Open Scholarship Initiative Proceedings, Volume 1, 2016
doi: http://dx.doi.org/10.13021/G89W24

Report from the Repositories & Preservation Workgroup

Joyce Backus, Robert Cartolano, Christina Drummond, Agathe Gebert, Brooks Hanson, James Hilton, Maryann Martone, Sarah Michalak, Richard Ovenden, Sarah Pritchard, Rita Scheman

Abstract

Repositories are a vital tool in modern information management and a key component of preservation and long-term availability. They are not well-suited, however, to the current challenges posed by our information-rich society and the multitude of stakeholders involved in the modern scholarly publishing system. Strengthening repositories and standardizing preservation processes are critically important. This challenge will require not only leading multiple stakeholder groups but also reforming multiple information systems, architectures, philosophies, practices, and more.

OSI2016 Workgroup Question

Are we satisfied with the current state of global knowledge preservation? What are the current preservation methods? Who are the actors? Is this system satisfactory? What role do institutional repositories play in this process? What does the future hold for these repositories (taking into account linking efforts, publishing company concerns about revenue declines, widespread dark archiving practices, and so on)? Would new mandates help (or do we simply need to tighten existing mandates so they actually compel authors to do certain things)? And how do versions of record figure into all of this—that is, how do archiving policies (with regard to differences between pre-journal and post-journal versions) affect knowledge accuracy and transfer? How can digital preservation advance open scholarship?

Not so long ago, a sense of order existed in the system of scholarly repositories and preservation. Scholars wrote articles and books, publishers published them, and libraries provided long-term access and preservation.

This concept has been severely challenged in recent years by a landscape of information flow that has evolved to become increasingly complex. Today there are thousands of isolated repositories, susceptible to multiple points of failure, ranging from technological breakdowns to organizational issues, as well as potential geographic and institutional catastrophes. There are countless new roles and stakeholders. Multiple new forms of scholarly artifacts are being enabled by digital technologies. Technologies have proliferated along with development communities and commercial players. There are islands of best practice, often unidentified by the repository community. Although efforts toward coordination have been attempted (e.g., the Distributed Digital Preservation

Framework Working Group and the Digital Preservation Network), sustained, continued coordination across these efforts has been lacking. Criteria for the curation of content and collections hosting varies among institutions and may not even include preservation as a priority. Terminology is inconsistent and used in different ways by different constituencies: the very terms "repository," "preservation," "access," and even "publishing," cannot be assumed to have a common point of reference in the many relevant discussions and applications.

What's at risk from this chaos? Everything. Without better organization in and between repositories, we risk entering a world where our ever-increasing flood of information is misfiled, disconnected, even lost. While Internet search engines give us the illusion that everything is findable and accessible, a large percentage of the content of repositories is not adequately discoverable with today's search engines. Strengthening repositories and standardizing workflows must be among our highest-priorities in scholarly publishing reform. But how can we even begin to think about improving such a complex system?

Defining the landscape

First, for our purposes here, the scope of this conversation is limited only to institutional or disciplinary repositories that are connected to scholarly publishing, such as the institutional repositories (or IRs) that university libraries maintain. In the scholarly publishing world, these repositories are storage boxes for information with multiple functions, workflows and relationships, and that—ideally anyway—operate under standards and other best practices in order to support the preservation of and access to the copious amounts of research information produced in the academic world today. In addition to possessing large amounts of durable storage space, these repositories are expected to safeguard this information through backups and quality controls, include metadata (data about the data, such as author, date of publication and file size/type) and provide at least some level of certification—for example, digital object identifiers (DOIs), and provenance or versioning information (giving the origin and history of a piece of information). They are also expected to remain well managed over a long period of time (meaning that they require professional stewardship and plans for long-term sustainability) while also supporting needs and expectations regarding access and interoperability (with other systems and repositories) as these needs and expectations evolve over time. To recap, repositories should:

- reliably store and backup information
- offer a browsing through the information stored
- operate according to best practices
- include metadata and digital identifiers
- certify information (by tracking history)
- be well-managed and well-funded (ensuring long-term sustainability)
- be responsive to the access needs and expectations of users and other repositories.

