JCOM

SPECIALISED PORTALS, ONLINE INFORMATION SERVICES, SCHOLARLY ONLINE NETWORKS: THE IMPACT OF E-INFRASTRUCTURES ON SCIENCE COMMUNICATION AND SCHOLARLY COMMUNITY BUILDING

Lots of data, lots of hurdles: aggregating usage information from distributed Open Access repositories

Ulrich Herb

Abstract

This article reflects the results of the project "Open Access Statistics", which was designed to collect standardized usage figures for scientific documents. The data gathered were primarily intended to provide impact values based on document usage for Open Access documents as these were excluded from databases used to provide citation based impact scores. The project also planned the implementation of more sophisticated procedures such as network analyses, but was confronted with complex legal requirements.

Keywords Scholarly communication

DOI https://doi.org/10.22323/2.17020305

The background

There is no lack of scientific communication infrastructures, but not all of them are well accepted by their scientific communities. Critical for their success is the ability to raise the reputation of users. In other words, the impact of an e-Infrastructure on science communication and scholarly community building depends largely on its own ability to provide impact. This contribution reports on an attempt to provide alternative, usage data-based impact in a network of distributed servers.

Considering the declaration of the Budapest Open Access Initiative in 2001 as the year in which "Open Access" was born, free access to scientific information was just five years old in 2006. In these days Open Access was for many scientists not a very attractive way of publishing as it did not provide any impact. Impact, or generally speaking the resonance of a scientific publication, at that time was only determined by citations numbers (e.g. by the Hirsch index) or citation rates (e.g. by the Journal Impact Factor) calculated by commercial databases such as the Web of Science (WoS) or Scopus. However, Open Access was excluded from these databases: repositories because they were not (and are not) mentioned in their selection criteria, journals because in most cases they were newly founded and because they were ignored by databases due to a lack of publication history and

citation heights. Although Google Scholar indexed Open Access material, it has only been available as a beta version since 2004 and was by no means considered to be a competitor to Scopus and the WoS. Especially its scope was hard to define, as Mayr and Walter stated, it "can be recommended only with some limitation due to a lot of inconsistencies and vagueness (...) in the data" [Mayr and Walter, 2007, p. 828]. However, since scientists are largely indifferent when deciding whether to publish Open or Closed Access and the reputation of a journal that manifests itself in impact is the most prominent decision criterion, Open Access was not an attractive option.

In an attempt to compensate for this handicap of Open Access, the later project partners came up with the idea of taking usage data from scientific documents as impact indicators. This was initiated by studies showing that Open Access publications were downloaded and cited more frequently than Closed Access documents [e.g. Brody, Harnad and Carr, 2006; Moed, 2005]. Even more elaborate application scenarios were presented by Johan Bollen and his colleagues: Bollen et al. [2003] proved that usage data could predict future research trends. Bollen et al. [2005] found out that usage data measure a different kind of resonance than citation, as it captures the behavior of readers, whereas citations only describe reuse by authors. Bollen et al. [2009b] demonstrated that the importance of individual scientific journals can be determined by means of clickstreams and social network analysis based on usage data. In addition, Bollen et al. [2009a] used a principal component analysis to prove that citations have comparatively little influence on impact: "Our results indicate that the notion of scientific impact is a multi-dimensional construct that can not be adequately measured by any single indicator, although some measures are more suitable than others. The commonly used citation Impact Factor is not positioned at the core of this construct, but at its periphery, and should thus be used with caution." The aforementioned network and clickstream analyses were considered to be more meaningful: "Usage-based measures such as Usage Closeness centrality may in fact be better' consensus' measures'." Even better: Bollen and Van de Sompel [2006] described the design of an architecture for collecting and processing usage data.

