

Enhancement of Data Quality in Distributed Open Access Repositories

Thomas Severiens^a, Manuel Klatt-Kafemann^b, Robin Malitz^b

^aUniversity of Osnabrück, ^bHumboldt University, Berlin

Summary

Building a comprehensive service layer on academic open access publication repositories is the goal of the project Open Access-Network. How these activities are intended to boost the acceptance and dissemination of open access publication as part of academic publication activities is being discussed. Also an overview of the technical implementation of the software infrastructure layer is being given.

1 Motivation

Over the last decade, self publishing and institutional publishing of academic results have become more and more an integral part of the research cycle. With the continuously accelerated progress in the research cycle (author, research, publication, distribution, reader, research, ...) the bottleneck of a slow publication and distribution process became more and more obvious. Online publishing on institutionally or self run services, in parallel to traditional journal publishing, has become an obvious solution for scientists from many fields over the last years.

One risk in this tendency of a growing number of publication repositories is a lack of quality control for the uploaded content. The filtering function, which is still the key role of the editorial board in any journal publication, needs to be implemented into the workflow for every repository, too. Otherwise scientific output would become a matter of arbitrariness instead of research results. The quality of the content offered by a service is the result of the weakest link in the chain of its workflow, from the author, via a repository into a global service of retrieval. To increase the quality of content offered via Institutional Repositories, well defined and documented workflows and collection policies are required to be developed in a standardized way.

To achieve these objectives as well as to harmonize the technical interfaces offered, DINI, the German Initiative for Network Information, developed the DINI-Certificate for Document and Publication Services (DINI 2007).

One of the motivations for authors to publish their articles in parallel in the Institutional Repository (of e. g. their university) is the impact of this way of publication. A high impact should result in many citations, which could boost the career of the authors. Any impact requires, that readers are able to find the articles they are interested in, which might be hard if the desired works are distributed across many repositories, that are not easily visible or searchable using generic web search engines.

“Open Access-Network” (OA-Network)¹ will act as a gateway service by handpicking certified repositories, collecting their content, process and clear their data and offer it to users and other services. The strict selection of repositories together with the technical clearing of the data is intended to keep the offered content on a high intellectual level. Fulltext and metadata retrieval as well as classification browsing and further add-on services are implemented on the enriched quality data and distributed globally.

1.1 DINI-certified Document and Publication Services

With the goal “to reach a higher level of scientific and scholarly communication both nationally and internationally” (DINI 2007) DINI developed in 2003 its certificate for “Document and Publication Services”, which is continuously being maintained. The certificate paper gives a detailed description of the state of the art in building up Institutional Repositories. It identifies a minimum and optionally an add-on set of features for repositories to fit into the academic publication workflow and network to set up global retrieval and publication platforms.

The DINI certificate gives a compilation of eight aspects, starting from more trivial ones like web access while ending at very detailed technical and organizational criteria. Starting from requirements on the availability and visibility of the service online, and the requirement of a written collection and operation policy, minimum standards on the support of authors, via legal aspects on hosting author’s copies and postprints on a web visible service, the certificate details criteria for the technical implementation, security and reliability of the service, including data integrity. To allow building up a subject browsing across multiple repositories, a subject indexing applying the Dewey Decimal Classification (DDC) in addition to some free text keywords is required. The build up and described model of an ideal repository is being completed by describing the metadata export as a complementary, documented, and standardized interface service, including OAI-PMH. Also generation and exchange of usage log files are a criterion, where the certificate paper gives some hints, regarding privacy and data protection regulations. To guarantee a persistent service, the certificate asks for a persistent linking of the objects, freeness of any DRM (digital rights management) constraints, and suggests cooperation with national libraries for the goal of long-term availability.

Every service provider may apply for the certificate by filling an online form. An authorized DINI working group evaluates if the service meets the minimum requirements of the certificate. This evaluation process is done as an all online expertise. If any of the criterias are not fulfilled, the evaluators are to guide and support the provider in fulfilling these requirements. After a successful evaluation process, the service may show the certificate signet on its homepage.

Goal of this process is to increase the number of certified servers, which can be embedded into comprehensive and quality controlled services like the OA-Network. The DINI Certificate and its criteria have had broad international effect. For example, the DRIVER Guidelines used part of the criteria and are based on this know-how, or the translation of the certification paper into Spanish language aiming at the development of a comparable certificate implemented by the REBUIN project as Alicia Lopez-Medina reported in (Lopez-Medina 2007).

1 OA-Network: <http://www.dini.de/oa-netzwerk/>

2 The Project OA-Network

OA-Network is a joint collaboration of the Humboldt Universität Berlin, Göttingen State and University Library, and of the University of Osnabrück. It is supported by the German Research Foundation (DFG) for a period from 2007 until September 2009.

It aims at developing a backbone infrastructure for integrated access to the content offered by all DINI certified repositories, while increasing the number of certified repositories in parallel.

2.1 Organizational Implementation and Data-Clearing

To make Open Access more accepted and widespread in scholarly and professional publishing, its impact internationally needs further development. This can be realized by embedding content from local repositories into international services, which may be field specific or general. While all repositories offer their content via an OAI-PMH interface, theoretically it is easy to develop a common and global service-provider, by just harvesting and offering all information offered – theoretically!

While OA-Network is building up an infrastructure and service layer for collecting, clearing and harmonizing the information offered by the distributed repositories, its collection policy is limiting harvesting to those repositories, which are certified by DINI. This strict policy may only be softened in exceptional cases for selected very high quality repositories, which are not yet certificated for miscellaneous reasons. This regulation of the collection policy is motivated by the target of developing a high quality service on professional scholarly content.

