

FINAL REPORT
JISC Usage Statistics Review

September 2008

Contact:

Email Christine.Merk@uni-konstanz.de

Phone +49 - 7531-88-2835

Executive Summary

The JISC Usage Statistics Review Project is aimed at formulating a fundamental scheme for repository log files and at proposing a standard for their aggregation to provide meaningful and comparable item-level usage statistics for electronic documents like e.g. research papers and scientific resources. The following elements describing usage events were agreed upon during the stakeholder workshop in Berlin, which was held in the context of the project:

Compulsory items:

- Who: Identification of user/ session
- What: Item identification
- What: Type of request performed (e.g. full-text, front-page, including failed/partially fulfilled requests)
- When: Date and time
- Usage event ID

Optional elements:

- From where: Referrer/ the referring entity
- Identity of the service

The thus described usage events should be exchanged in the form of OpenURL Context Objects using OAI. Automated access (e.g. robots) should be tagged. The definition of automated access has to be straightforward with the option of gradual refinement.

Users have to be identified unambiguously but without recording personal data to avoid conflicts with privacy laws.

With the JISC-funded Publisher and Institutional Repository Usage Statistics (PIRUS) and the DFG-funded Open-Access-Statistics there are two projects which will formulate standards for usage statistics and work on their implementation. To reach broad comparability national efforts should be bundled together. A central authority – which could for example be the Digital Repository Infrastructure Vision for European Research (DRIVER) – should aggregate the usage data. As the aggregator it would have to de-duplicate the items, which are available from more than one content provider, the tagging of non-human access would be its task, and it would have to engage in the quality control of the data recording.

Policies on statistics should be formulated for the repository community as well as the publishing community. Information about statistics policies should be available on services like OpenDOAR and RoMEO.

Table of contents

1	Introduction	1
2	Review of current practices.....	2
3	Formulation of the fundamental scheme	6
3.1	Data sources	7
3.2	The fundamental scheme.....	8
3.3	The normalization of the fundamental scheme.....	10
4	Exchange of the usage data	11
4.1	A common protocol for harvesting usage data	11
4.2	Mapping of the usage data.....	11
5	Comparison of the fundamental scheme with other practices	13
6	Legal constraints for recording and aggregating log files	15
7	Usage statistics policies.....	16
8	Future development of usage statistics services.....	18
	References	21
	Appendix	22

1 Introduction

The JISC¹ Usage Statistics Review Project is aimed at formulating a fundamental scheme for repository log files and at proposing a standard for their aggregation to provide meaningful and comparable item-level usage statistics for electronic documents like e.g. research papers and scientific resources.

The project is funded by JISC and it is conducted by a research consortium in which the Humboldt University Berlin (Computer and Media Service), the Göttingen University and State Library, the Library of the University of Konstanz, the Saarland State and University Library, and the Stuttgart University Library work together. It does not lie within the project's scope to test or even implement the derived recommendations. In the German context, the results will be taken further towards the implementation by the German Open Access Statistics project (OA-Statistics)², which is funded by the German Research Foundation (DFG)³. The project also intends to stimulate the attention of an international audience in order to attain international standards for the recording and aggregation of item-level usage statistics.

Core element of this review project is a stakeholder workshop. This workshop was held in Berlin on July 7th and 8th 2008. The participants were representatives of repositories and libraries as well as representatives of COUNTER, IRStats, JISC, LogEc, MESUR, OA-Statistics, and Network of Certified Open Access Repositories (OA-Network).⁴ During the workshop a fundamental scheme for the recording and the exchange of log files was developed; the normalization of the collected data was discussed and the means for the exchange of the raw data were looked at. Furthermore, legal constraints for the recording of personal data and copyright provisions for statistics were assessed. Finally, policies for the publication of the statistical data and the need for an aggregator were discussed. The results of the workshop were documented and were made available for comments from the participants and other stakeholders. The report at hand is the project's final report which is a synthesis of the workshop's results as well as a review of current practices and recommendations for the future of item-level usage statistics.

The report first gives a short overview of several projects, services and initiatives which are active in the field of online usage. This is followed by the description of the workshop's results dealing with the technical aspects and the mapping of the proposed scheme into OpenURL Context Objects. Third, the legal issues about the re-

¹ Joint Information Systems Committee, <http://www.jisc.ac.uk/>

² <http://www.dini.de/oa-statistik>

³ <http://www.dfg.de/en/index.html>

⁴ The list of participants is available in the appendix.

ording of personal data are described and finally policy issues and an analysis of the institutional needs are assessed.

2 Review of current practices

The DRIVER inventory study (Eijndhoven and Graaf 2007) of 114 repositories in 17 European countries showed that about 70 percent of those repositories logged the download and access data but only 30 percent offered item usage statistics for their end-users in 2006. About half of the repositories which had such a service were in the UK and overall a quarter was planning their introduction.

The benefits of usage statistics are especially their timeliness. Unlike citation-based statistics, the counts and rankings can be updated soon after the actual usage event. They are a measure for the visibility of electronic resources. The metrics make it more transparent for the authors to assess how visible their work is on a repository. Comparing different repositories, this supports the authors' choice of where to publish in order to have the highest impact. The users and the scientific community can evaluate an article or an author with the measures. Additional services like recommender systems, which are based on relational matrices, will increase the appeal of repositories to the users.

The counting and analyzing of online usage is done by various projects, initiatives and actors. In the context of this review the different definitions of a usage event, the de-duplication of multiple clicks, the identification of non-human access, and the aggregation of the usage data are of interest. We chose several prominent and also very different measurement concepts and had a look at their practices.⁵

There are initiatives dealing with scientific content like the British IRStats project or LogEc, which is designed for the economics database RepEc. In Germany, DINI recently started the projects OA-Statistics and Open Access Network (OA-Network). Furthermore, there are currently efforts in Australia: the Benchmark Statistics Service (BEST). The forerunner for structural measures of usage is the MESUR project conducted at the Los Alamos National Laboratories (LANL). In the repository community the Open Source analytical software package AWStats is often used. The widespread COUNTER code of practice only documents journal and database usage. The IFABC has formulated standards for the measurement of the use of page impressions of e.g. online services of newspapers.

