Confederation of Open Access Repositories

Promoting greater visibility and application of research through global networks of Open Access repositories

The Case for Interoperability for Open Access Repositories

Working Group 2: Repository Interoperability

July 2011
version 1.0

Abstract

Open Access repositories, whose number has been steadily rising, are an important component of the global e-Research infrastructure. The real value of repositories lies in the potential to interconnect them to create a network of repositories, a network that can provide unified access to research outputs and be (re-) used by machines and researchers. However, in order to achieve this potential, we need interoperability. The purpose of this paper is to provide a high-level overview of interoperability of Open Access repositories, identify the major issues and challenges that need to be addressed, stimulate the engagement of the repository community and launch a process that will lead to the establishment of a COAR roadmap for repository interoperability.

Open Access and Interoperability

'Open Access' refers to the practice of granting free access to research outputs via the Internet – free of charge and free of most licensing restrictions. Generally the goals for Open Access are to: increase access, visibility, and impact of research results; promote the progress and efficiency of science and spark innovation by using technology to further disseminate research and allow scholarship and research data to be mined, used and reused; and maximize the return of investment in science by making publicly-funded research freely and publicly available.

Open Access occurs via two methods:

1) "Gold" Open Access: Scholars publish in Open Access journals, i.e. peer-reviewed, scholarly journals that make all articles freely available.
2) "Green" Open Access: Scholars publish in any peer-reviewed journal and then deposit a copy of their article in an Open Access repository.

Today, Open Access repositories increasingly are being used to collect, archive, and disseminate all types of research outputs such as research articles, conference proceedings, dissertations, data sets, working papers and reports.

The research process is an international and distributed endeavour, involving a variety of stakeholders such as scientists as authors and grant recipients, research institutions, publishers, and research funding agencies – each with their own set of interests. Collaborations occur among scientists at various institutions, often working in countries around the world. The repository infrastructure mimics this international and distributed environment: most higher education institutions who are involved in supporting Open Access maintain their own institutional repositories in which they collect, archive, and disseminate the research output for their affiliated faculty and scientists. Some institutions also maintain large, discipline-based repositories such as arXiv.org and PubMed Central.

Each individual repository is of limited value for research: the real power of Open Access lies in the possibility of connecting and tying together repositories, which is why we need interoperability. In order to create a seamless layer of content through connected repositories from around the world, Open Access relies on *interoperability*, the ability for systems to communicate with each other and pass information back and forth in a usable format. Interoperability allows us to exploit today's computational power so that we can aggregate, data mine, create new tools and services, and generate new knowledge from repository content.

Students and researchers looking for information do not need to know where a specific item was published or where an article is housed. Instead, users rely on search engines to retrieve articles, and they are able to discover information they might not have otherwise located. Furthermore, by relying on search engines to provide a layer that knits together these disparate repositories, Open Access does not need to rely on a single repository to collect the research output of the world. The decentralized infrastructure and the value-added services and tools being built on top of repositories are all possible because of interoperability. The quality of these services depends on data provided by repositories and the standardization of that data.

Interoperability is the technical glue that makes this integration possible – and makes the goals of Open Access possible to achieve.

Interoperability in a Nutshell

Within repositories, interoperability occurs in multiple ways. At the system level, interoperability occurs when repositories are set up in such a way that they allow for data or digital objects to be passed into or out of repositories via external systems. One example, the Protocol for Metadata Harvesting (OAI-PMH), specifies certain criteria that must be met in order to allow external, third-party systems to access and harvest metadata from repositories.

Interoperability relying on the OAI-PMH protocol has become part of the standard implementation for repositories. Nearly all repositories meet this baseline for interoperability and allow for metadata to be harvested. Metadata records from many repositories are then aggregated and are able to be used in some new and different ways, most commonly in the creation of subject- or discipline-based portals and specialised search engines such as OAIster and BASE (Bielefeld Academic Search Engine).

Other protocols also try to harness the potential of interoperability by creating systems that are designed to facilitate sharing content (and not just metadata records) among systems. One such protocol, the Open Archives Initiative Object Reuse and Exchange (OAI-ORE), defines the standards for the description and exchange of aggregations of web resources, sometimes called compound digital objects, and allows items that have already been deposited into a repository to be copied into other collections.

Another protocol, SWORD (Simple Web-service Offering Repository Deposit), allows authors to deposit an article via a single interface and then route that item to multiple repositories. The objective of SWORD is to lower the barrier for deposits in order to make the process as seamless

and simple as possible for researchers. In order for any of these protocols to work, systems must be configured in consistent, interoperable ways.

However, this is truly only a first step in interoperability. With constantly-evolving technology and new tools being developed all the time, repositories need to be able to support far more complex and innovative services that rely on additional requirements for interoperability. Much of today's work in this area is related to another kind of interoperability: semantic interoperability. Semantic interoperability means ensuring that the precise meaning of exchanged information is understandable in consistent ways; it enables the faithful exchange of meaning among machines and people.

