Date submitted: 02/06/2009

**Measuring the Usage of Cultural Heritage Documents:
The German Project Open Access-Statistics**

Matthias Schulze
University of Stuttgart
Stuttgart, Germany

Meeting: **92. Statistics and Evaluation, Information Technology** and
Preservation and Conservation

Abstract

The German project "Open Access-Statistics (OA-S)" is a joint infrastructure project for the standardised accumulation of heterogenous web log data with an emphasis on institutional repositories. Project partners of the OA-S project are Georg-August Universität Göttingen (Göttingen State- and University Library), Humboldt-Universität zu Berlin (Computer- and Mediaservice), Saarland University (Saarland University and State Library), and Universität Stuttgart (University Library). The actions undertaken are linked with national and international cooperations among others DINI (Deutsche Initiative für Netzwerkinformation / German Initiative for Network Information), Digital Repository Infrastructure Vision for European Research (DRIVER), and Joint Information Systems Committee (JISC). In OA-Statistics the data providers for access data will be the participating repositories (Berlin, Göttingen, Saarbrücken and Stuttgart), and in the next stage of expansion all DINI-certified repositories. In the long run, connections to other repositories and also to different digitized content for example cultural heritage documents are to be expected. The infrastructure is planned to be open for national and international repository providers to join in and benefit from the data aggregating and processing services provided by a central service provider.

Introduction

"Document-access logs can indicate content relevance compared to other documents. Repository-access logs can be the qualitative / quantitative and technological basis for the evaluation of a document repository. Evaluation of a document-access log over the document's lifespan can give an indication of its continuous relevance. To enable comparisons of statistics of different Document and Publication Services, it is essential that standards exist and are adhered to. Today however, suitable standards for access statistics of institutional repositories do not exist. Differing protocols and access technologies create some of the problems of data acquisition and evaluation. No basis exists for a comparative evaluation of documents in different repositories, and of the Document and Publication Services themselves."[1]

This quotation from the chapter 4.7 "Logs and Statistics" of the DINI-Certificate in which the Electronic Publishing working group of the German Initiative for Network Information (DINI) laid down explanation and examples concerning the DINI-Certificate.[2]

This position can be seen as the background on which the project described here was initiated, the starting-point of Open Access-Statistics. The next sentences of this DINI text describes this yet much more precisely: "DINI observes international developments, e.g. the standardisation project COUNTER (Counting Online Usage of Networked Electronic Resources) and the project Interoperable Repository Statistics (IRS), and will make new recommendations if necessary. Other approaches, such as the evaluation of link-resolver logs, will also be followed."[3]

This article also as the project generally deals with Open Access repositories and documents stored in these repositories. So to speak the main focus are digital born materials like dissertations, electronic journals, articles etc. To which extent these materials could be subsumed under the label "digital heritage" is nowadays hard to say. The aspects and dimensions which furthermore occur as relevant for this project are preferential focused on open access materials and naturally on publications. But in the last chapter of this article the focus will be expanded and is explicitly aligned to the topic of cultural heritage documents and other choices for collecting data (as funding of digitisation projects, cooperations, political questions).

Measuring research impact has a long tradition. Especially the Journal Impact Factor (JIF or IF) created by Eugene Garfield is a well-known and respected measure of citations to scientific journal literature in some disciplines.[4] But the IF has also often been critisised.[5]

---

[1] DINI-Certificate. Document and Publication Services 2007, Working Group "Electronic Publishing", Version 2.0, September 2006, p. 32 <http://edoc.hu-berlin.de/series/dini-schriften/2006-3-en/PDF/3-en.pdf>.

[2] Here in general the "minimum standards" and "recommendations" for repositories are explained which are defined in the criteria section (for logs and statistics ibid., p. 16).

[3] Ibid., p. 32-33.

[4] Compare Richard Monastersky: The Number That's Devouring Science The Chronicle of Higher Education, 2005 <http://chronicle.com/free/v52/i08/08a01201.htm>.

[5] "Despite its merits the Impact Factor has shortcomings that steadily grow as the publication process changes (This is a problem every measurement method faces as behavioural patterns change over the years, e.g. IQ-Testing.): arbitrary sample composition, arbitrary observation period, journal level granularity, preference of publications in English, discrimination of non-journal content. These characteristics do not seriously hamper the IF's quality in certain fields like the established natural sciences. But as soon as a community leaves the path in any direction it risks its standing and therefore its funding due to declining IF values. It does not matter whether it is because of: individual publishing/archiving via your staff homepage (computer science), long reception cycles (mathematics), alternative publication modes (humanities) or language barriers (linguistics) the Impact Factor cannot compensate for these violations of its design prerequisites." Cited from an unreleased paper (draft status) by Björn Mittelsdorf: Metrics Proposal for Institutional Repositories as DINI recommendation, 2009 (this paper is not yet a paper from DINI but in this state a paper to DINI).

Especially for the field of open access new metrics are needed: "The fast-changing nature of scholarly communications makes us aware that traditional metrics might not be exclusively sufficient to describe research impact."[6]

Some international developments, e.g. the standardisation project COUNTER (Counting Online Usage of Networked Electronic Resources) or approaches such as the evaluation of link-resolver logs which were developed at Research Library of the Los Alamos National Laboratory should be mentioned here. The basic architecture to make logs from repositories and linkresolvers approachable was developed by Bollen/Van de Sompel.[7] These infrastructure developments were part of the background on which Open Access Statistics started. The DINI working group Electronic publishing initiated the project idea – as seen in the introduction by concerning the focuses of DINI's work – and animated the project partners to take part in this research project.

