Deutsche Initiative für NetzwerkInformation e.V.

WORKING GROUP ELECTRONIC PUBLISHING

DINI; c/o Staats- und Universitätsbibliothek Göttngen, Papendiek 14, 37073 Göttngen

Institutional Repositories and Enhanced and Alternative Metrics of Publication Impact

Report of an International Workshop held at Humboldt University Berlin, 20–21 February 2006

edoc.hu-berlin.de/series/dini-schriften/2006-8/PDF/8.pdf

urn:nbn:de:kobv:11-10063369

Frank Scholze (Stuttgart University Library) frank.scholze@ub.uni-stuttgart.de

Susanne Dobratz (Humboldt-University Berlin) dobratz@cms.hu-berlin.de

Introduction

Open Access is one of the most popular terms in the library and information science community. The definition of an open access publication has been largely agreed upon: The author or copyright holder grants to all users a free, irrevocable, world-wide, perpetual right of access to, and a licence to copy, use, distribute, perform and display the work publicly and to make and distribute derivative works in any digital medium for any reasonable purpose, subject to proper attribution of authorship, as well as the right to make small numbers of printed copies for their personal use. A complete version of the work and all supplemental materials, including a copy of the permission as stated above, in a suitable standard electronic format is deposited in at least one online repository that is supported by an academic institution, scholarly society, government agency, or other well-established organisation that seeks to enable open access, unrestricted distribution, interoperability, and long-term archiving.[1]

Widely and controversially discussed by experts is the question, which would be the most successful way to bring scientific authors to provide open access to their publications[2] [3].

An important factor in order to accomplish the open access idea within universities and research institutions is to enhance the visibility and usage of Institutional Repositories – IRs (so called green road to open access) and Open Access Journals (golden road to open access). Usage is employed in the passive sense of reading as well as in the active sense of citing a publication.

Workshop Background

DINI (the German Initiative for Networked Information) and DFG (German Research Foundation) – with support from OSI, SPARC Europe, SURF and JISC – organised a workshop to discuss enhanced and alternative metrics of publication impact given the fact that an increasing number of scientific publications is available open access in Institutional Repositories.[4] The idea for such a workshop originated during an informal meeting along the CERN workshop on Innovations in Scholarly Communication (OAI4) in Geneva[5] where representatives of the organisations mentioned before came together. It was decided to focus on three different aspects

- Alternative metrics of impact based on usage data (the LANL approach)[6]

- Interoperable and standardised usage statistics (Interoperable Repository Statistics and COUNTER)[7]
- Open Access citation information[8]

42 experts from 8 countries followed the invitation to this workshop. It was understood that the aim of this workshop was a pragmatic one. It should not serve as a forum for the development or redefinition of the concept of impact or scientific visibility. Starting from the well-known limitations of the current journal impact factors[9] a number of existing approaches was discussed which consider a wider basis of data and other algorithms to process this data in order to produce quantitative metrics of scientific visibility. Metrics was understood as processing of (eventually aggregated) raw data – impact or status is derived from rankings based on those metrics. Collection of raw data, aggregation and processing should be transparent in order to ensure the acceptance of possible rankings.

The selection of approaches presented in the workshop did not claim completeness but was regarded as promising for the future development of institutional repositories and the cause of open access to the scientific literature. Consequently there was an introductory session where Lars Bjørnshauge (Lund University Library) and Norbert Lossau (Bielefeld University Library) gave their views of the next generation of institutional repositories as a network within other networked services.

Institutional Repositories: Standards, Interoperability, Need for Professionalization

Lars Bjørnshauge gave a concise overview of the past and present situation of institutional repositories (IR) before he sketched his views for the future. However new the topic, there is already a kind of history of IRs. They began as an additional channel for dissemination of publications and rose to attention of a greater audience with the debates about open access to scientific publications. In his opinion it is important to keep in mind the different purposes for which IRs are run in a University or research institution. Besides the dissemination of publications already mentioned IRs are often meant to record the scientific published output of an institution. They can (and increasingly will) play a role in research assessment[10] and last but not least serve as input for scientist's CVs.