Usage information thus appeared interesting to the Open Access Community for three reasons:

- If usage frequencies can predict citation frequencies, the first consideration is that they capture impact in the same way as citations (but earlier than citations) and can be used as independent (and ideally free-of-charge) impact information. In short: usage impact can be used as an alternative impact source for scientific documents and thus compensate for the Open Access reputation deficit.
- Download statistics showing higher usage figures of Open Access compared to Closed Access and thus promise higher citation rates could seduce scientists to publish Open Access.
- The network analysis of usage data described by Bollen et al. [2009b] outlined possibilities to design sophisticated impact models. These methods were not based on counting banal absolute frequencies, but were methodically superior to citation counts or download counts.

Open Access statistics: the project

The prospect of promoting the acceptance of Open Access through usage-data-based impact approaches prompted the Lower Saxony State and University Library Göttingen (SUB Göttingen), the University Library of Stuttgart, the Saarland University and State Library (SULB) and the Computer and Media Service (CMS) of the Humboldt University of Berlin in 2006 to plan a project for the collection of standardised uasge data. In May 2008, the project "Open Access Statistics",[1] which was funded by the German Research Foundation (DFG) and officially named "Networked Repositories: services and Standards for Internationally Comparable Usage Statistics" was launched. The first phase of the project, which ended in December 2010, was followed by a further phase, which lasted from April 2011 to November 2013, and in which a new partner responsible primarily for technology joint the project, the Head Office of the Gemeinsamer Bibliotheksverbund (GBV).

Phase 1

The primary objective of the first project phase was to develop and establish a standard for access counts and usage statistics for publications in both Open Access repositories and Closed Access services such as e-journals. Usage data on the latter should be obtained by analyzing link resolver logs. These applications check on the fly during a search, e.g. in a database or a search engine, whether document access from a campus network is possible.

Both data types were to be collected to enable comparability of the access information. The download counts of Open Access documents were to be determined by analyzing web server logs; the abundance of this data is in principle very high, since access information can be logged in great detail. In addition, the data is usually provided by the repositories themselves. However, these logs have different forms, and their granularity depends on local system configurations. The link resolver logs in turn were to be harmonized with the web server logs. However, both sorts of logs are not only quite different in structure and granularity. In addition, the link resolvers are partly hosted by libraries themselves, partly by commercial providers. In the second case, agreements were needed in order to analyze the logs. Whether aggregation of web server and link resolver data would be feasible was considered as a research issue.

The intention was to achieve the greatest possible completeness of data on document use by combining these two methods. The link resolver and web server logs should be extracted from the link resolver services and repositories (the data providers) of the project partners and merged in a central database (the service provider). This required the definition and implementation of interfaces between the data providers and the service provider. The software used to log, store, and deliver access information to the data provider was to be designed generically so that it could be used in as many different systems as possible with as little effort for customization as possible. The service provider itself should provide several services and functions, including

1. detection of duplicates (accesses to identical documents on different data providers should be cumulated)

[1] https://dini.de/projekte/oa-statistik (DFG Grant ID 72662563).

2. usage analysis (document access as pure frequencies and clickstream data)
3. user identification (as a pre-condition of clickstream analysis)

Based on these functions, value-added services were to be developed in a second project phase. Point three refers to a central work package of the project, privacy. Furthermore a review of possible standards for measuring access to online sources was to be accomplished in order to determine which information the data providers should provide to fulfill these standards or a standard developed on the basis of the identified reference standards.

The standards

The project group identified three reference standards:

- COUNTER:[2] a procedure of science publishers to measure access to licensed documents.
- LogEc:[3] a procedure of the server network RePEc[4] (Research Papers in Economics) to measure access to Open Access documents.
- IFABC:[5] a procedure of the advertising industry for measuring the outreach of online advertising.

These differed mainly in the definition of double-click intervals[6] and methods for eliminating non-human access, e.g. by crawlers. None of them was designed for user identification, to create clickstreams or de-duplicate documents.

Regarding web server logs it turned out to be quite easy — given an appropriate configuration of the server environment — to meet the requirements of the standards. Subsequently, the technical prerequisites were set up to collect the data required from web server logs, extract them and store them locally. At the same time, interfaces were developed to exchange the data with a service provider run by the SUB Göttingen in test mode. At the same time, the procedure had to be approved by the privacy authorities of the participating universities.