The effort of producing and publishing a text is in vain if it is not read and noticed. Here there is no difference between scientific publications and works in the realm of fiction. The ease of access experienced with Open Access publications – lacking any need for authentication, financial transactions or personal identification makes it much easier to achieve a satisfying level of reception in a scientific community. This and similar hypotheses can be verified by empirical analysis.

Requests can be measured easily as webservers store most of the necessary pieces of information for internal management purposes. (Any approach which tries to surpass this level of insight has to face among others the problem that HTTP is a one way protocol. This makes it impossible on this level to learn anything definitive about dependent variables like success of file transfer or the time an user sticks to a document before moving on.)

Scientific publications cover a broad variety of publishers, hosts, business models, usage models, publication stages, logical and technical presentation. Therefore it is important to learn which portions of the publication space can be and which agents want to be included in the sampling. For those willing to participate only two aspects are relevant:

1. What data need to be gathered?
2. How can data be transferred to the statistics provider?

In this context the technical background of OA-Statistics and measuring research impact isn't the main focus of this article. So the following text will deal not very intensively with these topics. A good report concerning this topic is Frank Scholze: Measuring Research Impact in an Open Access Environment, 2007, p. 6-8 <http://elib.uni-stuttgart.de/opus/volltexte/2007/3234/pdf/liber_2007.pdf>.

[6] Ibid., p. 2. Compare also p. 2-3

[7] Johan Bollen / Herbert Van de Sompel: An architecture for the aggregation and analysis of scholarly usage data. In: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, 2006, pp 298 – 307, <http://doi.acm.org/10.1145/1141753.1141821>.

Project Description: Open Access-Statistics

The project Open-Access-Statistics (OA-S) is a joint research project addressing these questions. Starting in July 2008 an infrastructure will be built for the standardised accumulation of heterogenous web log data with an emphasis on institutional repositories. In tight cooperation with other Open Access projects various added value services will be made available to users.[8]

Figure 1: Open Access-Statistics Logo[9]

The "German Research Foundation" (Deutsche Forschungsgemeinschaft – DFG) funds OA-S in the start-up phase. Later on the project partners will supply it as a permanent service. The initial idea was - as mentioned before in the introduction - drafted by DINI (Deutsche Initiative für Netzwerkinformation / German Initiative for Network Information), more precisely by their workgroup Electronic Publishing.[10]

Project Partners of OA-S are Georg-August Universität Göttingen (State- and University Library), Humboldt-Universität zu Berlin (Computer- and Mediaservice), Universität des Saarlandes (Saarland University and State Library), and Universität Stuttgart (University Library). The actions undertaken are linked with national and international cooperations among others Digital Repository Infrastructure Vision for European Research (DRIVER)[11], Ligue des Bibliothèques Européennes de Recherche (LIBER), and Joint Information Systems Committee (JISC).[12]

---

[8] More information about Open Access-Statistics: <http://www.dini.de/projekte/oa-statistik/#c1203>. See also Ulrich Herb: OA-Statistik: Standardisierte Nutzungsanalysen als alternative Impact-Messungen wissenschaftlicher Publikationen. Workshop zu Standardisierung und Vernetzung von Open-Access-Repositorien für die Wissenschaft, Göttingen, 19.02.2009 <http://www.dini.de/fileadmin/workshops/oa-netzwerk-februar2009/herb.pdf>.

[9] Source: <http://www.dini.de/projekte/oa-statistik/>.

[10] More information about DINI <http://www.dini.de/english/>: "DINI is committed to developing information infrastructures and thus improving the information and communication services in institutions of higher education and in learned societies, namely regionally and supraregionally." Information about the DINI working group Electronic Publishing: <http://www.dini.de/english/ag0/e-pub0/>.

[11] "DRIVER the 'Digital Repository Infrastructure Vision for European Research' project is conducted by an EC funded consortium that is building an organisational and technological framework for a pan-European data-layer, enabling the advanced use of content-resources in research and higher education. DRIVER develops a service-infrastructure and a data-infrastructure. Both are designed to orchestrate existing resources and services of the repository landscape." Cited from "Driver Guidelines 2.0. Guidelines for Content providers - Exposing textual resources with OAI-PMH", November 2008, p. 9 <http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf>

