Towards Open Access Scholarship

Selected Papers from the Berlin 6 Conference

Edited by Cornelius Puschmann

and Dieter A. Stein

proceedings@berlin6.org

d|u|p düsseldorf university press

2010

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Towards Open Access Scholarship:
Selected Papers from the Berlin 6 Conference
Edited by Cornelius Puschmann and Dieter A. Stein

düsseldorf university press ISBN: 978-3-940671-43-1
1. Auflage 2010

Printed in Germany

Textgestaltung: Kim M. Barthel (Düsseldorf)
Umschlaggestaltung: Helge B. Unterweg (Erkrath)
Druck: docupoint GmbH Magdeburg

Contents

Timely or Timeless?
The Scholar's Dilemma. Thoughts on Open Access and the Social Contract of Publishing
DIETER A. STEIN AND CORNELIUS PUSCHMANN 5

Progress Towards Open Access at the EC
Speech Held at the Berlin 6 Open Access Conference Düsseldorf, 11 November 2008
HORST FORSTER 11

Access to Scientific Information in the Digital Age
European Commission Open Access Pilot
DEIRDRE FURLONG 21

Alternative Publishing Models
Exploring Costs and Benefits
JOHN W. HOUGHTON 27

SCOAP^3
A New Publishing Model for High-Energy Physics
ANNE GENTIL-BECCOT, SALVATORE MELE, JENS VIGEN 41

SciELO Program Publishes More than 500 Latin-American Open Access Journals
ABEL L. PACKER, ROGÉRIO MENEGHINI, SOLANGE M. SANTOS, FABIANA MONTANARI LAPIDO, ADRIANA LUCCISANO 51

Open Access Contributions in the Humanities?
The Approach of the Berlin-Brandenburg Academy of Sciences and Humanities

CHRISTIANE FRITZE 65

Peer Reviewed Data Publication in Earth System Sciences: Earth-System-Science-Data
A Data Publishing Journal

SÜNJE DALLMEIER-TIESSEN AND HANS PFEIFFENBERGER 77

Open Document Exchange Formats
Security, Protection and Experiences

CHRISTIAN ZIER 85

Timely or Timeless?

The Scholar's Dilemma. Thoughts on Open Access and the Social Contract of Publishing

Dieter A. Stein and Cornelius Puschmann

Heinrich-Heine-University Duesseldorf, Germany
stein@phil-fak.uni-duesseldorf.de
puschmann@phil-fak.uni-duesseldorf.de

Introduction

Some things don't change.

We live in a world seemingly over-saturated with information, yet getting it out there in both an appropriate form and a timely fashion is still challenging. Publishing, although the meaning of the word is undergoing significant change in the time of iPads and Kindles, is still a very complex business. In spite of a much faster, cheaper and simpler distribution process, producing scholarly information that is worth publishing is still hard work and so time-consuming that the pace of traditional academic communication sometimes seems painfully slow in comparison to the blogosphere, Wikipedia and the ever-growing buzz of social networking sites and microblogging services. How idiosyncratic does it seem in the age of cloud computing and the real-time web that this electronic volume is published one and a half years after the event its title points to? Timely is something else, you might say.

Dan Cohen, director of the Center for History and New Media at George Mason University, discusses the question of why academics are so obsessed with formal details and consequently so slow to communicate in a blog post titled “The Social Contract of Scholarly

Publishing". In it, Dan retells the experience of working on a book together with colleague Roy Rosenzweig[1]:

"*So, what now?" I said to Roy naively. "Couldn't we just publish what we have on the web with the click of a button? What value does the gap between this stack and the finished product have? Isn't it 95% done? What's the last five percent for?"*

We stared at the stack some more.

Roy finally broke the silence, explaining the magic of the last stage of scholarly production between the final draft and the published book: "What happens now is the creation of the social contract between the authors and the readers. We agree to spend considerable time ridding the manuscript of minor errors, and the press spends additional time on other corrections and layout, and readers respond to these signals — a lack of typos, nicely formatted footnotes, a bibliography, specialized fonts, and a high-quality physical presentation — by agreeing to give the book a serious read."

A social contract between author and reader. Nothing more, nothing less.

It may seem either sympathetic or quaint how Roy Rosenzweig elevates the product of scholarship from a mere piece of more or less monitizable content to something of cultural significance, but he also aptly describes what many academics, especially in the humanities, think of as the essence of their work: creating something *timeless*. That is, in short, why the humanities are still in love with books, why they retain a pace of publishing that is entirely snail-like, both to other academic fields and to the rest of the world. Of course humanities scholars know as well as anyone that nothing is truly timeless and understand that trends and movements shape scholarship just like they shape fashion and music. But there is still a commitment to spending time to deliver something to the reader that is a polished and perfected as one can manage. Something that is not rushed, but refined.

[1] http://www.dancohen.org/2010/03/05/the-social-contract-of-scholarly-publishing/.

Why? Because the reader expects authority from a scholarly work and authority is derived from getting it right to the best of one's ability.

This is not just a long-winded apology to the readers and contributors to this volume, although an apology for the considerable delay is surely in order, especially taking into account the considerable commitment and patience of our authors (thank you!). Our point is something equally important, something that connects to Roy Rosenzweig's interpretation of scholarly publishing as a social contract. This publication contains eight papers produced to expand some of the talks held at the Berlin 6 Open Access Conference that took place in November 2008 in Düsseldorf, Germany. While Open Access has successfully moved forward in the past eighteen months and much has been achieved, none of the needs, views and fundamental aspects addressed in this volume – policy frameworks to enable it (Forster, Furlong), economic and organizational structures to make it viable and sustainable (Houghton; Gentil-Beccot, Mele, and Vigen), concrete platforms in different regions (Packer et al) and disciplines (Fritze, Dallmeier-Tiessen and Pfeiffenberger) to serve as models, and finally technical standards to support it (Zier) – none of these things have lost any of their relevance.

Open Access is a timely issue and therefore the discussion about it must be timely as well, but “discussion” in a highly interactive sense is hardly ever what a published volume provides anyway – that is something the blogosphere is already better at. That doesn't mean that what scholars produce, be it in physics, computer science, law or history should be hallowed tomes that appear years after the controversies around the issues they cover have all but died down, to exist purely as historical documents. If that happens, scholarship itself has become a museal artifact that is obsolete, because a *total* lack of urgency will rightly suggest to people outside of universities that a field lacks relevance. If we don't care when it's published, how important can it be?

But can't our publications be both timely and timeless at once? In other words, can we preserve the values cited by Roy Rosenzweig, not out of some antiquated fetish for scholarly works as perfect docu-

ments, but simply because thoroughly discussed, well-edited and proofed papers and books (and, for that matter, blog posts) are nicer to read and easier to understand than hastily produced ones? Readers don't like it when their time is wasted; this is as true as ever in the age of information overload. Scientists are expected to get it right, to provide reliable insight and analysis. Better to be slow than to be wrong. In an attention economy, perfectionism pays a dividend of trust.

How does this relate to Open Access? If we look beyond the laws and policy initiatives and platforms for a moment, it seems exceedingly clear that access is ultimately a solvable issue and that we are fast approaching the point where it will be solved. This shift is unlikely to happen next month or next year, but if it hasn't taken place a decade from now our potential to do innovative research will be seriously impaired and virtually all stakeholders know this. There is growing political pressure and commercial publishers are increasingly experimenting with products that generate revenue without limiting access. Historically, universities, libraries and publishers came into existence to solve the problem of access to knowledge (intellectual and physical access). This problem is arguably in the process of disappearing, and therefore it is of pivotal importance that all those involved in spreading knowledge work together to develop innovative approaches to digital scholarship, instead of clinging to eroding business models. As hard as it is for us to imagine, society may just find that both intellectual and physical access to knowledge are possible without us and that we're a solution in search of a problem. The remaining barriers to access will gradually be washed away because of the pressure exerted not by lawmakers, librarians and (some) scholars who care about Open Access, but mainly by a general public that increasingly demands access to the research it finances. Openness is not just a technicality. It is a powerful meme that permeates all of contemporary society.

The ability for information to be openly available creates a pressure for it to be. Timeliness and timelessness are two sides of the same coin. In the competitive future of scholarly communication, those who get everything (mostly) right will succeed. Speedy and open publica-

tion of relevant, high quality content that is well adjusted to the medium and not just the reproduction of a paper artifact will trump those publications that do not meet all the requirements. The form and pace possible will be undercut by what is considered normal in individual academic disciplines and the conventions of one field will differ from those of another. Publishing less or at a slower pace is unlikely to be perceived as a fault in the long term, with all of us having long gone past the point of informational over-saturation. The ability to effectively make oneself heard (or read), paired with having something meaningful to say, will (hopefully) be of increasing importance, rather than just a high volume of output.

Much of the remaining resistance to Open Access is simply due to ignorance, and to murky premonitions of a new dark age caused by a loss of print culture. Ultimately, there will be a redefinition of the relativities between digital and print publication. There will be a place for both: the advent of mass literacy did not lead to the disappearance of the spoken word, so the advent of the digital age is unlikely to lead to the disappearance of print culture. Transitory compromises such as delayed Open Access publishing are paving the way to fully-digital scholarship. Different approaches will be developed, and those who adapt quickly to a new pace and new tools will benefit, while those who do not will ultimately fall behind.

The ideological dimension of Open Access – whether knowledge *should* be free – seems strangely out of step with these developments. It is not unreasonable to assume that in the future, if it's not accessible, it won't be considered relevant. The logic of informational scarcity has ceased to make sense and we are still catching up with this fundamental shift.

Openness alone will not be enough. The traditional virtues of a publication – the extra 5% – are likely to remain unchanged in their importance while there is such a thing as institutional scholarship. We thank the authors of this volume for investing the extra 5% for entering a social contract with their readers and another, considerable higher percentage for their immense patience with us. The result may not be entirely timely and, as has been outlined, nothing is ever truly time-

less, but we strongly believe that its relevance is undiminished by the time that has passed.

Open Access, whether 2008 or 2010, remains a challenge – not just to lawmakers, librarians and technologists, but to us, to scholars. Some may rise to the challenge while others remain defiant, but ignorance seems exceedingly difficult to maintain. Now is a bad time to bury one's head in the sand.

Düsseldorf, May 2010

* * * * * * * * * * *

This book could not have been printed without the tireless work of Kim M. Barthel, research assistant in Düsseldorf, Germany. Working from four different continents she gathered and prepared papers, tracked down authors and took care of correspondence to make this publication possible.

Progress towards Open Access at the EC

Speech held at the Berlin 6 Open Access Conference Düsseldorf, 11 November 2008

Horst Forster

European Commission
Javier.Hernandez-Ros@ec.europa.eu

1 Introduction

It is a great pleasure for me to meet you today in Düsseldorf, and a good opportunity to discuss the developments that are currently affecting the scientific information system (fostered notably by the Berlin Declaration on Open Access agreed in this particular event 5 years ago), and to convey to you steps taken recently by the Commission to foster access to scientific information.

2 EUROPEANA

Firstly, allow me to brief you very shortly on the forthcoming launch of the European digital library portal, EUROPEANA. I am happy to announce that EUROPEANA will be officially launched in Brussels on the 20th of November 2008, in the presence of President Barroso. EUROPEANA provides European citizens with a direct, multilingual access to our cultural heritage. At least two million digitised objects will be accessible at the launch: including film material, photos, paintings, audio, maps, manuscripts, books, newspapers and archival documents from Europe's museums, archives and libraries. This is just a beginning. We all expect that by 2010, EUROPEANA will encompass 10 million digital objects.

The creation of EUROPEANA is the result of a strong collaborative effort from the European cultural institutions organised through the

European Digital Library Foundation. EUROPEANA's content is a selection from what has been digitised and is already available in Europe's museums, libraries, archives and audio-visual collections. The target is simple and huge at the same time: to make Europe's cultural heritage accessible in the most user-friendly way. Key European cultural heritage organisations are members of The European Digital Library Foundation, which oversees the project. This body has been receiving financial support from the Commission through the eContentplus Programme.

The Commission's Digital Libraries initiative goes, as you know, well beyond EUROPEANA, which focuses on cultural heritage. Access to and preservation of scientific information is its second major role.

3 Importance of broad access to scientific information

More than ever, widespread and efficient access to scientific information is a must for research and innovation in a knowledge based society.

The benefits of better and quicker access to scientific information for the efficiency of research are considerable.

The Commission has a double role in this field as a policy making institution and as a funding organisation, both of research projects and of scientific information infrastructures. Commissioner Potocnik, responsible for research, and Commissioner Reding, responsible for the media industries, declared clearly that results of publicly funded research should be accessible to all easily.

4 The role of different stakeholders

Within the Commission's regular consultations and interaction with the relevant stakeholders in the dissemination of science, notably the research community and scientific publishers, we have witnessed that stakeholders' views, despite some joint action recently taken to reconcile positions, remain polarised.

On the one hand, scientific publishers defend their crucial contribution to the dissemination of scientific information, by organising and

ensuring adequate quality controls through the peer review system. They also claim that their scientific publishers' investments in recent years, for making their content available in digital format over the internet, have allowed for the provision and development of new services and that access has never been as good as it is today.

On the other hand, the research community, research funding organisations and librarians are supportive of alternative ways offered by new technologies to the traditional subscription journal model for disseminating research results. Open access advocates are adamant that it is beneficial to the research system, that there remains an important access problem today to scientific information and, highlight that the research community and citizens in general are entitled to open access on the Internet to the results of publicly funded research.

There is therefore a common understanding by all on the importance of widespread dissemination of research results, but disagreement remains on the way of achieving that common goal and on how to fund it.

5 Recent interesting developments OA Publishing

The Open Access system to scientific results, I would claim, is gradually gaining momentum.

Let us look on some recent facts and figures:

- arXiv[1], the physics and related fields repository hosted at Cornell University, reported last October that it had passed the milestone of half a million items (mainly pre-prints) being deposited in its repository. It claims not only that it is the primary daily information source for hundreds of thousands of researchers in its fields, but also journalists use the repository extensively to prepare scientific articles for the general public.
- PubMedCentral: it has been reported that the life sciences and life sciences repository operated by the National Institute on

[1] http://communications.library.cornell.edu/com/news/PressReleases/arXiv-milestone.cfm.

Health of the US, has 2.4 million items free in the repository for access[2].

- The number of OA journals is also on the rise. Recent figures published by the Directory of Open Access Journals[3] , indicate that the current number of OA journals is almost 3700, which represents a 15% increase since Sep 07.
- As regards the number of self archiving mandates (funded or unfunded) currently in place, the current figure (as of yesterday) is 57 which represents a non negligible 42% increase from last year. I shall come back later on to the impact of these mandates.

Established traditional publishers have been relying on a stable and comfortable business model based on subscriptions, but have at the same time been testing and experimenting in many different ways with open access publishing. In addition, and in response to the desires of the research community they now allow a certain degree of self archiving of pre/post prints in institutional repositories.

We are beginning to know more about the outcome of this experimentation. Open access advocacy has clearly had an effect on publishers' thinking. The proportion of publishers offering optional open access to authors has grown from 9% in 2005 to 30% in 2008. However, the take-up of the author pays open access option is very low. If we look at the Oxford University press, the latest figures from 2007 indicate open access papers constituted 6.8% of all papers published in Oxford Open journals[4]. Open Access publishing uptake amongst traditional publishers is well below this figure.

It seems, however, that the new emerging fully Open Access journals, are beginning to achieve financial sustainability and becoming quality journals in their respective fields.

---

[2] http://poeticeconomics.blogspot.com/.

[3] http://www.doai.org/ This Directory, managed by Lund University aims to be comprehensive and cover all open access scientific and scholarly journals worldwide.

[4] http://www.oxfordiournals.org/oxfordopen/charges.

For example, Hindawi Publishing Corporation, a rapidly growing academic publisher with already more than one hundred Open Access journals covering all major areas of science, or BioMed Central which was set up in 2000, today publishes some 193 open access journals with revenues of roughly € 20 million. Another important Open Access publishers based in the US is PLoS (Public Library of Science), which publishes 7 online peer-reviewed scientific and medical journals. Sources indicate that it expects to be in the black in the coming 2 years[5], and its journals are considered of very high quality in their respective fields. The impact factor figures provided by the Journal Citation Reports (JCR)[6], currently owned by Thomson Reuters[7] – which have a paramount influence on the scientific community, affecting decisions on where to publish, whom to recruit or the potential success of grant applications, signal that these journals are becoming increasingly important and influential. This concurs very nicely with reports that indicate an increasing number of paper submissions from researchers to these journals.

Traditional publishers have been experimenting with open access publishing with different degrees of commitment. In this respect I would like to highlight as forerider the BMJ (British Medical Journal) which decided that it was fully converting to OA after experimenting for some time with different business models.

Another very recent interesting development is Springer's purchase of BioMedCentral[8]. It seems that Springer, which publishes over 1700 journals, is confident that it can successfully develop BioMedCentral's OA business model8, which not long ago some considered experi-

[5] http://www.nature.com/news/2008/080702/full/454011a.html.

[6] The JCR is an annual publication by the Institute of Scientific Information, a division of Thomson Reuters, which provides citation information about academic journals in the sciences and social sciences. It was originally published as a part of Science Citation Index, and is compiled from the citation data found there.

[7] http://scientific.thomsonreuters.com/ Thomson Reuters is the world's leading source of intelligent information for businesses and professionals.

[8] http://www.stm-assoc.org/storage/Springer_BioMed%20Central_EN.pdf.

mental, and unsustainable in the long run from a financial perspective[9].

Whilst these facts and figures are encouraging, the overall figures of what is accessible on OA today are modest. We should ask ourselves why is there this limited uptake? There are reasons on both sides, on the publishers' side and on the research community side, not to speak of funding organisations. Rather than digging their heels into the soil and try not to move, all sides need to start moving.