Technical and legal hurdles

Regrettably, it soon became clear that in Germany — unlike in the U.S.A. where Bollen and his colleagues did their research — access to licensed content does not happen to a significant extent via link resolvers: hardly any data could be obtained from the link resolvers, nor was it always possible to identify users by analyzing the data provided. Furthermore, in cases where the link resolvers were not run by

[2] https://www.projectcounter.org/.
[3] https://logec.repec.org/.
[4] http://repec.org/.
[5] http://www.ifabc.org/.
[6] The double-click interval is the time period within which two hits on a document are interpreted as one usage event.

project partners themselves, it was difficult to gain access to the logs. The reason for this was data privacy: the service providers had not foreseen the case of the passing on of this data in their license agreements and refrained from frightening customers by this sensitive topic or to complicate the marketing of their product by involving privacy authorities.

However, data privacy also made it difficult to process web server logs. While some of the privacy authorities insisted only on the pseudonymisation of IP addresses, including salting and hashing,[7] others made more far-reaching specifications and discussed the pseudonymisation very controversially. In particular, the Central Data Protection Office of the Universities of Baden-Württemberg (Zentrale Datenschutzstelle der baden-württembergischen Universitäten, ZENDAS) questioned the project very critically, with the result that the final implementation of the privacy guidelines was still work in progress at the end of project phase one.

Preliminary conclusions

The results of the first project phase were:

- Link resolvers were not integrated into the architecture, partly because of unavailability or low quantity.
- Software was developed to collect, store, pseudonymise and provide local data that could satisfy the aforementioned standards.
- The service provider (in test mode) received the data and processed it.
- The terms of use of the local servers allowed the collection of data.
- Software and legal policies of the local repositories were available for reuse.
- Which information from the log files could be recorded, stored and passed on — yes, whether they could be passed on at all — was still subject to the investigation.

Phase 2

Phase two of the project had primarily the following goals:

- to increase the acceptance of Open Access through metrics and value-added services
- to initiate cooperation for internationally comparable usage statistics
- to provide a sustainable service infrastructure

Point one referred to offering elaborate metrics or value-added services such as clickstream-based recommender services. For both functions, data from the distributed servers had to be aggregated. It was assumed that the clarification of

[7] In this scenario, the IP is enriched by a string (the salt) and replaced by a obscured value (the hash).

the privacy issues would allow clickstream analyses (and the pseudonymisation required for this purpose), at least to a certain extent.

If the pseudonymisation should not be possible, other features had been planned, e.g.

- GeoIP analysis: for each document it should be displayed where one is interested in its content.
- A standardized display of document downloads based purely on usage counts.

Not prominently mentioned, but very important was the role of GBV as a new project partner, who was to run the service provider operated as a test system by SUB Göttingen.

Evaluation of standards

The evaluation of the standards was based on expert interviews. Of the three standards put up for discussion, the IFABC procedure was considered unsuitable for measuring access to scientific documents. The best rated was LogEc, which was superior to COUNTER both regarding the double-click interval, which was criticized as being too short for COUNTER, and the identification of non-human accesses. However, the experts recommend the use of COUNTER rather than LogEc, as the latter was considered to be too unknown to find acceptance.[8]

Privacy

The result of the privacy audit [Zentrale Datenschutzstelle der baden-württembergischen Universitäten ZENDAS, 2011] provided the project partners with recommendations on the storage and processing of the data and confronted them with a major strategic problem. In many respects, there was a certain interpretation range between a very narrow and presumably inviolable interpretation of the legal norms and a potentially, but not certainly, risky interpretation. The first one allowed hardly innovative functionalities, but meant a high degree of compliance with privacy laws; the latter enticed with rich features and metrics, but exposed every institution using the service to legal uncertainties. The project group opted for legal certainty, assuming that the presumption of legal problems could fundamentally damage the attractiveness of the service — not to mention possible concrete lawsuits.