[12] JISC Usage Statistics Workshop Berlin, July 7-8, 2008, <http://www.dini.de/veranstaltungen/workshops/jisc-workshop-2008/>. See also JISC Usage Statistics Review. Final Report, September 24, 2008 <http://ie-repository.jisc.ac.uk/250/>.
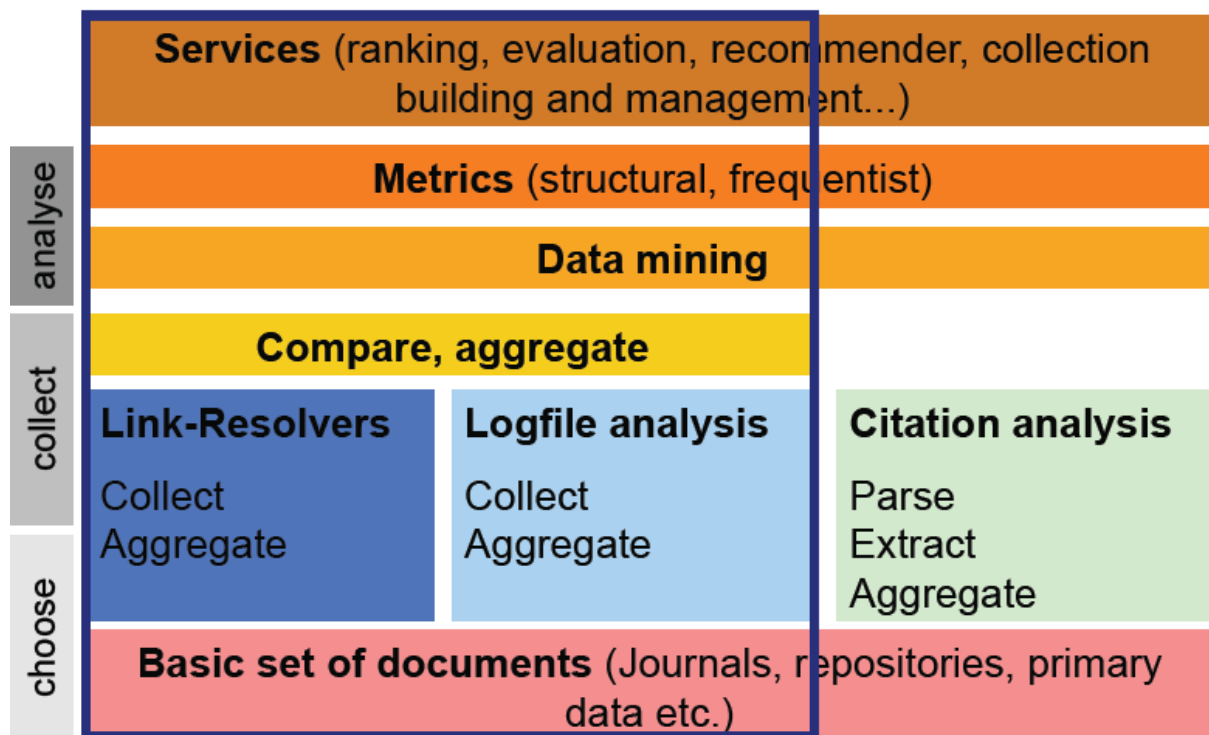
Figure 2: Measuring Publication Impact: The Elements (schematic)[13]

From the perspective of the central service/statistics provider, various data providers are sources for access data. In OA-Statistics the data providers for access data will be the participating repositories (Berlin, Göttingen, Saarbrücken and Stuttgart), and in the next stage of expansion all DINI-certified repositories.[14] In the long run, connections to other repositories are to be expected. The infrastructure is planned to be open for national and international repository providers to join in and benefit from the data aggregating and processing services provided by the central service provider. Not only do different repositories use different software solutions (DSpace, OPUS, edoc), but there are also qualitative differences between the information gathered on a server actually hosting documents and the information horizon on a link-resolving (SFX, Ovid) or license-controlling (HAN) server.

---

[13] Taken from Frank Scholze: Measuring Research Impact in an Open Access Environment, 2007, p. 6 <http://elib.uni-stuttgart.de/opus/volltexte/2007/3234/pdf/liber_2007.pdf>.
[14] DINI-certified server: <http://www.dini.de/no_cache/service/dini-zertifikat/zertifizierte-server/>.
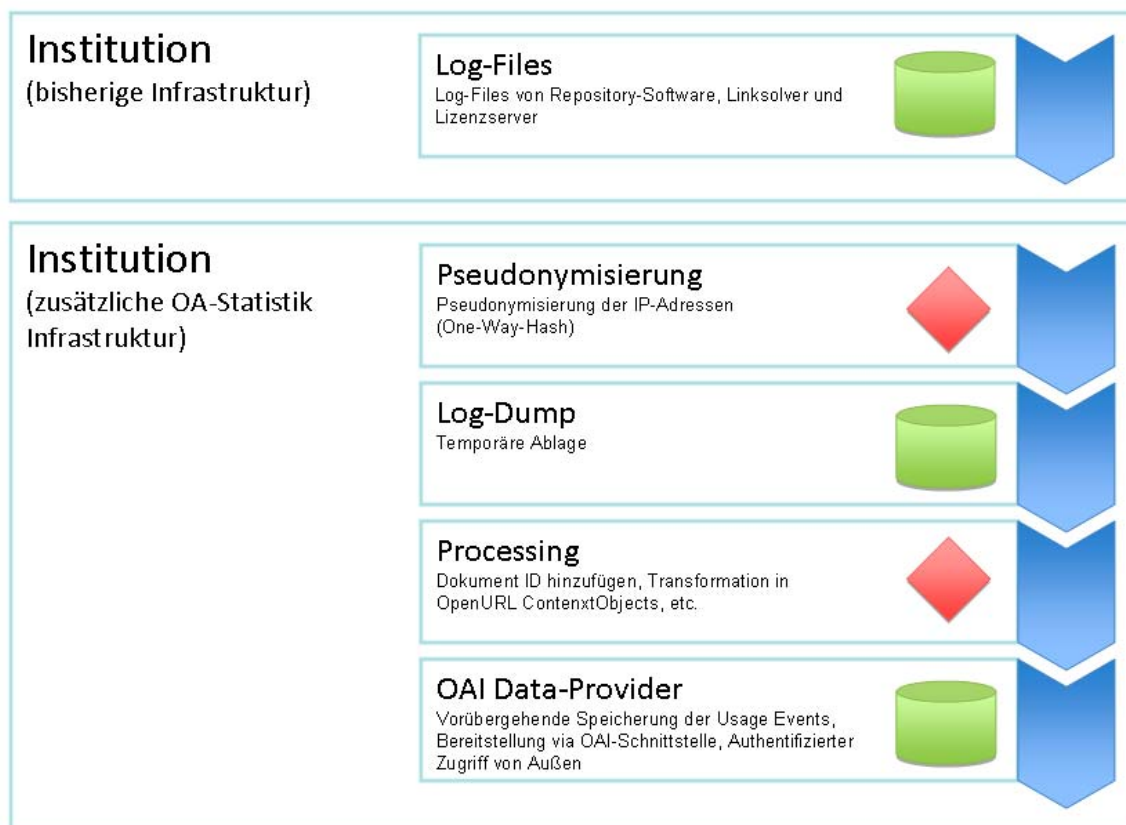
Figure 3: Preparation of the data exchange