6 Self Archiving Developments

As regards recent developments with the other avenue that enables open access, that of self archiving, there is certainly at present no shortage of mandates. The Registry of Open Access Repository Material Archiving Policies or ROARMAP[10] indicates that today 57 self archiving mandates are in place worldwide.

Among those, is the very much awaited and expected one at the National Institute of Health in the US, which earlier this month reported an increase in compliance rate from research authors from April to August 2008, that tripled the previous figures of preceding years, reaching 60% compliance rates.

In a similar fashion the UK Welcome's Trust has also reported that 27% of their papers were accessible. It seems that further efforts have to be undertaken to make the additional 40 and 63% of papers respectively accessible, as it is required in the grants agreement signed between research and the funding organisation. These are encouraging figures.

However, mandated deposit rates are not as rosy as it may seem. In many other institutions with mandates in place, the deposit/compliance rates are below the above mentioned figures. Again, we should ask ourselves why mandates do work in some institutions and not in others? I would claim that researchers and their funding

[9] http://www.nature.com/nature/focus/accessdebate/3.html.

[10] http://www.eprints.org/openaccess/policysignup/ This Registry which is managed by the University of Southampton tracks down and follows issues regarding self archiving mandates.

institutions are not pursuing this goal vigorously enough. The implications of both researcher and funding institutions are paramount for sharing the output of their work, to provide greater transparency to their institution, and consequently for the development of Open Access. They need not look outside of their institutions for reasons of lack of accessible information. They simply need to take care of this task themselves. I would like to highlight again that at present almost all publishers allow for self archiving.

7 Institutional repositories

Whilst the number of repositories increases every day, I believe that there is a need to take stock of what is available at present, and build gradually on solid ground, rather than unsystematically increase their numbers, and their costs.

1.

Let me start with metrics and impact. Whilst there are some ongoing limited initiatives (for example the research assessment exercise in UK universities) that are analysing the content of repositories in terms of impact and analysis[11], efforts should be undertaken to systematically evaluate institutional repositories; to assess their impact, to analyse the downloads and usage of their content. It is only by measuring its impact, knowing who is using its services, that one can estimate whether it is providing the right service or not, and consider whether corrective measures need to be taken.

2.

Another issue that I am concerned about is the quality and long term sustainability of repositories. If we have a look at the Registry of Open Access Repository[12], I am impressed by the first 10% of the list, but if we look further down the list in more detail, the number of

[11] http://www.rae.ac.uk.

[12] The ROAR service provides a quality-assured listing of open access repositories around the world, which is being developed and maintained by the University of Southampton.

records/items held by the repositories of well known institutions begins to diminish considerably at a preoccupying pace. The number of repositories, for example, in Germany which has less than 100 items is as large as 33. These poor figures lead one to reconsider whether many of these repositories have reached a minimum critical mass of content and whether they will be sustainable in the future. Some rationalisation of these infrastructures is be considered.

3.

In a similar way the issue of the quality of the repositories comes into question. Whilst efforts are being undertaken to set minimum quality standards by the Deutsche Initiative für Netzwerk Information (DINI)[13] for example, which institutions or bodies are thoroughly responsible for defining a set of minimum standards for repositories and its operators in a systematic way? Who takes responsibility for the monitoring of security, authenticity and data integrity, indexing, metadata, long-term availability of content, etc? Who is taking care of all these issues not just personally, but is there a proper infrastructure in place to oversee these issues which are relevant for making the repository appealing for regular users.

4.

The above mentioned concerns also apply to the issue of preservation. We need to actively preserve the digital material our institutions produce, to prevent it from being lost for future generations and to keep it accessible and usable over time. This important task has been undertaken efficiently by traditional publishers and libraries. As repositories begin to undertake this task, they should be aware of the many challenges they will need to surmount[14].

---

[13] German Initiative for Networked Information) is a Coalition of German Higher Education Infrastructure- or Service-Institutions: Libraries, Computing Centres, Media Centres and the Scientific Community, creating recommendations for standardized and interoperable information services and communication networks in and between universities.

[14] http://ec.europa.eu/information_society/activities/digital_libraries/doc/scientific_information/council_conclusions_nov2007.pdf.

I believe that these issues need to be addressed and properly corrected, so that repositories can become a stable, systematic, efficient, solid and reliable vehicle for facilitating access and making available scientific information.

8 Commission's actions

Let me finally briefly report on what the Commission is doing, operationally, in relation to scientific information.

The Commission's objective in the scientific information field is to maximise the benefits of information technologies for better access to and easier use of scientific knowledge. This requires a fair remuneration for those who invest to make the system work. The Commission put the topic on the political agenda in 2007 and ministers discussed the matter at several occasions.

The Commission as funding organisation conducts a pilot initiative on open access to peer reviewed research articles in its Seventh Research Framework Programme (FP7). In this pilot, open access to articles resulting from research funded in areas participating in the pilot should be achieved within a specified time period. This initiative covers 37% of the research budget. The exercise has specific embargo periods (6-12 months) for different areas of funded research, so that scientific publishers have the possibility to recoup their investments. I ask everybody to support this large scale experiment and actively participate in it. The pilot will run until 2013 and will be monitored and evaluated at regular intervals and the results will serve as input to determine dissemination policy for our future research programmes.

We will continue interacting with all relevant stakeholders, notably through the High Level Group on Digital Libraries. This High Level Group, which includes representatives of European research organisations, scientific publishers and European universities and is chaired by Commissioner Reding, aims at sharing experiences and good practices and works towards shaping a common vision for the future. A visible and concrete result of the High Level Group is the PEER (Publishing and the Ecology of the European Research) project. This joint

initiative of leading research organisations (Max Planck, European Science Foundation) and scientific publishers will jointly and transparently monitor the effects of systematic archiving of published articles in open repositories over time. This notable joint effort between key stakeholders will help collect much needed evidence on the impact of OA and help shape common, future strategies for the dissemination of scientific information. This project is funded by our eContentplus programme.

Our funding programmes, notably the 7th Framework Programme for research and development and the Competitiveness and Innovation programme and Science and Society Programme will continue to support new actions and experimentation in the field of dissemination of scientific information, for example through testing new business models in particular research fields, developing the establishment and consolidation of e-infrastructures that enable the development of e-science and data mining services, etc.

9 Conclusions

Let me conclude as follows: change is taking place in the way scientific information is accessible. This is an irreversible process.

Open Access to scientific information is a promising route, but it is still in its early stages. It needs wide support and uptake by the research community as a whole, and concrete action needs to be taken and addressed to make things happen. Let's do it. Let's not ask for others to do our job.

Strong cooperation between all stakeholders is a must. We all acknowledge that quality labelling is a must in this world of “over-information”. If publishers do not organize the system, who will? Let's benefit from their competence and know how.

We are in a transitional phase on many different fronts, and we need to allow for experimentation. The Commission will do its bit to facilitate this process and we aim to improve access to scientific information across Europe.

Thank you for your attention.

Access to Scientific Information in the Digital Age

European Commission Open Access Pilot

Deirdre Furlong

European Commission, Research Directorate-General, Science, Economy and Society Directorate Governance and Ethics Unit
RTD-open-access@ec.europa.eu

Abstract

In order to become an increasingly competitive knowledge-based economy, Europe must not only improve the production of knowledge but also its dissemination and application. For this reason the European Commission has been analysing over the past years the new opportunities for the dissemination of research results offered in the digital age. After extensive consultations with the numerous stakeholders, the European Commission launched in August 2008 an open access pilot project for its research funding programme to experiment with open access as a means to enhance the dissemination of research findings and maximise the returns on investment in R&D.

1 Improving access and dissemination of results from publicly-funded research

The European Commission (EC) outlined its interest and motivation for improving access to and dissemination of the research results produced from EU funding. As a public funding body the Commission has a duty to maximise its return on investment and ensure that the socio-economic benefits of scientific research are optimised for the good of all. The advent of the internet has provided unprecedented opportunities for the dissemination of scientific information and these

opportunities should be fully exploited. As research builds on former research, it is crucial that efficient, timely and reliable access to this research is available in order to advance scientific discovery and promote innovation. Moreover, enshrined in the legal foundations of the European Community Treaty is the responsibility of the European Community for the “dissemination and optimisation of the results and activities in Community research”[1]. The Lisbon agenda which aims to make Europe the most competitive and dynamic knowledge-based economy in the world further strengthens this responsibility. The future Lisbon treaty given the EU “the objective of strengthening its scientific and technological bases by achieving a European research area in which researchers, scientific knowledge and technology circulate freely….”[2] The EU's i2010 programme[3] promotes the positive contribution that information and communication technologies (ICT) can make to the economy, society and personal quality of life. Through its digital libraries initiative[4] it encourages the wide dissemination and preservation of scientific information.

2 Roles of European Commission

The role of the EC on the topic of access to scientific information is threefold, firstly it acts as a policy making body by launching a policy debate at European level and encouraging and inviting coordinated actions at member state level. Secondly it acts as a public research funding by setting access and dissemination rules for EC-funded research through the Framework Programmes, currently the seventh Framework Programme (FP7). Thirdly it acts as a capacity-building and supporting body providing for digital infrastructures and relevant research and networking activities.

---

[1] http://eur-lex.europa.eu/en/treaties/dat/12002E/htm/C_2002325EN.003301.html Title XVIII.

[2] Art 179, Lisbon Treaty, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2008:115:0047:0199:EN:PDF.

[3] http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm.

[4] http://ec.europa.eu/information_society/activities/digital_libraries/index_en.htm.

3 Policy-making body

In its policy making capacity, the publication of the "Study on the economic and technical evolution of the scientific publication markets in Europe"[5] in 2006 was the first step by the EC to look at the scientific publications market from a research perspective. The objective of this study was to access the evolution of the market for scientific publishing and to discuss European level recommendations to improve the system governing access to and dissemination of scientific publications. This study proposed a series of policy recommendations the principle of which was to provide open access to the results of publicly funded research. After its publication, a public consultation on this study took place providing all stakeholders with the opportunity to comment on its results. Responses were received from individual researchers, academic organisations, libraries and information organisations, and scientific publishers. This study marked the first contribution to the policy debate on the functioning and efficiency of the scientific publication system, launching a rigorous policy process at EU level and to date the EC has been actively engaging with stakeholders on these issues. The next important milestone in this policy process, was a communication on "Scientific Information in the digital age: access, preservation and dissemination"[6] published by the EC in February 2007. This policy document examined how new digital technologies can be better used to increase access to research publications and data as an important driver for innovation in our increasingly knowledge-based economy. It provided an overview of the current state of play in Europe regarding scientific publishing and the preservation of research results, examining organisation, legal, technical and financial issues. It also stated the intention of the EC to improve and enhance access to scientific information and proposed experimenting with open access in FP7. The communication was presented at a high level conference in Brussels, on 15 and 16 of February, entitled "Sci-

[5] http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf.

[6] http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf.

entific Publishing in the European Research Area: Access, Dissemination and Preservation in the Digital Age"[7]. The goal of this conference was to bring together stakeholders concerned with access, dissemination and preservation issues in connection with scientific publication and data in an effort to provide policy options for scientific publishing under FP7 and in the European Research Area. Both the Commissioner for Science and Research, Mr. Potočnik and the Commissioner for Information Society and Media, Ms. Reding spoke at the event. Topics discussed by stakeholders included publishing business model, e-infrastructure and preservation, copyright and digital rights management and quality assurance and excellence.

The member states provided their support to the efforts of the EC in this area, by producing Council Conclusions[8] on the Communication, from the Competitiveness (Internal market, Industry and Research) Council meeting under the leadership of the Portuguese presidency in November 2007. These Council Conclusions invited the member states to reinforce their national strategies and structures for access to and dissemination of scientific information, to enhance their co-ordination among each other on access and dissemination policies and practices and to ensure the long term preservation of scientific information. They invited the EC to experiment with open access to scientific publications resulting from projects funded by the EU Research Framework Programmes, to support experiments and infrastructures for access to and preservation of scientific information and to contribute to improved policy co-ordination between Member States and to a constructive debate between stakeholders.

4 Research funding body

It its capacity as a research funding body the EC manages and implements the EU Research Framework programme, currently FP7, which began in 2007 and will run until the end of 2013 with a budget of

[7] http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public. topic&id=1720&CFID=3260087&CFTOKEN=a379240cfdfedf01-650921FD-E0CF-FB2A-F704FC49B59C1B4C&jsessionid=b101d7703173379267e02c1853571b66574bTR.

[8] http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/intm/97236.pdf.

over € 50 billion. Since the beginning of FP7, the 'gold' (paid) route to open access is supported by the EC, under this 'gold' route, researchers can claim 100% reimbursement for publication costs, including open access publishing. The legal reference for this possibility can be found in the FP7 Model Grant Agreement Article II.16 on the upper funding limits[9].

As a further experiment with open access, the EC launched it Open Access Pilot in FP7[10] in August 2008. This pilot supports the 'green' (self archiving) route to open access and covers approximately 20% of the FP7 budget covering the areas of Energy, Environment, Health, Information and Communication Technologies (cognitive systems, interaction, robotics), e-Infrastructures, Science in Society and Socio-economic Sciences and Humanities. Under this pilot, grant agreements in these areas signed after the launch of the open access pilot will contain a special clause requiring beneficiaries:

- to deposit articles resulting from FP7 projects into an institutional or subject based repository;
- to make their best efforts to ensure open access to these articles within six months (Energy, Environment, Health, Information and Communication Technologies, Research Infrastructures) or twelve months (Science in Society, Socio-economic Sciences and Humanities).

This pilot will run until the end of FP7, the results of which, along with the results of other initiatives funded by the Commission in this area, will feed into a broader access and dissemination policy for FP8.

5 Capacity-building / supporting body

In its role as a capacity building and supporting body, the EC is funding other initiatives in this area including the project e-SciDR[11] which is

[9] ftp://ftp.cordis.europa.eu/pub/fp7/docs/fp7-ga-annex2-v2_en.pdf pg 18, point 4.

[10] http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=1680.

[11] http://www.e-scidr.eu/.

a study of European repository infrastructure, OAPEN[12] which looks at open access publishing in European networks focusing on social science and humanities and PEER[13] which is a pilot programme investigating the effect of the deposit of author manuscripts on publishing and DRIVER[14] which aims to set up a federation of digital repositories for publications and data.

6 Conclusions

It can be seen that there has been steady progress by the EC in this area since 2006. European policy makers have recognised the importance of access and dissemination for research and technology policies and these issues have been firmly placed on the European policy agenda. The next steps for the EC include mobilising the member states for coordinated activities in an effort to promote a European policy in this area, monitoring ongoing initiatives in order to move towards an access and dissemination policy for the 8th Framework Programme (from 2014) and to continue supporting initiatives and activities which will strengthen and promote wide dissemination of publicly funded research results and the development of sustainable e-infrastructure in all Member States.

[12] http://www.oapen.org/.
[13] http://www.peerproject.eu/.
[14] http://www.driver-repository.eu/.

Alternative Publishing Models

Exploring Costs and Benefits

John W. Houghton

Centre for Strategic Economic Studies, Victoria University, Australia
John.Houghton@pobox.com

Abstract

This paper reports on the approach to a UK JISC funded study into the economic implications of alternative publishing models. It outlines the underlying issues and take-off point for the project, the approach to identifying scholarly communication life cycle activities and costs, and to activity-based costing. Key issues arising during the research and some preliminary findings are also discussed.

1 Introduction

In late 2006, Colin Steele, Peter Sheehan and I produced a report for the Australian Department of Education, Science and Training (DEST) on *Research Communications Costs, Emerging Opportunities and Benefits* (http://dspace.anu.edu.au/handle/1885/44485). The aim of the project was to identify all the costs associated with scholarly communication and to explore the potential benefits of more Open Access to research findings (focussing primarily on publications).

The Australian study generated quite a bit of interest, particularly in the method that we used to quantify the impacts of Open Access (OA). As a result of that interest, I am currently nearing the completion of a similar project for the Joint Information Services Committee (JISC) in the UK, in which we are estimating the economic implications of alternative scholarly publishing models. I am also working, through the European Knowledge Exchange Group, on similar pro-

jects elsewhere in Europe. These include a SURF funded project in the Netherlands with EIM, preliminary work on a smaller project in Denmark with the Danish Electronic Research Library, and a proposal for a project in Germany with Frankfurt University, in which we are bringing the German *National Licensing Program* into the mix of alternative models.

The take off point for the original Australian study was a sense of frustration with the discussion of the economics of scholarly publishing, which focuses almost entirely on costs. From an economic perspective, the aim is to have the most cost-effective system, not (*necessarily*) the cheapest: and however much you study costs, you will never know which is the most cost-effective system until you examine both the costs *and* the benefits. So, we looked at research communication costs *and* benefits.

2 JISC project in the UK

This year I have been working on a project for the Joint Information Services Committee (JISC) in the UK, in collaboration with Loughborough University's Departments of Information Science and Economics, and the statistical unit (LISU).[1] We have also been working closely with Bo-Christer Björk of the Hanken School in Helsinki, and Donald King of the University of North Carolina, who is part of the project advisory group.

The aim of the project is to look at costs *and* benefits, and in so doing explore the institutional, budgetary and wider economic implications of the major alternative models of scholarly publishing. The models we are focusing on are:

1. Subscription or toll access publishing;
2. Open Access publishing; and
3. Self-archiving.

[1] See http://www.jisc.ac.uk/whatwedo/programmes/reppres/economicsscholarlypublishing.aspx.

Being a JISC funded project, the emphasis is on the economic implications for UK universities, but our perspective is wide ranging.