[8]COUNTER relies on a somehow arbitrary and short robot list because the standard was developed to measure access to Closed Access content. LogEc, on the other hand, uses not only a more sophisticated robot list but also data mining techniques to identify non-human accesses.


Tendenz | Diagramm | Relevante Dokumente

Titel	relative Abrufhäufigkeit*
Anwendungsmöglichkeiten scientometrischer Methoden in Wissenschaft und Forschung exemplarisch dargestellt am Beispiel der Nanotechnologie	15,33
Alte Hüte und neue Konzepte : Qualitätssicherung, Qualitätsmessung und Zitationshäufigkeiten	2,16
A scientometric method to analyze scientific journals as exemplified by the area of information science	5,30
OpenAccess Statistics: alternative impact measures for Open Access documents? : an examination how to generate interoperable usage information from distributed Open Access services	2,3
Open Access, zitationsbasierte und nutzungsbasierte Impact Maße: Einige Befunde	4,68
Die Zukunft der Impact-Messung - Social Media, Nutzung und Zitate im World Wide Web	21,62
Zur Evaluation wissenschaftlicher Publikationsleistungen in der Sportwissenschaft	17,3

*Durchschnittliche Zugriffe pro Tag multipliziert mit 100
Zugriffszahlen erhoben nach COUNTER-Standard
Die Daten unterliegen der Lizenz des Projektes Open-Access-Statistik

Figure 1. Recommender feature in SciDok.

In fact, however, this meant that theses aims had to be abandoned:

1. the aggregation of pseudonymised data across servers
2. point one also made clickstream-based metrics and recommenders impossible. The same was true for eliminating multiple accesses to identical documents on distributed servers within the double-click interval.
3. Furthermore, according to the ZENDAS report, the evaluation of other information from the logs appeared to be risky, e.g. the referrer or the GeoIP information.

If one adds the evaluation of the link resolver logs, which had already been cancelled in phase one, a further goal is given up: the comparison of usage figures of Open Access and Closed Access documents, which, combined with citation information, could have proven to be valuable for scientometric research.

Résumé

Besides the evaluation of the standards, in phase two the service provider was launched in an operational mode by the GBV. The local repositories provide anonymized data to it, the service provider processes it according to COUNTER specifications and returns it to the giving repositories (the data providers) that store the usage counts as metadata and display it together with the corresponding document. The limitation of the COUNTER robots list was counteracted by an extension of this list in coordination with other projects. In some cases, value-added services have been developed locally, e.g. on the repository SciDok of SULB: in a recommender function,[9] similar documents including their usage counts are displayed when viewing a record (see Figure 1).

[9]In this case the similarity is assessed by a keyword analysis.

Closing considerations

It is difficult to assess the project: on the one hand, the goal formulated in the project title ("Networked Repositories: services and Standards for Internationally Comparable Usage Statistics") was achieved: there is a service for reporting standardized access counts on repositories. However, the goal of creating an infrastructure for clickstream-based metrics and recommenders, which was not formulated in the title of the project, but was nonetheless envisaged, could not be achieved. The decisive factor was the necessity to weigh up between functionalities and legal certainty.

In the light of today's booming Altmetrics services one can certainly consider it a very early attempt to measure impact in an alternative way. At the same time the project may also be considered a little bit old-fashioned today, because the presentation of the impact information is very varying and, compared to the catchy Altmetrics visualizations, e.g. in the form of the metrics donut from the provider Altmetric.com, it seems a little clumsy.

It should be noted that the project has collected a lot of information on the legal and technical feasibility of collecting data for impact assessment purposes. This is perhaps its greatest merit, even if the findings are somewhat disillusioning: in Germany the collection of such data would have been legally simpler if the project had pursued a commercial goal.[10] In addition, one could learn the lesson that data should not be collected by the project itself, but to use external sources that — due to national laws — know fewer privacy barriers, as the Altmetrics services do when using the API of a web services such as Twitter, for example.

