Open Scholarship Initiative Proceedings, Volume 2, 2017, issn: 2473-6236
doi: http://dx.doi.org/10.13021/G8osi.1.2017.1291

Institutional Repositories Workgroup

Geraldine Clement-Stoneham, Najko Jahn, Catherine Mitchell, Jake Orlowitz, Dave Ross, William Simpson, and Andrew Tein

Summary

Our task in the second OSI convening of the institutional repository workgroup was to propose a way forward for repository and infrastructure solutions: detailing what's needed before action can be taken, what this action should look like, and what actors should be involved.

Our main recommendation is directional: repositories must evolve and move toward interoperability and sustainability.

- Repositories should be diverse, decentralized, interoperable networks across the world.
- It is time for repository staff to shift focus more towards interoperability (policy-driven, research-relevant, and standards-based) and less on supporting content.
- The scholarly communication community should be incentivized to make choices related to repositories that are more sustainable.

The scope and power of OSI lies in clarifying what this means and coordinating (or suggesting coordination) among existing stakeholders. OSI is not currently able to sustain, support, or itself build the solutions.

Background

Institutional repositories are not a new phenomenon in open scholarship; they have been in use at academic institutions for nearly two decades. According to Peter Suber's seminal work on Open Access, institutional repositories are online databases of open access works, which aim to host the research output of an institution. This includes, but is not limited to, self-archived copies of peer-reviewed journal articles, books, book chapters, technical reports, theses, digital collections, research data or scientific code from all subjects represented at an academic institution. Institutional repositories differ from disciplinary repositories such as ArXiv or PubMed Central, which serve research outputs from specialized academic fields. They also vary from output-specific repositories such as systems designed for research data.

Institutional repositories can be found worldwide. In April 2017, more than 3,000 institutional repositories were listed in the Registry of Open Access Repositories. In total, we identified 109 countries with institutional repositories. Although nearly 20% of the institutional repositories are operated in the US, our data suggest that institutional repositories are global phenomena in use throughout Asia, Australia, Europe and the Americas (see Figure 1).

This document reflects the combined input of the authors listed here (in alphabetical order by last name) as well as contributions from other OSI2017 delegates. The findings and recommendations expressed herein do not necessarily reflect the opinions of the individual authors listed here, nor their agencies, trustees, officers, or staff.

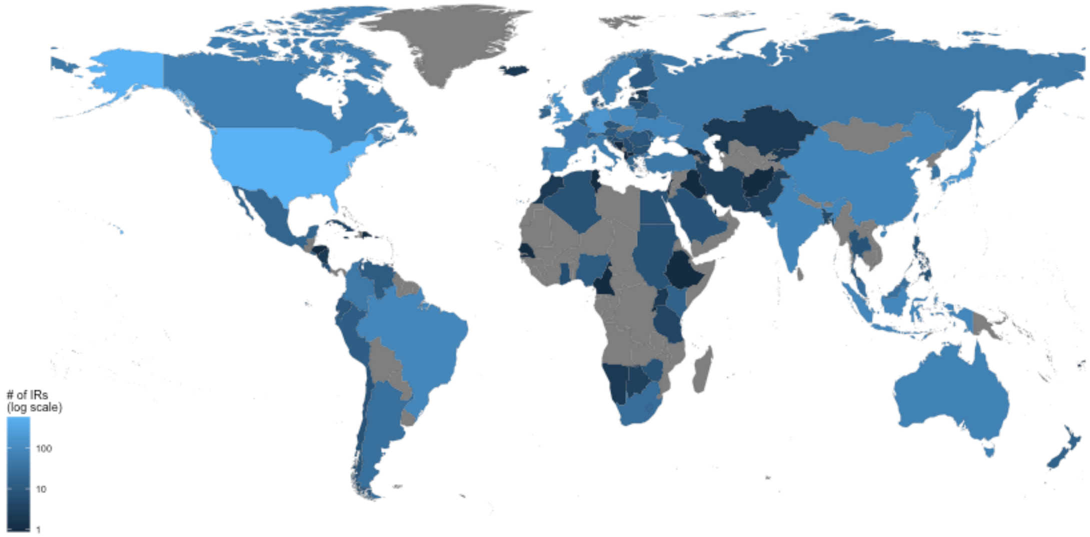

Figure 1: Global distribution of institutional repositories per country. For countries that are colored grey, no institutional repository could be found. Data were gathered from the Registry of Open Access Repositories (ROAR), April 19, 2017.