Preservation is a function of some (but not all) digital repositories and its purpose is the long-term protection of an object to ensure its integrity and accessibility for future use. The preservation function:

- implies that the content of the preservation archive has been chosen by a knowledgeable curator
- usually implies that if the archive is not dark (i.e. access is either limited to certain individuals or completely restricted to all) most of the contents will be free and openly accessible.

OSI2016 asked the workgroup to consider challenges of "repositories" and "preservation" in tandem, although preservation of scholarly materials also occurs outside of repositories. For the purposes of OSI2016, our team limited its discussion to open preservation repositories for scholarly research output. Some of the scholarly research content that is currently included in these repositories is listed on the following page (see table 1), along with notations regarding whether preserving this content is currently mandatory for ensuring integrity and reproducibility; is in the process of being identified for inclusion by funder guidelines and community and discipline standards; or is not currently being considered for inclusion.

The challenges

Stakeholders in the broad scholarly publishing system—authors, researchers, universities, libraries, publishers, and others—are often not clear about where they fit into this landscape of preservation and repository or about what their roles are supposed to be. Indeed, these stakeholder groups often have unique and conflicting viewpoints.

Our workgroup identified a small subset of the challenges and opportunities related to the preservation of the scholarly record and the roles of repositories. Juxtaposed against what we described earlier as the ideal of a scholarly research repository, the current overlapping and conflicting environment instead leads to an utter lack of coordination, which results in the loss of key data and software, a failure to ensure long-term preservation, and a lack of dependable means to retrieve information in these repositories.

This chaos is evidenced through three main symptoms of dysfunction:

1. Redundancy of effort
2. A lack of coordination and standards
3. A lack of sustainability.

The first symptom, redundancy, needs little explanation. There is a high level of unintended redundancy of content between repositories, resulting in vast amounts of duplicated effort in categorizing and re-categorizing the same information artifact across multiple storage locations.

The second symptom, a lack of coordination and standards, is evidenced in a variety of ways:

- There are no standards for what basic kinds of information repositories should capture, what priority this information should have, and what structure repositories should use for storage and retrieval (repositories differ from one another based on software choice—Fedora or D-Space, for example).
- Collections themselves are driven by the unique goals and values of individual institutions. Therefore, content choices are usually local and themes and formats of content will vary widely among and between repositories.

Table 1: Deposit imperatives in scholarly research

		Deposit imperative			
Scholarly research output	**Specific content**	**Required or high priority**	**Currently being identified**	**Not yet required**	**Notes**
Peer-reviewed publications (papers and monographs)	Version Of Record (VOR)	x			Includes material with DOI, errata, etc.
	Earlier VORs			x	
	Accepted articles post peer review (AAM) (for papers)	x			Required output by agencies
	Important metadata associated/linking	x			
	Peer review comments			x	Sometimes connected to VOR, sometimes not
Public Drafts (preprints)		x			Being preserved by some repositories now
Software associated with publication (w/ metadata)		x			Often included as supplemental, but not always
Data associated with publications (w/ metadata)			x		Including multimedia
Data or software products (maps, software packages)				x	Might have IP
Materials (mice, reagents, samples)	Metadata around materials	x			If can't include material, at least include metadata about it.
Integrated research data sets	Funded datasets, funder mandate	x			
	Special collections, such as cultural heritage or web/email/social media collections			x	
	Web-based multimedia scholarly products			x	
	Comments, related to VOR			x	
	Private Drafts (drafts research group may be working on)			x	
Isolated data				x	May be leftover for analysis
"Raw" data coming off instruments				x	
Records of research (notes, emails, correspondence with collaborators, scientific exchange)				x	Some are Freedom of Information Act items
Grey literature (blogs, YouTube, Twitter, etc.)				x	

