Preprints 2022 der Zeitschrift *BIBLIOTHEK – Forschung und Praxis*, BFP.2022.0050 **Shearer**

Kathleen Shearer

Repository Networks: Past, Present and Future

Abstract: Repositories represent critical infrastructure for open science / open scholarship. They preserve and provide access to a wide range of valuable research and educational resources, including articles, pre-prints, research data, images, software, and so on. However, the value of any individual repository is greatly enhanced when it becomes part of a distributed network, enabling resources to be more discoverable, linked with other related content, and part of the larger international corpus of international research. This article discusses the evolution of repository networks since the early 2000s and presents the current vision of the COAR Next Generation Repository Initiative to significantly enhance the networking capabilities of repositories.

Keywords: Next generation repositories, aggregators

1 Introduction

Repositories represent critical infrastructure for open science / open scholarship. They preserve and provide access to a wide range of valuable research and educational resources, including articles, pre-prints, research data, images, software, and so on. And when resourced properly, repositories are sustainable and long-lived services because they are, for the most part, collectively managed by research institutions and their libraries.

While digital repositories have been part of the landscape for decades, it was only in the early 2000s that the notion of the institutional repository (IR) was proposed, setting the scene for a significant increase in the number of open access repositories across the world. IRs were positioned as tools that would both transform the scholarly communication system, as well as offer a window to the output of an institution. Cliff Lynch, in his seminal piece about institutional repositories wrote,

> The development of the institutional repository emerged as a new strategy that allows universities to apply serious, systematic leverage to accelerate changes taking place in scholarship and scholarly communication, both moving beyond their historic relatively passive role of supporting established publishers in modernizing scholarly publishing through the licensing of digital content, and also scaling up beyond ad-hoc alliances, partnerships, and support arrangements with a few select faculty pioneers exploring more transformative new uses of the digital medium. (Lynch 2003)

Early on in the evolution of institutional repositories, it was clear that the value of any individual repository could be greatly enhanced when it becomes part of a distributed network, enabling resources to be more discoverable, linked with other related content, and part of the larger international corpus of international research. A 2011 report by the Confederation of Open Access Repositories (COAR) articulated the case for the networked repository, "The real value of repositories lies in their potential to become an interconnected repository network – a network that can provide unified access to an aggregated set of scholarly and related outputs that machines and researchers can work with in new ways." (COAR 2011). However, despite this value proposition, the networking aspect of repositories has advanced relatively slowly.

2 The technical evolution of the network

2.1 OAI-PMH

To date the main 'networked' aspect of repositories has been made possible through metadata harvesting using the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)[1], a Open Archives Initiative developed by Herbert Van de Somple and colleagues at Los Alamos National Laboratory. OAI-PMH, which is still widely used today, supports the aggregation of metadata across distributed repositories and other content providers by harvesting services, enabling searching and discovery of repository items via these metadata aggregations.

There are currently a number of aggregation /services in the scholarly communications landscape that use this approach. In Europe and Latin America, in particular, several countries maintain their own national aggregators in order to better track research outputs. International aggregators such as BASE[2], CORE[3] and OpenAIRE[4], as well as domain aggregators, also offer a variety of discovery and other services to the community. These network services then identify connections and relationships across records in the aggregation, creating what is commonly referred to as a research graph, enabling better discovery and analysis of various aspects of the research ecosystem. These research graphs are further enhanced by the availability of full text content (openly available on the web) that enables aggregators to text and data mine for other characteristics/relationships that may not be available in the metadata.

Repository aggregator services offer a variety of services that add value to the individual repository collections. The most basic use case is to improve discovery, by enabling users to search and browse aggregate records collected from disparate data providers (repositories, journals, indexes, etc.). More

[1] Open Archives Initiative Protocol for Metadata Harvesting: https://www.openarchives.org/pmh/.
[2] Bielefeld Academic Search Engine: https://www.base-search.net/.
[3] CORE: https://core.ac.uk/.
[4] OpenAIRE: https://www.openaire.eu/.

recently we have seen the development of more value added services, such as the CORE recommender service, which provides a plugin for repositories, journal systems and web interfaces to suggest similar articles; comparable metrics services, such as the IRUS UK[5] service which aggregates usage statistics, enabling UK universities to share and compare information about usage of items in their institutional repositories based on

the COUNTER[6] standard. Dashboards are also a relatively new service offering such as the OpenAIRE dashboards for content providers, research communities and funders. These dashboards offer specialized views of a subset of the content in their research graph (aggregation) related to the specific organization or community.