Lars defined a number of key issues on which IR development has got to focus on its way from experiment to service. He pointed out, that because the development of institutional repositories has been very diverse, there is a low degree of uniformity in standards, definitions, formats and protocols. This is the reason why it is difficult to develop secondary services based on institutional repositories at the moment.

Those key issues or critical success factors for IRs are:

- the adoption of standards,
- the modelling and deployment of workflows,
- the long term availability of the electronic documents contained in an IR,
- the question of impact and visibility of electronic documents in IRs and
- filling the IRs with content.

Norbert Lossau (Bielefeld University Library) presented an overview of the European project DRIVER which will most probably run from June 2006 till August 2008. It intends to build a European IR infrastructure. As a supporting action a series of strategic studies will be performed that intend to bring more detailed information and analysis about the IR situation in

Europe than the CNI-JISC-SURF Amsterdam 2005 report[11] could which was the very first of its kind.

DRIVER emphasises that national programmes for IR development and standardisation should be set up. DRIVER will support and co-ordinate such national efforts as far as possible. DRIVER itself is intended to serve as a demonstrator and testbed for transnational access to existing standardised IRs. Its main aims are personalised access to virtual collections, quality of services, interoperability, extensibility and integration of local services. DRIVER should also serve as a basis for other services which are built on top.

Alternative metrics of impact based on usage data

Johan Bollen (Los Alamos National Laboratory) presented the work done at LANL and California State University with respect to alternative metrics of journal impact based on usage data. Usage data (the more precise term is access data as usage can not be defined precisely) is collected through link resolvers. In order to do that it is essential to create a digital library infrastructure that enables linking servers to record the biggest part of usage. IRs should be able to become part of that infrastructure i.e. they should at least be OpenURL enabled. Linking server logs are then serialised as OpenURL ContextObjects[12] and exposed by an OAI-PMH data provider. The repository retains full control of what is exposed and how, e.g. anonymization of user IDs.

A trusted third party (or federation thereof) can harvest and aggregate logs from a range of repositories. Next the aggregated logs are subjected to datamining techniques which derive item networks from access sequences recorded in the logs under the assumption that similar items are accessed by similar users. These networks can for example be used to construct recommender services useful to both local institutions as well as third-party aggregators.

As a last step, different metrics can be applied to the aggregated data. In addition to the well-known and popular frequentist metrics (e.g. used to calculate the journal impact factor) there is also the possibility to derive structural metrics of quality from the generated item networks leading to more complete, fine-grained and reliable evaluation of scholarly communication. Log data furthermore is free from publication delays and can be used to track immediately contemporary trends in science. At California State University nine major institutions participated in a field experiment from November 2003 to August 2005. Collected journal usage data was processed with a weighted PageRank algorithm resulting in a different list of journal ranking compared to ISI impact factors.[13]

Interoperable and standardised usage statistics

Tim Brody (University of Southampton) gave an overview of the project Interoperable Repository Statistics (IRS) which plans to investigate the requirements for UK and international stakeholders, design an API for gathering download data, build distribution and collection software for IRs and develop generic analysis and reporting tools. There has been a report on stakeholder requirements derived from interviews with domain experts.[14] The aim is not to have theoretically exact usage data but comparable usage data. However there is no draft or RFC at the moment which summarises IR requirements for comparable usage data.

Workshop participants strongly felt the need for supporting actions in this field incorporating COUNTER and Project IRS activities in order to develop criteria and working definitions that can be tested on a select number of IRs.