Phase I of the project is about the identification of costs and benefits. In Phase I, we describe the three models of scholarly publishing in such a way as to characterise key differences between them, and then identify all the costs and potential benefits of each of the publishing models. Phase II is about the quantification of costs and benefits. It seeks to quantify the costs and benefits identified, and compare the costs and benefits for the three models.

3 JISC Project: Phase I

We started the project with a wide ranging literature review and found two distinct approaches. The majority of writers focus on publishing and discuss the functions and costs involved. Others explore the wider context, seeing publishing as part of a wider system of knowledge creation and dissemination. To my mind, however detailed, studies that focus on publishing activities alone will tend to overlook areas in which costs are shifted around the system, confuse cost shifting with cost reduction, and not take account of the full system costs. So, we adopted a systems perspective, and our cost model includes activities related to publishing, and those related to funding research, performing research, and the dissemination of findings.

3.1 The scholarly communication process

We adopted a formal process modelling approach to provide a solid foundation for the identification of activity costs, and we use the IDEF0 standard process model which is typically used for business process re-engineering.

We have development of the life-cycle model that Bo-Christer Bjork (2007) outlined, expanding it to include five main activity elements:

1. Fund research and research communication;
2. Perform research and communicate the results;
3. Publish scientific and scholarly works;

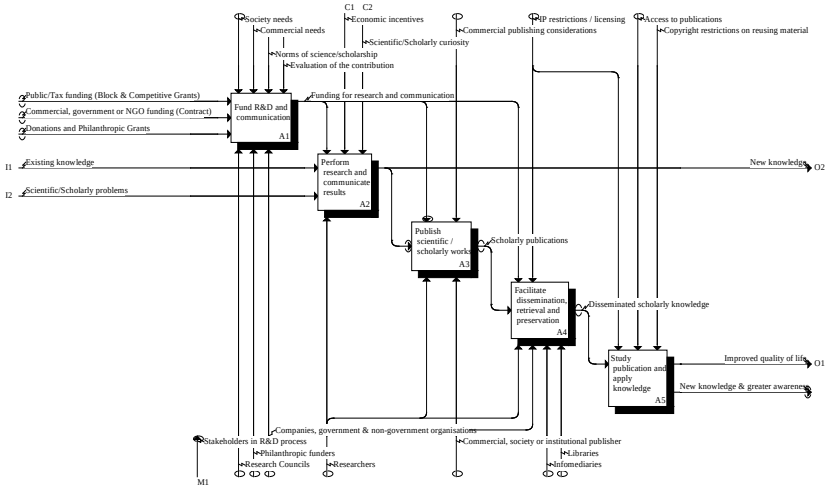

Fig. 1. JISC Project: Scholarly communication process model, http://www.cfses.com/EI-ASPM/SCLCM-V7/. Source: Houghton, J. W. *et al*. (forthcoming) *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits*, Report to The Joint Information Systems Committee, London and Bristol.

4. Facilitate dissemination, retrieval and preservation; and
5. Study publications and apply the knowledge (Fig. 1.).

In its current form, the model includes more than 50 diagrams and around 200 activities.[2]

One interesting feature of the AIOWin software we are using is that it has an integrated activity cost module. So if we had the data, it would be possible to populate the entire process model and explore activity costs and value-adding throughout the system.

3.2 A matrix approach to costing

We have adopted a matrix approach to activity costing, based on the process modelling, so we can use the activity costings to explore costs for actors, objects and functions. The aim is to be able to break

[2] Details of the entire model in 'browseable' form can be found on the Web at: http://www.cfses.com/EI-ASPM/SCLCM-V7/.

down, and re-assemble, the value chain along any of these dimensions.

So, for example, a costing of the time spent by university staff doing peer review can contribute to our understanding of the costs:

- To various actors (*e.g.* universities);
- Of various activities (*e.g.* peer review);
- Of various objects (*e.g.* journal articles); and
- Of various functions (*e.g.* certification and quality control).

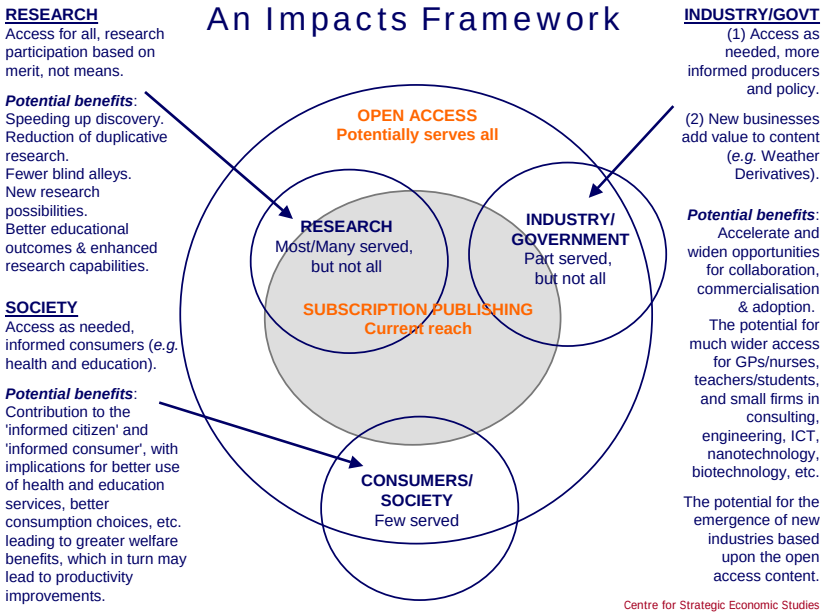

Fig. 2. JISC Project: An impacts framework. Source: Houghton, J.W. *et al*. (forthcoming) *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits*, Report to The Joint Information Systems Committee, London and Bristol.

3.3 Identifying the impacts

In some ways the costs are relatively easy. The next step is to identify the dimensions of impact, and we have further developed the impacts

framework that we used in the Australian study – in which we examined the implications of more Open Access for research users, industry and government users, and the general public (Figure 2). It is a useful starting point, but it did not fully cover the production-side impacts. So we went back to the drawing board.

The key issues in Open Access are access and permission, where:

- *Access* includes accessibility, in the sense of ease and affordability of access (in terms of time and cost); and
- *Permission* refers to permission to use the material, in terms of what is permitted and the time and cost involved in obtaining permission.

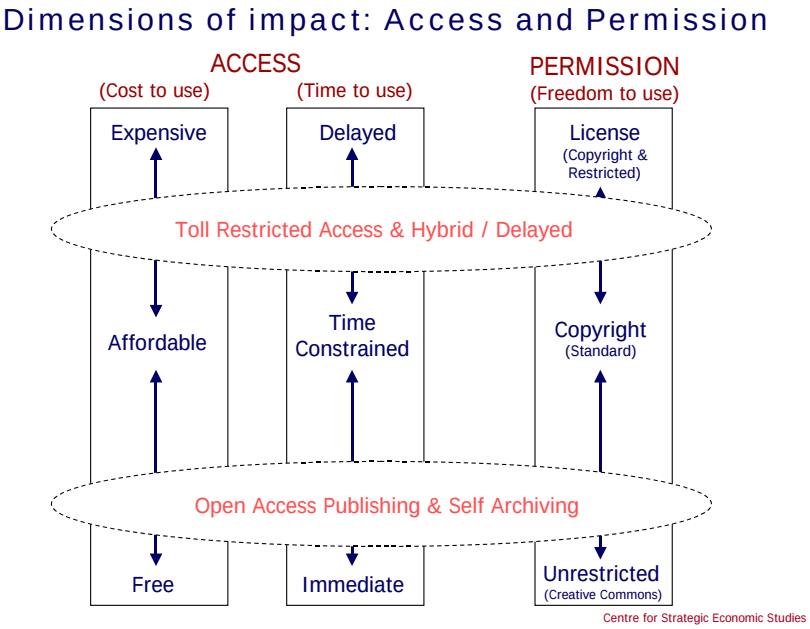

Fig. 3. JISC Project: Access and Permission. Source: Houghton, J.W. *et al*. (forthcoming) *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits*, Report to The Joint Information Systems Committee, London and Bristol.

That suggested an analysis along the overlapping dimensions of access and permission, mediated by cost in terms of both money and time (Figure 3). In essence, we are exploring the impacts of the three publishing models on the time and cost involved in accessing and using scholarly works, however and whenever required, and for whatever purpose – setting the three models against the ultimate goal of OA for free, immediate and unrestricted access.

4 JISC Project: Phase II

We adopted a staged approach to Phase II that tackles it from the bottom-up (as case studies and scenarios) and the top-down (in a simple economic model).

1. First, we explore the costs of the process elements and system costs, and from that we can see cost differences and direct savings.
2. Then we present cases and scenarios, to explore the cost savings resulting from alternative publishing models – looking, for example, at cost impacts on search and discovery, library handling costs, etc. From that we can explore the indirect cost differences (*i. e.* seeing cost savings as first order benefits).
3. Then we approach it from the opposite end and model the impacts of changes in accessibility and efficiency on returns to R&D.

In the first of these steps, we produced detailed costings of activities relating to: funding research, performing research, publishing and dissemination. The focus is on cost differences between the publishing models.

The key problem is to separate the cost impacts of publishing models from those of format, so that we can identify the cost differences between toll and Open Access independent of differences between print and electronic formats. Our approach is to estimated costs for print, dual mode and electronic only (e-only), for toll and Open

Access, and then to compared toll and Open Access models as if they were all e-only.

Not surprisingly, the cost difference between print and electronic formats is often greater than the difference between toll and Open Access. Figure 4 is an example: it shows estimated UK higher education library handling costs, and the difference between toll access (TA) print and e-only is much greater than the difference between toll and Open Access (OA) in e-only form.

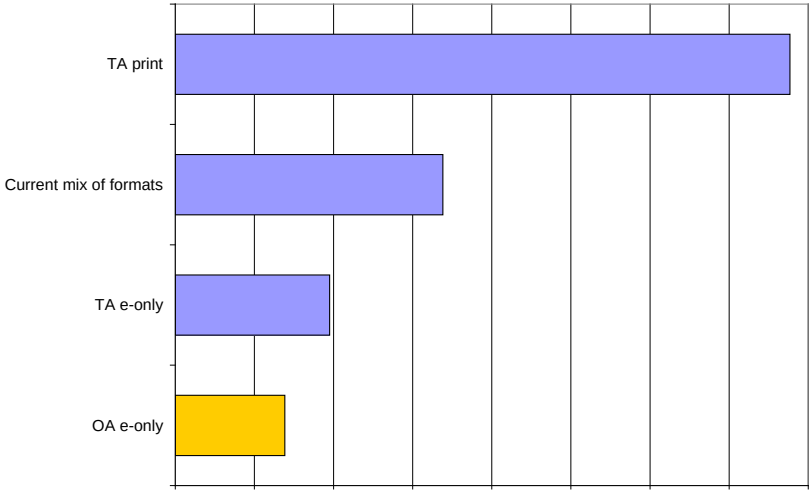

Fig. 4. JISC Project: SCONUL library handling costs by mode and model (UK Higher Education. Source: Houghton, J. W. *et al*. (forthcoming) *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits*, Report to The Joint Information Systems Committee, London and Bristol.

Then we summed the costs of the three publishing models through the main phases of the scholarly communication life-cycle, to highlight system cost differences. For example, estimated system costs per journal article in UK higher education, including research production, publisher, and library and dissemination costs, are shown in Figure 5. From this we estimated cost savings from alternative publishing mod-

els for UK research nationally and UK higher education, for articles and books, and for articles alone.

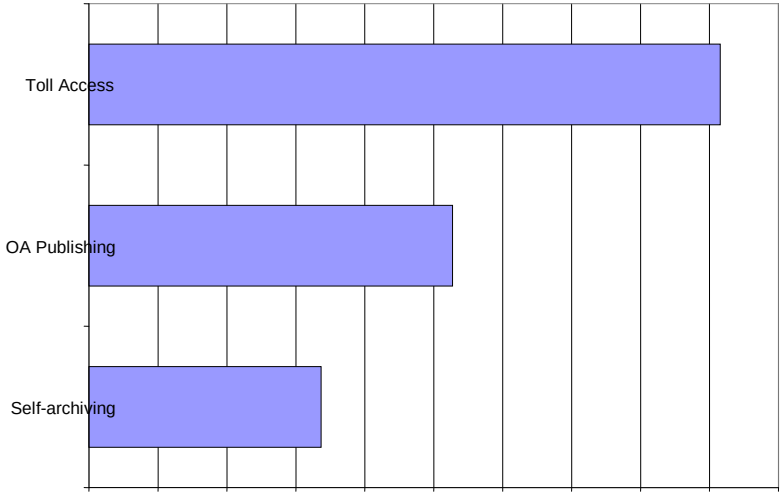

Fig. 5. JISC Project: Estimated system costs per article. Source: Houghton, J.W. *et al*. (forthcoming) *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits*, Report to The Joint Information Systems Committee, London and Bristol.

Then, in the third step, to estimate the impacts of changes in accessibility and efficiency on returns to R&D we modified a basic Solow-Swan model. There is a vast literature in Economics that focuses on estimating the rate of return to R&D. While highly varied, a characteristic finding is that the rate of return is high and typically in the region of 20-60% a year (Arundel and Geuna 2004).

I will not go into the details of the model, but the standard neoclassical approach makes some key simplifying assumptions. It assumes that:

- All R&D generates knowledge that is useful in economic or social terms (*efficiency of R&D*);
- All knowledge is equally accessible to all entities that could make productive use of it (*accessibility of knowledge*); and

- All types of knowledge are equally substitutable across uses (*substitutability of knowledge*).

A good deal of work has been done to address the fact that the substitutability assumption is not realistic. Clearly research is often application specific. Much less has been done on the other two, equally unrealistic assumptions, which is where we focus.

Basically, what we do is introduce *accessibility* and *efficiency* as negative or friction variables, to reflect the fact that there are limits and barriers to access and to the efficiency of production and usefulness of knowledge, and then we look at the impact on returns to R&D of reducing the friction by increasing access and efficiency.

HERD 6,062	Rate of return to R&D				
	20%	30%	40%	50%	60%
Per cent change in accessibility and efficiency	Recurring annual gain from increased accessibility & efficiency (million)				
1%	24	37	49	61	73
2%	49	73	98	122	147
5%	*124*	186	249	311	373
10%	255	382	509	637	764

Fig. 6. JISC Project: Impact estimation ranges (UK Higher Education Expenditure on R&D 2006). Source: Houghton, J.W. *et al*. (forthcoming) *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits*, Report to The Joint Information Systems Committee, London and Bristol.

We produced range estimates, looking at rates of return from 20% to 60% and increases in access and efficiency of 1% to 10%. So, we produce a table for each category of R&D expenditure (Figure 6). The ranges are quite large, but for the purposes of discussion, based on a review of the literature, we assume a 20% social return on public sector R&D and suggest that a 5% increase in accessibility and efficiency might be plausible.

We have devoted many weeks of work and many pages in the report to why 20% and 5% make sense, and despite limitations in models of this type I think we have now got to a point where our model parameters well grounded and, if anything, we are erring on the conservative side.

We have based the percentage changes in accessibility and efficiency on metrics relating to:

- The share of publications in general, and journals in particular, in the research Stock of Knowledge;
- The share of the research Stock of Knowledge potentially available to Open Access;
- A number of proxy measures of accessibility, including: UK research library subscriptions, conservative estimates of the Open Access citation advantage, and estimates derived from repository downloads; and
- A number of estimates of the potential efficiency implications of access limitations, such as poorly informed and duplicative research, and of relaxing those limitations, such as speeding up the research and discovery process, and facilitating greater collaboration.

In terms of returns to R&D, we present a summary of the literature from neo-classical, new growth and evolutionary perspectives, and we adopt a very conservative consensus value. And in view of the, so called, attribution problem, we stay at the ends and do not stray into the middle. So it is based on cases and scenarios, on the one hand, and aggregates, such as Higher Education Expenditure on R&D (HERD) and Government Expenditure on R&D (GovERD), on the other.

These are preliminary estimates, and rates of return vary considerably, so the further one moves from the aggregate the larger the range of uncertainty. Nevertheless, just to give an idea: for the major categories of research expenditure in the UK in 2006, and given a 20% rate of return to public sector R&D, a 5% increase in accessibility and efficiency would have been worth:

- £172 million a year in increased returns to public sector R&D;
- £124 million a year in increased returns to higher education R&D (HERD); and

- Around £33 million a year in increased returns to UK Research Council (RCUK) competitive grants funded research.

These are recurring annual gains from one year's R&D expenditure. So, if the change that brings the increases in accessibility and efficiency is permanent, they can be converted to growth rate effects.

4.1 Comparing costs and benefits

It is not possible to compare toll with Open Access publishing directly at the national level, because they perform very different roles: toll access publishing seeks to provide UK subscribers with access to worldwide research (to the limits of affordability); whereas Open Access seeks to provide worldwide access to UK research. These are very different things.

So, we approach the question from both sides, and we try to explore the lower and upper bounds. First, we explore the cost-benefit implications of simply adding Open Access publishing and self-archiving to current activities, all other things remaining the same (*i.e.* ceteris paribus scenarios); and then we explore the implications of Open Access publishing and self-archiving as alternatives to current activities by adding the estimated savings to estimated returns (*i.e.* net cost scenarios).

I am not in a position to release any of the results, and its preliminary work at the moment anyway, but at the moment the benefits of more Open Access exceed the costs in all cases except the nonsensical one of parallel publishing everything produced in UK higher education in both subscription and Open Access journals simultaneously. However, that is the only exception. So, for example, the benefits of parallel self-archiving (*i. e.* Green OA), even without cancellation of subscriptions, would outweigh the costs.