We were also looking for initiatives in Norway and the Netherlands. In Norway there are currently no projects dealing with usage statistics. In the Netherlands on the other hand, there will be a project about collecting and providing usage data initiated by the SURFfoundation. Transnational, there were talks about usage statistics within Knowledge Exchange in 2007. Its report contained strategic challenges and issues

⁵ This does not claim to be a comprehensive list of all the existing efforts in the field of usage statistics.

connected with usage statistics at that time (Knowledge Exchange 2007). Knowledge Exchange is a platform for the co-operation of DEFF (Denmark's Electronic Research Library), DFG (German Research Foundation), JISC (UK), and SURFfoundation (Netherlands) in developing the ICT infrastructure for higher education and research.⁶

IRStats (Interoperable Repository Statistics)⁷ is a British project at the University of Southampton aimed at the design of a usage statistics module for repositories using eprints or Dspace. It was funded by JISC for 24 months starting in June 2005. The result of the project is a pilot version of the statistics tool at the eprints-repository at the School of Electronics and Computer Science at the University of Southampton.⁸ The programme excludes multiple clicks within 24 hours and uses the AWStats-robots list to delete non-human access. So far, the package cannot be implemented on a broad level and can only be operated with AWStats as a basis.

LogEc is a free online service which complements the metadata aggregator RepEc, which is specialized in economic literature. LogEc provides the statistics for items available from the participating services of RepEc: IDEAS, EconPapers, Socionet Personal Zone and Inomics.⁹ It also creates rankings of these items by number of abstract views and download frequency. The statistics for every available service i.e. abstract view or download for each item can be accessed via its front page. This free online service is hosted by the Swedish Business School at Örebro University. Multiple clicks are excluded using the IP-addresses for re-visits within one month. The log analyzer is run locally by the participating services and the usage data is then uploaded to LogEc. Non-human access is excluded in a two step process: Firstly, robots.txt¹⁰ sets rules for the robots about which directories they are allowed to index and known hosts of robots are excluded. Secondly, users who access more than 10 percent of all the items on RepEc within a month are eliminated.¹¹

DINI e.V.¹² (German Initiative for Networked Information) supports the co-operation and standardization of information and communication services in Germany. For repositories it sets standards with its DINI-Certificate. So far, its criteria for the certification do not include a comprehensive checklist for repository statistics; but with the

⁶ <http://www.knowledge-exchange.info/>

⁷ <http://irs.eprints.org/proposal.html>

⁸ http://eprints.ecs.soton.ac.uk/index_stats.html

⁹ <http://repec.org/docs/RePEcIntro.html>

¹⁰ <http://www.robotstxt.org/>

¹¹ <http://logec.repec.org/about.htm> and additional information from Sune Karlsson.

¹² <http://www.dini.de/ueber-dini/>

projects OA-Statistics and the OA-Network it is currently working on the implementation of a pilot version for a network of Open Access repositories.

Australian Benchmark Statistics Service (BEST)¹³ is a project which was conducted by the Australian Partnership for Sustainable Repositories (APSR) between September and December 2007. Their aim was to “identify an approach and initiate the design of a pilot service, providing the framework for further development” (Benchmark Statistics Project 2007: 3) for usage statistics. Its scope is very much like this project’s. It makes recommendations for the introduction of item-level statistics in the Australian context. For the exchange of the usage data they incorporated OAI-PMH. Its postulations for the exclusion of non-human access are similar to LogEc’s criteria but add the AWStats-list it: the application of robot.txt, the use of a dynamic access criterion like more than 10 percent of the content. They also launched a prototype for two data sources, which can display item statistics as well as aggregated measures.¹⁴

MESUR (Metrics from Scholarly Usage of Resources)¹⁵ is the most advanced project in the field of usage statistics for scholarly work. This research project makes the step from frequentist measures to structural measures. It uses a database with about 1 billion usage events to test different usage metrics and to generate a network of journals and a network of items. The tracking of the users’ click stream from document to document creates relational data. Based on this data, relationship matrices were then used to design the prototype of a recommender system on the article-level (Bollen/ Van de Sompel 2006: 304).

The data from different sources – i.e. publishers, aggregators, and institutions – was collected between 2002 and 2007. These are the required entries for the raw data: an identification of the event, an identification of the user or the session, a persistent identifier of the item and the time of the event (Bollen et al. 2007: 5). For the cases where a unique identifier is not available, MESUR uses a ‘bag of identifiers’-approach to de-duplicate the journals. It starts out with the ISSN; then the record is compared to the bag of identifiers, which contains versions and abbreviations of journal titles. The items below the journal level are matched using the year of the publication and the first 25 characters of the title. Non-human access is identified by its typical behavioural characteristics. Part of the data – the linking server log files from the California State University system – was packed into OpenURL Context Objects and harvested via OAI-PMH (Bollen et al. 2007: 5).

¹³ <http://www.apsr.edu.au/best/index.htm>

¹⁴ <https://devel.apsr.edu.au/cosi/best/reports/index.php>

¹⁵ <http://www.mesur.org>

The Open Source analytical software **AWStats**¹⁶ is widely used in the repository context. Multiple clicks within one hour are excluded via the IP-address. The click stream is not recorded and the statistics are only available for the respective repository. Non-human access is excluded using AWStats' own robots list. This list sets some kind of a standard, as it is also used by IRStats but it is a quite ambiguous benchmark because different versions of the list circulate.