The aggregates derived by the statistics provider from the access data generated locally will be hosted on a central server, referred to as Service Provider. Local repositories will be able to create added value services by integrating statistics into the documents' frontdoors. Another currently popular example would be a recommendation system based on click stream analysis. Initially repositories will probably focus on the portion of data which describes their digital objects such as reliable usage frequencies distinguishing between local and international visitors.

The DINI-Certificate[15] will be extended by proposals concerning and supporting data collection, presentation of statistical information and integration of the own repository into the OA-N networking infrastructure in order to propagate the results of this project. Guidelines, technical documentation, and software implementations will be provided for interested repositories. Thus they can - with a reasonably small effort - realise and concentrate on services which support the user in his quest for knowledge.[16]

The project OA-Statistics aims at two targets. First objective is the establishment of a permanent infrastructure – connected to the network of Open Access Repositories (OA-Network) – for the collection and processing of usage data is sought. The use of electronic information resources are collected at various levels. In addition different repository software products for the different project partners (DSpace, OPUS and e-doc) so- called link resolver (SFX, Linkresolver if necessary) and license server (HAN) are considered, too.
A second objective of the OA-S project is to develop and establish a binding standard for measuring traffic and statistics to seek for publications in institutional repositories. Therefore

---

[15] For information about the DINI-Certificate look here: <http://www.dini.de/english/dini-certificate/>.

[16] Open Access-Statistics received several letters of support, among others from the LIBER Access Division.

the already existing standards and rule sets such as COUNTER, LogEc or IFABC are assessed and taken into consideration.[17]

The usage data is collected by the participating project partners and subsequently processed. While preparing the usage data, it will be transferred into a standardized and uniform format (OpenURL Context Objects). This standardised data is provided on a OAI-interface (Figure 4).
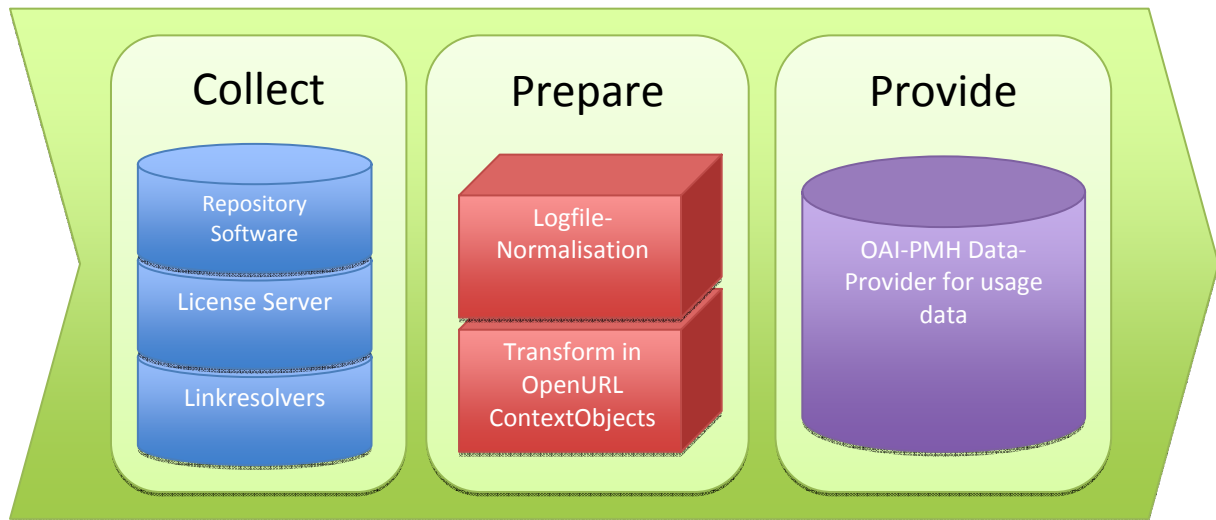