Of course, a major issue in comparing costs and benefits over a period is whether to model the transitional period or a steady-state hypothetical alternative system: and it makes a big difference. There is a lag between R&D expenditure and the realisation of returns to the research, so the impact is lagged and the value must be discounted accordingly. Put simply, in the transitional period, this has the effect

that, over 20 years, we are comparing 20 years of costs with 10 years of benefits. In a hypothetical alternative 'steady-state' system, the benefits of historical increases in returns would be realised from year one. So it would be comparing 20 years of costs with 20 years of benefits.

We took the view that it was more realistic and interesting, and of more immediate concern, to model the transition, but it must be emphasised that a transitional model returns significantly lower benefit/cost ratios than would a hypothetical alternative 'steady-state' model. Actually we "forced fed" our transitional model by putting returns into year one to see what difference it made, and it returned benefits around 10 times higher than in the transition.

4.2 Frequently asked question (FAQs)

We also explore some frequently asked question (FAQs), and we are still thinking of things to explore. Currently we have three cases.

The first is the issue of the diversion of research funding to author-pays fees and the implied loss of research. We look at this issue from both sides, asking: (*i*) if, for example, Wellcome Trust or RCUK's current spending on author-pays fees is beneficial; and (*ii*) what is the maximum percentage of research funding that could be diverted before exhausting the benefits derived from enhanced access. We can use the latter to "reverse" calculate what the author-fee would have to be to exhaust the benefits. Again, no announcements, but there appears to be plenty of scope to support author-fees – with the loss of research from the marginal diversion of funds more than compensated for by system cost savings and increased returns.

The second case explores the impact of delayed OA, estimating the impact of a one year delay, such as an OA embargo, on returns to R&D over 20 years. The third case explores the impact of speeding up the research and discovery process, estimating the impact of a one year reduction in the lag between R&D expenditure and its economic impact, from self-archiving pre-prints, for example.

5 Summary

We are currently working within the project team to complete and agree the base data, so we can finalise the report. We hope to make the models available for use online, so people can explore the issues for themselves, using different parameters. With permission, we may even try to record what people explore and which parameters they change – mainly for interest, but perhaps also to guide further work.

6 References

1. ARUNDEL, A., Geuna, A.: Proximity and the use of public science by innovative European firms. Economics of Innovation and New Technology 3 (6), 559-580 (2004).
2. BJÖRK, B-C.: A model of scientific communication as a global distributed information system. Information Research 12 (2) paper 307 (2007); Available http://informationr.net/ir/12-2/paper307.html.
3. HOUGHTON, J.W. et al.: Economic implications of alternative scholarly publishing models: Exploring the costs and benefits. Report to The Joint Information Systems Committee. London and Bristol: (forthcoming).
4. HOUGHTON, J.W., Sheehan, P.J. and Steele, C.: Research Communications Costs, Emerging Opportunities and Benefits. Department of Education, Science and Training (DEST), Canberra (2006); Available http://dspace.anu.edu.au/handle/1885/44485.
5. JISC Project; http://www.jisc.ac.uk/whatwedo/programmes/reppres/ economicsscholarlypublishing.aspx
   http://www.jisc.ac.uk/news/stories/2008/08/podcast55johnhoughton.aspx.

SCOAP^3

A New Publishing Model for High-Energy Physics

Anne Gentil-Beccot, Salvatore Mele, Jens Vigen

CERN, European Organization for Nuclear Research
Anne.Gentil-Beccot@cern.ch, Salvatore.Mele@cern.ch, Jens.Vigen@cern.ch

Abstract

The High Energy Physics (HEP) community has explored Open-Access for decades, though its pervasive "pre-print culture" and is now spearheading the road for an innovative model for the transition of its scholarly publishing to Open Access. The Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP^3) aims to be a central body to finance peer-review service rather than the purchase of access to information as in the traditional subscription model, with all articles in the discipline eventually available in Open Access. Sustainable funding to SCOAP^3 would come from libraries, library consortia and HEP funding agencies, through a re-direction of funds currently spent for subscription to HEP journals. This contribution presents the SCOAP^3 model and the current status of this project.

1 Introduction

The Open Access debate has spread to all areas and actors of scholarly communication, and there is growing awareness that the scientific community does not only have to advocate and support Open Access, but has the responsibility to take tangible steps towards Open Access. Open Access models are now being proposed by most publishers and several definitions are actively debated. This paper will present one concrete and specific Open Access model tailored to the needs of the High-Energy Physics (HEP) community: SCOAP^3 (Spon-

soring Consortium for Open Access Publishing in Particle Physics). HEP has a long tradition of innovations in scholarly communication and Open Access, and the lessons learned from the momentum gathered by the SCOAP^3 initiative in this field can inform the evolution of Open Access publishing in other fields.

HEP scientists aim to attain a fundamental description of the laws of physics, in order to explain the origin of the Universe. To reach these scientific goals, experimental HEP scientists team in collaborations to build very large instruments that reproduce on Earth the energy densities of the Universe at its birth [1]. Theoretical particle physicists collaborate to formulate hypotheses and theories, based on complex calculations, to explain and predict experimental findings. In total, the HEP community numbers around 20,000-30,000 active scientists, spread all over the word, all with a rooted culture of international collaboration to bring down technical and conceptual barriers.

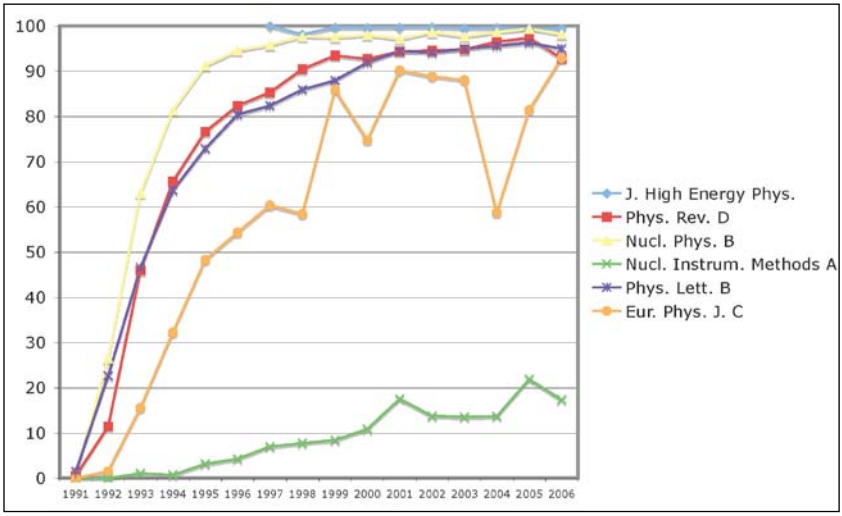

Fig. 1. Fraction of articles published in the six most important High-Energy Physics journals between 1991 to 2006 which is available on arXiv (Source SPIRES).

2 Scientific communication in HEP

Open Access is not new for HEP scientists: it has for a long time been the guiding principle of their scientific discourse [2], [3]. Indeed, in the paper era, HEP scientists used to mass-mail their preprints around the world in order to disseminate their research results rapidly and avoid the delays involved in the publishing process [4]. With the growth of the online world, this practice evolved, and soon after its creation, arXiv became the most important repository in HEP, and later in other disciplines [5].

Thanks to this preprint culture, HEP is today an almost entirely "green" Open Access discipline; a discipline where authors self-archive their research results in repositories which guarantee their unlimited circulation. Posting an article on arXiv, even before submitting it to a journal, is common practice. Even revised versions incorporating the changes due to the peer-review process are routinely uploaded. Almost all publishers allow such self-archiving and many encourage it, as a way to accelerate submission of manuscripts: often only the arXiv number is all is needed [6]. The single largest publisher in the field, the American Physical Society, even hosts a mirror of arXiv [7].

It is interesting to note that this success of "green" Open Access in HEP happens without mandates and without debates: very few HEP scientists fail to take advantage of the tremendous opportunities offered by the discipline repository of the field, and the linked discovery and citation-analysis tools of community services such as SPIRES [8].

The SCOAP^3 initiative has its deepest foundations in these simple observations: while the community has a vital need for the peer-review process guaranteed by the published journals, these are not any longer a vehicle of scientific dialogue. The system is therefore unstable, as libraries facing increasing subscription costs and decreasing budgets may cancel journals and put the entire peer-review system at risk. Moreover, the implicit link of libraries purchasing journals to support the peer-review service rather than provide access should be made explicit, shedding clarity on their crucial role in the process of scholarly communication.

It is important to emphasize that the call for Open Access journals in HEP does not only originate from the library community, which strongly supports it, but from the scientific community. At the beginning of 2007, the four experimental collaborations working at the CERN LHC accelerator, ATLAS, CMS, ALICE and LHCb, with a total of over 5,000 scientists from 54 countries, declared: "We, […] strongly encourage the usage of electronic publishing methods for [our] publications and support the principles of Open Access Publishing, which includes granting free access of our publications to all. Furthermore, we encourage all [our] members to publish papers in easily accessible journals, following the principles of the Open Access paradigm" [11]. These scientists followed up to their principles by publishing Open Access the seminal articles describing the multi-decade effort in the construction of their experimental facilities [12]. This decision enabled over 60,000 downloads of these articles in the first few months online, a number larger than the practitioners of the field, showing how Open Access can bring science back to the general public. In 2008, scientists from the CMS collaboration went one step forward and declared that they will favor "SCOAP^3 friendly journals" for their articles.

3 The SCOAP^3 model

The SCOAP^3 model was proposed in 2007 following a two-year debate involving libraries, funding agencies, research organizations and publishers involved with the HEP community [13]. It aims to convert high-quality peer-reviewed HEP journals to Open Access pursuing two goals:

- To provide open and unrestricted access to all HEP research literature in its final, peer-reviewed form.
- To contain the overall cost of journal publishing by increasing competition whilst assuring sustainability.

In practice, transition to Open Access will be facilitated by the fact that no more than 6,000 HEP articles are submitted to arXiv annually, and subsequently published in peer-reviewed journals. Most articles are published in just six peer-reviewed journals from four publishers [14,

15]. Five of these six journals carry a majority of HEP content. These (and their publishers) are:

- Physical Review D (American Physical Society)
- Physics Letters B (Elsevier)
- Nuclear Physics B (Elsevier)
- Journal of High-Energy Physics (SISSA/IOP, as of 2010 SISSA/Springer)
- European Physical Journal C (Springer)

The aim of the SCOAP^3 model is to assist publishers to convert these core HEP journals entirely to Open Access. In addition, two more 'broadband' journals carry HEP content: Physical Review Letters (American Physical Society) and Nuclear Instrument and Methods B (Elsevier), with 10% and 25% of HEP articles, respectively. It is the aim of SCOAP^3 to sponsor the conversion of this fraction to Open Access. This list of journals is not exhaustive and the SCOAP^3 initiative is open to all high-quality journals carrying HEP content.

The essence of this model is the formation of a consortium to sponsor HEP publications and make them Open Access by redirecting funds that are currently used for subscriptions to HEP journals. Today, libraries (or the funding bodies behind them) purchase journal subscriptions to implicitly support the peer-review and other editorial services and to allow their users to read articles, even though in HEP the scientists mostly access their information by reading preprints on arXiv. The SCOAP^3 vision for tomorrow is that funding bodies and libraries worldwide would federate in a consortium that will pay centrally for the peer-review and other editorial services, through a re-direction of funds currently used for journal subscriptions, and, as a consequence, articles will be free to read for everyone. This evolution of the current "author-pays" Open Access models will make the transition to Open Access transparent for authors, who will not have to pay publications fees. In addition, it will not imply additional costs for libraries, as it is based on the redirection of current subscriptions.

In order to calculate the annual budget for the transition of HEP publishing to Open Access, the SCOAP^3 Working Party considered several indicators: most publishers quote a first-copy price in the range of 1,000–2,000 Euros per published article, the total number of HEP publications in high-quality journals is between 5,000 and 7,000, according to the definition of the field. Therefore, the annual budget for the transition of HEP publishing to Open Access would amount to a maximum of 10 million Euros per year.

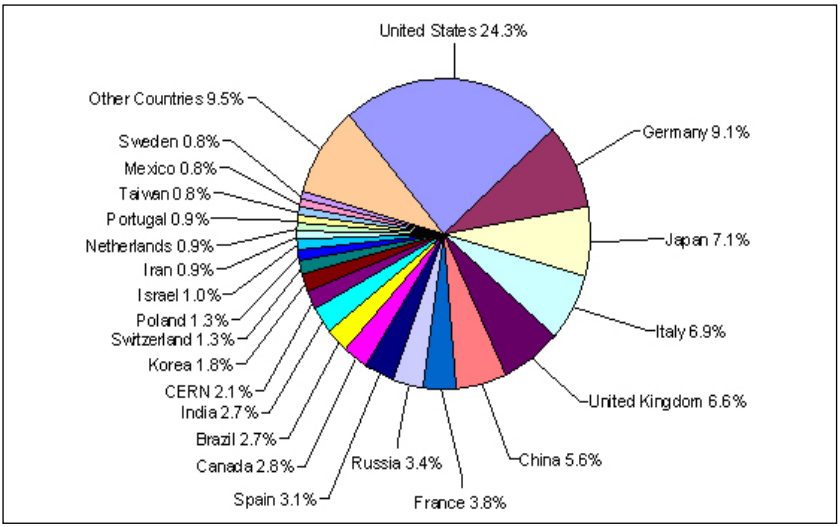

Fig. 2. Distribution of HEP articles by country, average 2005-2006. Krause *et al*., 2007 [15].

The costs of SCOAP^3 will be distributed among all countries according to a fair share model, based on the distribution of HEP articles per country, as shown in Figure 2 [15]. In practice, as the model will support peer-review, the costs will be distributed according to how much each country uses the scholarly communication system. To cover publications from scientists from countries that cannot be reasonably expected to make a contribution to the consortium at this time, an allowance of not more than 10% of the SCOAP^3 budget is foreseen.

4 Current status

SCOAP^3 is now collecting Expressions of Interest from partners worldwide to join the consortium. In Europe, almost all countries have pledged their contribution to the project. In the United States, Over 150 leading libraries and library consortia have pledged a redirection of their current expenditures for HEP journal subscription to SCOAP^3. In addition, Australia, Canada, Israel and Turkey have also joined SCOAP^3. In total, SCOAP^3 has already received pledges more than 2/3 of its budget envelope, as presented in Figure 3. Advanced negotiations are in progress in the countries that have not yet joined the consortium, in Europe, Asia and the Americas [13].

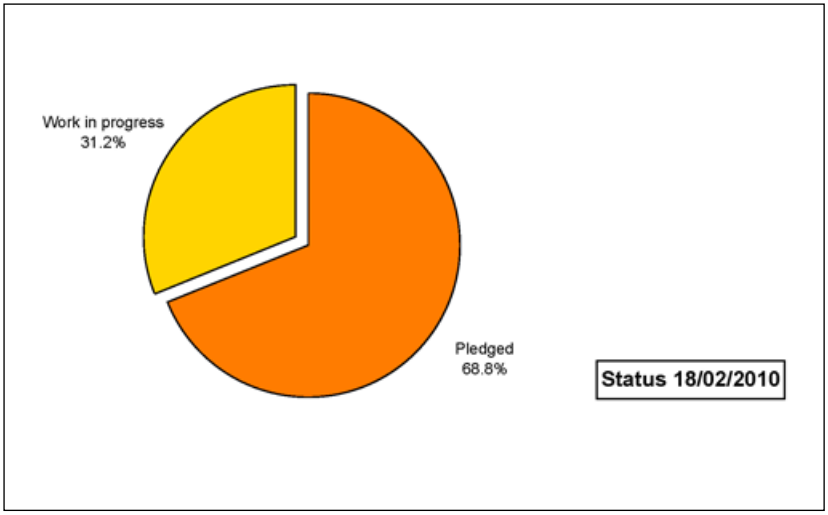

Fig. 3. SCOAP^3 fundraising status as of February 2010 [16].

Once it has reached a critical mass, and thus demonstrated its legitimacy and credibility, the SCOAP^3 consortium will be formally established, and its international governance put in place. This governing board will shepherd SCOAP^3 to issue a call for tender to publishers, who will answer by quoting a price per article for the peer-review and other editorial services and Open Access dissemination, in addition to

agreeing with other conditions such as unbundling Open Access journals from existing subscription packages, and the corresponding reduction of subscription prices. This call for tender will for the first time link price, quality and volume in the scientific publishing market place for this discipline. The governing board, if satisfied with the outcome of the tendering process, will then move forward with negotiating contracts with publishers. Eventually, partners will commit to the SCOAP^3 consortium through a Memorandum of Understanding.

As the fundraising process is being completed, the next phases of the project are being prepared. While waiting for this new era in scientific publishing in HEP, it is interesting to note that the HEP publishing scenario is already moving towards Open Access, with all publishers already showing support for this and other initiatives, attentive to the requests of the HEP community.

The APS has for a long time supported author self-archiving of publisher-formatted post-peer review articles via the authors' institutional repository, and its support for discipline repositories extends to even hosting a mirror of arXiv. Furthermore, since 1998, the APS has published Physical Review Special Topics Accelerators and Beams, an Open Access instrumentation journal based on sponsorship. SISSA since 2007, offers an Open Access institutional membership model for its Journal of High-Energy Physics (SISSA/Springer) as well as the Journal of Instrumentation (SISSA/IOP). This model allows member institutions to publish all their papers in Open Access form.

Since November 2007, publishing fees are waived for all HEP experimental papers to be published as Open Access in the journal European Physical Journal C (Springer) [17]. In September 2008, Elsevier also announced that it will join Springer and SISSA/IOP in publishing articles describing the physics results of the LHC as Open Access without any author fees [18]. In November 2008, EPL announced a similar policy [19].