The **COUNTER** (Counting online Usage of Networked Electronic Resources)¹⁷ code of practice for journals and databases is the very well established standard for journal usage statistics. In August 2008 its third release was published, which has to be implemented by August 2009 by the vendors to be COUNTER-compliant. The major changes in this release, which are relevant in the context of this report, are the now obligatory use of SUSHI for the harvesting of the COUNTER reports, the requirement to provide the reports in XML-format and the exclusion of non-human access from the reports (COUNTER 2008: 2). For the latter, COUNTER's current list contains 36 robots and crawlers (COUNTER 2008: Appendix K) but is open for ongoing updates. The list is also only a minimum standard, while it is not clear whether the additional exclusion of hosts has to be documented by the vendors. Access from federated searches and LOCKSS¹⁸ accesses also have to be excluded or at least itemized separately. Multiple clicks for a HTML-document by a single user are not counted if they happen within 10 seconds and for PDF-documents the time span is 30 seconds. For the multiple-click-filter the users are either identified by their IP-address or another kind of identification like a session cookie. Counted are only successful item-requests i.e. with the return codes 200 and 304 (COUNTER 2008: 32; Appendix D).

COUNTER uses journals as the lowest level of granularity for its statistics. Metrics for the item-level are more meaningful especially for documents on repositories. In an interdisciplinary repository aggregated statistics can tell something about its visibility and popularity but they provide no information on whether it is especially important within one subject area. COUNTER is currently working on the extension of its standards to the item-level for journal articles. The Project PIRUS (Publisher and Institutional Repository Usage Statistics) unites the publishers and the repository community in an effort to formulate common standards for the recording of item-level usage statistics which are applicable within both contexts. Results can be expected by the end of 2008.¹⁹

¹⁶ <http://awstats.sourceforge.net/>

¹⁷ <http://www.projectcounter.org/>

¹⁸ Lots of Copies Keep Stuff Safe: www.lockss.org

¹⁹ Information on PIRUS was provided by Paul Needham and Richard Geyde.

The **United Kingdom Serials Group (UKSG)**²⁰ as an interest group for publishers and librarian is also involved in PIRUS. Besides that, COUNTER-Executive Peter Shepherd (2007) did a stakeholder survey on the introduction of a Usage Factor (UF) for journals in 2007 commissioned by the UKSG. The UF puts the journal usage from the COUNTER-reports in relation to the number of articles published online; both variables are in reference to a specified period (Shepherd 2007: 4). In the near future, the UKSG plans to issue a request for proposals for a framework for the UF.²¹

The **International Federation of Audit Bureaux of Circulations (IFABC)**²² sets standards for the measurement of the usage of online content from not research related commercial providers. Though its standards do not deal with scholarly content it is included in this overview because its guidelines represent one prominent way of measuring internet traffic. Page impressions and visits per domain are counted, here; a page impression is “(a) file, or a combination of files, sent to a USER as a result of that USER’s request...”²³ Besides that, usage of other kinds of online content is recorded e.g. chat impressions or stream impressions.

The IFABC sets minimum standards; it gives guidelines for different measurement approaches e.g. users can be identified via their IP-address and the user agent, a cookie or a registration ID; it is only important that a visit can be reconstructed. More concrete are the standards of the German member organization of the IFABC – the IVW.²⁴ They prescribe the use of tracking bugs. Multiple clicks are excluded within a time span of 30 minutes (IVW 2008: 6). The technical implementation of the theoretical construct of user identification is not strongly regulated. This is similar to the COUNTER code of practice. The British ABC electronic²⁵ implements the IFABC rules and uses a list of robots for the exclusion of non-human access. The documentation of the robots exclusion and the harvesting of the usage data is not publicly available because the IFABC and its members are commercial services which provide documentation only for their members.

3 Formulation of the fundamental scheme

Ideally, usage statistics will be event based in the future: The users’ click stream is traced through the online content. These paths set up a relational network of the

²⁰ <http://www.uksg.org>

²¹ <http://www.uksg.org/usagefactors>

²² <http://www.ifabc.org/who.asp>

²³ <http://www.ifabc.org/standards.htm>

²⁴ Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V. – Interest group for the identification of circulation, www.ivw.de

²⁵ Audit Bureaux of Circulation electronic: www.abce.org.uk

items; those which attracted the same audience are densely connected. But repositories are not yet able to provide such data. There are several steps which have to be taken before rankings and other metrics like for example the prestige of an article or its betweenness centrality can be calculated. One step on the way towards this aim is delivering statistics which are comparable to the well-institutionalized COUNTER statistics for e-journals. But unlike COUNTER²⁶ they will be on the item-level instead of on the journal-level. This chapter mainly contains the results of the discussions at the stakeholder workshop in Berlin.

3.1 Data sources

The request for a document in itself says nothing about its actual usage i.e. reading or citing it; more accurately it should be called access data. Nevertheless, the term 'usage data' is predominantly used. At the workshop, it was agreed upon the definition of an access as the successful request of an item or its front page. An item's appearance in the result list of a search does not constitute a usage event because it does not clearly enough indicate the user's interest in the item.

Log files come in different formats depending on the source. There are differences between web servers in general but also between licence servers, link resolvers, and repository software packages. In order to make them comparable they have to be parsed and converted into a normalized format. The workshop's participants agreed upon the XML-format in the shape of OpenURL Context Objects as syntax for the normalized format. XML is also used by COUNTER, IRStats and MESUR.

The access to web pages is primarily recorded by web server logs but can also be derived from link resolver logs, licence servers and repository software packages. All four kinds of log files are viable options and can be combined, but not every repository software package logs the requests.

Linking servers can be used to record usage data. 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. This is not the case in many libraries or research institutions in Europe. There are other obstacles, e.g. the linking server log-files might not be available for free. Whereas licensed publications can be tracked elegantly content from repositories is not always included in the resolver systems. Furthermore, documents cannot only be accessed via link resolvers but also directly via a persistent identifier on a publisher site for example.

Web server logs have a higher i.e. full coverage because they are always produced when an item is requested. But on the other hand, they are not always available to the institution which provides the licence for the access but does not deliver the content. These usage events can be logged by the link resolver. The inherent problems

²⁶ www.projectcounter.org

of these two options make a combined approach the most promising. This means that the usage data from web server logs and from link resolvers is merged.

