Open Access Statistics:
An Examination how to generate Interoperable Usage Information from Distributed Open Access Services

Publishing and bibliometric indicators are of utmost relevance for scientists and research institutions. The impact or importance of a publication (or even of a scientist or an institution) is mostly regarded to be equivalent to a citation-based indicator, e.g. in form of the Journal Impact Factor (JIF) or the Hirsch-Index (h-index). Both on an individual and an institutional level performance measurement depends strongly on these impact scores. The most common methods to assess the impact of scientific publications show several deficiencies, for instance:

- The scope of the databases that are used to calculate citation-based metrics (Web of Science WoS respectively the Journal Citation Reports JCR and Scopus) is restricted and more or less arbitrarily defined.
- The JIF and h-index are showing several disciplinary biases (exclusion of many document types, the two years timeframe of the JIF, etc.).
- Both JIF and h-index are privileging documents in English language.

Even though in principle citation-based metrics provide some arguments pro open access[1], they mostly disadvantage open access publications – and by that reduce the attractiveness of open access for scientists. Especially documents that are self-archived on open access repositories (and not published in an open access journal) are excluded from the relevant databases that are typically used to calculate JIF-scores or the h-index.

Open access journals on the other hand may have a JIF-score and indeed some of them even have an impressive Impact Factor. Nevertheless they are often discriminated by the JIF-formula and the scope of the JCR:

- Since many open access journals are quite new, they are lacking the citation history a journal needs to be indexed by the JCR and to reach an attractive JIF-score.
- Open access journals are published above average in developing countries. Due to its unbalance towards the English language these journals usually attain minor JIF-scores – if they are indexed at all by the JCR.

Assuming that the motivation to use open access publishing services (e.g. a journal or a repository) would increase if these services would convey some sort of reputation or impact to the scientists, alternative models of impact measurement are discussed. Prevailing research results indicate that alternative metrics based on usage information of electronic documents are suitable to complement or to relativize citation-based indicators. Following Bollen et al. (2005, 2009) not only the bare frequency of citation or usage may be meaningful, but also the structure of citation networks or usage networks of scientific documents.

To test, evaluate and produce such alternative indicators based on document usage it needs a sophisticated infrastructure to generate and exchange interoperable usage data within a network of several different servers, especially if the data shall contain information on the context of document usage. For example this includes the logging of usage events on open access repositories that are indexed by legions of robots and that contain multi-file documents and duplicate documents (maybe

[1] Scientific documents that can be used free of charge are significantly more often downloaded and cited than Toll Access documents are (Harnad & Brody, 2004; Lawrence, 2001). Moreover the frequency of downloads seems to correlate with the citation counts of scientific documents (Brody, Harnad & Carr, 2006)

in different file formats) as well as a an data structure that makes clickstream analysis possible. An infrastructure like that faces all the problems known from weblog analysis in digital libraries as reported for instance by Jamali, Nicholas & Huntington (2005).

The project *Open Access Statistics* (OAS)[2] tried to create an infrastructure that meets the requirements mentioned. OAS is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)[3] and was initiated by the Electronic Publishing working group of DINI (Deutsche Initiative für Netzwerkinformation/German Initiative for Network Information)[4]. Four project partners[5] collaborate together to pursue the following aims:

1. Develop a common standard to exchange usage data between different services.
2. Implementation of an infrastructure to collect, process and exchange usage information between different services.

OAS collaborates tightly with two associated projects. While Open Access Statistics addresses usage description, *Open Access Citation* (or Distributed Open Access Reference Citation Services – *DOARC*[6]) aims at tracking citations between electronic publications. *Open Access Network*[7] intends to build a network of repositories and it will also bundle the results of OAS and DOARC in one user interface[8]. It also offers services for DOARC and OAS, e.g. the deduplication of documents which is based on asymmetric similarities of full text documents.

OAS has implemented an infrastructure to collect and exchange usage information between different services (e.g. open access repositories, licence servers, linkresolvers) and to process this information according to the standards of COUNTER, LogEc and IFABC. This allows comparing hits from different services.
The OAS infrastructure is two-tier. Firstly, the data providers generate logs about document usage and pseudonymize user information (e.g. IP addresses). In the following step they process usage information (add a unique document ID, transforms data into OpenURL ContextObjects etc.) and finally offer the information via OAI-PMH. Secondly, the central service provider collects the usage events from each single data provider and processes this data. It deduplicates documents (e.g. it sums up the hits on files with the same content on different servers) and also deduplicates users, so it is possible to create download graphs or to conduct clickstream analysis. It also processes the data according to the three standards mentioned beforehand (including the removal of non-human access and considering standard-specific parameters like double-click spans). After the calculation the usage data will be retransferred to the distributed services (the data providers) and to the Open Access Network service.