These win-win agreements are giving the expected outcomes: in December 2009, the first LHC physics results, from the ALICE Collaboration, have indeed been published in Open Access and under the terms of the Creative Commons Attribution Noncommercial License by

the European Physical Journal C [20]. Shortly after, the CMS Collaboration published its first physics results in the Journal of High Energy Physics also Open Access and under a Creative Common license [21].

Publishers show therefore willingness to work on Open Access solutions together with the community, while waiting for the SCOAP^3 model to become operational.

5 Conclusions

It appears at first glance to be an ambitious enterprise to organize a worldwide consortium of research institutes, libraries and funding bodies that cooperates with publishers in converting the most important journals of a scientific field to Open Access. Since the previous Berlin conference, in September 2007 in Padua, the SCOAP^3 initiative had made impressive progress, reaching pledges covering 69% of its projected budget envelope, and expanding its consensus basis to include a large fraction of its North American contribution, through Canada and numerous partners in the U.S., with countries in Australasia and the Middle East also pledging support to the initiative. This success proves the potential of this sustainable Open Access alternative to the subscription model, which meets the expectations of researchers, funding agencies, libraries and publishers. SCOAP^3 also demonstrates that Open Access solutions can be implemented through worldwide consensus and international co-operation. It represents a successful transfer of knowledge from the HEP community and CERN to the Open Access debate, promoting international co-operation and consensus building in addressing global issues in scholarly communication.

6 References

1. CERN Public webpages, http://public.web.cern.ch/public/en/Science/Science-en.html.
2. HEUER, R., Holtkamp, A., Mele, S.: Innovation in Scholarly Communication: Vision and Projects from High-Energy Physics, Inf. Servi. Use 28, 83-96 (2008) arXiv:0805.2739.

3. AYMAR, R.: Scholarly communication in High-Energy Physics: past, present and future innovations, 2008. Eur. Rev. 17 (1), 33-51 (2009).
4. GOLDSCHMIDT-CLERMONT, L. 1965, Communication Patterns in High-Energy Physics, http://eprints.rclis.org/archive/00000445/02/ communication_patterns.pdf.
5. GINSPARG, P.: First Steps Towards Electronic Research Communication. Computers in Physics 8, 390 (1994).
6. JHEP submission page, http://jhep.sissa.it/jhep/help/helpLoader.jsp?pgType=author#submission.
7. APS arXiv mirror, http://aps.arxiv.org.
8. The history of SPIRES is recounted in L. Addis, 2002, http://www.slac.stanford.edu/spires/papers/history.html.
9. GENTIL-BECCOT, A. et al.: Information Resources in High-Energy Physics: Surveying the Present Landscape and Charting the Future Course. J. Am. Soc. Inform. Sci. Technol. 60 (1), 150-160(2009) arXiv:0804.2701.
10. GENTIL-BECCOT, A., MELE, S. and BROOKS, T.C Citing and Reading Behaviours in High-Energy Physics. How a Community Stopped Worrying about Journals and Learned to Love Repositories. Scientometrics, 2009. arXiv:0906.5418.
11. Statement of the LHC collaborations in support of Open Access, http://www.scoap3.org/community/lhc.html.
12. JINST Home Page, http://www.iop.org/EJ/journal/jinst.
13. S. BIANCO et al.: Report of the SCOAP3 Working Party, http://www.scoap3.org/files/Scoap3WPReport.pdf (2007).
14. MELE, S. et al.: Journal of High-Energy Physics 12 (2006) S0, arXiv:cs.DL/0611130.
15. KRAUSE, J. et al.: Quantitative Study of the Geographical Distribution of the Authorship of High-Energy Physics Journals (2007) http://scoap3.org/files/cer-002691702.pdf.
16. SCOAP3 website, http://scoap3.org.
17. Open Access Publishing in the European Physical Journal (EPJ), http://www.epj.org/open_access.html.
18. Elsevier statement on Open Access for LHC papers, http://www.elsevier.com/framework_products/promis_misc/nuclearadcolour80x210RGB.pdf.
19. EPL statement on Open Access for Experimental and Theoretical HEP Articles, http://www.iop.org/EJ/journal/EPL.
20. ALICE Collaboration, Aamodt, K. et al., European Physical Journal C 65 (2009) 111.
21. CMS Collaboration, Khachatryan, V. et al., Journal of High-Energy Physics 02 (2010) 1.

All links last visited February 25th, 2010.

SciELO Program Publishes More than 500 Latin-American Open Access Journals[1]

Abel L. Packer, Rogério Meneghini, Solange M. Santos, Fabiana Montanari Lapido, Adriana Luccisano

Latin-American and Caribbean Center on Health Sciences Information (BIREME/PAHO/WHO)
solange.santos@bireme.org

Abstract

SciELO (Scientific Electronic Library on Line, www.scielo.org) is an eleven-year old public funded Open Access program aimed to improve quality of developing countries' national and regional scholarly journals by increasing their visibility, accessibility, use and impact. The program manages and operates the online publishing of a network of national and thematic collections of quality scientific journals. SciELO collections include also mechanisms to follow up the performance of the collections and the individual journals through the online control and publication of use and impact through indicators of access and citations. In 2008, SciELO network comprises eight national and two certified collections, totaling more than 500 journals, 191 thousands full text articles and 3.5 millions of bibliographic citations. SciELO has made a remarkable contribution to the improvement of the quality of the journals as measured by editorial and publishing parameters, number of access to articles and number of citations received by the journals. Under the general coordination of BIREME/PAHO/WHO in cooperation with the State of São Paulo Research Foundation (FAPESP), the SciELO collections are managed and operated in a decen-

[1] This text reflects the presentation done by Solange M. Santos in the Open Access for Development, Open Acces around the World panel on the Berlin 6 Open Access Conference, which was based on the studies cited in the references and statistics, methodologies and technologies available on the SciELO website – www.scielo.org.

tralized way by national research institutions following common methodologies and technologies.

1 Introduction

The mainstream scientific literature (as expressed by the journals indexed by the products of the Thomson-Reuters Scientific, former Institute for Scientific Information (ISI) does not accurately reflect the scientific productivity in less developed countries. This is a statement that most of the scientists from developing countries would agree upon, even those that usually publish in mainstream journals. Several studies have already addressed this problem, and also in Latin America and Caribbean this issue has long been discussed.[1,2]

In this context, the SciELO Program (Scientific Electronic Library on Line) was implemented with the regular operation of the SciELO Brazil collection, as a result of the SciELO Project partnered since 1997 by scientific editors, the State of Sao Paulo Research Foundation (FAPESP, www.fapesp.br) and the Latin American and Caribbean Center on Health Sciences Information (BIREME-PAHO-WHO, www.bireme.org), a Pan American Health Organization/World Health Organization specialized center. As of 2002, the National Council of Scientific and Technological Development (CNPq, www.cnpq.br) also began supporting the operation and development of the SciELO Brazil collection[3].

SciELO was born in 1997 as an Open Access collection of Brazilian high quality journals. In 1998, it was adopted by Chile's National Science and Technology Commission (CONICIT), starting the development of the SciELO Network. Over the years, SciELO has acquired an important presence and role in the international Open Access movement devoted to a great extent to scientific literature excluded from mainstream. In fact, SciELO is one of the first Open Access initiative operating collections of journals, introduced five years before the Budapest and six years before Bethesda and Berlin declarations, and presently, one of the largest ones[4].

Fig. 1. Timeline of Open Access initiatives. Source: Folder of International Seminar on Open Access for Developing Countries, by BIREME.

2 SciELO Model

The SciELO model for electronic publishing is primarily focused on improving scientific journals from developing countries, with an emphasis on Latin America and the Caribbean, strategically aiming to contribute to the advance of scientific research and publishing of its results.

The SciELO model is also decentralized and strongly based on developing and strengthening capacities and alliances among national and international institutions, forums and stakeholders related to scientific communication (including communities of researchers, editors, scientific and technological institutions, research support agencies, universities, libraries, scientific and technological information centers, etc.) to disseminate, improve and ensure their sustainability and develop national capabilities in scientific communication.

The three main objectives of the program are [5]:

1. To publish online national collections of best journals in Open Access mode. This would bring a wider national and international visibility to these journals. This aim was fully achieved initially by the Brazilian and Chilean collections, encompassing, respectively, 213 and 77 journals, and more than 100 thousand and 17 thousand articles by the end of 2008. After the adoption of the initiative by Chile and other Iberian countries (www.scielo.org), the SciELO network covers over 500 journals and 160 thousand articles.

2. To improve the quality of the journals in the countries that adopted SciELO with respect to several attributes like relevance of the articles, accuracy in the methodology, care in presentation and assessment of articles by *ad hoc* referees. All these requirements are judged for each journal by a special *ad hoc* panel.
3. To create a bibliometric/scientometric database, producing indicators similar to those provided by ISI-JCR, for scientific and technological studies which had not been possible using the international databases only.
4. The specific objectives to apply the model are to increase visibility, accessibility, quality, credibility, use and impact of scientific journals from developing countries and, therefore, of researches whose results are published in national or regional journals. By accomplishing these objectives, SciELO has decisively contributed to overcome the phenomenon known as the lost science in developing countries[6].

3 SciELO Network

The SciELO Network has been operating for ten years already and continues to expand with the national collections from Latin America and Caribbean countries, Spain and Portugal. The well-succeeded operation and expansion of SciELO in countries with higher scientific production in Latin America and the Caribbean has occupied an important space in the international Open Access scientific production movement.

In November 2008, the SciELO Network operates ten certified collections of online journals and there are six in development. National collections are required to accomplish a set of criteria to acquire and maintain the certification of SciELO collection. As shown in table 1 and 2 , the national certified collections are from eight countries: Argentina, Brazil, Chile, Colombia, Cuba, Portugal, Spain and Venezuela. The thematic collections include Public Health and Social Sciences English Edition. There are over 500 titles of certified journals and more than

191 thousand online full-text articles, including original scientific articles, review articles, editorials and other types of communication.

Certified Collections			
Collection	**start**	**titles**	**articles**
Argentina	2004	47	5145
Brazil	1997	213	111908
Chile	1998	71	20586
Colombia	2004	57	5780
Cuba	2001	20	9257
Spain	2001	36	11637
Portugal	2004	20	2547
Venezuela	2000	41	7446
Public Health	2000	11	16388
Social Sciences	2006	29	378
Total		**545**	**191072**

Table 1. SciELO Network – certified collections status in November 2008. Source: SciELO Portal (www.scielo.org) October 2008.

Development Collections		
Collection	**start**	**titles**
Costa Rica	2000	9
Mexico	2003	26
Paraguay	2007	3
Peru	2004	23
Uruguay	2005	7
Total		**69**

Table 2. SciELO Network – Development collections status in November 2008. Source: SciELO Portal (www.scielo.org) October 2008.

The collections SciELO Brazil and Chile are in an advanced stage and, in the first semester of 2008, had a monthly average of 8.6 and 2.3 million visits, respectively.

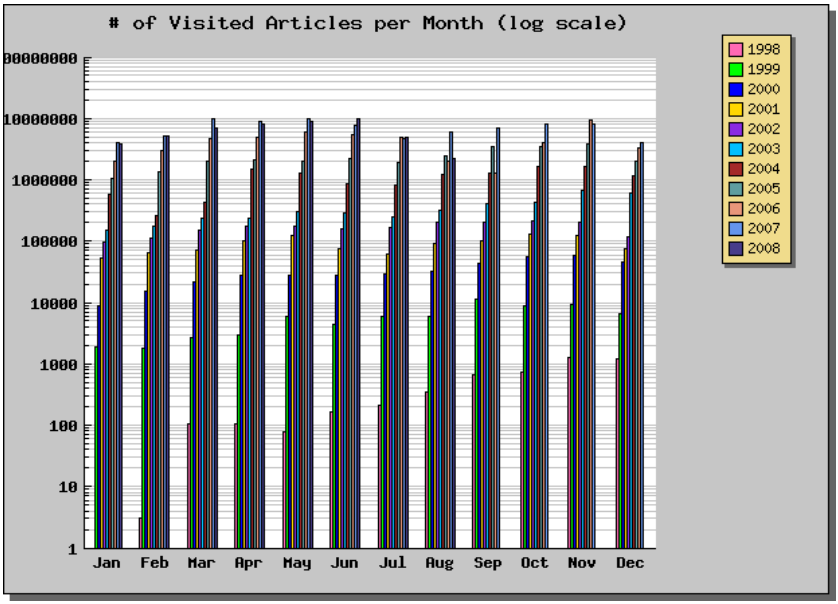

Fig. 2. SciELO Brazil: articles visited per month. Source: SciELO Brazil site usage report (http://www.scielo.br/scielo.php?script=sci_stat&lng=en&nrm=iso).

The SciELO Network publishes, in Open Access, the most credible scientific journals from Latin America and the Caribbean. It also includes a progressive interoperation among the journal collections in order to enhance even more the synergy environments and scalability promoted by the SciELO collections, so as to include quality journals from developing countries as full part of the international scientific communication flow. Thus, it contributes to render the research and its publishing more visibly, with higher quality, more reliably and used in the development of science and technology.

The collections at the SciELO Network, both national and thematic, operate the same methodology and follow the same inclusion and permanence criteria for the collection titles. Both the management and operation of the collections have been improving and reflect the progressive improvement of the editorial capacity of the countries that operate SciELO to produce high quality scientific journals on the Internet.

4 SciELO Platform – Resources and Features

The SciELO Platform offers several resources to enhance online publishing scholarly journals, such as: dynamic links with main national and international databases and systems, using standard protocols such as OAI-PMH and SOAP, publishing in more than one language, ahead of print publication of articles as soon as they are approved by peer review and edited; articles identified with DOI (Digital Object Identifier); use and impact indicators per journal and collection, online submission, etc.

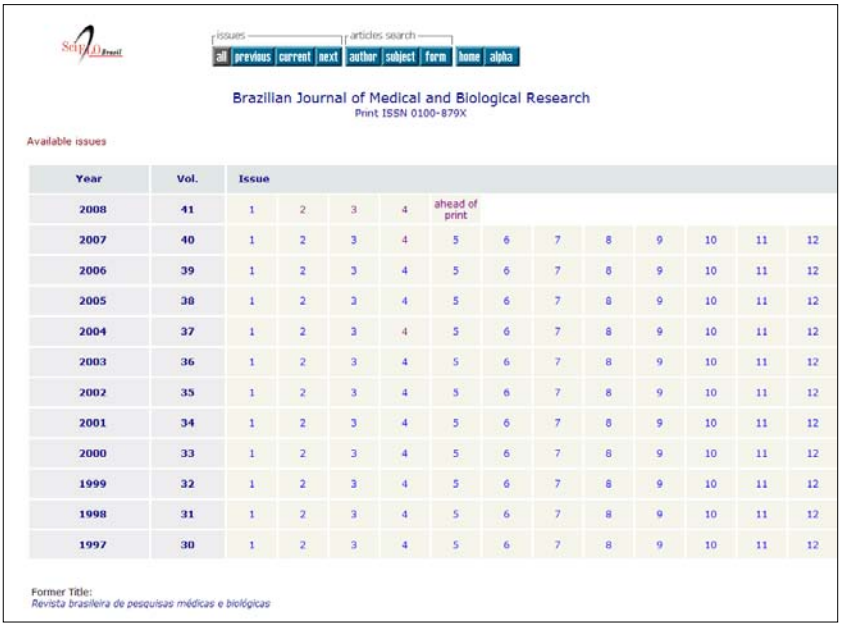

Fig. 3. Ahead of print publication.

Fig 4. Publishing in more than one language.

Fig 5. Use and impact indicators per journal.

5 SciELO Portal

The SciELO Network Portal (http:www.SciELO.org) was reformulated and a new version with several new services and options was launched in August of 2007 during the celebration of the 10th anniversary of the SciELO project. The Portal offers integrated access to the national and thematic collections allowing identification of articles cited and similar articles of the SciELO Network. In addition to the traditional search mechanisms, other retrieval options are also available, such as: lexical similarity, relevance based on thesaurus concepts, and integration with Google Scholar, etc. Moreover, the graphic interface was redesigned to become more attractive and dynamic, providing services on demand based on user profile (My collection, articles of my profile, My Alerts, My Links, My Comments, etc.).

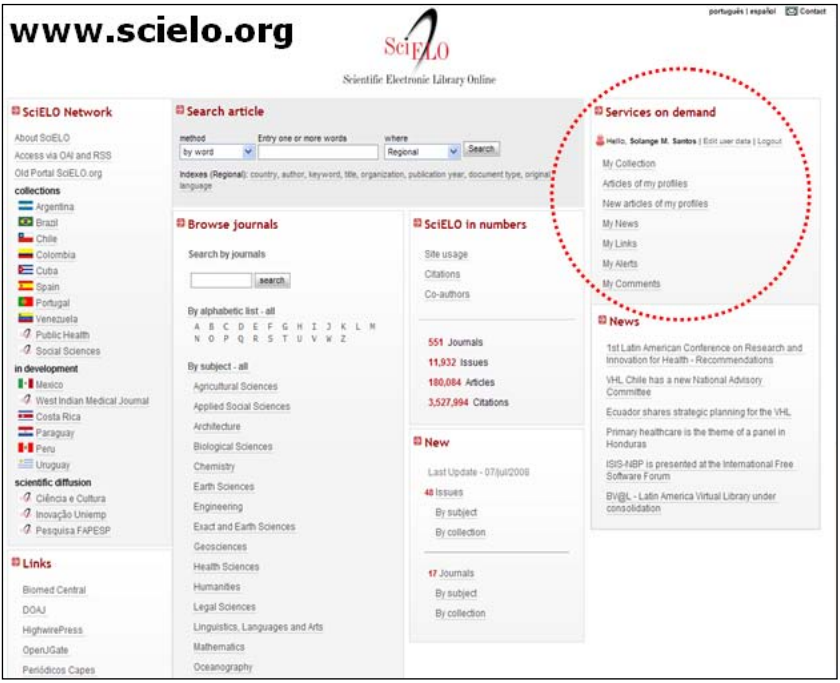

Fig. 6. SciELO Network Portal (www.scielo.org).