Sebastian Mundt (Hochschule der Medien Stuttgart) gave an update on COUNTER activities. Currently there are two so called Codes of Practice operable, one for journals and databases (Release 2)[15] and one for books and reference works (Release 1)[16] . Usage data is processed at the source (publisher site or IR) and accuracy of the usage reports is tested by an external auditor within a certain tolerance (-8% to +2%). It became obvious that open access sites like IRs have to cope with additional problems like automated web-crawler access in normalising their usage data. The question how to exclude spider and robot access, a problem especially occurring in Open Access Repositories, was raised and regarded an important field of action in future. A specific code of practice for IRs has not the highest priority within COUNTER at the moment. Workshop organisers and participants will actively engage in closer co-operation with COUNTER and IRS to ensure that these important issues will be solved and promoted (e.g. in the DINI Certificate Document and Publication Repositories.)[17]

Open Access citation information

Jeff Clovis (Thomson Scientific) gave an overview about the Web Citation Index (WCI). This new service started in November 2005 to cover material available in IRs. WCI is a commercial product to become fully available later in 2006. Currently 38 IRs are indexed. An additional 500 have been selected by Thomson Scientific content editors for inclusion already. OpenDOAR[18] and other IR listings have been taken as a starting point. The published version of an article is indexed in Web of Science, the preprint or postprint version accessible in an IR is indexed in WCI – with links between the two services.

Workshop participants agreed within the discussion that there is a need to standardise citation information within IRs, so they can become a solid basis for commercial services like WCI as well as open public services.

Ralf Schimmer (Max Planck Society) gave an update of open access activities within the Max Planck Society and of the co-operation with Thomson for WCI. The MPG eDoc-Server (MPG's institutional repository) is not indexed in WCI yet but this is planned for April 2006. eDoc contains about 10.000 documents at the moment of which about 3.400 are open access. President, Vice-President and Nobel Laureates of the MPG shall be persuaded to deliver their published work open access to create a “me too” effect within the MPG comparable to the “cream of science” effort in the Netherlands. Tests with the WCI showed that there are still issues concerning the linking between different versions of a document (preprint, postprint in IR, postprint on publisher site).

Tim Brody (University of Southampton) could assert the versioning problem when he reported about Open Access Citation Index Services. However extraction and parsing of reference data in IRs could serve at least two functions:

On the one hand it would allow controlling of references during the upload of new documents. This would enable authors to correct citations especially if reference data could be checked against existing services like CrossRef, Web of Knowledge, Google Scholar etc. The catchword for such a service could be “Click & Canonical“. Reference parsers are already deployed in IRs for rather homogeneous research communities like High Energy Physics (e.g. in CDSWare).[19] These solutions can not easily be applied to a broader spectrum of disciplines or a general solution for reference parsing.

On the other hand references in IRs could be displayed and exported as metadata. This also needs well structured references and a standardised form of exchange. This exchange could be well realised as an extension of the OAI-PMH using XML ContextObjects.[20]

For both scenarios the situation can be improved however by supplying author tools to support the structure of scientific writing and citing (like RefWorks[21], Endnote, the BibTeX format) and thus increasing the quality of input data.

Breakout sessions

Discussions in the breakout sessions on the second day of the workshop led to the conclusion that all three aspects covered by the workshop are relevant for the visibility of scientific publications.

1 - Metrics of impact based on usage data

Open URL techniques (eventually metadata based identification of documents if no persistant identifiers are available) were regarded as promising and the possibility to discriminate documents by subjects was demanded. So standardised link server logs, which could even be derived from apache web server logs (although this is a rather noisy approach) would provide the basic data. Participants agreed, that linking servers are crucial factor of DL concepts. This means creating a digital library infrastructure that enables linking servers to record the biggest part of usage. Therefore all digital library services (catalogue, online databases, document delivery, repositories etc.) have to be involved. Deployment of linking servers in Germany is incomplete in two ways. Quite a number of library systems do not use linking servers at all and some of those which use them do not connect all services exhaustively.

Linking server logs can be serialised as XML-ized OpenURL ContextObjects and exposed by an OAI-PMH repository. The repository retains full control of what is exposed and how, e.g. anonymization of user IDs. A trusted third party (or federation thereof) can harvest and aggregate logs from a range of repositories. The aggregation of usage data should be performed on different levels: local, global and community based. Data mining and analysis must become end user services. It was agreed that there should be demonstrators for data mining techniques and results.

*2 - Interoperable and standardised usage **statistics***

Within this breakout group it was discussed, what is needed from an infrastructural perspective of IRs to achieve global interoperable usage statistics and which services can be built on standardised usage statistics.