- In terms of access policies, some repositories are configured for long-term preservation with strong access expectations, while others may be dark or configured for access only under certain specified "trigger" conditions. Interestingly, institutional repositories are not necessarily configured for long-term preservation: Their primary purpose is to provide access to the publications of a university's faculty and researchers.
- There are no universally accepted standards for preservation (although some best practices are emerging among certain large hubs and in many instances beyond the US).
- With the lack of standards as well as the absence of a comprehensive system of quality control comes potential repository failure and loss or corruption of important content.
- There is little successful coordination or collaboration among repositories due to both technical and philosophical differences.

The third symptom of dysfunction is a lack of sustainability. Repositories are often funded on a project-by-project, collection-by-collection basis. When project funding disappears, planned work-flows end and inconsistencies develop.

The path forward

How can we begin, then, to improve the outlook for repositories and preservation? The first step is to recognize the vital role of repositories in scholarship and to establish a set of guiding principles around which reform can be built. To wit:[1]

- Scholarly knowledge, key research outcomes, and digital and material scholarly resources are a world heritage and should be credited, preserved, open, and made accessible as soon as possible and should be preserved for the benefit of future generations (with the caveat that certain research will remain private or inaccessible due to reasons such as privacy, business competition, or national security).
- Preservation should be accomplished in ways that optimize quality, discoverability, interoperability, provenance, and ease of adoption and use.
- Preservation is the responsibility of society but especially of scholars, data stewards, libraries, and sponsoring institutions.

The next step is to agree on specific goals that move beyond simply reducing the current levels of dysfunction. These broad goals should include developing a coordinated approach to:

- enhance access via aggregated search and discovery
- raise broad awareness of the distinctions between functions of hosting and preservation
- improve distribution access control and workflows to make scholarly content available in human and machine-readable formats
- develop and achieve sustainable funding models beyond, or in addition to, grants
- leverage economies of scale for discipline and domain repositories
- improve curation and quality control via data formats, metadata standards, import/export, and forward migration services
- replicate the scholarly record in multiple, distributed, certified repositories

- replace random redundancy with the planned redundancy of preserved content
- develop guidelines and criteria for what is preserved (beginning by limiting discussions to the scholarly research record)
- establish the mechanisms of coordination within and across existing organizations to pool resources and avoid siloes or duplicated efforts.

Within this framework for action, the specific actions to be taken will vary depending on where reform efforts gain traction. What is abundantly clear is that the repository and preservation system needs to be funded in a planned and sustainable manner in order to ensure quality, consistency, and uninterrupted preservation. Such funding and planning cannot be achieved in a vacuum, but requires a broad plan for collaboration and coordination. It is the responsibility of the entire preservation repository community to work together on this new future, avoid the continuation of current, haphazard practices and decisionmaking, and encourage a rigorous adherence to standards and best practices as purposeful components of a comprehensive access and preservation plan.

The specific—and we believe achievable—action items that we propose are:

- **Clarify** opportunities for UNESCO and WSIS to engage in this effort
- **Coordinate** action among meta-organizations (e.g., COAR, CLIR/ DLF)
- **Raise funds** for improved sustainability and stewardship through investments and endowments in repositories
- **Support** aggregation driven by preservation concerns, such as:
    - Electronic legal deposit (UK)
    - Portico, Chronopolis, APTrust, and DuraSpace
    - DPN, MetaArchive Cooperative, CLOCKSS
- **Build** workflows and an ecosystem in order to ensure long-term access and preservation.

How might the Open Scholarship Initiative fit in with this effort? There are many major initiatives and discipline-based efforts, some with decades of experience and hard-fought lessons resulting in best practices across fields. However, to date these efforts have had few opportunities to coordinate. Affecting the status quo will require the informed alignment of these efforts, coupled with additional resources, leading toward systemic change across a wide range of stakeholder groups, including governments, academia, scholarly societies and associations, research libraries, and for-profit and non-profit organizations. OSI may be well-positioned to help push these stakeholders as a community towards an agenda that will move progress forward on these issues. Coordination on this initiative with other efforts, such as the Scholarly Commons effort of Force11 and the Research Data Alliance, is highly recommended (see Appendix).