3.2 The fundamental scheme

Before the aggregation the records have to be made interchangeable and comparable between different web servers and between different repositories. Therefore, the recording has to be harmonized; this is called the 'fundamental scheme' throughout the report. During the workshop the participants agreed on the necessary entries in the log files. These are the identification of the user, the session, and the item, the type of request, and the date and time of the request. Those elements should be an integral part of the log files. They provide information about the basic questions 'Who?' 'What?' and 'When?' But beyond those entries the scheme should be open to extensions. Especially the identity of the service, the referrer, and the referring entity would lend themselves to extend the format. But for basic services those would not be needed.

User/ Sessions/ Usage event ID

The user, who requests a resource, has to be identified to make the exclusion of multiple clicks viable and to facilitate the tracking of the click stream through the online content. The first option for this is the IP-addresses; but as those can be tracked back to the user, privacy issues arise. It is questionable whether the IP-addresses should be recorded at all. In the very restrictive German legal context the processing of IP-addresses or any other data, which enables the identification of the user, is restricted. Another problem is proxy servers or a network of computers which share the same IP-address; they make it almost impossible to distinguish between two distinct users.

There are other strategies which render the recording of IP-addresses unnecessary and make more sophisticated end user services possible. For the services in mind the user can be identified via an automatically assigned session-ID instead of the IP-address. Both of the following approaches should be more appropriately called session statistics.

The first alternative is session-IDs for every visit i.e. a UUID (Universally Unique Identifier) for a continuous click stream. This allows controlling for multiple requests of a document by a user during one session. It also shows the user's path from one document to another when the requests for one session-ID are ordered according to their time stamp.

The second option is the use of session cookies or cookies. Those make the tracking of users even easier. A session cookie is deleted when the browser is closed. A cookie, which expires after for example one month, makes revisits within this time span identifiable and it extends the length of the recorded click stream. A disadvantage is the negative selection of users, who disable cookies. In 2005 about 5 percent

of users rejected cookies.²⁷ Another downside of cookies is the high implementation costs because they have to be configured for the various software programmes in use. This might exceed the budget of the repositories. Without a fairly large aggregator the implementation might not even be worth the effort: users visit small repositories only to download a single item. Only if the users' sessions can be tracked across several content providers, i.e. a consortium of repositories, cookies can be useful. The implementation of cookies for the access to PDF-documents is more complicated but overall not problematic: the request can be redirected to a tracking script which tracks the request before redirecting the users' request to the PDF.

A unique identifier for every event or session is also a necessary element. The identifier facilitates the de-duplication and prevents the double counting of events, which were triggered by a link resolving event and therefore recorded twice. So far, the technical implementation of session-IDs for linking servers is difficult and rarely available. The MESUR project used an extension for SFX. Whether other link resolver services are technically able to implement session-IDs is unclear.

From the IP-addresses the request's country of origin can be identified. This is an interesting piece of information as it shows the geographical spread of the users and therefore tells more about the impact of a document as well as a repository in general. But this evokes privacy issues, which will be discussed in chapter 6.

Item

The item has to be identifiable for the aggregation, the de-duplication of items and - if event-IDs are not available - for the de-duplication of events. Multiple hits for HTML-documents which contain images and are therefore recorded multiple times in the log file have to be controlled for, too. The items have to be identified on the basis of their metadata.

Ideally, every document is assigned a unique identifier like a DOI (Digital Object Identifier), a URN (Universal Resource Name) or similar persistent identification solutions. Unfortunately, the various repositories adhere to different standards for unique item identification; a unique identifier for all items has to be made a prerequisite for the exchange of usage data. In the cases where this is not possible fuzzy matching based on the metadata is an option to aggregate the events for an item (Bollen/ Van de Sompel 2005: 305) like it is also done by MESUR and OA-Network. This approach is also viable for the identification of documents which are not just part of one repository and have been assigned different unique identifiers. The de-duplication of items and events has to be addressed by the aggregator in order to be coherent.

²⁷ http://www.web-analytics.org/index.php/webanalyse-artikel/weiter/ohne_cookies_kein_korrektes_tracking/

3.3 *The normalization of the fundamental scheme*

Log files are not only a record of human access but also of the activity of spiders or web crawlers; those distort the usage data and make it invalid. Log file entries of spiders can be excluded using different strategies, which have already been mentioned in chapter 2:

- robots.txt
- Lists like the AWStats- or the IFABC robots list
- Access of a certain percentage of the content during a specified time period
- Exclusion of users with atypical usage patterns

Several participants argued against the deletion of the records of non-human access. On the one hand could this data tell more about the behaviour of robots and it does on the other hand open up alternative method of robot exclusion like implemented by MESUR. The click stream for a human user is supposed to be a lot shorter and to follow a meaningful pattern while automated access randomly retrieves all the documents. This avoids the dependence on potentially incomplete lists of robots. The log file entries from non-human access could either be tagged on the local level or by the aggregator. The tagging on the local level would incur the additional effort of controlling the repositories' practices before finally analyzing the data. The deletion or tagging of non-human access is therefore most effectively done on a centralized level. This would allow for the consistent exclusion of robots.

This approach raises another issue with non-human access: the amount of storage its recording requires. Tim Brody reported that about 90 percent of the hits can be ascribed to non-human access. In July 2007 robots accounted for 74 percent of the abstract views on RePec.²⁸ The size of the log files including the robots' accesses exceeds the storage capacities of many repositories. They delete them quickly because they can hardly keep them for a longer time. A solution would be the consideration of alternative storage like for example Amazon S3²⁹, which offers relatively cheap storage. Another option would be to harvest the usage data very often and to allow the repositories to delete the log files once they are aggregated.

The overall consent is that there will always be non-human access and that not all of it can be identified. But it is essential to have common guidelines for the exclusion in order to make the data comparable; otherwise the metrics are meaningless. The identification should be pragmatic and easy to implement on a wide variety of systems (very much like the example of RepEc). It should also be possible to refine and adapt the identification gradually.