Participants agreed upon the fact that there is still a lack of standards. Even in projects like COUNTER or IRS there are currently no adequate standards yet that could easily be adapted to IRs. Furthermore the number of IRs complying to standards (like the DINI certificate) is too small.

This group came up with a problem description, indicating that the document space covered by current services is not appropriate. This means, that services do not include enough or the proper documents from most subject specific perspectives. Because there are different views on the information or subject sectors, different needs for investigating those document spaces appear. The participants agreed that there is a necessity to have 1) linking statistics, 2) citation statistics and 3) usage statistics and even to use combinations of different metrics. This would allow to have different analysis for documents e.g. when used for teaching or for specific research communities. There is no urgent need to be cited regarding for example a lecture book, (in other words to gain prestige), but rather that students read it (which means to gain popularity). In this example, it would be more appropriate to collect access statistics rather than citation statistics.[22]

The group pointed out that there is a need for agreed definitions and that the purpose of statistics for different stakeholders should be defined within a scope statement of the service offered.

Further research of the meaning of statistics was deemed necessary, especially regarding the mathematical and technical background, interpretation (e.g. correlation with other factors), relational statistics and context based statistics.

In order to put those ideas into practise, it was suggested to use existing organisations and to take an existing proposal (e.g. COUNTER) and check against the criteria that need to be developed and against a good selection of example repositories.

3 - Open Access citation information

The group's answer to the question what is needed from an infrastructural perspective of IRs to support citation indexing and analysis was interoperability. It was seen necessary to integrate open access and commercial citation data, to be able to receive the citation data of all resources, not only of a few.

IRs should ideally provide a possibility to import references of articles and an online-tool to add references to an article. The participants agreed that only added value can really convince users to deposit their documents in IRs. Such added value could be that the IR automatically checks for correct references and standardises them (auto-find DOIs etc.).

The group discussed, how reference parsing tools can be integrated. An online service was suggested, which parses for references and allows a parallel upload of BibTeX or Endnote references. In order to integrate reference resolving services, already existing services, like CrossRef, Web of Science or Google Scholar should be approached.

The group also discussed, how storage and communication of reference data can be standardised. Standardisation of a Reference Exchange Protocol, e.g. realised as an extension of the OAI-PMH using XML ContextObjects seemed to be reasonable and possible.