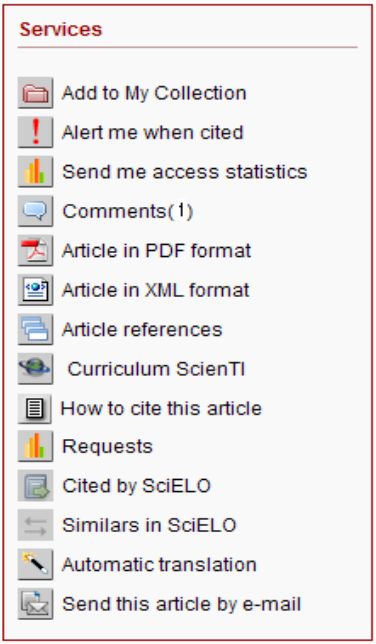

Fig. 7. SciELO Article services box.

6 Latin American journal indicators and indexing Structure

Many Ibero-American countries have successfully developed their infrastructures of research in different dimensions such as, human and financial resources, and in the formulation, review and development of research projects and the publication of their results, preferably in Thomsom-Reuteurs JCR journals. However, these countries face difficulties in publishing their own quality journals[5].

The success of SCIELO is indicated by a range of metrics: increased use (numbers of visits and downloads; cf. fig. 2), raised impact factor (fig. 9), both in the SciELO collection and in JCR Thomson-Reuters, increased numbers of submissions and rejections of articles, and finally more titles added to international indexes.

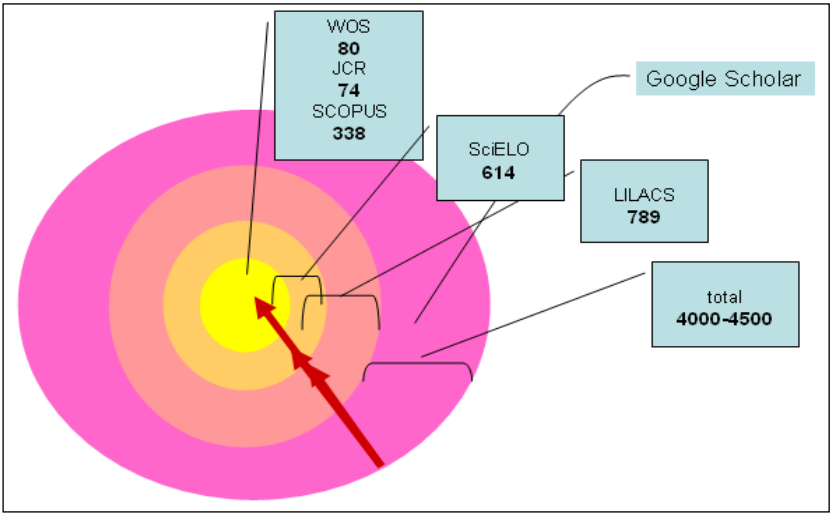

Fig. 8. Indexing structure of journals in AL&C. Source: BIREME/PAHO/WHO, 2008.

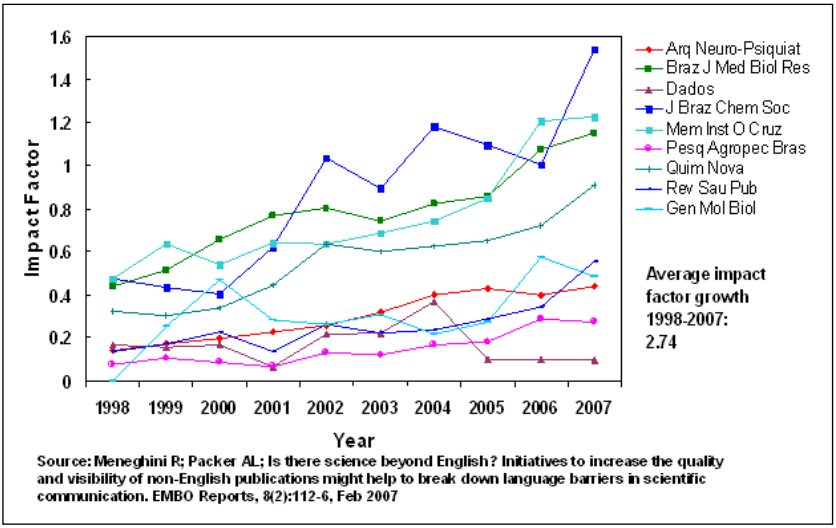

Fig. 9. Impact factor of the Scientific Electronic Library Online (SciELO) journals from 1998 to 2005. Source: Thomson Scientific Journal Citation Reports.

Most SciELO Brazilian journals that are also indexed at the JCR Thomson-Reuters had a dramatic increase in number of citations received as well as in the impact factor[7]. In this context, SciELO represents one of the most important international initiatives that brought greater impact to strengthening journals in Brazil and other Latin American countries.

SciELO database also allows systematic generation of data and documents that may be important to subsidize political decisions in the scientific arena. In a recent study using SciELO and JCR databases[8] it was found that the SciELO/Brazil journals could be classified into two categories:

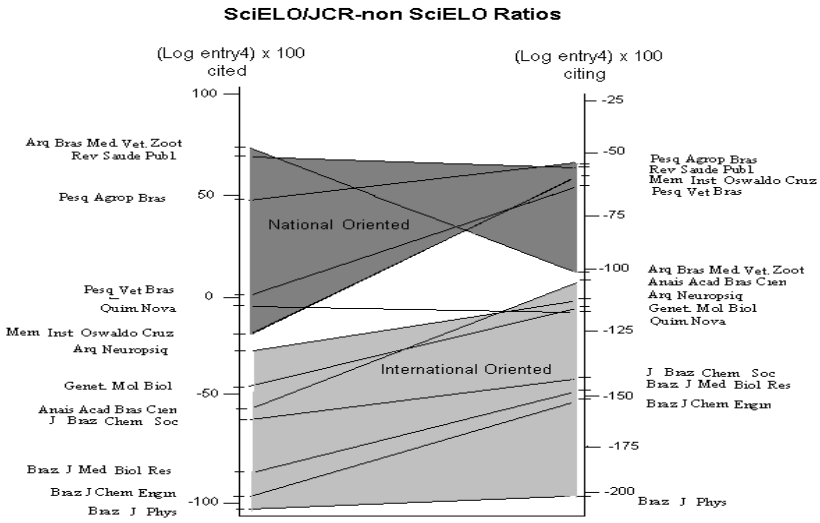

Fig. 10. Brazilian Journals classified as national or international oriented. Source: Meneghini R, Mugnaini R, Packer AL (2006) International versus national oriented Brazilian scientific journals. A scientometric analysis based on SciELO and JCR-ISI databases. Scientometrics 69: 529–538.

One in which the journals have a tendency to cite and to be cited by authors of the national scientific community. The other one had a more accentuated trend to seek international visibility, both in terms of citing and being cited. Journals representative of the first category

belonged to the areas of agriculture, animal sciences, health sciences and tropical medicine. On the other hand, journals dealing with the basic sciences, physics, chemistry and biology, were more internationally oriented.

This is an important issue since the Brazilian funding agencies have been very much inclined to privilege the Thomson Scientific indicators for their decisions, regardless of the scientific area. It seems clear that in certain areas scientific information exchange is prevalent among national scientists, as measured by the flow of citations. This trend and the need to make it sustainable can now be considered by funding agencies so as to avoid the risk of bringing about a decline in the information exchange in the Brazilian context. It is likely that other developing countries face the same challenges and could benefit of similar studies.

7 Conclusions

SciELO model has been consolidated as the main path to promote and enhance the national scientific communications, enlarging in an extraordinary way the national and international visibility and accessibility of the main scientific journals published in the Latin American and Caribbean countries.

SciELO indexation turned out to draw the interest of most Latin American journals, emulating the attractiveness that JCR and related indexes generate in the international context. In fact, national research and education funding and evaluation institutions from Latin American countries are progressively ranking favorably SciELO indexed journals among national publications. In this context, SciELO has created a virtuous circle that promotes the quality improvement of the national scientific journals.

8 References

1. VELHO, L.: Indicadores Brasileiros. Intersciencia 15, 139–144 (1990).
2. SPINAK, E.: Los análisis cuantitativos de la literatura científica y su validez para juzgar la producción latinoamericana. Boletín de la Oficina Sanitaria Panamericana 120, 139–147 (1996).
3. PACKER, A. L. *et al*.: SciELO: uma metodologia para publicação eletrônica. Ciência da Informação 27 (2), 109-121 (1998).
4. GIBBS, W. W.: Lost science in the Third World, Scientific American 273 (2), 92-99 (1995).
5. MENEGHINI, R.,PACKER, A. L.: Is there science beyond English? Initiatives to increase the quality and visibility of non-English publications might help to break down language barriers in scientific communication. EMBO Reports. 8 (2), 112-116 (2008).
6. MENEGHINI, R., PACKER, A. L., Mugnaini, R.: International versus national oriented Brazilian scientific journals. A scientometric analysis based on SciELO and JCR-ISI databases. Scientometrics 69 (3), 529-538 (2006).

9 Links

1. UHLIR, P. F.: Issues in Providing Open Availability to Public Scientific Information Resources in Developing Countries. InterAcademy Panel on International Issues (2005), Available at: http://www4.nationalacademies.org/iap/iaphome.nsf.
2. PACKER, A. L., MENEGHINI, R.: Learning to communicate Science in Developing Countries. Interciennncia 32 (9), (2007). Available at: http://www.scielo.org.ve/scielo.php?script=sci_arttext&pid=S0378-18442007000900014&lng=es&nrm=iso.

Open Access Contributions in the Humanities?

The Approach of the Berlin-Brandenburg Academy of Sciences and Humanities

Christiane Fritze

Berlin-Brandenburg Academy of Sciences and Humanities
christianefritze@googlemail.com

Abstract

The paper illustrates contributions to the Open Access movement in the humanities approached by the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). It describes first the general role that German academies play within the scientific landscape of the country and analyses then the specific considerations and the approach of the BBAW towards Open Access publications. In the second part, two case studies of separate but related Open Access initiatives of the BBAW are employed to illustrate this approach. The first example describes the Telota Archive-Editor, which is a generic tool to compile resources of highly structured, distributed and independent research data. The second example portrays the cumulative concept of the Deutsches Textarchiv (German Text Archive) and its mission of being an active archive. The implementation of both projects has raised a couple of issues and challenges concerning pioneering methods in the humanities. These are discussed in the final part of this chapter.

1 Introduction

Dissemination of research results is broadly discussed among scholars, institutions, publishing houses and libraries, and Open Access is the “hot topic” of the current discourse. While sciences on the one

hand seem to have a very clear opinion on that, the situation in the humanities looks differently. Gersmann (2007: 78) says that "[a]nders als in den STM-Fächern (Science, Technoloy, Medicine) wissen nur wenige Fachvertreter im Detail, was die Forderung nach Open Access bedeutet, geschweige denn, dass sie dem Ruf nach einer Archivierung ihrer Texte auf Hochschulservern oder in fachspezifischen Repositorien Folge leisten würden." (Gersmann 2007: 78) From the point of view of a humanities scholar (to be precise: a digital humanist), I suppose that these differences originate from culture and tradition in the respective disciplines. This paper discusses the approach, the Berlin-Brandenburg Academy of Sciences and Humanities has undertaken to contribute to the Open Accesgs movement in the humanities.

In the first part of this paper, I describe the situation of the academies within the scientific landscape of Germany which are a counterpart to other non-university institutions like the Fraunhofer-Gesellschaft or the Helmholtz Association of German Research Centres. I will then come to the role of my home institution — the Berlin-Brandenburg Academy of Sciences and Humanities — and their approach towards Open Access publications. I am illustrating this with two examples: The Telota Archive-Editor and the cumulative approach of the German Text Archive, two independent but related Open Access initiatives of the BBAW. Finally I will raise some questions concerning the challenges while working with pioneering methods in the humanities.

2 The Berlin-Brandenburg Academy of Sciences and Humanities and their Open Access requirements

A good year ago, the German Academy of Sciences Leopoldina was appointed the National Academy of Sciences — by the way a controversially[1] adopted election[2]. Besides this organisation which works traditionally in the fields of science and medicine, and since its desig-

[1] Focus Online 18.02.2008, sueddeutsche.de 14.07.2008, and ZEIT online 14.07.2008.

[2] Presseerklärung der Gemeinsamen Wissenschaftskonferenz, Bonn, 14.07.2008 <http://www.gwk-bonn.de/fileadmin/Pressemitteilungen/pm2008-09.pdf>.

nation as national academy is responsible for “science-based recommendations and statements which are addressed to political institutions and the interested general public”[3] Germany has eight more federal state’s academies[4]. “Scientific academies in Germany are fellowships of scholars, elected for distinction and achievement in their disciplines. Thus, an academy is a scholarly society that offers to its members the opportunity for regular interdisciplinary discussion of their work and results.”[5] In contrast to the Leopoldina consisting exclusively of academy members, the federal state’s academies are equipped with employees carrying out active research, namely arts and humanities research. In order to better represent their common interests, the federal state’s academies joined together to the Union of the German Academies of Sciences and Humanities.

In the not university related research landscape of Germany this union builds the complement to the associations like the Fraunhofer-Gesellschaft or the Helmholtz Association of German Research Centres with their STM departments.

The Union of the German Academies of Sciences and Humanities represents “more than 1,600 scientists and academics who are outstanding representatives of their disciplines, both nationally and internationally, and whose work covers a broad range of subjects.”[6]

The BBAW hosts different kinds of projects which produce a wide variety of research results. The projects are in the field of research in lexicography (e. g. Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, Goethe-Wörterbuch), historical and critical editions (e. g. Leibniz-Edition, Marx-Engels-Gesamtausgabe), prosopography (Prosopographie der mittelbyzantischen Zeit), history of art (e. g. Corpus

[3] http://www.leopoldina-halle.de/cms/en/academy/recommendations-and-statements.html.

[4] Akademie der Wissenschaften in Hamburg; Akademie der Wissenschaften und der Literatur, Mainz; Akademie der Wissenschaften zu Göttingen; Bayerische Akademie der Wissenschaften; Berlin-Brandenburgische Akademie der Wissenschaften; Heidelberger Akademie der Wissenschaften; Sächsische Akademie der Wissenschaften zu Leipzig; Nordrhein-Westfälische Akademie der Wissenschaften und der Künste.

[5] http://www.akademienunion.de/akademiegeschichte/english.html.

[6] http://www.akademienunion.de/union/english.html.

Vitrearum Medii Aevi), bibliography (e. g Jahresberichte für deutsche Geschichte, Kant's gesammelte Schriften), and philology (e. g. Corpus Coranicum, Corpus Medicorum Graecorum/Latinorum). Despite the large number of scientific enterprises, the academy focuses its mission on three main aspects: "on research projects reconstructing cultural heritage through long-term ventures"[7] , "on interdisciplinary and transdisciplinary projects devoted to the identification and tackling of scientific and social problems of the future"[8], (e. g. Visual Cultures) and "on promoting the dialogue between the scientific community and the general public"[9]. In this area falls the Electronic Publishing Working Group.

The BBAW was one of the very early movers towards Open Access. Consequently, the BBAW was one of the initial signatories of the "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities" in 2003.[10] Even before the launch of the Berlin Declaration, the BBAW had begun to implement activities according to an Open Access strategy. The Electronic Publishing Working Group's report for the year 2000 paved the way for the Telota initiative (The Electronic Life of The Academy) which aims at providing a sustainable, interactive and transparent way to inspire activities in supporting research, communication and presentation with electronic media.[11] The initiative promotes national and international standards as guidance for documentation, presentation, and use of scientific research results. It comprises all activities which support research, communication and presentation by electronic means. One of the main goals of the Telota initiative is to improve the visibility and impact of the academy's research projects.

Having in mind these goals, it soon became obvious that Open Access has to be regarded in a comprehensive sense and should not be restricted to the publication and dissemination of scientific papers but

[7] http://www.akademienunion.de/bbaw/print.html.
[8] http://www.akademienunion.de/bbaw/print.html.
[9] http://www.akademienunion.de/bbaw/print.html.
[10] http://oa.mpg.de/openaccess-berlin/signatories.html.
[11] Czmiel/Fritze/Neumann 2007, p 101.

the provision of raw data as well as source material. The Berlin Declaration covers this aspect by stating that "[...] open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material."[12] Research projects of the BBAW show this broad variety of output formats as can be seen within the projects of the Telota initiative. Telota can hence be regarded as an excellent test-case for the implementation of the Berlin Declaration.

So far, however, the availability of such resources for future use in research is still very limited. In the past decade, research in the humanities created a huge amount of data, for example in databases or through xml-documents on the web. These resources can be regarded both as the production of texts as results of research on their own which might be published in a traditional way i.e. a printed edition or a journal paper, but also, as the fundamental starting point for new research. Disseminating data in such a way which seems to have been common practice in the sciences for quite a while is, however, a new paradigm for the humanities and might lead to a changing scholarly communication in the knowledge society.

For the near future, we will encounter a considerable increase in the presentation of research results in digital form as well as in the usage of digital resources. A suitable means to deal with this changing working practice is the creation, provision and maintenance of sustainable and of course openly accessible data repositories. We have to consider not only the repositories themselves, but more importantly, also the tools that are and were used to generate and present the data. We consequently regard it as our duty to publish the source-code of the software-tools the Telota working group has developed and will develop in future.