²⁸ <http://logec.repec.org/about.htm>

²⁹ <http://www.amazon.com/gp/browse.html?node=16427261>

4 Exchange of the usage data

4.1 A common protocol for harvesting usage data

The two competing approaches for the harvesting of usage data are OAI-PMH and SUSHI. The harvesting is not just done in one direction from content provider to aggregator but also vice versa. The content provider (e.g. a repository) can harvest aggregated statistics or even metrics from the aggregator in order to supply its users with the information needed. Local institutions would act as OAI data providers and the central aggregator as OAI service provider as well a OAI data provider.

The SUSHI-protocol is very well defined but the experiences with its implementation are ambiguous and it is not as well institutionalized as OAI-PMH in the repository community. SUSHI is the standard for exchanging statistics in the publishing community. The workshop's participants were in favour of OAI-PMH. There will have to be crosswalks between the two options if OAI-PMH instead of SUSHI is implemented. This is particularly important on the semantic level. Therefore, it is highly appreciated that the results of this workshop are fed into the development of COUNTER regarding the extension of its standards of the item-level for journal articles with the PIRUS project.

The more basic approach is the data exchange with FTP (File Transfer Protocol); this was discussed, too. Its technical implementation is less complicated but the harvesting is less accurate than with OAI-PMH. Files which have already been harvested cannot be tagged and the deletion of files is not uniformly recorded as FTP does not provide a naming standard for the files.

It was agreed upon the usage of OpenURL Context Objects as containers for the XML payload. The advantages of OpenURL Context Objects are that they can be easily extended and that the data is highly compressible.

4.2 Mapping of the usage data

The mapping for the OpenURL Context Objects for usage data has already been proposed by Bollen and Van de Sompel (2006). We adapt their proposal and slightly adjust it to the needs of the fundamental scheme consulting with the OpenURL Standard (ANSI/NISO Z39.88-2004). Starting out with the entries in a log file in the left-hand column the respective elements within a Context Object are identified in Table 1 in the right-hand column.

Table 1: Mapping of usage data to the data structure of OpenURL Context Objects

Log file entry	OpenURL Context Objects
document identifier	<i>Referent</i>
time of access	<i>Timestamp</i>
event identifier	<i>Context Object Identifier</i>
IP-address	<i>Requester</i>
session identifier	<i>By-Value metadata of the requester</i>
user agent	<i>Referrer</i>
http status code	<i>ServiceType</i>

A Context Object can contain six different kinds of entities: the *Referent*, the *ReferringEntity*, the *Requester*, the *ServiceType*, the *Resolver* and the *Referrer*. Those can be described by an *Identifier*, *By-Value Metadata*, *By-Reference Metadata* and *Private Data*.

The *Referent* is the requested item, which is to be identified by some kind of persistent identifier, which would be the URN for example. *By-Value Metadata* can be added to the *Referent* in case the items have to be de-duplicated. This would increase the size of the payload considerably. By using the *By-Reference Metadata* instead the event data could be merged back with the metadata if necessary. The time of the access and the event identifier are part of the header of the XML Context Object.

The user's IP-address or an encoded version of it can be mapped onto the *Requester* field. A separate entry for a session-ID does not exist but it can be included into the *Requester's* metadata. The format for this entry would yet have to be defined (Bollen/ Van de Sompel 2006: 301). The user agent is written into the *Referrer*-field.

The mapping of the HTTP status code onto the *ServiceType* is not possible yet. Possible values for the *ServiceType* entity are 'fulltext', 'abstract', 'citation', 'holdings', 'III', and 'any' according to the OpenURL registry.³⁰ To include the information from the http response code new values would have to be defined. The information whether the *Referent* is a full-text or an abstract should be included into the Context Object, too, though it might not be available from the log files. The content provider has to provide this information and either add it to the log files or write it into the

³⁰ info:ofi/fmt:xml:xsd:sch_svc

ServiceType by-Value metadata. In order to have both kinds of *ServiceTypes* the Context Object can be filled with two *ServiceType* entries.

Context Objects are harvested using OAI-PMH but as they are in the XML-format they should also be harvestable with SUSHI. The OAI-header has to be replaced by the SUSHI header. SUSHI itself does not require the XML payload to have a specified format (ANSI/NISO Z39.93-2007). Technical interoperability cannot be assessed further at this point but the effort should be made in the future to stay in line with developments from the publishing community.

5 Comparison of the fundamental scheme with other practices

The characteristics of the scheme drafted at the workshop and those of COUNTER, LogEc, and the IFABC are summarized in Table 2. It will be shortly assessed how this report's proposal would have to be adjusted to be comparable to statistics from COUNTER, LogEc or the IFABC and whether it would be meaningful at all.

A comparison between the workshop proposal and COUNTER measures according to the third release of their Code of Practice is not viable. The granularities are too different at this point. COUNTER only captures journal or database usage and the workshop proposal is aimed at the item-level and is focussed on repository content; even in an aggregated form the comparison would not be meaningful. As COUNTER is working on the extension of its code to the article-level, the new developments from the PIRUS project should be closely watched in order to adapt to those.

The LogEc scheme is just slightly different as two aspects of the fundamental scheme are not defined in more detail yet: the normalization of the data and the time criterion for the exclusion of multiple clicks. In the German legal context the identification of a revisiting user within one month might not be possible even when a one-way hash is used. LogEc does not harvest with a pull-mechanism but the locally analyzed usage events are uploaded using FTP. But this is not a significant difference for the metrics as it does not affect the definition of what is counted.