This type of interoperability allows end users to look at data from multiple repositories and combine it in meaningful ways. Using some common terminology from repository to repository allows researchers to have a more consistent way to find and retrieve relevant items through searches. For example, using a consistent term for document types would allow searchers to more easily find the items for which they are looking and could allow for the creation of a tool that aggregates video collections from a set of repositories. Semantic interoperability standards such as the Resource Description Framework (RDF) are being implemented to express digital objects relationships in a machine understandable way, allowing machines to create sophisticated services over the global representation of knowledge distributed across repositories and other systems, to make cross-discipline connections, and to combine disparate findings to arrive at new insights.

Current Issues & Challenges with Interoperability

While the repository community has developed tools that address harvesting and content-ingest processes, a great deal of further work is still needed to address other existing issues as well as new challenges arising from the continually-evolving e-Research global infrastructure.

The distributed environment has created several challenges on both the technical level as well as administrative and organizational levels. Some current challenges where interoperability guidelines could lead to more consistency in data from different repositories and would provide support for repository managers to handle challenges in consistent ways:

Technical Challenges

- **New content types**. In addition to the texts of articles, new content types are being added to Open Access repositories: multimedia, data sets, software to process this data, simulations, computations models, graphics, automated analysis, etc. The content of repositories is increasingly becoming heterogeneous and dynamic, but none of the major interoperability guidelines address these new content types.
- **Software and systems**. While most repository installations use one of a handful of open source software systems (DSpace, ePrints, Fedora, or Invenio), the systems handle standard processes in different ways, leading to interoperability challenges. Furthermore, Open Access repositories are only one piece of the e-Research infrastructure – repositories should be interoperable with Open Access journals, research information systems (CRIS), Virtual Research Environments (VRE), e-Science infrastructures (i.e. grid computing), learning object repositories, repositories of Open Educational Resources

(OERs), and other types of digital collections such as library-run digital archives. As we move towards a truly global Open Access environment, should interoperability lead repository software development?

- **New service layers**. With an increase in Open Access mandates and policies at the institutional, research funding agency, and national levels, scholars are going to demand new services that make repository processes become more integrated into their workflows. As the corpus of Open Access materials becomes greater, scientists are also going to want to analyze and manipulate the body of data in new ways.
- **Usage data**. To allow for useful comparisons of systems, measure impact of individual contributions or develop new research metrics or monitoring tools, there is a need to aggregate and exchange consistent usage information (e.g. how many times particular items have been downloaded, viewed, and cited).
- **Consistent identification and terminology**. Consistency and interoperability regarding identification and naming (authors, items, institutions, funding agencies and grants) is a fundamental requirement to create real information spaces from distributed repositories.
- **Language challenges**. As Open Access becomes more prominent throughout the world, challenges with languages and different types of scripts are becoming more plentiful. Interoperability has the potential to support search and discovery through translation services.

Administrative and Organizational Challenges

- **Global context**. Currently, no global set of recommendations exists for exposing repository content or aggregating content from repositories on a global level.
- **Long-term sustainability of guidelines and standards**. Some guidelines have emerged from research projects, but once a funded research project is completed, guidelines become "orphans" without an organization to take continued responsibility for updates or maintenance. Since the e-Research environment is dynamic and rapidly changing, it is imperative that guidelines stay current and have an organization responsible for their long-term oversight.
- **Support for implementing guidelines**. As Open Access repositories become more tied to national policies and research funding, it will be necessary for smaller organizations with limited resources to implement and maintain repositories. While among COAR members there are good examples of institutional, national or regional support structures in place such as in Japan, China, Latin America and Europe, no such structure exists to support interoperability of repositories on a global level. Furthermore, following guidelines and standards, particularly when they are not built into repository systems, can become cost-prohibitive and extremely time-consuming as guidelines change and evolve.

Open Access repositories are a dynamic piece of the e-Research infrastructure being developed throughout the world. In order for this decentralized infrastructure to best support researchers and disseminate scholarship, repositories and other systems must follow interoperability guidelines. The need for interoperability will continue to grow as we move into the future.

COAR and Interoperability

COAR has identified "facilitating the discussion on interoperability among Open Access repositories and as part of a wider e-research infrastructure" as a key objective for 2010-2011. A working group has been created and charged with supporting this objective. The main steps envisioned are:

- Collect and analyze stakeholder input about the current interoperability environment. Sample questions: What are the major challenges not addressed by current interoperability guidelines? What problems exist for the research community that might be addressed by interoperability in the future? What problems related to interoperability act as barriers to progress? What should be the scope be for Guidelines issued, supported, and/or maintained by COAR?
- Conduct an environmental assessment. Review current guidelines, outdated and orphaned guidelines, guidelines under development, and technology changes that might have an impact on interoperability in the near future.
- Establish a COAR roadmap for repository interoperability based on stakeholder input, trying to maximize the (re)use of existing guidelines and best practices.