Figure 4: Workflow Data Provider

A service provider has access via the OAI-interface to the available data and collects this data from the project partners. As the usage data is available in a uniform format, the service provider is able to generate a variety of statistics and metrics from the data. In addition, a deduplication of usage data will be carried out. The aggregated data will be made available to the project "OA-Netzwerk". In addition, it is possible that the repositories get back the statistics of their resources to integrate them in their local installation (Figure 5).
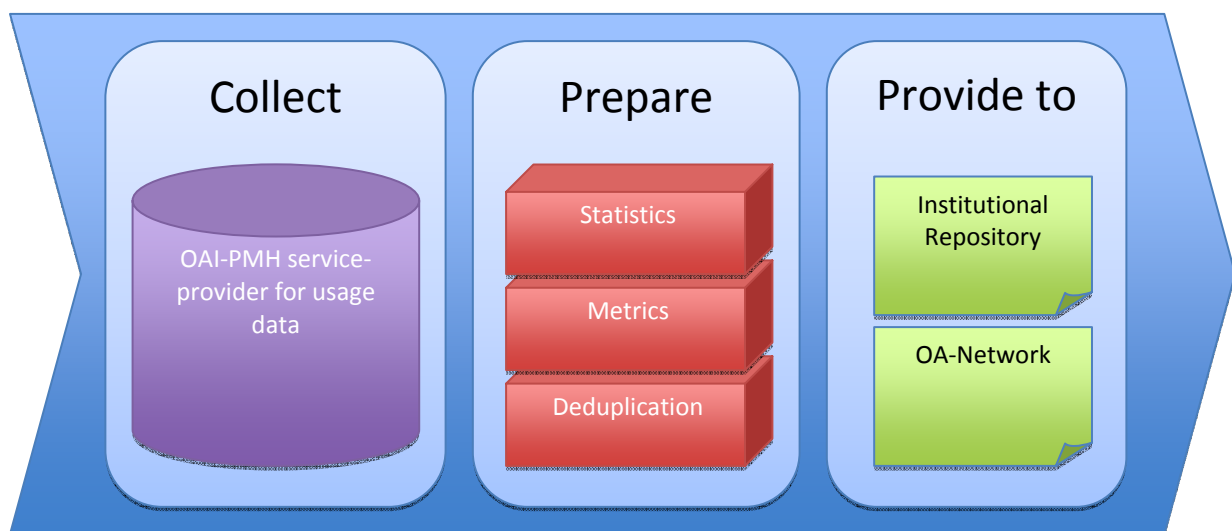
Figure 5: Workflow at the Service-Provider

---

[17] Concerning "Use of usage statistics exchange" see the Annex of the "Driver Guidelines 2.0. Guidelines for Content providers - Exposing textual resources with OAI-PMH", November 2008, p. 132-135 <http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf>.

The project objectives include the design (and construction) of various value-added services, which users can be offered, for example, on the level of institutional repositories. In both cases, the user of the services gets additional value. Such value-added services enrich the current offers by providing new possibilities for selection and display of electronic documents. Furthermore additional data to every document is available, e.g. how frequently it has been viewed locally but also nationally. This will be an attractive service for scientists, universities and research and funding institutions. It indicates standardized and reliably way how scientific publications are used and what impact free access has on the usage and visibility of the publications.

In order to receive the widest possible dissemination of project results on the one hand recommendations will be added to the DINI-Certificate will promote the standardised collection and provision of usage data and statistics as well as the connections to OA-Network. On the other hand guidelines, help documents, and software packages will be made available for repository managers, so they simply use standardized data and provide descriptive statistics. This enables them to offer their users – for a reasonable price – value-added services (Figures 6-8).[18]
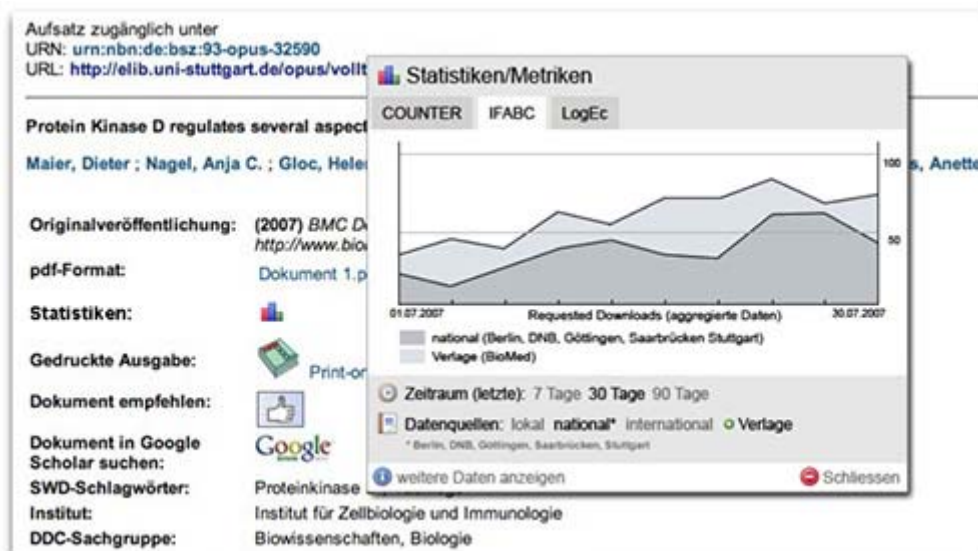
Figure 6: Value-Added Service "Statistics / Metrics" for Repositories (Draft)

---

[18] Figures 6-8 with exemplary realisation (drafts) of value-added services for repositories are taken from <http://www.dini.de/projekte/oa-statistik/#c1190>.