Although or maybe because of the huge variety of data produced within the projects, quality assurance is an important factor. Telota therefore applies established data standards, for example text encod-

[12] http://oa.mpg.de/openaccess-berlin/berlindeclaration.html.

ing compliant to the guidelines of the Text Encoding Initiative (TEI)[13]. In the same way, we encode metadata in the current metadata standards METS/MODS[14] and Dublin Core[15].

In the remainder of this paper, I would like to illustrate these general design principles of our Open Access approach by employing two examples from the ongoing work at the BBAW and paying particular attention to previously cited passage of the Berlin Declaration. One particularity of these projects is the form of their presentation. In contrast to classic publication by articles in scientific journals, these results are mainly evident in heterogeneous data collections. It is in the nature of editing activities, for instance, that research results will not lead in a classic journal article but in an edition which takes usually years for completion.

The first example is taken from two projects with a focus on historical personal data. The Prosopography of middle Byzantine period (Prosopographie der mittelbyzantinischen Zeit) is a “Who is who” of the byzantine empire from 641 to 1025 AD. So far, it has been published only in printed form. The Prussian State and Culture (Preußen als Kulturstaat) is a newly set up project that interprets archive files to examine culture within the Prussian state.

As one can easily imagine — because of the large amount of data, the complexity and the rather high structuring, and the permanent increase of the prosopographic data base — printed books are quite an inappropriate form to represent data such as prosopographic records. Besides this, a digital edition opens the scholar much more possibilities for his research. To find a better way for data acquisition and representation, the Archive-Editor was developed by the Telota initiative.

The Telota Archive-Editor[16] is a tool for historians to support the structured indexing of historical files. It was developed in agile exchange with the project members of the long term academy project

[13] http://www.tei-c.org/.
[14] http://www.loc.gov/standards/mets/, http://www.loc.gov/standards/mods/.
[15] http://dublincore.org/.
[16] http://www.telota.de/telota-lab/personendb/.

The Prussian State and Culture. The essential components of this tool are an interface for editing and indexing activities, and a data base implemented as client server architecture. The main parts of the editor are freely configurable by the scholars who work with it. All the scholarly work is done offline in the archives and can be synchronized afterwards with the online database. Thus, each project member has access to the current version of the records, including recent changes by his colleagues. The database can be accessed and queried using a web browser. The Telota Archive-Editor runs on any operating system which supports the Java Runtime Environment in version 5 or higher. The Archive-Editor is not only limited to the indexing of prosopographical material: Because the conceptual design of the tool follows an event driven approach, the editor can be used for all structuring purposes as long as they are based on events.

Second example: The German Text Archive (Deutsches Textarchiv, DTA)[17] is a project in the field of linguistics and funded by the German Research Foundation. Its aim is to establish a core corpus of the most important works in the German language dating from the beginning of letterpress printing until present. The first project phase from July 2007 to June 2010 focuses on the digitising of about 750 texts (circa 200,000–250,000 pages) dating from between 1780 and 1900. The forthcoming second phase deals with texts originating from the 1650 until 1780. Texts originating from the 20th century were incorporated by a co-operation with the academy project Digital Dictionary of the German Language (Digitales Wörterbuch der deutschen Sprache), a TEI P5 XML linguistically annotated corpus of one billion tokens.

The selection of the initial 750 works for the corpus of the German Text Archive is based upon recommendations given from three groups of scholars: firstly, the editors of the famous Deutsches Wörterbuch by Jacob Grimm and Wilhelm Grimm, secondly, the academy fellows and thirdly, the members of the consortium Arbeitsgemeinschaft Sammlung Deutscher Drucke (six libraries collaborating to build a compre-

[17] http://www.deutsches-textarchiv.de/.

hensive collection of printed literature published in German-speaking countries).

As a whole the German Text Archive comprehends most important German texts of different genres such as fiction, science, technology, medicine, philosophy and law. We digitise almost exclusively the first editions of these works.

The texts will be highly structured, allowing a word by word alignment of the images with the full-text and vice versa. To achieve this detailed structure for all of the texts within the DTA corpus, we employ a customisation of the TEI P5 XML format in which the reference between image and text will be based on the encoding of coordinates for each character for most of the corpus texts. Afterwards, they will pass computational linguistic routines such as tokenisation, lemmatisation, morphologic analysis, part-of-speech tagging and orthographical mapping.

On the descriptive level, metadata sets based on METS/MODS or Dublin Core will be derived from the TEI P5 XML format to allow, on the one hand, interchange with the libraries whose books we digitised. On the other hand, they are needed for potential future co-operation such as the integration into German research portals. To guarantee the persistent identification of the digital facsimiles and the associated full-text the usage of Uniform Resource Names (URNs) is envisaged.

Another aim of the German Text Archive is to become an “active archive” which allows not only to search and browse the text corpus but also to work on the texts themselves. This collaborative conception can be thought of under two aspects: firstly, as an enlargement of the basis of the archive to extend the corpus with further texts. Secondly, the collaborative aspect can be expressed as a structural enrichment of the documents by contribution of scholars, i.e. additional annotations made by registered third party users. For instance, it is imaginable that one scholar annotates proper nouns, another tags geographical information and a third one enhances the basic text structure by a mark up of a metrical analysis of poems. This leads to added value in some respects which can in turn be made available to the scholarly community.

Like the Telota Archive-Editor, the realisation of the DTA corpus follows vastly the principles of the Berlin Declaration:

1. The texts of the German Text Archive will be available for download in different formats as long as they do not fall under any copyright restrictions.
2. Developed software will be published under an open source licence.
3. To allow contributions to the German Text Archive, the content of the archive has to be freely accessible.
4. And, consequently, the contributions, too, must be available for the public.

While the first three aspects are quite obvious, especially the last facet raises a couple of issues, some of them not yet adequately addressed.

One is the question of quality assurance: while the system can assure that contributions by third-party scholars are flawless in form due to the underlying data standards, the quality of the contributions regarding the content is a different thing.

This question arises not only with regards to the content but to the contributors themselves: who is eligible to contribute?

How do you credit the contributions? Traditional scholarly publication in form of articles is evaluated using impact factor and citation analysis. But these procedures are not yet suitable for the data repositories and digital resources as seen before. Other means of evaluation and quality assurance must be found.

Another interesting aspect can be seen in the humanities disciplines themselves. Does humanities research change its nature by this new and direct access to raw data? Can, for instance, the availability of the original full-text of “The history of the Hellenic clans” by Otfried Müller[18] — only to arbitrarily name a book title in the corpus —

[18] Müller, Karl Otfried: Geschichten Hellenischer Stämme und Städte, Breslau: Max, 1819-1824.

stimulate new ideas in Hellenic studies? Or does the comprehensive access to elaborate prosopographic data allow new inter-disciplinary approaches?

Enabling scholarly contributions to a living full-text archive of important works, developing of standards for digital editions with scalable architectures, or comprehensive access of prosopographic databases: The next challenge for the digital humanities will be the answer to the question on how to preserve, guarantee Open Access and finally evaluate such a variety of raw research data.

3 References[19]

1. CZMIEL, A., Fritze, C., Neumann, G.: Mehr XML – Die digitalen Projekte an der Berlin-Brandenburgischen Akademie der Wissenschaften. In: Braungart, G., Gendolla, P., Jannidis, F. (eds.) Jahrbuch für Computerphilologie 8, pp. 101-125. Paderborn: mentis (2007), <http://computerphilologie.tu-darmstadt.de/jg06/czfrineu.html>.
2. GÜNZEL, C.: Hort der Weisheit. In: Focus Online 18.02.2008, <http://www.focus.de/wissen/campus/tid-8643/nationalakademie_aid_234544.html>.
3. GERSMANN, G.: Open Access in den Geisteswissenschaften. In: ed. Deutsche UNESCO-Kommission, Open Access. Chancen und Herausforderungen – ein Handbuch, pp. 78-79 (2007), <http://www.unesco.de/fileadmin/medien/Dokumente/Kommunikation/Handbuch_Open_Access.pdf>.
4. SCHULTZ, T.: Gesellschaft der Gelehrten. In: sueddeutsche.de (14.07.2008)
5. <http://www.sueddeutsche.de/jobkarriere/535/448269/text>.
6. o. A. [die Redaktion], Der Aufstieg der Leopoldina, in: ZEIT online (14.07.2008), <http://www.zeit.de/online/2008/29/leopoldina-nationale-akademie>.

Links

1. Berlin Declaration, http://oa.mpg.de/openaccess-berlin/berlindeclaration.html
2. Berlin Declaration Signatories, http://oa.mpg.de/openaccess-berlin/signatories.html
3. Deutsches Textarchiv, http://www.deutsches-textarchiv.de/
4. Dublin Core Metadata Initiative (DCMI), http://dublincore.org/

[19] Last visit of all links: 2009-06-30.

5. Gemeinsame Wissenschaftskonferenz (GWK),
http://www.gwk-bonn.de/fileadmin/Pressemitteilungen/pm2008-09.pdf
6. Leopoldina,
http://www.leopoldina-halle.de/cms/en/academy/recommendations-and-statements.html
7. Metadata Encoding and Transmission Standard (METS),
http://www.loc.gov/standards/mets/
8. Metadata Object Description Schema (MODS),
http://www.loc.gov/standards/mods/
9. Text Encoding Initiative (TEI),
http://www.tei-c.org/
10. Telota initiative,
http://www.telota.de/
11. Telota Archiv-Editor,
http://www.telota.de/telota-lab/personendb/
12. Union der deutschen Akademien der Wissenschaften,
http://www.akademienunion.de/union/english.html.
13. Union der deutschen Akademien der Wissenschaften, Geschichte,
http://www.akademienunion.de/akademiegeschichte/english.html
14. Union der deutschen Akademien der Wissenschaften, BBAW,
http://www.akademienunion.de/bbaw/print.html

Last visit of all links: June 30th, 2009.

Peer Reviewed Data Publication in Earth System Sciences: Earth-System-Science-Data[1]

A Data Publishing Journal

Sünje Dallmeier-Tiessen and Hans Pfeiffenberger

Sünje Dallmeier-Tiessen
c/o Alfred-Wegener-Institut für Polar- und Meeresforschung,
Helmholtz Gemeinschaft, Am Handelshafen 12,
27570 Bremerhaven, Germany
suenje.dallmeier-tiessen@awi.de

Abstract

A new concept in data publishing in Earth System Sciences facilitates data reuse by employing the traditional means of scientific publishing. We created a new type of journal – an Open Access data journal – which offers quality assurance on datasets by way of peer review and better and more comprehensive data documentation. Moreover a data journal provides countable incentive – namely countable publications – for the authors to make data available for reuse.

1 Introduction

Today, research data and published papers are both an essential basis of research. Researchers, publishers, librarians etc all use (different parts of) a common way to publish the research results: This happens mostly via peer reviewed paper publications, which in the last decades were mostly published online as well or only. The process of peer review is an established quality control mechanism in most

[1] http://www.earth-system-science-data.net/, Open Access journal by Copernicus Publications.

disciplines. The resulting papers (or rather, the number of papers) are used to evaluate research in most research organizations.

However, when it comes to research data, things are still quite different: As of yet, there is no general consensus on data publishing.

For many researchers, the fear of plagiarism or “free riding” is the most important concern: With no consensus on data publishing in general, there is also no common understanding on data citation which can result in perceived (data) plagiarism. DOIs[2] have been introduced successfully as a tool in data citation [1]. However, providing a DOI is not (yet) equivalent to a scholarly reference.

The fear of misuse of a dataset is not the only motivation for researchers to not publish the data they consider their own. They do not see immediate incentives of data publishing in relation to the extra work they have – this is mostly due to the fact that research evaluation is so far only based on published text providing the interpretation of data. Since no common workflow in data publishing exists so far, it has been very difficult to assess published data and use it for research evaluation.

Quality control mechanisms in publishing data do not exist in many disciplines. The missing general consensus on data publishing is also due to the wide range of opportunities that exist: one can choose between the researchers’ website with an excel spreadsheet or publish data in tables as a supplement to the peer reviewed article or put data into a disciplinary datacentre – just to name few options out of many.

As diverse as these opportunities are the methods of quality assessment: Who checks the quality of a dataset put on a website? Who checks the data submitted along with the 30 page article as a supplement?

In this short paper we show a new way in data publishing that offers and links incentives, quality control and citability by combining traditional text publication with making data available through reliable repositories. Current data publication habits will be compared to the innovative data journal publication workflow. Finally, first experiences

[2] DOI: Digital Object Identifier, persistent identifier: http://www.doi.org/.

with such a data journal in the field of Earth System Science will be shown.

2 Current data publishing habits

Today there are many different options to publish data - they vary depending on the discipline and the specific characteristics of research data. Researchers initially have to decide whether they want to publish their data (and spend some extra time on it) or not as there are only few mandates or policies that tell them to do so and how to do it. Secondly, they need to decide how and when to publish their data – will they use a community platform or wiki and publish their data only within the community or do they want it to be publicly available on the internet? The question of the timing of the data publication is crucial and is related to the fear of data misuse which might interfere with one's own traditional paper publication. These are important decisions to make in view of their careers.

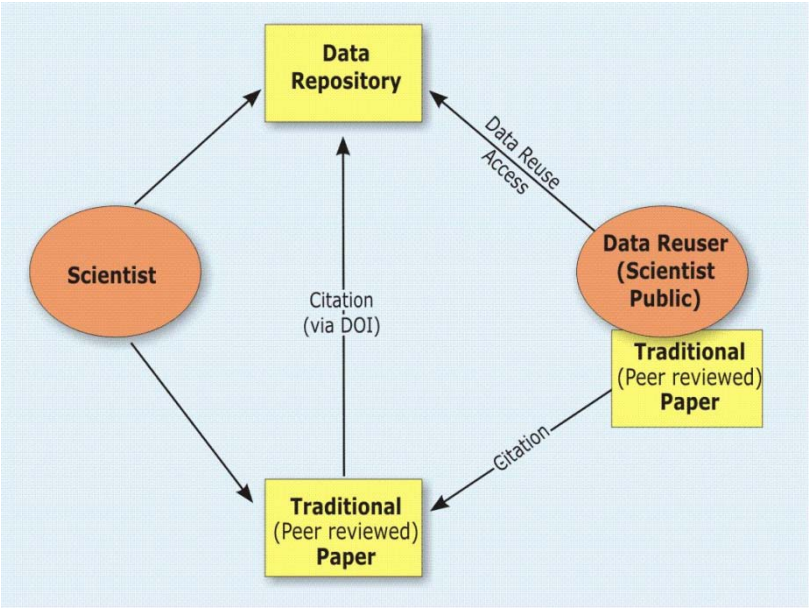

Fig. 1. Exemplary workflow in data publishing and data citation today.

A possible scenario in Earth System Science: The author may choose a data repository to publish their data while or before the text publication interpreting the data has been accepted. By using data repositories such as Pangaea[3] the authors get a DOI (a persistent identifier) pointing to the dataset which can be included in the article, e. g. in the methods and materials section (or in a supplement). The reviewer – if (s)he had the time to do so – could check the data quality. The author receives no immediate incentive for making the data available, but by using a DOI the dataset is citable and permanently linked to the paper. A reader is now able to reuse the data as there is a link to it from the paper (Figure 1). Unfortunately, in most cases (s)he cannot be sure if the dataset available and its documentation have been checked for good quality. Moreover, in most cases a full documentation is missing, thus (s)he has to rely on the mostly brief methods and materials section in the text publication.

3 Data publishing workflow with a data journal (e.g. Earth-System-Science-Data, ESSD)

By publishing research data with a data publishing journal an easily recognized workflow can be established that provides incentives for data creators and quality assurance for data users. The data journal itself does not hold the data but refers to a reliable data repository via persistent identifiers (such as DOIs). The papers published describe the dataset and the methodology in sufficient detail within the text or referring to peer reviewed literature. ESSD aims at a fully peer reviewed dataset and data documentation based on the classical journal characteristics and environment. Scientists do not have to familiarize themselves with a new publishing format and thus will easily see their advantages of such a workflow.

A possible workflow scenario in Earth System Science could be as follows (see Figure 2):

3 Pangaea – Publishing Network for Geoscientific and Environmental Data, www.pangaea.de. An award-winning data repository that uses DOI for data identification, publishing and citation.

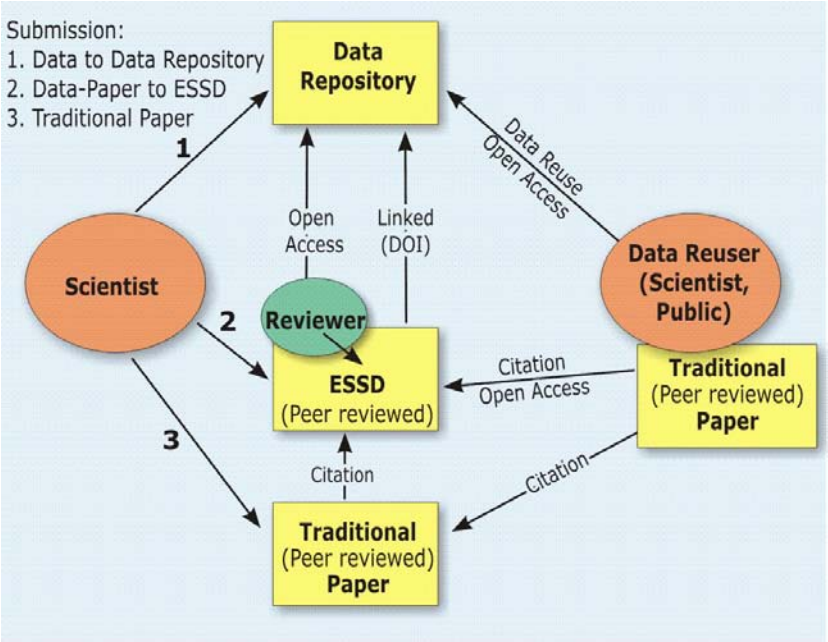

Fig. 2. Data publishing workflow with a data journal like Earth System Science Data (ESSD). The journal ESSD and the dataset in the repository are published Open Access. The original author (left) receives 2 publications (indicated by 2&3) and citations for publishing the dataset and the interpretation (right).