Additionally, OAS outlined further services for repositories based on usage information and developed implementation guidelines which make it easy for other services to join the OAS infrastructure.
The usage data produced by OAS may be used

- from an *user perspective* as a criterion to estimate the relevance of a document (e.g. rankings)
- from an *author perspective* as an indicator for the dissemination of a concept
- from a *repository perspective*:

[2] http://www.dini.de/projekte/oa-statistik/english
[3] http://www.dfg.de
[4] http://www.dini.de
[5] Georg-August-Universität Göttingen (Goettingen State- and University Library), Humboldt-Universität zu Berlin (Computer and Media Service), Saarland University (Saarland University and State Library), and the Universität Stuttgart (University Library)
[6] http://doarc.projects.isn-oldenburg.de/
[7] http://www.dini.de/projekte/oa-netzwerk/
[8] http://oansuche.open-access.net/findnbrowse

- as additional metadata for search engines, databases etc.
- as a recommender service
- as additional metadata for users

Data providers have to fulfil rather light-weight requirements to take part in the OAS infrastructure. Their web servers have to use a defined but easy to handle configuration, they must pseudonymize user information and isolate the local document identifier and as a last step they have to offer the information as OpenURL ContextObjects containers (with the elements *referent, referring entity, requester, service type, resolver* and *referrer*) via an OAI-PMH-interface to the service provider or aggregator service. DSpace or OPUS repositories may even use modules developed by OAS, other products can easily be configured to be OAS-ready[9].

Some lessons OAS learned by now are on the one hand that linkresolver logs are hard to integrate in the framework. Some services (OVID) do not offer suitable information while the information form other services (SFX) seem very heterogeneous. On the other hand the deduplication of documents appears very difficult for several reasons. For instance a document may have more than one ID or even more than one persistent identifier due to multiple deposits on different repositories. Visa versa, two documents with exactly the same content may use different sorts of persistent identifiers. The formal publication in a journal may have a persistent identifier in form of a DOI, while the postprint in a repository has a persistent identifier in form of an URN. Another problem is that a given document may have several splash pages on different servers pointing to one single file on one server due to metadata harvesting.

By now OAS strives for a second funding phase to solve the beforehand mentioned challenges and pursue new goals for a second period. Some of the main issues for OAS-2 will be:

- the extension and the integration of new contributing services/data providers (in form of journals or repositories),
- the standardisation of indicators that are based on the absolute frequency of document usage,
- the implementation of added-value services for repositories based on usage data,
- the evaluation of indicators that are more complex (mostly using techniques of usage data network analysis) than pure usage frequencies of documents and
- the internationalisation of the project.

Especially internationalisation and standardisation need an intense exchange of information with other projects tackling related issues as SURE, COUNTER, PIRUS, NEEO, PEER or OAPEN and Knowledge Exchange, a cooperation of Denmark's Electronic Research Library (DEFF), the DFG, the Joint Information Systems Committee (JISC) and the SURFfoundation.

Note:
This proposal is part of a coordinated effort by several German open access-related projects[10] to present and discuss their work at the Open Repositories Conference 2010. We see the conference as an ideal opportunity to discuss patterns of strategic and everyday collaboration and to open up to international partners. We would suggest to present these papers in a joint session.

[9] For more information see „Specification: Data Format and Exchange for OA Statistics". http://www.dini.de/fileadmin/oa-statistik/projektergebnisse/Specification_V5.pdf
[10] OA-Network, OA-Subject Repositories, OA-Statistics, DOARC, CARPET

Literature

Bollen, J., Van De Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS one*, *4*(6), e6022.
doi: 10.1371/journal.pone.0006022.

Bollen, J., Van De Sompel, H., 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.
doi: 10.1016/j.ipm.2005.03.024.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier Web Usage Statistics as Predictors of Later Citation Impact. *Journal of the American Association for Information Science and Technology*, *57*(8), 1060-1072.
doi: 10.1002/asi.20373.

Harnad, S., & Brody, T. (2004). Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine*, *10*(6).
doi: 10.1045/june2004-harnad.

Jamali, H. R., Nicholas, D., & Huntington, P. (2005). The use and users of scholarly e-journals: a review of log analysis studies. *Aslib Proceedings*, *57*(6), 554-571.
doi: 10.1108/00012530510634271.

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature*, *411*(6837), 521. Nature Publishing Group.
doi: 10.1038/35079151.