In addition to the selection restriction on subscribed repositories, a clearing, enrichment, and harmonizing of data will be implemented into the OA-Network backbone software layer. This will include:

- parsing and validation of the XML datastream offered by the repositories,
- plausibility check of information like dates, pages (format), classification, type etc.,
- harmonizing of date (timestamp) information,
- bundling of objects (duplicates, formats and versions of the same object offered by one or distributed servers),
- crosswalks for classified objects to DDC, in addition to the original classification,
- automatic try to classify texts (DDC).

Repositories or single objects, which do not comply with the minimum technical quality requested by the DINI certificate, will be automatically recognized and logged, so the administrative contact of the repository can be informed about the state - and hopefully the existing bugs can be fixed, soon.

The technical implementation of the infrastructure is an enabling layer, which registers modules to communicate with a database of objects. The role of the enabling layer is to register and synchronize services. All services communicate with the database via SOAP or REST interfaces. This allows every registered module to be run on different machines in different locations and to easily add more services. The only necessity is a good internet connection.

While several core modules are developed within the OA-Network project itself (searching, browsing along the DDC etc.), additional services are developed by the related projects OA-Statistics and OA-Citation-Analysis (to be started in April 2008). Additional or improved modules may be developed by everyone interested, plugged into the network via SOAP.

2.2 Technical Implementation of the Network

Basically, OA-Network offers a comprehensive open platform access to the cleared and harmonized data, hosted by the repositories, on an API implemented in SOAP. For this, a harvesting and indexing service collects data and updates from the repositories and stores objects into the database. Though ordinary changes of objects should result into new objects at the repositories, updates are frequent, so the used data-model is able to handle this scenario. If repositories disappear on objects are being removed from the repository, they will persist in the OA-Network database with a marker as being obsolete. By an aggregation tool, metadata (DCMI simple) is being extracted from the OAI-PMH data, other metadata formats are being translated into DCMI simple, as far as possible. One module tries to guess the fulltext link, which often requires some additional human knowledge in a learning phase of the software.

The “spectra of service modules” in the open OA-Network infrastructure

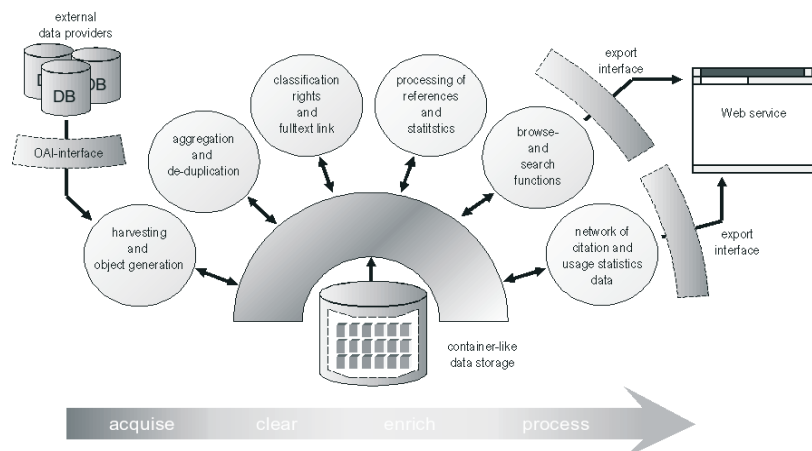


Figure: Architecture overview of enabling layer of OA-Network.

Based on the original OAI-PMH stream, the metadata extracted and the fulltext of the object itself, further services like data enrichment, clearing and harmonisation of the object metadata are implemented. Steps taken to achieve this are: to bundle similar object (duplicates, versions, formats), to check and harmonise timestamps, to convert and add classification information.

For bundling of similar objects, a shingle based algorithm is used, which is similar to the one described by Monica Henzinger in her article "Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms" (Henzinger 2006).

The cleared and enriched database is made available via the API for further services on the one hand, and as an OAI-PMH data-provider on the other hand. This will enable other services to process the data itself and develop advanced services.

3 Outreach and Vision of the OA-Network

To enhance the visibility and impact of Open Access publishing in Institutional Repositories, exporting of the comprehensive content into international services like Google® Scholar or Thomson's ISI® Web of Knowledge® Citation Database is one of the goals of OA-Network in cooperation with OA-Statistics and OA-Citation-Analysis. This added value of being a DINI certificated repository, should motivate further institutions to certify their repository, which includes to develop authoring support and to professionalise the workflow of documents into the repository.

One of the close cooperation partners of OA-Network is the European DRIVER² project. DRIVER is building up a more general and fully distributed infrastructure layer of Institutional Repositories, which technically is fully compatible with the built up OA-Network implementation.

4 References

- DINI (2007): DINI-Certificate for Document and Publication Services.
<http://nbn-resolving.de/urn:nbn:de:kobv:11-10075687>
- Henzinger, M (2006): Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms.
<http://ltaa.epfl.ch/monika/mpapers/nearduplicates2006.pdf>
- Lopez-Medina, A. (2007): in DRIVER-Wiki.
<http://www.driver-support.eu/pmwiki/index.php?n=Main.Spain#repositories>

Contact Information

Thomas Severiens
Department for Mathematics and Computer Science, University of Osnabrück
Albrechtstraße 28a, 49069 Osnabrück, Germany
severiens@mathematik.uni-osnabrueck.de

Manuel, Klatt-Kafemann
Computer- and Media-Service, Humboldt Universität
Unter den Linden 6, 10099 Berlin, Germany
manuel.klatt@cms.hu-berlin.de

Robin Malitz
Computer- and Media-Service, Humboldt Universität
Unter den Linden 6, 10099 Berlin, Germany
malitzro@cms.hu-berlin.de

² DRIVER: <http://www.driver-repository.eu>