To position institutional repositories in open scholarship is difficult because there are multiple stakeholders in the repository ecosystem, which leads to a diverse landscape of repository implementations in general, and various conceptions about the role and perspectives of institutional repositories. More specifically, we mapped the following repository stakeholders:

- Governments
- Funders
- Publishers
- Institutions
- Libraries
- Disciplines
- Scholars

Crucially, the incentives that drive decision-making by these stakeholders vary by group and often don't even overlap. Even the repositories themselves are not homogenous or monolithic; there are multiple different types of repositories, as evidenced in the rough typology below:

IR	(campus-based, research org based, consortial)
Subject Preprint	(arXiv, SocArXiv, BioaRxiv)
Discipline	(Humanities Commons, MLA Core, PMC)
Funder	(Gates Open Research, Welcome Open Research)
National	(CRIStin - Norway National Library)
International	(SciELO)
For-profit	(Academia.edu, ResearchGate - interdisciplinary)
Long-tail	(Zenodo)
Data	(Dryad - multidisciplinary)
Networks	(SHARE, OpenAire, LaReferencia, HAL)

Current motivations and challenges for institutional repositories

During our workshop, we determined that the several, sometimes conflicting, motivations for institutional repositories discussed in the literature and among practitioners and policy makers must be clarified. Institutional repositories not only vary by type, but also by the function they have in open scholarship. Accordingly, institutional repositories contain a multitude of goals:

Shop window

Since the advent of institutional repositories, one of the rationales for these archives has been to provide a single point of access to the intellectual output of an academic institution. Many institutional repositories therefore aim at demonstrating the unique value of the institution by providing unified access to the scholarly publications of their faculty and students. Consequently, operators of institutional repositories often share metrics about activity, media coverage and usage. One example is MIT's institutional repository, which prominently presents media coverage of discussion papers and other open access content being made available via DSpace@MIT. Harvard's DASH repository shares user stories and usage statistics online.

Preservation

An essential role of institutional repositories is to preserve publications and thus the intellectual output of academic institutions. Standardized technical and organizational means for making content available in the long-term exist both within and across institutions. In the latter case, national libraries as well as lightweight preservation networks based on the LOCKSS technology, such as the international SAFE-PLN network, address at scale institutional repositories' mission for long-term preservation.

Open Access Policy Implementation and Assessment

Open access policies from academic institutions often require the deposit of publications in institutional repositories and funders' mandates often rely on these online archives to make research outputs freely accessible. One prominent example is the European Union's (EU) research and innovation framework HORIZON 2020, wherein grantees are not only required to deposit their EU-funded publications into eligible repositories, but the EU also funds the Open Access Infrastructure for Research in Europe (OpenAIRE).

OpenAIRE is a network of repositories and other scholarly communication services aimed at the implementation and assessment of EU's open access policies. Institutional repositories also participate in this network on the basis of shared standards and services at the European level. Another example is the UK's Higher Education Funding Council for England (HEFCE), which from April 1st, 2016, requires that all research articles published by UK-based researchers be deposited in the relevant IR and made OA (respecting any embargo periods) with discoverable metadata if they are to be considered for the periodic Research Assessment Framework.

Alternative publishing platform

Institutional repositories can provide faculty with alternative means to publish their research. The most common examples of primary publications via institutional repositories include robust OA journal publishing programs, as well as working paper series. The journal programs provide support for publications that don't fit neatly into traditional publishing venues: those within emergent fields, cross-disciplinary domains, disciplines that include non-academic practitioners, etc. Journal programs also support publications that seek local control of the editorial process and are, frequently, committed to Open Access. The motivation behind working paper series is early and rapid dissemination of research findings in lieu of the long time lag from submission to journal publication.

Discoverability

Because institutional repositories are globally distributed, a growing number of mechanisms have been developed to unify access to OA works deposited in these repositories. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), first introduced in 2002, has had a huge impact within the repository community, sharing metadata about repository collections. It has motivated national and continental repository networks as well as global discovery solutions such as the Bielefeld Academic Search Engine (BASE) that indexes more than 100 million scholarly records from open access sources. In recent years, it has become more important for institutional repositories to adapt new web technologies in order to make content discoverable through large search engines. Google Scholar, for instance, indexes institutional repositories when they satisfy technical as well as content-related criteria. Recently, open data corpuses of repository collections have enabled new discovery solutions such as the Open Access Button and Unpaywall.

Data sharing

Against the background of the increasing call to share and archive research data and code, some institutional repositories have also started to provide services for these research outputs. These services mainly address long-tail research data, i.e., data that may not be covered by existing disciplinary data repositories or data within disciplines that have not yet established domain-specific repositories.

Research Corpus

Given the various content types and multidisciplinary coverage of institutional repositories, well-curated, standardized, and interconnected institutional repositories have the potential to become a research corpus for a broad range of scholarly studies. These coordinated repositories could complement existing literature databases with selective indexing coverage such as the Web of Science or Scopus as well as full-text corpora for text and data mining. However, so far little evidence about the coverage of institutional repositories in comparison to these databases exists.

Our proposal to move forward

The repositories workgroup explored the ideal mode of institutional repository interoperability, given the worldwide distribution, broad group of stakeholders, and sometimes disparate goals of these repositories. In thinking through these challenges to interoperability, we borrowed a framework from network theory to envision what the future interconnectedness of libraries might and should look like. Imagine a spectrum: on one end is a fully centralized network with a single main node to which all others connect; on the other end is a fully distributed network where no node has any more connections than another. In between is a decentralized network with multiple key nodes, which have more connections than others, and which connect among themselves as well.