2.2 The Next Generation Repository

In 2016, COAR launched the Next Generation Repositories (NGR) Working Group to identify which called for new levels of web-centric interoperability. The vision outlined by the NGR working group was to position repositories as the foundation for a distributed, globally networked infrastructure for scholarly communication, on top of which layers of value added services will be deployed. The report goes on to explain,

> The vision rests on making the resource, rather than the repository, the focus of services and infrastructure. Rather than relying on imprecise descriptive metadata to identify entities and the relationships between them, our vision relies on the idea inherent in the Web Architecture, where entities (known as "resources") are accessible and identified unambiguously by URLs. In this architecture, it is the references which are copied between systems, rather than (as at present) the metadata records. (COAR 2017)

In the next generation repository worldview, the existing networking approach of harvesting metadata will be greatly enhanced by introducing bi-directional links as a result of an interaction between resources in different repositories.

Building on the recommendations of the Next Generation Repository Initiative, in July 2021, COAR repositories. This would introduce bi-directional links and add another dimension to the networking potential of repositories by enabling platforms (and individual resources in those platforms) to interact with each other directly, with no reliance on centralized services, such as aggregators. As outlined in the document that outlines this conceptual model for this approach, *Modeling Overlay Peer Review Processes with Linked Data Notifications*,

[5] IRUS UK: https://core.ac.uk/services.
[6] COUNTER: https://www.projectcounter.org/.

> The resource-oriented nature of the Web is well suited to an environment which places value in the fact that control of resources is distributed across a large number of repositories. In such an environment, it makes sense to take a pass-by-reference approach to interaction between different networked services, rather than relying on machine or human mediated processes to pass copies of resources around the network. (Walk et al. 2020)

Linked Data Notifications are a W3C standard protocol that "describes how servers (receivers) can have messages pushed to them by applications (senders), as well as how other applications (consumers) may retrieve those messages. Any resource can advertise a receiving endpoint (Inbox) for the messages. Messages are expressed in RDF and can contain any data" (WC3 2017). This resource-oriented approach will allow any compatible service (e.g. repository) to communicate with any other compatible service (e.g. peer review service) enabling the linking of related resources across distributed platforms. Currently, the COAR Notify Project[7] is working with a number of implementing partners to adopt linked data notification functionality and define a standard vocabulary for exchanging information across different platforms. The aim is to support a wide variety of use cases, but with an initial focus on connecting articles in repositories and archives with peer reviews undertaken by open peer review services and overlay journals.

3 Networks are more than technologies

Repository networks can be about a lot more than just technologies and services. While the specific technologies adopted by repositories are critical for the network, there is also an important non-technical component that should not be overlooked. Building a community around the services and technologies can significantly increase the value and impact of the network. Some repository networks, such as OpenAIRE, have recognized this important human aspect to the networks and actively seek to nurture the repository community through a variety of activities. A recent COAR-SPARC Expert Group[8] developing a vision for a US network of repositories identified a large number of possible 'features' of a repository network, including services such as discovery and shared infrastructure, but also a host of other characteristics such as communities of practice, common standards, collective investments, and advocacy. In particular, it is extremely helpful to have a compelling vision across the network about shared understanding about its role in the ecosystem. This can bring the community together and act as an incentive for repositories to engage more actively in the network services.

Given that many aggregators are heavily reliant on metadata, high quality and standardized metadata

[7] COAR Notify Project: https://www.coar-repositories.org/notify/.
[8] Catalyzing the creation of a repository network in the US: https://sparcopen.org/news/2021/catalyzing-the-creation-of-a-repository-network-in-the-us/.

contributes to the quality of services this type of network can offer. This includes agreeing on a common format for exposing the metadata, the introduction of Permanent Identifiers (PIDs) and the use of controlled vocabularies, which ensure the same word is used to mean the same thing. New communities are being formed to collaborate on metadata and data curation activities across institutions. The Data Curation Network (DCN), for example, is a group of "professional data curators, data management experts, data repository administrators, disciplinary scientists and scholars that represent academic institutions and non-profit data repositories" (Data Curation Network 2022). DCN brings together a pool of data curation experts to curate a wider variety of data types, bringing in subject expertise and expanding curation efforts beyond what any single institution might be able to offer.