All stakeholders are invited to participate in the coming months by giving input to identify existing needs and challenges, work with COAR to help develop the roadmap, discuss the need for COAR guidelines, and participate in the process. We look forward to your participation.

Acknowledgements & Contributors

Editors

- Eloy Rodrigues (Universidade do Minho, Porgtual)
- Abby Clobridge (Clobridge Consulting, United States)

Experts & Reviewers

- Theo Andrew (EDINA, University of Edinburgh, United Kingdom)
- Andi Aschenbrenner (Goettingen State and University Library, Germany)
- Magchiel Bijsterbosch (SURF, The Netherlands)
- Stefan Buddenbohm (Goettingen State and University Library, Germany)
- Peter Burnhill (EDINA, University of Edinburgh, United Kingdom)
- Leonardo Candela (CNR, Italy)
- Pablo de Castro (Universidad Carlos II de Madrid, Spain)
- Donatella Castelli (CNR, Italy)
- Neil Jacobs (JISC, United Kingdom)
- Alicia López Medina (UNED, Spain)
- Norbert Lossau (Goettingen State and University Library, Germany)
- Anja Oberländer (University of Konstanz, Germany)
- Robin Rice (EDINA, University of Edinburgh, United Kingdom)
- Birgit Schmidt (Goettingen State and University Library, Germany)
- Friedrich Summann (University Bielefeld, Germany)
- Alma Swan (Key Perspectives Ltd., United Kingdom)

Spanish Translation

- Malgorzata Lisowska (Universidad del Rosario, Colombia)

Appendix 1: Resources for Further Reading about Open Access

- A Very Brief Introduction to Open Access by Peter Suber or Peter Suber's Open Access Overview
- Budapest Open Access Initiative (2001) – One of the original statements in support of Open Access, now an open document that can be signed by the public.
- Berlin Declaration of Open Access to Scientific Knowledge (2003) – A major international statement in support of Open Access. Organizations that sign the document commit to implementing this definition of Open Access.
- Directory of Open Access Repositories (OpenDOAR) – A directory of academic Open Access repositories.
- Open Access Scholarly Information Sourcebook (OASIS) – Practical steps for implementing Open Access.
- SHERPA JULIET – Summaries of funding agencies' grant conditions on self-archiving of research publications and data.
- Registry of Open Access Repositories (ROAR) – Information about the growth and status of repositories throughout the world.
- Registry of Open Access Repository Material Archiving Policies (ROARMAP) – Links to institutional and funder mandates and repositories throughout the world.

Appendix II: Key Interoperability Guidelines, Protocol, Projects & Resources

DINI Certificate for Document and Publication Services (2010) – A certificate that describes the technical, organizational, and legal aspects (including interoperability) that should be considered in setting up a scholarly repository service.

DL.org Community – Digital Library Interoperability, Best Practices and Modelling Foundations.

DRIVER Project – A two-phase project that established both the organizational and technical infrastruture of a repository network in Europe, and produced several studies and the DRIVER Guidelines.

euroCRIS – The European Organization for International Research Information has been working on interoperability between CRISs (Current Research Information Systems, designed to store data about current research activity, grants, researchers, etc. and Institutional Repositories.

Knowledge Exchange Usage Statistics Guidelines – Guidelines for the aggregation and exchange of usage data.

Open Archives Initiative – Object Reuse and Exchange (OAI-ORE) – Defines standards for aggregation of compound digital objects.

Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) – Protocol by which repositories allow for their metadata to be harvested and collected by other systems.

Publishing and the Ecology of European Research (PEER) – Project to "investigate the effects of large-scale, systematic depositing of authors' final peer-reviewed manuscripts."

PersID – Project to build a persistent identifier infrastructure for digital publications and electronic resources.

Resource Description Framework – A standard model for web-based data interchange.

Scholarly Output Notification and Exchange (SONEX) – An initiative supported by JISC to identify and analyze use cases for deposit of research papers into repositories.

About COAR

COAR, the Confederation of Open Access Repositories, is a young association of repository initiatives launched in October 2009, uniting over 80 members and partners from 24 countries from throughout Europe, Latin America, Asia, and North America. Its mission is to enhance greater visibility and application of research outputs through global networks of Open Access digital repositories. More information about COAR and its members is available at the COAR web site <http://www.coar-repositories.org>.

Contact Information

COAR Office
Dr. Birgit Schmidt
Katharina Müller
c/o Goettingen State and Univesity Library
Platz der Gottinger Sieben 1
37073 Goettingen
Germany
Phone: +49 551 39-22215
Fax: +49 551 39-5222
coar-office@sub.uni-goettingen.de