Making the statistics comparable to the IFABC standards is a little more difficult as the robots list is not freely available. The definition of what is counted would have to be reduced to the full-text access as splash pages, which are the equivalent to an abstract view, are not counted by the standards of the IFABC.³¹

³¹ <http://www.ifabc.org/standards.htm>

Table 2: Comparison of the JISC usage statistics review proposal with relevant parts of other schemes

Criteria	Fundamental scheme	COUNTER for Journals^a	LogEc^b	IFABC^c
Granularity	Item-level	Journal-level	Item-level	Page impression, visit
Definition of usage event	Successful abstract views and downloads	Successful full-text requests	Abstract views and downloads	Number of page impressions per domain, number of visits
De-duplication of multiple clicks	No specification	For HTML 10 seconds, for PDF 30 seconds	One month	1 hour
Identification of non-human access	No specification	Robots list	Robots.txt, dynamic criterion	Robots list
Harvesting	OAI-PMH	SUSHI	Upload via FTP of the locally analyzed usage data	-

^a COUNTER 2008^b <http://logec.repec.org/about.htm> and additional information from Sune Karlsson.^c <http://www.ifabc.org/standards.htm>

6 Legal constraints for recording and aggregating log files

The legal session at the workshop was moderated by Michael Seadle from the Institute for Library and Information Science at the Humboldt University Berlin and by Hannes Obex from the Institute for Information, Telecommunication and Media Law at the University of Muenster.³²

There are different legal concerns connected with usage statistics: (1) the privacy of the user (2) the privacy of an author/ publisher whose pieces' usage statistics are published (3) the copyrights for the metrics.

Privacy laws can be infringed upon by the recording and processing of IP-addresses or the use of cookies, but the regulations vary strongly between countries. Privacy laws are nationally determined and they are subject to European legislation. The EU legislation is stricter than the US regulations but it is less strict than the German laws. The situation in the United Kingdom is mainly dominated by EU regulations.

In the US, the Federal Privacy Protection Act (1974) applies only to public authorities. The Telecommunications Act (1996) was mainly designed for the context of access provision; it does also not apply to content providers. Users' privacy protection relies heavily on self-regulation; privacy protection laws for the private electronic communication sector do not exist in the US. This low level of regulation makes the data exchange with actors within the EU difficult. The transfer of personal data from the EU to the US is not permitted under European law. The only exception is companies which belong to the safe harbour system. The participating US companies have to comply with EU standards. The data transfer within the European Union is permitted.

Germany has very restrictive regulations for the privacy protection. The constitution grants the right of informational self-determination. The recording and handling of personal data is very restricted. Personal data in the context of usage statistics means the IP-address, which makes the user identifiable. The German Telemedia Act (TMG) defines the boundaries: personal data must not be collected or processed unless it is for the purpose of providing the service or billing. The provider of a free service is therefore not allowed to process IP-addresses (TMG §15 (1)). Whether this applies to the parsing of the country code from the IP-addresses before the data is pseudonymized or anonymized is unclear. Service providers are allowed to generate pseudonymized user profiles for market research, advertisement, and the demand-oriented design of the service. The country of origin can be interpreted as a viable element of a user profile and might therefore not raise legal issues as long as the data is pseudonymized afterwards. The merging of pseudonymized data with the personal data is prohibited.

³² This section about the legal constraints relies strongly on Obex's and Seadle's talks and their expertise. Their slides are available at http://www.dini.de/fileadmin/workshops/JISC-workshop/Legal_Session.pdf

The user's right to object the recording of the access data is generally unaffected by the regulation. The content provider has to inform the user beforehand about the extent and the intention of the data recording and processing. If the offered service is unique and not available otherwise the service must not be denied to the user even if the consent to the recording of the data was not given. The usage data has to be anonymized before the exchange or aggregation and the user has to be informed. It is not yet clear whether the data exchange between the members of a consortium falls within the legal definition of data exchange. The use of cookies is subject to the same restrictions unless they do not contain personal data.

In the German legal context the recording of IP-addresses is strongly restricted and the current interpretation of the act is ambiguous. To avoid legal problems it would be best to pseudonymize IP-addresses shortly after the usage event or not to use IP-addresses at all but to promote the implementation of some sort of session-identification, which does not record IP-addresses.

The privacy of authors and publishers on the other hand should pose no legal problems for the publication of the statistics; they do not constitute personal data; the same applies to publishers. But there might be social norms in different disciplines which might raise strong objections against the publication of usage statistics. Publicly available usage statistics also might interfere with the economic interests of publishers.

Copyright for statistics is a debatable issue: the basic question is whether statistics have enough inherent originality to have them protected. If they lack originality, they would be categorized as facts and would as such not be protected by copyright laws. Copyright laws are not automatically enforced; the rights holder has to press charges against an infringement. The likelihood of such a claim increases with the perceived or actual economic value which could be drawn from the statistics.

7 Usage statistics policies

The policy session at the workshop dealt with the publishing of usage statistics, quality standards, and the institutional support of usage statistics. Furthermore, it is of interest what has to be done to initiate the implementation of the proposed usage statistics.

It was discussed whether usage statistics should be open access and if yes to what extent. There was a common understanding that the raw data should not be publicly available as privacy might easily be breached. Only strictly regulated access for research should be possible.

Less consent was reached about the status of the usage statistics. Many repositories are on the one hand part of the Open Access movement and therefore do not want to contradict its ideals. On the other hand, the infrastructure for the services has to be

financed. Usage statistics would be a valuable service. They can be used for research evaluation and they are the precondition for the introduction of recommender systems. A third option besides a freely available or a fee-based service is a partially publicly available service. Basic measures can be made open access while the access to more sophisticated measures and recommender systems can be restricted. This is also a way to adjust a repository's services to its available resources especially during the first stages of the introduction of usage statistics. In this layered approach the content providers can decide what level of statistics they offer. They can either only provide COUNTER statistics on the item-level or go beyond simple counting and offer structural measures.