4 Interoperability

Interoperability across repositories is a fundamental characteristic of a repository network. Since its inception in 2009, a major priority for COAR has been to advance repository interoperability. A 2012 COAR report argues that to fully optimize repositories, they must be interoperable, "specifically, that repositories follow consistent guidelines, protocols, and standards which allow them to communicate with each other; connect with other systems; and transfer information, metadata, and digital objects between each other" (COAR Working Group 2: Repository Interoperability 2011). Specific priority areas for repository interoperability were further articulated in the 2015 COAR report, *Future Directions for Repository Interoperability*, prepared by Freidrich Summann and Kathleen Shearer (COAR Working Group 2: Repository Interoperability 2015)

In a similar sense repository aggregators should also be interoperable. Current aggregator services differ in their scope, operations and functionalities, and often have distinct requirements based on their specific user communities and jurisdictions. However, in order to support the global nature of research, collaboration and dialogue across networks is critical. As such, COAR has also been working to foster greater cooperation between repository networks by identifying common principles and areas of collaboration that can lead to concrete actions and the development of global services. The aim is to ensure a general level of interoperability, while still supporting the unique functions and use cases of each network.

5 Discussion

Despite the obvious value proposition for networks in the repository landscape, there are a number of challenges that have contributed to their relatively slow progression. Because institutional repositories were initially conceived as "digital collections that capture and preserve the intellectual output of university communities" (Crow 2002). This may have led to the

emphasis on the value of IRs to the institution, as opposed to the repository as a node in a broader network. In addition, the technical architecture of repositories has likely contributed to the stand alone nature of the services. Repositories were originally designed as digital libraries and content management systems, rather than resources-oriented, and therefore did not reflect the highly networked architecture of the web. As such, the networking functionalities of repositories have only been incorporated incrementally into the platforms over time, leading to relatively slow advancement over the past 20 years. Contributing to this issue, is that the communities of practice around repositories are mainly based on the software platforms, creating software siloes in the repository landscape, making interoperable networking functionalities more difficult to implement.

The role of COAR in bridging these silos has become apparent and is increasingly crucial to help advance the role of repositories in the ecosystem, and enhance the networking capacity of repositories in an interoperable manner. This has been demonstrated by the widespread adoption of recommendations of the COAR Next Generation Repository Initiative in the different repository platforms, and subsequently with the widespread community interest in adopting the linked data notification technologies as proposed through the COAR Notify Project.

Undoubtedly, technologies will continue to evolve, as will the ways in which repositories are networked with each other. To that end, COAR will seek to maintain its role in progressing platform agnostic interoperability and the networking of repositories worldwide.

This article is dedicated to Freidrich Summann in thanks for his many contributions to the repository community throughout his career. Through his work with BASE and COAR, he has played a significant role in advancing the role of repository networks and repository interoperability over the last two decades.

References

COAR Next Generation Repositories website 2017. *Confederation of Open Access Repositories*. https://ngr.coar-repositories.org/ (accessed on February 1, 2022)

COAR Working Group 2: Repository Interoperability 2011. The Case for Interoperability for Open Access Repositories. *Confederation of Open Access*

*Repositories*https://www.coar-repositories.org/files/A-Case-for-Interoperability-Final-Version.pdf

COAR Working Group 2: Repository Interoperability 2015. *Future Directions for Repository Interoperability. Confederation of Open Access Repositories*

https://www.coar-repositories.org/files/Roadmap_final_formatted_20150203.pdf (accessed on February 1, 2022)

Crow, Raym. 2002. The Case for Institutional Repositories: A SPARC Position Paper. *SPARC*. https://ils.unc.edu/courses/2014_fall/inls690_109/Readings/Crow2002-CaseforInstitutionalRepositoriesSPARCPaper.pdf (accessed on February 1, 2022)

Data Curation Network website. 2022. https://datacurationnetwork.org/ (accessed on February 1, 2022)

Lynch, Cliff. 2003. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. Association of Research Libraries, 226.

W3C. 2017 Linked Data Notifications W3C Recommendation 2 May 2017. https://www.w3.org/TR/ldn/ (accessed on February 1, 2022)

Walk, Paul, Martin Klein, Herbert Van de Sompel, & Kathleen Shearer. 2020. Modeling Overlay Peer Review Processes with Linked Data Notifications. *Confederation of Open Access Repositories.*

https://www.coar-repositories.org/files/Modelling-Overlay-Peer-Review-Processes-with-Linked-Data-Notifications.pdf (accessed on February 1, 2022)

Kathleen Shearer

0000-0001-8617-5781