Open Scholarship Initiative Proceedings, Volume 1, 2016
doi: http://dx.doi.org/10.13021/G89W24

OSI2016 Repositories and Preservation Workgroup

Joyce Backus, Associate Director for Library Operations, National Library of Medicine, National Institutes of Health

Robert Cartolano, Vice President for Digital Programs and Technology Services, Columbia University

Christina Drummond, Director of Strategic Initiatives, Educopia Institute

Agathe Gebert, Open Access Repository Manager, GESIS-Leibniz-Institute for the Social Sciences

Brooks Hanson, Director of Publications, American Geophysical Union

James Hilton, University Librarian and Dean of Libraries, Vice Provost for Digital Education and Innovation, University of Michigan

Maryann Martone, Director of Biosciences, Hypothes.is, and President, FORCE11

Sarah Michalak, Associate Provost for University Libraries and University Librarian, University of North Carolina Chapel Hill (UNC)

Richard Ovenden, Bodley's Librarian, Bodleian Libraries, University of Oxford

Sarah Pritchard, Dean of Libraries, Northwestern University

Rita Scheman, Publications Director, American Physiological Society

Appendix: Select Preservation Communities

Belmont Forum (https://www.belmontforum.org/)
Confederation of Open Access Repositories (COAR) (https://www.coar-repositories.org/)
COPDESS (http://www.copdess.org/)
DuraSpace (http://www.duraspace.org/)
Force11 (https://www.force11.org/)
Open Repositories, annual conference (http://or2016.net/)
Preservation and Archiving Special Interest Group (PASIG) (http://www.preservationandarchivingsig.org/)
Research Data Alliance (https://rd-alliance.org/)
SPARC (http://sparcopen.org/)
SPARC Europe (http://sparceurope.org/)

References

Halbert Martin, Katherine Skinner, and Christina Drummond, "Vertically Integrated Research Alliances: A Chrysalis for Digital Scholarship, A White Paper for Community Discussion," Educopia Institute, 2015, (note: see stakeholder section, bibliography), as of June 9, 2016: https://educopia.org/sites/educopia.org/files/deliverables/Draft_Vertically_Integrated_research_Alliances_A_Chrysalis_for_Digital_Scholarship_0.pdf

Maron Nancy L., "A Guide to the Best Revenue Models and Funding Sources for your Digital Resources," Strategic Content Alliance, Ithaka S+R and JISC, 2014, as of June 9, 2016: http://www.sr.ithaka.org/wp-content/uploads/2015/08/Jisc_Report_032614.pdf

Maron Nancy L., Jason Yun, and Sarah Pickle, "Sustaining Our Digital Future: Institutional Strategies for Digital Content," Strategic Content Alliance, Ithaka S+R and JISC, 2015, as of June 9, 2016: http://www.sr.ithaka.org/wp-content/uploads/2015/08/Sustaining_Our_Digital_Future.pdf

Skinner Katherine,Christina Drummond, and Martin Halbert, "Chrysalis: Moving Forward Collectively," (white paper), Educopia Institute, 2014, as of June 9, 2016: https://educopia.org/sites/educopia.org/files/deliverables/Draft_Chrysalis_Moving_Forward_Collectively.pdf

"WSIS + 10 Outcome Documents," ITU World Summit on Information Society, International Telecommunication Union (ITU), Geneva, 2014, as of June 9, 2016: http://www.itu.int/net/wsis/implementation/2014/forum/inc/doc/outcome/362828V2E.pdf

Notes:

[1] The following recommendations have been adapted from the "Data Management and Research Policy" Position Statements of the American Geophysical Union (AGU); as of June 9, 2016: http://sciencepolicy.agu.org/agu-position-statements-and-letters