Besides the already mentioned status of the statistics and the description of the involved actors, a repository policy should contain guidelines for the collection and processing of the data. The Berlin Principles on Ranking of Higher Education Institutions³³ can be taken as an orientation for the creation of a policy on statistics. The Berlin Principles have been formulated by the UNESCO European Centre for Higher Education, the CHE University Ranking (Germany) and the Institute for Higher Education Policy (USA) in 2006. Especially the guidelines on the design and weighting of indicators, the collection and processing of data and the presentation of the ranking results apply to the context of usage statistics. Derived from those principles the policy should state transparently how the data is normalized as well as how the indicators are constructed and weighted.

The aggregation and processing of the data as well as the quality assurance has to be done by trusted and neutral third parties. The centralization of the normalization and the processing of the data minimizes the threats to the reliability and the validity of the statistics; it rules out many sources of error due to divergent local practices. The centralized authority should also install an auditing process to ensure the ongoing high quality of the data recording as it is already practiced by COUNTER or the IFABC. This central authority should have the power to make ad-hoc changes in the policy if needed. Already existing institutions which lend themselves to the role of the aggregator are for example DRIVER, DINI or JISC Collection. The choice will in the end depend on the overall scope of the initiative: it can either result in national or international usage statistics.

³³ http://www.che.de/downloads/Berlin_Principles_IREG_534.pdf

8 Future development of usage statistics services

In the United Kingdom, the next step towards comparable item-level usage statistics will be taken by the recently initiated PIRUS project. Its aim is to formulate a COUNTER-compliant standard for publishers and repositories for the measurement of the usage of journal articles. The standard will be designed to also be applicable in the repository context. It is funded by JISC from September to December 2008; it marks a joint effort of publishers and repository representatives. The implementation of usage statistics lies not within its scope.

As an aggregator and an initiator of further development in Great Britain JISC is probably the most suitable actor because it has the necessary resources. The UKSG on the other hand mainly represents libraries and publishers; it is not an ideal first advocate of usage statistics in the repository context but it is an important partner for PIRUS and COUNTER.

In the German context the results of this report will be primarily taken further by the DFG-funded project OA-Statistics. It develops a pilot version for a statistics service for the project partners' repositories; their focus mainly lies with the design of the infrastructure less with the actual metrics. In one of the last work packages, OA-Statistics will analyze their data with network analytical methods and will identify possible end-user services. Other important tasks, like the de-duplication of articles, are done centrally by OA-Network. Those two projects cooperate strongly and plan to develop a common user interface for searching repository content and for delivering usage statistics (figure 1). The projects started in mid-2008 and the final results can be expected by the last quarter of 2009. The services provided by OA-Network are to be made available and adjustable beyond the project's duration. Part of the project is to formulate a business model for the operation; this will most likely include an ongoing engagement of DINI in the provision of the service.³⁴

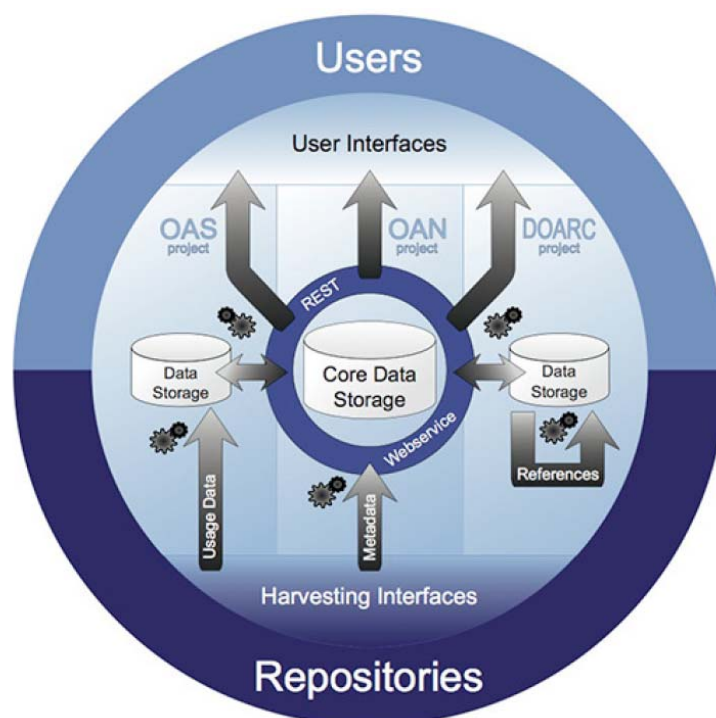
The right-hand side of figure 1 shows the DOARC project³⁵ (Distributed Open-Access Reference Citation services). DOARC is planned as a complementary service to OA-Network and OA-Statistics, which analyzes and links referenced works. The project has not started yet.³⁶

³⁴ <http://www.dini.de/projekte/oa-netzwerk/projektbeschreibung/>

³⁵ <http://doarc.projects.isn-oldenburg.de/index.php?site=doarc>

³⁶ At the time of the final version of this report the implementation of DOARC was still uncertain.

Figure 1: Structure and interrelation of the German OA-Statistics, OA-Network, and DOARC projects



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

The current various efforts in Germany, the Netherlands, and the UK are important initiatives but to make more efficient use of the resources and to foster a common European standard those projects should work as closely together as possible. The JISC Usage Statistics review workshop was an important event for the promotion of the dialog and the co-operation between the projects.

Already within the Knowledge Exchange in 2007 the national organizations DEFF, DFG, JISC, and SURF shared their views on usage statistics; this should be continued.³⁷ Even more important is the uptake of the national projects' results in the DRIVER context. OA-Network is the German contribution to the DRIVER project. Its modules – and therefore also OA-Statistics – are designed to be compatible with the DRIVER-platform. The integration of OA-Network and OA-Statistics' results in DRIVER are therefore highly desirable and also feasible. The adaption of the OA-Statistics guidelines for DRIVER might even be possible; respective contacts have already been established. Cooperation between the national projects would also guarantee a broad support within the DRIVER community.