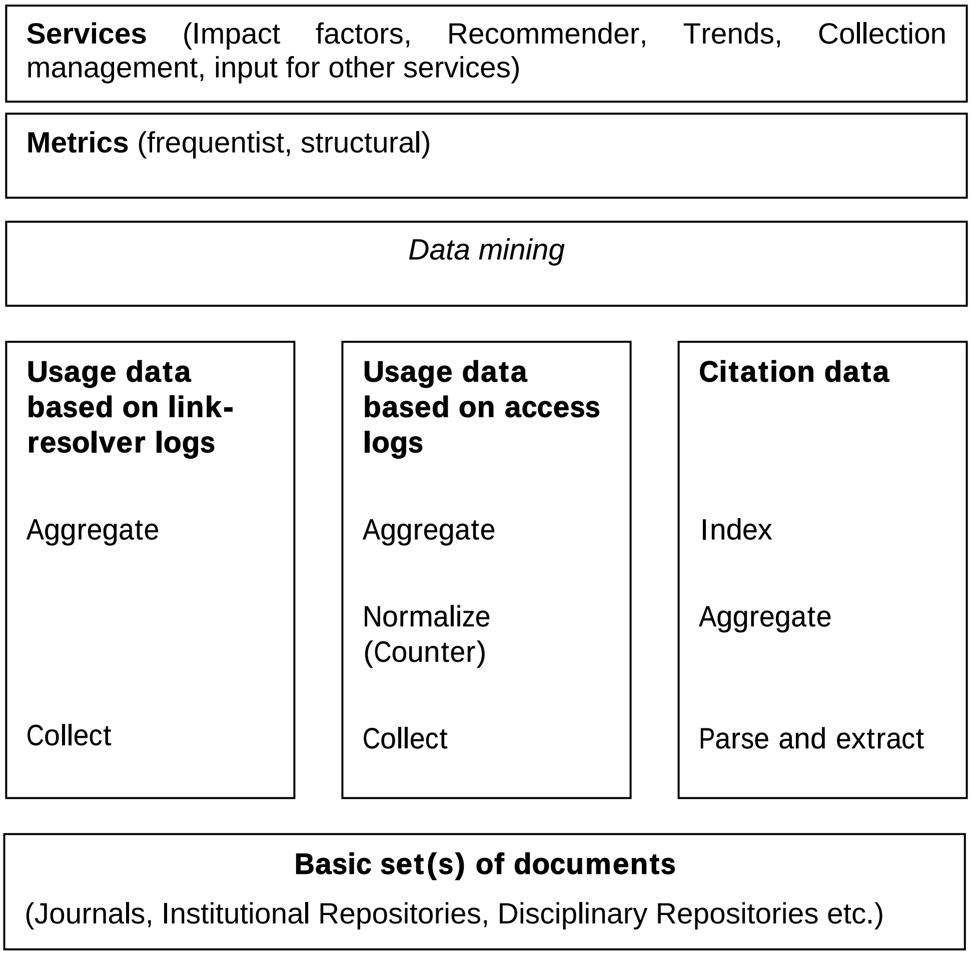
Summary and Conclusion

Breakout sessions reported their findings to the final plenary session. The need to fill the repositories was strongly highlighted throughout the discussion, in order to reach a critical mass. It was agreed that all three topics covered within the workshop pose different advantages and disadvantages and need different lines of action.

Usage data collection based on link resolver systems has been successfully performed in a huge field trial at CalState and LANL. It has to be tested under different basic conditions in Europe. The focus lies on the development of a suitable digital library infrastructure and on the question which organisation(s) should aggregate and process the data.

In the field of usage data based on access logs standardisation and collaborative efforts have to be intensified in order to come to comparable usage analyses for IRs and publisher sites. Workshop participants strongly felt the need for supporting actions in this field incorporating COUNTER and Project IRS activities in order to define criteria and working definitions that can be tested on a select number of IRs. Workshop organisers and participants will actively engage in closer co-operation with COUNTER and IRS to ensure that these important issues will be solved and promoted (e.g. in the DINI Certificate Document and Publication Repositories.)[23]

Workshop participants agreed that there is a need to standardise citation information within the IRs, so they can become a solid basis for commercial services like WCI as well as open public services. Efforts in citation analysis have to be focused on author tools, reference resolving and exchange. Results have to be complementary to already existing solutions and players in the field (like WCI, Google Scholar etc.). It is important to keep in mind however that there is a variety of options to calculate and rank the importance and visibility of scientific publications (Fig. 1). Collecting data, aggregating and processing it have become more separated compared to past times. One striking example is the different metrics applied to ISI citation data. Johan Bollen and his group compared the "classic" impact factor with a weighted PageRank algorithm and proposed a combination of both (the Y-factor).[24]



Services (Impact factors, Recommender, Trends, Collection management, input for other services)

Metrics (frequentist, structural)

Data mining

Usage data based on link-resolver logs

Aggregate

Collect

Usage data based on access logs

Aggregate

Normalize (Counter)

Collect

Citation data

Index

Aggregate

Parse and extract

Basic set(s) of documents

(Journals, Institutional Repositories, Disciplinary Repositories etc.)

Figure 1: Collecting and processing quantitative data of scientific visibility (schematic overview)

In order to follow all aspects relevant for the visibility of scientific publications workshop participants agreed to form three working groups taking on board additional experts. Discussions have to be continued on an international level, especially with the European Science Foundation[25] and the Knowledge Exchange Office[26]. Based on the results of the workshop these working groups will formulate requirements and implementation details in all three fields more thoroughly. Thus they will guarantee that this successful workshop will have sustainable consequences.