Last but not least Open Access Statistics was more modest than the Altmetrics services, since it was a declared goal to only provide standardized data that should be used for impact metrics and not to provide its own metrics, whose methodological foundation in the case of Altmetrics is more than questionable [Herb, 2016a; Herb, 2016b]. Moreover the data provided by Open Access Statistics are standardized and freely accessible, qualities that Altmetrics services still do not have.

References

Bollen, J. and Van de Sompel, H. (2006). 'An architecture for the aggregation and analysis of scholarly usage data'. In: *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries – JCDL '06*. New York, NY, U.S.A.: ACM Press, p. 298. https://doi.org/10.1145/1141753.1141821.

Bollen, J., Luce, R., Vemulapalli, S. and Xu, W. (2003). 'Detecting Research Trends in Digital Library Readership'. In: *Research and Advanced Technology for Digital Libraries*, pp. 24–28. https://doi.org/10.1007/978-3-540-45175-4_3.

Bollen, J., Van de Sompel, H., Smith, J. A. and Luce, R. (2005). 'Toward alternative metrics of journal impact: A comparison of download and citation data'. *Information Processing & Management* 41 (6), pp. 1419–1440. https://doi.org/10.1016/j.ipm.2005.03.024.

Bollen, J., Van de Sompel, H., Hagberg, A. and Chute, R. (2009a). 'A Principal Component Analysis of 39 Scientific Impact Measures'. *PLoS ONE* 4 (6), e6022. https://doi.org/10.1371/journal.pone.0006022.

[10] Article 96 of the German Telecommunications Act (Deutsches Telekommunikationsgesetz TKG) allows the logging of usage data explicitly for marketing purposes and for processing customer specific information (e.g. customer numbers), which are incompatible with Open Access services.

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A. and Balakireva, L. (2009b). 'Clickstream Data Yields High-Resolution Maps of Science'. *PLoS ONE* 4 (3). Ed. by A. Ruttenberg, e4803. https://doi.org/10.1371/journal.pone.0004803.

Brody, T., Harnad, S. and Carr, L. (2006). 'Earlier Web usage statistics as predictors of later citation impact'. *Journal of the American Society for Information Science and Technology* 57 (8), pp. 1060–1072. https://doi.org/10.1002/asi.20373.

Herb, U. (2016a). 'Altmetrics zwischen Revolution und Dienstleistung: Eine methodische und konzeptionelle Kritik'. In: Soziologie in Österreich – Internationale Verflechtungen. Kongresspublikation der Österreichischen Gesellschaft für Soziologie. Ed. by H. Staubmann, pp. 387–410. https://doi.org/10.15203/3122-56-7.

— (2016b). 'Impactmessung, Transparenz & Open Science'. *Young Information Scientist*. https://doi.org/10.5281/zenodo.153831.

Mayr, P. and Walter, A.-K. (2007). 'An exploratory study of Google Scholar'. *Online Information Review* 31 (6), pp. 814–830. https://doi.org/10.1108/14684520710841784.

Moed, H. F. (2005). 'Statistical relationships between downloads and citations at the level of individual documents within a single journal'. *Journal of the American Society for Information Science and Technology* 56 (10), pp. 1088–1097. https://doi.org/10.1002/asi.20200.

Zentrale Datenschutzstelle der baden-württembergischen Universitäten ZENDAS (2011). 'Datenschutzrechtliche Bewertung des Projekts "Open-Access-Statistik"'. URL: https://dini.de/fileadmin/oa-statistik/gutachten/ZENDAS_Gutachten_2011.pdf.

Author

Ulrich Herb, Sociologist and Information Scientist, is working for the Saarland University and State Library and as a freelance consultant.
E-mail: u.herb@sulb.uni-saarland.de.

How to cite

Herb, U. (2018). 'Lots of data, lots of hurdles: aggregating usage information from distributed Open Access repositories'. *JCOM* 17 (02), C05.
https://doi.org/10.22323/2.17020305.

ISSN 1824-2049. Published by SISSA Medialab. jcom.sissa.it