Beyond the technical and organizational implementation, it is important to provide repositories and authors with information and support in dealing with usage statistics; this will increase the acceptance of usage statistics as well as the compliance with

³⁷ <http://www.knowledge-exchange.info>

the standards. So far, the OpenDOAR³⁸ policy tool helps repositories to formulate policies on metadata, data, content, submission, and preservation. This tool could be extended to provide possible policies for usage statistics. OpenDOAR is an authoritative directory of academic open access repositories. It could in the future also give an overview of the different usage statistics policies of the repositories. The users could easily find out whether the repository publishes usage statistics and what their granularity or format is. In addition, the SHERPA/RoMEO³⁹ service listing the Open Access Policies of publishers could be extended to display information about which statistical data is available from publishers under which conditions.

³⁸ <http://www.opendoar.org/>

³⁹ <http://www.sherpa.ac.uk/romeo.php>

References

- ANSI/NISO Z39.93-2007 (2007). *The Standardized Usage Statistics Harvesting Initiative (SUSHI) Protocol*. www.niso.org/workrooms/sushi
- ANSI/NISO Z39.88-2004 (2004). *The OpenURL framework for context-sensitive services*. www.niso.org/standards/resources/Z39_88_2004.pdf
- Benchmark Statistics Project (BEST) (2007). *Guiding Principles: Version 4*. Australian Partnership for Sustainable Repositories (APSR). http://www.apsr.edu.au/best/best_gp_4.pdf
- Bollen, Johan, Rodriguez, Marko A. & Herbert Van de Sompel (2007). *MESUR: usage-based metrics of scholarly impact*. Prepared for the JCDL '07. http://www.mesur.org/Documentation_files/JCDL07_bollen.pdf
- Bollen, Johan & Herbert Van de Sompel (2006). An architecture for the aggregation and analysis of scholarly usage data. *JCDL '06: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, Chapel Hill, NC, USA. 298-307.
- COUNTER – Counting Online Usage of Networked Electronic Resources (2008). *The COUNTER code of practice: Journals and Databases – Release 3*. <http://www.projectcounter.org/r3/Release3D9.pdf>
- Eindhoven, Kwame van & Maurits van der Graaf (2007). *Inventory study into the present type and level of OAI compliant Digital Repository activities in the EU*. <http://www.driver-support.eu/documents/DRIVER%20Inventory%20study%202007.pdf>
- IVW (2008). *Anlage 1 zu den IVW-Richtlinien für Online-Angebote: Definitionen und technische Erläuterungen, Version 2.1*. http://daten.ivw.eu/download/pdf/Online_RichtlinienV2_1_Anlage1.pdf
- Knowledge Exchange (2007). *Institutional repositories workshop strand report, strand title: Usage statistics*. <http://www.knowledge-exchange.info/Default.aspx?ID=164>
- Shepherd, Peter T. (2007). *Final report on the investigation into the feasibility of developing and implementing journal usage factors*. <http://www.uksg.org/usagefactors/final>

Appendix

Agenda

July 7th 2008 (11:00-17:30)

11:00-11:15 Welcome and introduction

11:15-12:00 **Usage data:** Workshop objectives from the perspective of the DINI - DFG - JISC projects (Moderators: Frank Scholze, Stuttgart University Library; Nils K. Windisch, Goettingen State and University Library)

12:00-12:15 Project MESUR (Johan Bollen, Los Alamos National Laboratories; LANL)

12:15-12:30 Project COUNTER & Sushi (Peter T. Shepherd, Director of COUNTER)

12:15-13:15 **Usage data:** Participants' perspective (short statements by the participants)

13:15-14:00 *Lunch*

14:00-15:30 **Legal session:** copyright status of statistical data, privacy issues (Moderators: Michael Seadle, Institute for Library and Information Science, HU Berlin; Hannes Obex, Institute for Information Law, Telecommunication Law and Media Law, Muenster University)

15:30-16:00 *Coffee*

16:00-17:00 **Technical session I:** 'fundamental format' for exchanging repository usage statistics (Moderators: Frank Scholze, Nils K. Windisch)

17:00-17:15 *Break*

17:15-17:30 Wrap-up (Moderator: Frank Scholze)

July 8th, 2008 (09:00-14:00)

09:00-10:30 **Technical session II:** 'fundamental format' for exchanging repository usage statistics (Moderators: Frank Scholze, Nils K. Windisch)

10:30-11:00 *Coffee*

11:00-12:00 **Policy session:** policy on statistics, should statistics be open access? (Moderator: Werner Stephan, Stuttgart University Library)

12:00-12:15 *Break*

12:15-13:00 Wrap-up: Do we have all requirements for a common 'fundamental format' for exchanging repository usage statistics? (Moderator: Frank Scholze)

13:00-14:00 *Lunch and farewell*

Table1: List of participants with affiliation

Name	Affiliation
Chris Armbruster	Max Planck Digital Library
Johan Bollen	MESUR
Tim Brody	IRStats
Susanne Dobratz	Computer and Media Service, Humboldt University Berlin
Fred Friend	JISC
Helmut Hartmann	Graz University Library
Sabine Henneberger	Computer and Media Service, Humboldt University Berlin
Ulrich Herb	University and State Library of the Saarland
Sune Karlsson	Swedish Business School, Örebro University
Anja Kersting	Library of the University of Konstanz
Frank Lützenkirchen	Duisburg-Essen University Library
Ross MacIntyre	MIMAS
Christine Merk	Library of the University of Konstanz
Björn Mittelsdorf	University and State Library of the Saarland
Sebastian Mundt	Stuttgart Media University
Hannes Obex	Institute for Information Law, Telecommunication Law and Media Law, Muenster University
Peter Schirnbacher	Humboldt University Berlin
Frank Scholze	Library of the University of Stuttgart
Michael Seadle	Humboldt University Berlin
Peter Shepherd	COUNTER
Jan Steinberg	Library Service Baden-Württemberg
Werner Stephan	Library of the University of Stuttgart
Adrian Stevenson	JISC
Fred Vos	NEEO
Nils K. Windisch	Göttingen State and University Library