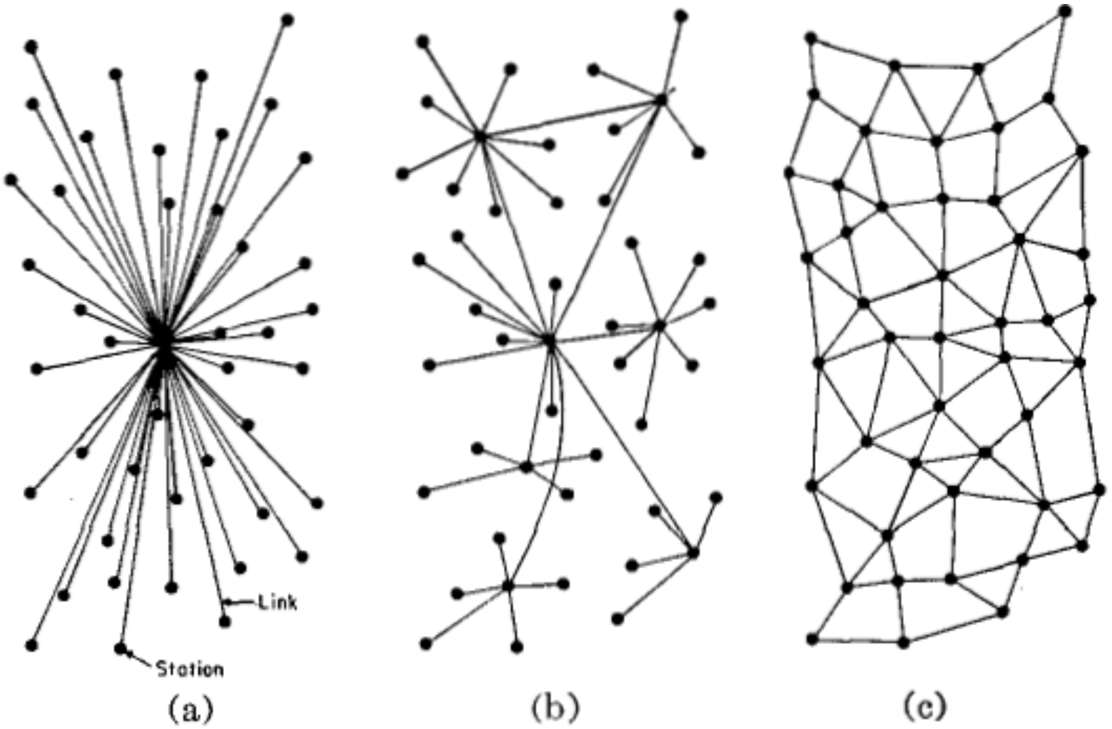

Fig. 1—(a) Centralized. (b) Decentralized. (c) Distributed networks.

Illustration from *On Distributed Communication Networks*, Paul Baran (1964).

We view a centralized repository network as politically and technologically difficult. They are also vulnerable to a single point of failure. Alternatively, the fully distributed network lacks centers of influence and dissemination, and requires full interoperability. Our "Goldilocks" version is the decentralized model with hubs and spokes that allow for some efficiency while still recognizing the importance of quasi-local forms of centralization. Given this proposed model, it is essential, we believe, for the Open Scholarship Initiative to identify the potential (sub-) networks as well as the nodes in a network of repositories.

For the next step, we advocate to "convene the conversation" with major stakeholders at the table: e.g., COAR, HathiTrust, Publishers, Libraries, Funders, Researchers, etc.

- Key questions to address include:
- What problems are repositories trying to solve?
- What repository behavior would we like to see? Why? How can we work together to incentivize it?
- How can we attend to different scholarly communication needs across different fields?
- How can we make everyone accountable: publishers, libraries, funders, and researchers?
- How can we achieve a sustainable, decentralized, networked system while gaining efficiency through higher levels of aggregation?
- How do we minimize waste and maximize value in the repository ecosystem?

We thus recommend that a meeting of the willing be held, under UNESCO's authority, to which umbrella organizations (e.g. COAR), publishers (commercial and scholarly), academic library consortia, and non-academic information producers (e.g. Wikimedia, Open Knowledge) are invited. We also assert that geographically diverse research organizations such as the Global Young Academy and representatives from the Global South must be involved to reflect the expansive landscape of repositories.

Such a meeting seems a necessary first step in affecting change within the world of repositories, many of which languish individually with insufficient resources but could, in concert, create a powerful and efficient worldwide hub of openly discoverable and accessible information.

Institutional Repository Workgroup

Geraldine Clement-Stoneham, Knowledge and Information Manager, Medical Research Council, RCUK
Najko Jahn, Scholarly Communication Analyst, University of Gottingen
Catherine Mitchell, President, Library Publishing Coalition and Director, Access & Publishing Group, California Digital Library
Jake Orlowitz, Head of The Wikipedia Library, Wikimedia Foundation
Dave Ross, Executive Director, Open Access, SAGE Publishing
William Simpson, Associate Librarian and Institutional Repository Librarian, University of Delaware
Andrew Tein, Vice President, International Government Partnerships, Wiley