Figure 7: Value-Added Service "Dissertation Ranking" for Repositories (Draft)

Figure 8: Value-Added Service "Usage Relevance" in Repositories (Draft)

Networking: German OA-Projects and the International Context

Digital Publications networks flourish only in their real potential in terms of their global visibility and use. Though networking repositories are thus much more attractive for researchers regarding the decision to provide works in repositories. Networking activities exist at disciplinary and cross-disciplinary (in relation to the project presented here it is referred solely to the cross-disciplinary activities) as well as national and international levels. In particular, two initiatives are worth mentioning: at the German level, the above-mentioned

project "Network of Open Access Repositories (OA-Network)" and at the European level, the project "Digital Repository Infrastructure Vision for European Research (DRIVER)".

The DINI-initiated and DFG-funded national project "OA-Network"[19] aims at the virtual networking of all DINI-certified document and publication services. For this purpose, OA-Network makes available a portal with various end-user services (e.g. specialist browsing, export functions), provides its data room to re-use, and collaborates with other projects to expand the service spectrum, for example concerning usage statistics and citations.

Moreover, OA-Network supports repository operators during the DINI certification process and thus contributes to increasing the number of DINI-certified repositories. Currently twenty-nine German repositories hold the DINI-Certificate, another five are in the application process.[20] The certificate is internationally recognized and available in English, German and Spanish versions. Together with the "DARE Guidelines" the DINI Certificate served as a basis for the development of the DRIVER Guidelines for Content Providers.[21] Therefore, all DINI-certified repositories simultaneously are DRIVER-compliant and after passing are integrated within both networks.

In comparison with a collection of printed stocks digital repositories offer a much higher diversity of opportunities to develop value-added services and to provide these new services. This applies in particular to the collection of webometric data such as usage statistics or citation. In addition to the DFG project "Open Access-Statistics" the recently launched DFG project "Distributed Open Access Citation Reference Services (DOARC)"[22] is worth mentioning, too. In this project the technical requirements for the collection and analysis of citation coherences in Open Access offers should be made available. Motivated by the fact that – as mentioned before – the scientific landscape in the past used impact models which often do not cover Open Access offers and furthermore were faulty, both projects should develop alternative approaches measuring impact.[23] OA-Statistics therefore can be seen as part of DINI's initiative to build a network of certified repositories across Germany (Figure 9).[24]

---

[19] OA-Netzwerk: <http://www.dini.de/projekte/oa-netzwerk/>.

[20] For a list of DINI-certified servers see: <http://www.dini.de/no_cache/service/dini-zertifikat/zertifizierte-server/>.

[21] "Driver Guidelines 2.0. Guidelines for Content providers - Exposing textual resources with OAI-PMH", November 2008: <http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf>.

[22] DOARC: <http://doarc.projects.isn-oldenburg.de/>.

[23] Another procject in this context is the project "CAPRPET - Community for Academic Reviewing, Publishing and Editorial Technology". This project aims to support the efficient use of electronic tools and services for scientific publishing. More information about CARPET <http://www.dini.de/projekte/eng/carpet/>.

[24] For this topic and the general theme Open Access in Germany see: Birgit Schmidt / Karin Ilg-Hartbecke: Open Access in Deutschland – Erweiterte Perspektiven für die Wissenschaft. In: GMS Medizin - Bibliothek - Information Volume 9, Issue 1, 2009 (forthcoming), <http://www.egms.de/de/journals/mbi/index.shtml>.
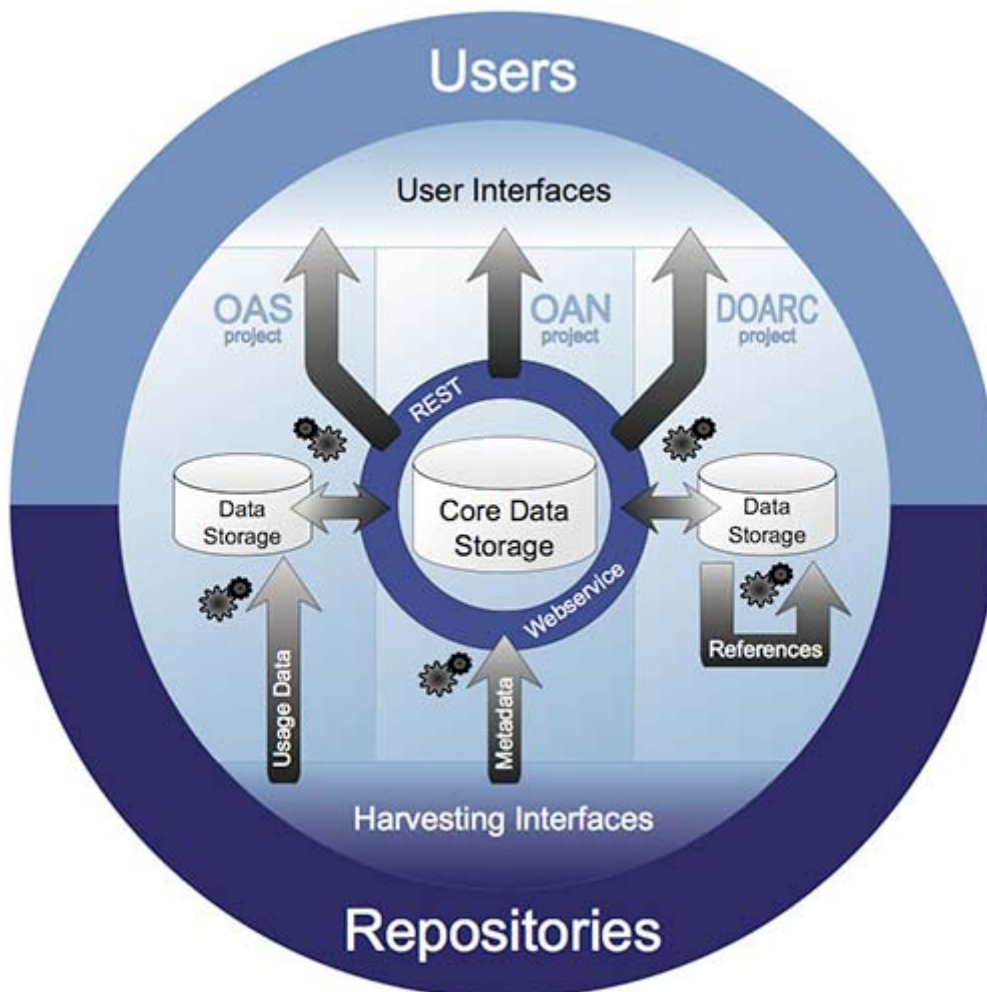
Figure 9: Structure and interrelation of the German OA-Statistics, OA-Network,
and DOARC projects[25]

