metadata from eight national nodes aggregating from about 200 repositories.[5] In 2013, the SHARE initiative was launched to bring together information about the publication output in *Northern America*. SHARE, a joint effort by three large stakeholders, creates an openly available data set about research activities across their life cycle by collecting, connecting, and enhancing metadata that describes research activities and outputs.[6]

Several pilot collaborations are being set up in order to show how repository alignment can help moving towards a global knowledge network. Concretely, the networks of LA Referencia, OpenAIRE, and SHARE are engaged consensually to discuss synergies and potential areas of collaboration. In particular, a number of specific areas were identified between regional networks and COAR in which all commit to collaborate on:

- Regular data exchange: Exchange data and develop agreements around jurisdictional harvesting and aggregation leading to greater coverage and efficiencies across regions.
- Common metadata and vocabularies: Work towards consensus about key metadata elements and common vocabularies to express funders and institutional affiliations, open access status, and project IDs. This will contribute to the COAR-CASRAI work already underway aimed at developing common metadata elements and support repository managers in better exposing their collections.[7]
- Common technological services: Assess the feasibility of adopting common broker/router technologies and other services.
- Ongoing dialogue: Meet regularly to share approaches and perspectives about technical and strategic challenges.[8]

2.2. Interoperability between OpenAIRE and LA Referencia

A regional workshop focusing on interoperability between LA Referencia and OpenAIRE (Nov 2015/Rio de Janeiro/Brazil)[9] aimed to set up closer collaboration

[4] http://www.openaire.eu

[5] http://lareferencia.redclara.net/rfr/

[6] http://www.share-research.org/

[7] See https://www.coar-repositories.org/activities/advocacy-leadership/working-group-global-interoperability-of-oa-repository-networks/

[8] See https://www.coar-repositories.org/activities/advocacy-leadership/aligning-repository-networks-across-regions/collaboration-on-data-exchange-technological-development-and-metadata/

between these two networks in order to enhance the usability and visibility of the collective content in the networks and enable the development of value added services across the two regions.

Specific agreements were made:

- The national nodes of LA Referencia will adopt the OpenAIRE Guidelines and participate in the development of the upcoming guidelines versions.
- LA Referencia will develop a strategy and launch communities of practice to facilitate sharing of expertise across participating countries and to support implementation of guidelines at local institutions.
- LA Referencia, OpenAIRE and COAR will partner to develop a blended learning course to build capacity in managing repositories in Latin America.
- LA Referencia and OpenAIRE will provide validators that will enable repositories to assess their level of compliance with the guidelines.

2.3. Engaging with Other Regions

Amidst the different levels of network ties and acknowledging the diversity of approaches and capacities across different regions, COAR is seeking possibilities to collaborate worldwide and to involve regions with less network structure. In this spirit, COAR promotes international alignment with other regions and countries in Africa, Australia, Canada, China, Japan and India based on dialogue and presentations with various organizations and players in those regions. The most recent development is the launch of a regional Open Access initiative in Asia with COAR in March 2016.

3. Conclusion

There is already significant repository infrastructure in many countries and regions of the world, which are connected through national and thematic networks. These regional and national networks are in the process of further aligning their practices globally through the COAR Aligning Repository Networks initiative, making their collections more valuable by enabling new services to be built on top of their aggregated contents. Through a number of strategic and pragmatic activities, COAR is working on different aspects of alignment between repositories and repository networks. A special focus is given to three regions worldwide represented by SHARE, LA Referencia, and OpenAIRE. The alignment work is being undertaken via several mechanisms including bi- and multilateral meetings and agreements between the major networks, international working groups, and informal conversations. COAR aims to expand these activities to include a greater number of organizations and regions. Finding the right balance between local and regional needs and a truly global agreement for repositories will be the key element in this process.

[9] See http://lareferencia.redclara.net/rfr/sites/default/files/docs_publicos/informeworkshop25-26noviembre2015.pdf

References

Shearer, K. (2014) Towards a Seamless Global Research Infrastructure, Report of the Aligning Repository Networks Meeting, March 2014. Available at: https://www.coar- repositories.org/files/Aligning-Repository-Networks-Meeting-Report.pdf (Accessed: 1 March 2016).

Shearer, K. (2015) Promoting Open Knowledge and Open Science, Report on the current state of Repositories, May 2015. Available at: prod/assets/documents/international/COARStateOfRepositories.pdf (Accessed: 1 March 2016).

Shearer, K., Müller, K. (2015) Aligning Repository Networks, Report of Strategic Meeting, April 2015. Available at: https://www.coar-repositories.org/files/Aligning-Repository-Networks-2015-Milestone-Meeting-Report.pdf (Accessed: 1 March 2016).

F. Loizides and B. Schmidt (Eds.)

Subject Index

This page intentionally left blank

Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

Author Index

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank