OR2024 Presentation Submission

The 19th International Conference on Open Repositories, June 3-6th 2024, Göteborg, Sweden

How to make repository content indexed and discoverable

Petr Knoth, KMi, The Open University, petr.knoth@open.ac.uk

Martin Klein, Los Alamos National Laboratory, mklein@lanl.gov

George Macgregor, University of Glasgow, george.macgregor@glasgow.ac.uk

Matteo Cancellieri, KMi, The Open University. matteo.cancellieri@open.ac.uk

Paul Walk, Antleaf Ltd., paul@paulwalk.net

Abstract

Millions of users access scholarly indexes each month. A solitary repository, if not correctly indexed, will languish alone. According to the Confederation of Open Access Repositories (Rodriguez, 2011), 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. However, even today, many repositories are not yet configured in a way that would enable their resources to be comprehensively indexed in scholarly infrastructures.

Ensuring the discoverability of repository research outputs is crucial for:

- Maximizing the impact and reach of scholarly works
- Making research visible to the general public
- Facilitating research collaborations
- Delivering on the open science mission
- Enabling systematic literature reviews
- Monitoring compliance with funding agency policies
- Promoting reproducibility
- Fostering innovation

We propose a panel session that will provide a set of practical wide-ranging recommendations for repositories to enable, validate and monitor the indexing of their repository content. By implementing the wide-ranging recommendations and principles discussed in this panel session, repositories will be able to markedly improve their content's discoverability.

Keywords

Discoverability, indexing, scholarly search, harvesting, interoperability, aggregation, machine access

Audience

Repository managers, developers

Proposal

Even in 2024, the content in most repositories is not yet comprehensively included in the widely used scholarly indexes. This makes repository content insufficiently discoverable and thus the knowledge contained within these repositories is not disseminated to its full potential.

The objective of this panel is to discuss the broad range of actions and principles that repositories can employ to ensure their content is accurately indexed within widely used scholarly infrastructures.

Former efforts in this area, presented at previous iterations of the Open Repositories conferences, were typically limited to providing guidelines on the appropriate use of metadata schemas within repositories. While we acknowledge that compliance with metadata standards is crucial, it is, on its own, not sufficient to guarantee that content from repositories can be comprehensively indexed. In some countries, e.g. in the UK and US, there is a perception that repository resources are poorly discoverable. Some repositories try to address this by minting article-level PIDs (often DOIs) for repository records, partially shifting the discoverability responsibility to 3rd parties, especially Crossref and DataCite. Others fundamentally disagree with this approach, on the basis that it results in potentially multiple DOIs minted for the same scholarly resource, or find this to be an expensive, time-consuming or undesirably centralised undertaking for repositories.

To make repository resources discoverable, it is essential to make them friendly to machine agents (robots, crawlers, harvesting systems, etc.). This goes beyond just supporting a particular metadata schema or minting article-level PIDs. It encompasses a range of aspects including:

1. Discoverability of the repository and its affordances: repository registration, FAIRiCat
2. Machine agent access to the repository: e.g. robots.txt, meta-tags
3. Prioritizing ongoing improvements to repository user experience as measured by machine agents
4. Support and adoption of harvesting protocols: e.g. OAI-PMH, ResourceSync, Sitemaps
5. Linking metadata to research outputs (full text, dataset, software and other resources): e.g. Signposting, Rioxx v3
6. Validation and monitoring: Continuously ensuring that any given repository setup remains functional for machine agents. Available tools include the CORE Repository Dashboard.

The objective of this panel is to provide a forum at the Open Repositories conference to discuss the wide range of actions that repositories can undertake to rapidly increase their chances for comprehensive indexing by scholarly infrastructures.

The panel will bring together the following experts, each of whom has a unique perspective on this problem:

Martin Klein: a researcher at the Los Alamos National Laboratory who was instrumental in the creation of several widely used repository standards, such as Signposting, ResourceSync and COAR Notify, will reflect on his experience of how these standards should be used by the community to expose resources in an interoperable manner.

Petr Knoth: Head of CORE, a widely used repository aggregator, will cover the recent creation of the CORE Data Provider's Guide (https://core.ac.uk/documentations), which aims to assist repository managers in configuring their repositories for successful harvesting.

George Macgregor: Assistant Director - Digital Library, University of Glasgow will reflect on his extensive experience as a repository manager on what repositories can do to achieve comprehensive indexation in scholarly search engines.

Paul Walk: The founder of "Rioxx: The Research Outputs Metadata Schema", will reflect on his experience of designing metadata standards in a way that makes the indexing of repository content feasible for machine agents. He will also mention his efforts towards building a global community-governed repository registry.

Matteo Cancellieri: Lead developer at CORE, will introduce validation and monitoring tools available to repository managers. He will argue that proactive validation and monitoring are essential for the successful indexing of global repository content.

Each of the panelists will initially reflect and give insights on a specific aspect of this problem providing concrete guidance to repository managers. The group will then discuss the challenges that individual repositories face and will highlight gaps that still need to be addressed at the level of the open repositories community.

References

1. Knoth, P., & Zdrahal, Z. (2012). CORE: three access levels to underpin open access. D-Lib Magazine, 18(11/12), 1-13.
2. Knoth, P., Herrmannova, D., Cancellieri, M. et al. CORE: A Global Aggregation Service for Open Access Papers. Nature Scientific Data 10, 366 (2023). https://doi.org/10.1038/s41597-023-02208-w
3. Eloy Rodrigues and Abby Clobridge. 2011. The case for interoperability for open access repositories. Working Group 2: Repository Interoperability. Confederation of Open Access Repositories (COAR).s/rrs.html