The above-mentioned EU project "DRIVER", with the aim on building a network of European repositories, deals in its current second phase with testbed originated in Phase I, which will be expanded to a consolidated technology infrastructure. In addition DRIVER aims to build a sustainable organisational structure, a "Confederation of European Repositories", and includes explicitly national and regional organisations (DINI, SHERPA, etc.), technical initiatives, developers of repository software, as well as other stakeholders (e.g. SPARC Europe, LIBER). Furthermore, DRIVER investigates in its current phase recent technical developments and tests the enrichment of publications ("Enhanced Publications") as well as the connection of a long-term archive for repositories which store such complex publications.[26]

Outlook: Open Access-Statistics II and Integration of Digital Heritage Documents

From the start the project Open Access-Statistics was designed to enable a subsequent application. The first application focuses primarily on the conception and design of an infrastructure. In a subsequent application one main goal could be to integrate other

---

11

Institutional Repositories into the infrastructure which was developed in the first phase. For a targeted second project phase[27] (the first phase will end in February 2010) several plans exist. In comparison to the current project which is primarily a research project, the OA-S II projects intension is to develop and establish an infrastructure. Further tasks could be to examine the collected and aggregated data with respect to the local features that could be developed and offered on the basis of this data, e. g. recommender features based on affinity relations or on similarity. In addition some other themes could be of interest and may be part of the new application: new designed metrics should be developed and scientifically evaluated, building up a website with information concerning the projects matters (guidelines, FAQs, contact, help offers etc.), cooperation with other players, commercial players too (projects like PEER[28] for example).

Figure 10: Goobi. Usage Statistics on Digitized Repositories[29]

Another important part in the second phase of the project could be the integration of digitised material which is not digital-born. In this context usage statistics of digitised heritage documents will be automatically integrated. The OA-S project partner Göttingen State and University Library is involved as a core developer of "Goobi",[30] a software package which includes different modules supporting digitisation processes in libraries. The focus of Goobi is on two tasks: to support high quality (mass-)digitisation and to meet the needs of researchers concerning the presentation of the digitised objects. Special emphasis is collecting and handling of metadata: Goobi consistently uses the METS[31] data model and is a "open source workbench for the digital library" (see Fig. 11).