[1] IFLA definition http://www.ifla.org/V/cdoc/open-access04.html

[2] Guédon Jean-Claude. 2004. The "Green" and "Gold" Roads to Open Access: The Case for Mixing and Matching. Serials Review 30, no. 4 (1201): 315-328

[3] Harnad Stevan.2005. "Fast-Forward on the Green Road to Open Access: The Case Against Mixing Up Green and Gold", Ariadne no. 42, http://www.ariadne.ac.uk/issue42/harnad/

[4] http://www.dini.de/veranstaltung/workshop/oaimpact/

[5] http://www.cern.ch/oai4

[6] Bollen Johan, Sompel H van de, Smith J A, Luce R. 2005. "Toward alternative metrics of journal impact: A comparison of download and citation data" Information Processing & Management 41(6): 1419-1440

[7] http://irs.eprints.org/

[8] http://eprints.ecs.soton.ac.uk/11536/

[9] Dong P, Loh M, Mondry A. 2005. "The "impact factor" revisited" Biomedical Digital Libraries 2(7), doi:10.1186/1742-5581-2-7

[10] E.g. during the UK Research Assessment Exercise 2008, cf. http://irra.eprints.org/

[11] http://www.surf.nl/download/country-update2005.pdf

[12] NISO OpenURL standard Z39.88-2004 http://www.niso.org/standards/standard_detail.cfm?std_id=783

[13] http://www.cni.org/tfms/2005b.fall/abstracts/PB-bx-bollen.html

[14] http://irs.eprints.org/report/

[15] COUNTER: Counting Online Usage of Networked Electronic Ressources 2005b, *COUNTER code of Practice for Journals and Databases: Release 2,* http://www.projectcounter.org/r2/COUNTER_COP_Release_2.pdf

[16] COUNTER: Counting Online Usage of Networked Electronic Ressources 2005a, *COUNTER Code of Practice for Books and Reference Works: Release 1,* http://www.projectcounter.org/cop/books/cop_books_ref.pdf

[17] http://www.dini.de/documents/Zertifikat-en.pdf

[18] http://www.opendoar.org/

[19] Claivaz Jean-Blaise, Le Meur Jean-Yves, Robinson Nicholas. 2001. "From Fulltext Documents to Structured Citations: CERN's automated Solution", High Energy Physics Library Webzine, issue 5, http://doc.cern.ch/heplw/5/papers/2/

[20] Lagoze Carl, Sompel Herbert Van de, Nelson Michael, and Warner Simeon. 2002. The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0. http://www.openarchives.org/OAI/openarchivesprotocol.html, cf Herbert van de Sompels presentation at the Conference about current and future uses of the proposed OpenURL

Framework Standard Z39.88-2004, Washington DC, USA, 29 October 2003, http://www.ariadne.ac.uk/issue38/apps-rpt/

[21] http://www.refworks.com/

[22] Bollen et al. 2005 make the same point concerning popularity versus prestige regarding review journals which are widely read but less cited. Philip Ball „Prestige is factored into journal ratings 'Y-factor' measures quality as well as quantity of citations“ Nature 439, 770 - 771 (16 February 2006); doi:10.1038/439770a cf. Bollen Johan, Rodriguez Marko A., Sompel Herbert Van de "Journal Status" preprint http://www.arxiv.org/abs/cs.GL/0601030

[23] Deutsche Initiative für Netzwerkinformation (DINI) AG Elektronisches Publizieren (2003): *DINI*-Certificate Document and Publication Repositories. URL: urn:nbn:de:kobv:11-10046073 or http://www.dini.de/documents/Zertifikat-en.pdf
Deutsche Initiative für Netzwerkinformation (DINI) AG Elektronisches Publizieren 2006, *DINI-Certificate Document and Publication Repositories 2nd edition,* in Press

[24] Ball Philip. 2006. "Prestige is factored into journal ratings 'Y-factor' measures quality as well as quantity of citations" Nature 439, no. 7078: 770-771, doi:10.1038/439770a. The importance of this topic is stressed e.g. by the UKSG call for studies about usage based measures of journal quality http://www.uksg.org/rfp.pdf

[25] http://www.esf.org/

[26] Knowledge Exchange Office accessible via http://www.bs.dk/