Data is sent to a data repository such as Pangaea, the author receives a DOI for the dataset.

Along with publishing the data in the repository a first data documentation article is submitted to a data journal such as Earth System Science Data. Peers will review the dataset in the data repository and will assess the data documentation in the article.

An article with an interpretation of the dataset will be published in another traditional peer reviewed journal referring to the article in the data journal (rather than pointing to the repository via a DOI).

Reuse (Figure 2, right) is now easily facilitated as the dataset is accessible via the data repository and the articles (both are Open Access). Moreover, data users can be assured that the dataset is of

good quality as it passed the peer review process. Future reusers will discuss the outcomes in their own articles which refer to the article in the data journal (for reusing the dataset) and the traditional article (for the interpretation). Thus, the original data publisher will not only publish two peer reviewed papers, but is also likely to receive a considerable number of citations.

4 First experiences with the data publishing journal Earth System Science Data

Serving the Earth System Science community the data repository Pangaea was established as early as 1992. Within Pangaea, DOIs have been introduced to facilitate data citation. The journal Earth System Science Data, launched in cooperation with Copernicus Publications in 2008, employed both to publish its first peer reviewed dataset. In the meantime more cooperations with further "trustworthy" data repositories have been established with more to come. The assessment regarding trustworthiness is based on state of the art data repository certification (e. g. including the usage of persistent identifiers).

The first article appeared in discussion form in autumn 2008 and was published in a final revised version in late 2008 with a dataset from the Antarctic. The article starts with a short section called "Data coverage and parameter measured" in which some key facts for the dataset are given, for example the repository reference (Figure 3).

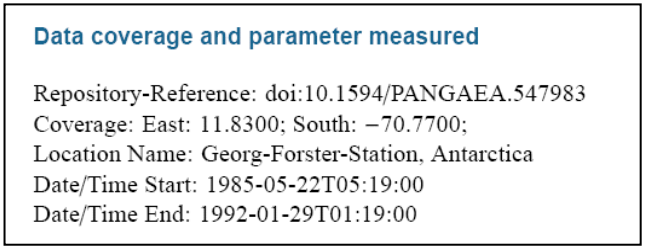

Fig. 3. Using persistent identifiers for the link to the data repository. Snapshot of the first paper in ESSD.

In the introduction of the paper the author describes the “history” of the dataset. Such a dataset history includes for example a short description of the campaigns or projects in which the dataset has been produced and the aims behind these undertakings.

The second part of the paper is about the instrumentation. Here, the author describes the methodology and instruments used in sufficient details. Methods etc. that are state of the art should reference other peer reviewed or acceptable grey literature (if persistently available online).

The third part explains the structure and provenance of the dataset. Moreover, additional and related datasets are mentioned. This is intended to enable reviewers as well as users to compare and correlate the new with existing datasets and put it into perspective.

The fourth part of the paper deals with the data access. Here, access to the dataset is described again in sufficient details. If necessary, the source of required software can be described here as well.

The paper has to pass a thorough peer review process, namely the open peer review process as described by Poeschl [2]. For the journal ESSD the review criteria had to be adapted to the needs of data publishing: The reviewer is expected to assess the article and the dataset. The three key criteria are: originality, significance and data quality. For example, (s)he has to evaluate if the data or the methods are new and thus if the datasets are of additional value for the scientific community. It is the reviewers’ task to decide if the data is unique, useful and complete and if the quality of the data and the article is at least state of the art.

5 Conclusions – Significance of Open Access

A new innovative way for publishing research data has been presented which offers considerable incentives for the authors (publications and citations) and facilitates future reuse of data through quality assessment. An exemplary workflow from the field of Earth System Science has been outlined. It shows that many stakeholders and roles are involved that are familiar from the traditional publishing process. Such a workflow and future reuse of research data in general requires

Open Access to the data and materials involved. The data journal ESSD itself is published Open Access; the datasets published are required to be Open Access by the aims and scopes of the journals. The open peer review system can function only if all materials are freely available to the reviewers and to the public.

6 References

1. KLUMP, J., BERTELMANN, R., BRASE, J., DIEPENBROEK, M., GROBE, H., HÖCK, H., LAUTENSCHLAGER, M., SCHINDLER, U., SENS, I., WÄCHTER, J.: Data publication in the Open Access initiative. Data Science Journal 5, 79-83 (2006) http://dx.doi.org/10.2481/dsj.5.79
2. PÖSCHL, U., KOOP, Th.: Interactive Open Access publishing and collaborative peer review for improved scientific communication and quality assurance. Information, Services and Use 28, 105-107 (2008)

Open Document Exchange Formats

Security, Protection & Experiences

Christian Zier

Federal Office for Information Security
christian.zier@bsi.bund.de

Abstract

This article is about open document exchange formats (ODEF), their properties and why there is a need for such formats. In the introduction I compare the way we store data nowadays using computers with the methods used before electronic data processing was available. Afterwards the properties of standards and open standards are briefly discussed, before the advantages of open document exchange formats in general and details of the formats ODF and OOXML in particular are presented. Of special interest is the aspect of security and protection against malicious software. Finally some experiences made by the Federal Office for Information Security (BSI) when migrating from Windows to Linux are presented.

1 Introduction

To guarantee Open Access to papers and documents they not only have to be available, but also to be stored in an open document exchange format. To display the data in a human readable form we depend on applications like editors and word-processing software. In order not to become dependent on the software producer of such applications, we should take care that we can access and display our data also using other independent applications.

Just for illustration I want to compare this with the way data were stored in the past. For example, how did the poet Heinrich Heine store

the poem “Loreley” (1824), which was set to music by Friedrich Silcher in 1837? Both used the open standards of German handwriting and music notes to write down the poem and the composition, respectively. Everybody who had a copy could access the information contained in these documents. But how are documents stored today? Usually in binary (machine readable) and proprietary formats. This is as if Heine and Silcher would have given the poem and music to a translator who had noted them down in a different language, which only the translator himself was able to understand. Even if the authors would have liked to read their own documents, they would have had to ask the translator to translate them back into their own language and present the translations to the authors. So they would depend on this translator, having lost access to and control over their own documents. Now, standards are helpful in the way that at least various translators can translate the documents and communicate with each other, i.e. they can work in an interoperable way.

2 Standards

A technical standard is an established norm or requirement. It is usually a formal document that establishes uniform engineering or technical criteria, methods, processes and practices and thus provides a basement for various interoperable implementations of the same standard. A concise definition of standards is given by the British Standards Institute: “A standard is a publicly available technical document, developed in cooperation with interested parties, which is based on scientific results and technical experiences. Its intention is to improve the public welfare.”

In contrast to *de jure* standards, which correspond to the guideline “a standard defines a product”, de-facto standards also exist. These standards (formal or informal) have achieved a dominant position, by tradition, enforcement, or market dominance. They have not necessarily received formal approval by way of a standardization process, and may not be an official standard document. Such standards can rather be described in terms of the guideline “the product defines the standard”.

Standards allow various subsystems to communicate via standardized interfaces such that complex systems can be built efficiently. They provide the basis for interoperable products which implement the standard and thus promote competition between these products.

Therefore, multiple competing standards for the same purpose actually question the meaning of standards. While reducing the competition between implementations and applications, they increase the expenses for software producers to implement multiple standards for the same purpose in one application. For example, if a manufacturer wants to support in an office application the format ODF besides OOXML and Microsoft Office binary formats, all standards have to be implemented, increasing the costs and size of the application. Moreover, this application is likely to be more error-prone due to its enhanced complexity and also vulnerable to exploits of all the implemented standards. Generally, not all standards for the same purpose will be implemented in one product and therefore multiple standards reduce interoperability. Competing standards are especially disadvantageous for customers, as might be known for instance in the context of entertainment electronics. Customers who bought a VCR playing beta-tapes could not play VHS-tapes and vice versa. Just recently this has been repeated in a very similar way with the different standards HD-DVD and Blue-ray for the successor of DVDs. However, sometimes competing standards for the same purpose can even cause serious problems. For example, the Mars Climate Orbiter mission with a budget of about 125 million dollars failed in 1999 because different involved research teams used different units in their calculations (SI-units and American units). This emphasizes that competition should take place between various implementations of the same standards, but not between the standards themselves.

3 Open Standards, the basement for ODEF

There is no clear difference between a standard and an open standard. Open standards stress the availability and accessibility of the standard and want to make sure that there are no obstacles in their implementation. They are independent of certain implementations and

manufacturers and facilitate the access to their specifications. Hence, more manufacturers are enabled to implement the standard and the competition is stimulated between the implementations rather than the standards, resulting in less market segmentation. The customer profits in terms of an increased interoperability between different implementations and less dependency on certain products. Open standards also help to avoid vendor lock-ins, which can become very costly for consumers, as for example in case of printers and cartridges. The development of Free/Libre Open Source Software (FLOSS) also becomes easier using open standards and ensures sustainable access to archived data: if necessary a new application can be developed, implementing the same standard, to access the data if the old software is not available or usable on current operating systems any more. Moreover implementations are extensible to additional features, like ODF, which is XML based. Using open standards for word-processing software allows authors to access and control their own documents.

An Open Standard is publicly available and has various rights to use associated with it. However, many different definitions of open standards exist, some of which (used in academics, by the EU, governments of Denmark, France and Spain) preclude requiring fees for use, while others permit patent holders to impose “reasonable and non-discriminatory” royalty fees and other licensing terms on implementers and/or users of the standard. For example, the definition of an open standard given by the European Union is:

- The standard is adopted and will be maintained by a not-for-profit organization, and its ongoing development occurs on the basis of an open decision-making procedure available to all interested parties (consensus or majority decision etc.).
- The standard has been published and the standard specification document is available either freely or at a nominal charge. It must be permissible for everybody to copy, distribute and use it without a fee or at a nominal fee.
- The intellectual property – i. e. if patents are present – of (parts of) the standard is made irrevocably available on a royalty-free basis.

- There are no constraints on the re-use of the standard.

4 Open document exchange format

Open document exchange formats are based on open standards and therefore provide the same advantages which are offered by such standards. These formats are developed in an open process and are independent. Their documentation has to be sufficiently detailed in order to allow any interested party to implement these formats on the basis of their specification.

The user of open document exchange formats benefits from a larger selection of implementations by different software producers he or she can choose from because of an enhanced software diversity and competition. The resulting interoperability between the different products allows the user to exchange and share documents with others who chose another implementation of the open document exchange format. In addition the authors of text stored in this format retain free access to their documents and retain control over them. For instance, documents stored in the Open Document Format (ODF) can also be read with a regular text editor, although the formatting of the text is lost in this way.

Another advantage of open document exchange formats is their adaptability and the possibility for extensions in order to meet certain requirements and to include additional features. Also important is a future proof archive security. Suppose the producer of your word-processing software does not support it anymore with the most recent operating system or the company does not even exist anymore. The files stored in an open document exchange format can still be accessed with other implementations, which also could be adjusted to certain requirements, if necessary.

Furthermore, such formats enhance security because they strongly facilitate automated scanning of the documents and filtering out certain parts while keeping the document readable (see next section). Hence, the properties of open document exchange formats also meet the requirements of E-Government for the internal exchange of documents between offices as well as external exchange between offices,

citizens and companies. Thus, in the process of adopting these formats it is actually not a question of whether this will happen, but rather of how it will happen.

4.1 Security & Protection

Infection of computers with malware is a big problem. Recent evaluations show that a major part of attacks on IT-systems is carried out via manipulated office documents which include binary code that is not displayed. This code is executed in the background without the user knowing about it. For these attacks almost exclusively binary document formats by Microsoft are used because of the widespread distribution of the corresponding applications and the known exploits they offer. Well organized groups with good technical knowledge carry out targeted attacks, stressing the need to be able to fend off such attacks. To make this possible, it is essential that the format allows scanning the documents in order to detect and isolate malicious code. In case of a critical vulnerability, the only alternative to protect one's own system might actually be to completely block binary/proprietary documents.

Both open document exchange formats, ODF and OOXML, provide much better possibilities than the widely used proprietary formats to analyse the documents and prevent attacks of malicious software. The structure of these documents, based on XML document formats, enable a complete and transparent analysis. Thus, the detection of malicious software is strongly improved while at the same time possibilities to hide malware somewhere in the document are reduced considerably. Meta data containing information, e.g. about the author and changes made in the document, are immediately visible. Because of the much clearer structure a potentially dangerous code that is embedded in the document, e.g. macros, pictures or videos, can be isolated much more efficiently from the actual document. If in doubt, the suspicious content can be filtered out without necessarily losing the information of the entire document.

4.2 Open Document Format (ISO 26300)

The Open Document Format (ODF) was originally developed by Sun and specified as standard by OASIS (2005). Later, in 2006, it was published as ISO standard 26300. The members of the OASIS consortium decide how and what work is undertaken through an open, democratic process. Numerous organizations, which can be small end users, local government agencies, trade groups and universities, contribute to the standards development efforts that directly affect their work. Foundational sponsors are IBM, Microsoft, Oracle, Primeton and Sun.

As can be expected from an open document exchange format, many different and independent implementations of ODF exist such as OpenOffice/StarOffice, NeoOffice, KOffice, AbiWord and more. Implementations are also available for many platforms, e.g. Linux, Microsoft Windows and Mac OS X. Because of the various platform-independent implementations, this format also fulfills the interoperability requirements of the E-Government, so that documents can be exchanged with internal and external partners. Through the past years ODF has been analysed and extensively tested with the result that it also meets the security requirements for E-Government, like the ability that the format can be scrutinised.

An ODF-file can be a single file or more files collected in a compressed archive. It allows to directly access and edit the XML-files and ensures future-proof access to archived data. However, it has been criticized that the standard does not define how to store formulae in tables. Therefore, calculations in tables might not yield the same result in different implementations. This will be taken care of in version 1.2, where a uniform standard will be defined for this purpose.

4.3 Office Open XML (ISO 29500)

The Office Open XML format (OOXML) was developed by Microsoft and approved as standard by Ecma international in 2006. Later, in 2008, OOXML became ISO standard 29500 in a somewhat controversial process. The ISO has officially published the standard on November 19th, 2008.

Until now no product exists, including Microsoft Office, that implements the standard ISO 29500. Actually even no reference implementation existed when OOXML was approved as ISO standard. It is still not clear, when ISO 29500 will be implemented in Microsoft Office or other word-processing software, so that implementations of this standard could not yet have been analysed and tested for their security.

Nonetheless, the specification of the standard alone suggests that it will probably be more elaborate and costly to perform security scans and, if necessary, to isolate malicious code, in comparison to ODF. For instance, the definition of the same properties, like the representation of text color and alignment, makes use of different XML-tags in text documents, presentations and excel documents. Consequently, the different types have to be analysed separately and knowledge about one type cannot necessarily be applied to another. Also, the extensive size of the specification of the standard, more than six times as much as that for ODF, indicates a larger complexity and therefore a more complicated security analysis of the standard itself as well as of the corresponding documents. Furthermore, the execution of macros can be crucial for the security. While in ODF (ISO 26300) the handling of macros is defined via tags, there is no such specification in OOXML. Apart from a probably reduced interoperability between different implementations, this also complicates the detection of potentially malicious software and consequently its automated isolation.

Another point that should be considered is that a more complex standard will be more difficult to implement in a product. Therefore, it can be expected that there will be only very few independent implementations of this standard. The fact that the main developer of this standard is also the only implementor is not very desirable and could imply that the implementation will not be platform independent.

5 Migration at the BSI

Within the past few years the BSI has migrated 50% of desktop systems from Windows to Linux. Therefore, interoperability in the heterogeneous system is a basic requirement. This has been achieved by replacing the groupware Exchange with the Kolab server, using Kon-

tact as Linux client and Outlook as Windows client. In May 2008, StarOffice has been installed on all desktop systems, including Windows boxes, implying a migration to ODF. Only a few terminal servers still support Microsoft Office to maintain interoperability with external documents when needed. The default document formats used for the exchange of documents are ODF and PDF.

Not surprisingly it turned out that the more recent the software is the less troublesome is the migration: Over the past years since the migration started, software has become much more mature and hardware support has improved significantly. The packaging of the software and its rollout is much more comfortable with Linux. Bugs can be identified more easily and fixed faster while the distribution of the patches is strongly facilitated. In the beginning there were some problems with printers when Debian Woody was still in use. Some printers were not detected, and sending of files in PDF-format was problematic. However, these problems could be solved, also by using more recent software.

Before migrating to ODF and StarOffice, existing templates and documents have been analysed for parts which might be critical or elaborative to convert, like macros. Only few such documents have been found. Eventually the most frequently used templates were converted, for instance a much used document generator. The migration was supported by a training for StarOffice as well as by a person that could be contacted to help solving problems with StarOffice.

In summary, it can be said that most people accept only a few drawbacks and that the everyday workflow has to function at a level of at least 90%. Because many colleagues extensively use templates, the administration of the templates was quite important. It turned out that the similarity between StarOffice and Microsoft Office was very helpful. The success of migration also depends on the motivation of the users to engage in new software. Those with previous knowledge about Linux were more motivated to migrate to Linux, while Windows-only users with long-time experiences with Windows tend to avoid to get used to new applications.

Just recently the migration of the desktop systems has been increased to almost 100%. In conclusion it can be said that the migration can generally be considered as unproblematic and successful so that the question is actually not if, but how to migrate!