---

[27] This chapter deals with themes which are only intentions. The points mentioned here are not dicussed finally or part of the application.
[28] PEER – Publishing and the Ecology of European Research: <Publishing and the Ecology of European Research>. The EC-project is a pioneering collaboration between publishers, repositories and researchers: "PEER, supported by the EC eContentplus programme, will investigate the effects of the large-scale, systematic depositing of authors' final peer-reviewed manuscripts (so called Green Open Access or stage-two research output) on reader access, author visibility, and journal viability, as well as on the broader ecology of European research." Ibid.
[29] The Goobi slides (Fig. 10-12) are prepared and by courtesy of Ralf Stockmann, SUB Göttingen.
[30] Goobi: Digital Library Modules: <http://www.goobi.org/>.
[31] METS - Metadata Encoding & Transmission Standard: <http://www.loc.gov/standards/mets/>.
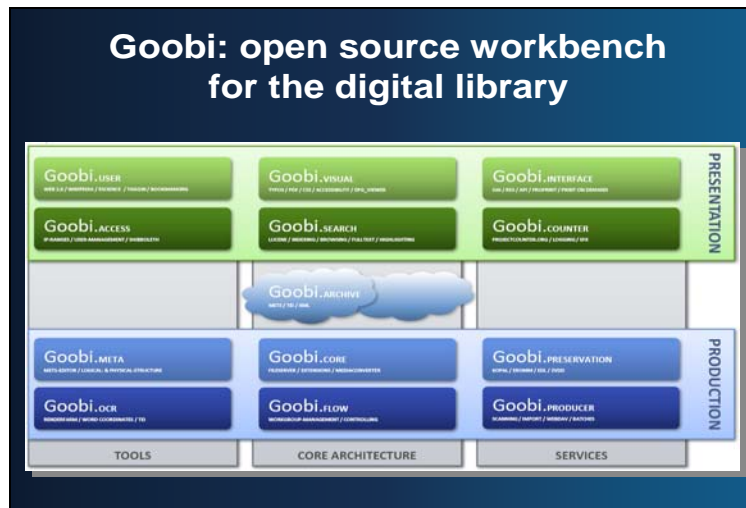
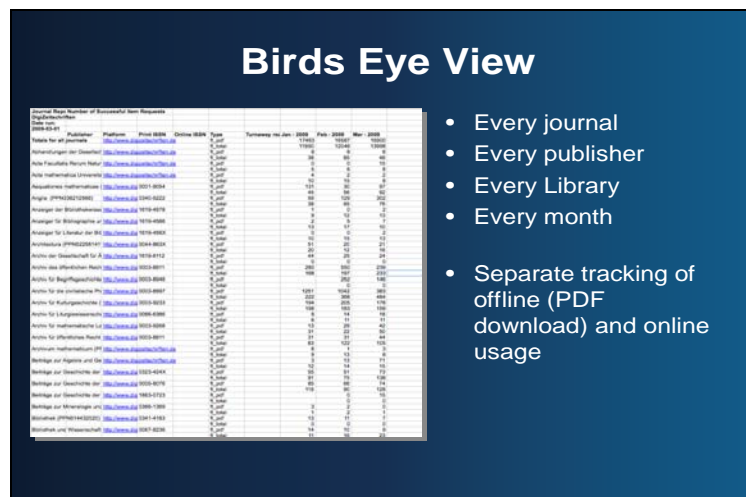Figure 11: Visualisation of the different Goobi Workbench Layers

Figure 12: "Birds Eye View" as part of the Usage Statistics created by Goobi

One of the Goobi modules in the section "presentation" is "Goobi.COUNTER" (Fig. 11). Goobi has extensive operating experience with Counter (an example of usage statistics presentation is shown in Fig. 12), so a collaboration between "Goobi" and the OA-S project, which also analyses COUNTER[32] for the purposes of usage statistics, could be very promising in this field. At the Center for Retrospective Digitization, Göttingen (GDZ) more than 5 Million pages have been digitised up to now.[33]

Conclusion

In an article published recently it is clearly pointed out, that digitisation can help reading ancient texts: "In a 21st-century version of the age of discovery, teams of computer scientists, conservationists and scholars are fanning out across the globe in a race to digitize crumbling literary treasures. In the process, they're uncovering unexpected troves of new finds, including never-before-seen versions of the Christian Gospels, fragments of Greek poetry and commentaries on Aristotle. Improved technology is allowing researchers to scan ancient texts that were once unreadable -- blackened in fires or by chemical erosion, painted over or simply

---

[32] COUNTER website: <http://www.projectcounter.org/index.html>.
[33] For more information see: Center for Retrospective Digitization, Göttingen (GDZ) <http://gdz.sub.uni-goettingen.de/index.php?id=2&L=1>.

too fragile to unroll. Now, scholars are studying these works with X-ray fluorescence, multispectral imaging used by NASA to photograph Mars and CAT scans used by medical technicians."[34] Considering this statement the number of digitised (cultural heritage) documents subsequently will increase in the future. And so the usage of this digitised content will appear as an important fact in the future, too. For example for collection management, project proposals concerning digitisation projects, and third party funding.

Maybe the infrastructure, the value-added services and metrics as are developed in the Open Access-Statistics project could give some contributions to usage data statistics which are helpful for both purposes: On the one hand for the usage statistics needs of researchers or users of (Open Access) materials.[35] And on the other hand e. g. for collection managers or digitisation funders who are interested in usage statistics from a different perspective.

Presenter/Contact:
Dr. Matthias Schulze
University Stuttgart - University Library
Holzgartenstraße 16
70174 Stuttgart
Tel: +49 (0) 711 - 685 82509
Fax: +49 (0) 711 - 685 83502
E-Mail: matthias.schulze@ub.uni-stuttgart.de

Biographical information:
Matthias Schulze studied medieval and modern history, political science and communication science in Göttingen, Germany. He holds a Master degree and a Ph. D. in history. Futhermore, he studied Library and Information science at Humboldt University in Berlin and holds a Master degree in Library and Information Science. From 2004 to 2007 he worked in several digital library projects at Göttingen State and University Library. At the Helmut-Schmidt-University – University of the Armed Forces in Hamburg he worked at the University Library as a subject librarian for miscellaneous humanities and social sciences and gave lectures in information literacy. Since 2009 he is working at University Stuttgart for the University Library as subject librarian especially for several social sciences and is project coordinator of several open access and other library projects. His interests include all aspects of the digital library, especially the topics long-term preservation, open access and electronic publishing.

---

[34] Alexandra Alter: The Next Age of Discovery. In: The Wall Street Journal, May 8th 2009, <http://online.wsj.com/article/SB124173896716198603.html>.

[35] And also for the enhancement or complement of existing metrics like the Journal Impact Factor (JIF).