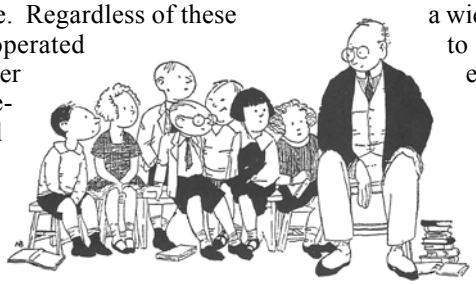
ATG Special Report — On Institutional Repositories, "Beyond the Repository Services," their Content, Maintainers, and Stakeholders

by **Don Brower**, **Sandra Gesing**, **Rick Johnson**, **Natalie Meyers**, **Jessica Meyerson**, and **Mikala Narlock**

Institutional repositories (IRs) have proliferated over the past two decades. University, disciplinary, and professional society users depend on IRs for preservation and dissemination of scholarly research objects. Yet, IR growing pains are well known and vociferously lamented.[1] Under-resourced repositories easily become vulnerable and difficult to upgrade if their code-base or feature sets obsolesce. Regardless of these pain points, whether developed and operated in-house, on the cloud, or delivered under platform as service (PAAS) contracts, repositories are evolving. IRs have moved beyond end-of-life preservation toward transparently supporting the entire research data lifecycle. IRs now play important organizational roles[2] as preprint services, data repositories and distinct sociotechnical systems reflecting institutional standards and norms. Thus, the IR writ broadly continues to be a strategic investment, especially when viewed in the context of "Next Gen Repositories,"[3] "Scholarly Communication Resources"[4] and "Beyond the Repository"[5] services.



To What Purpose

Asking if institutional repositories have the same purpose as when they were first established is a red herring. **Clifford Lynch** described institutional repositories in 2003 as "...a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members."[6] Since then, that set of services has varied for each institution over time, and it isn't possible to make an absolute judgement on how well our own or others' older repositories have served their purpose over the last twenty years. Early developers and adopters have expanded the formats they accept or made their content more findable. Others who implemented IR platforms using second or third wave repository frameworks that leverage community infrastructure[7] readily support a multiplicity of deposit formats with robust, contemporary data citation features. Still others have separate repositories for papers, data and images. Some prefer an all-in one solution.[8] General purpose IRs with broad collecting and preservation policies, like ours at **Notre Dame** (*Curate.ND.edu*), can offer researchers a single, simple place for grant-funded research outputs that might not have a disciplinary best-fit repository. We refer out to best-fit solutions that complement our repository's strengths and don't concern ourselves with whether our IR "competes" against discipline specific (ICPSR,[9] Neurovault[10]), general (Dataverse,[11] Dryad,[12] Figshare,[13] Mendeley Data,[14] Zenodo[15]) or special purpose repositories like SuAVE.[16] Instead, we consider the whole as a rich scholarly ecosystem.

There are still hurdles. Convincing researchers to self-deposit can be an uphill climb. Supplementing self deposit with mediated workflows, pre-generated DOIs, and offering prompts against harvested citations to encourage postprint deposits breaks down barriers and lowers data entry burdens. Concurrently, information maintainers are repairing, caring for and documenting a wide variety of knowledge systems beyond IRs to facilitate access and optimize user experience. These can be complementary software systems: library catalogs, federated discovery systems, identity management solutions, 3rd party repositories, digital humanities and media-arts projects, knowledge management systems, intranets, or documentation stores. For an IR maintainer, these activities may overlap and can include ontology updates, or include prompting individual users or crowd-source project participants to accept and enhance auto-harvested or machine learning-generated content. IRs continue to evolve in novel ways, through broad collecting policies as well as filling important roles as reliable preservation systems remain strategically important to researchers.

IRs as Stabilizing Force in a Rapidly Evolving Landscape

At first glance, researchers continue to trust IRs in ways that appear to contradict CNI round table conversations about the future of IRs.[17] In the *PresQT Needs Assessment,* we were surprised to find that researchers identified their institutions above funders and journal publishers when asked about who has the infrastructure required to provide long-term public access to research data (See Figure 1).[18]

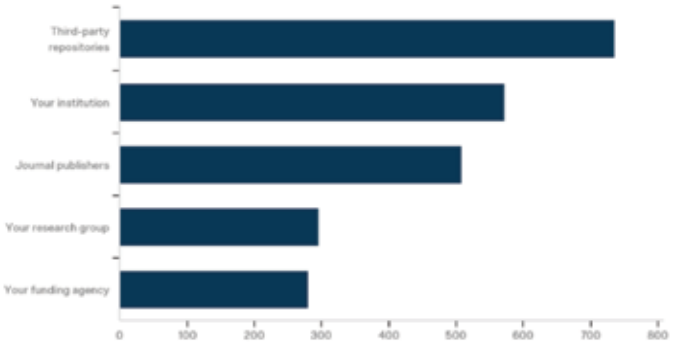


FIGURE 1. PresQT Needs Analysis Responses to Question: In your estimation, which of the following currently have the infrastructure required to provide long-term public access to your research data? [19]

continued on page 71

In the subsequent 2019 Research Data Alliance (RDA) Exposing Data Management Plans survey,[20] respondents again indicated high trust in institutional repositories, selecting General Institutional Repo as their repository of preference when asked "What would be your preferred mechanism/method(s) for publishing DMPs?" (See Figure 2.)

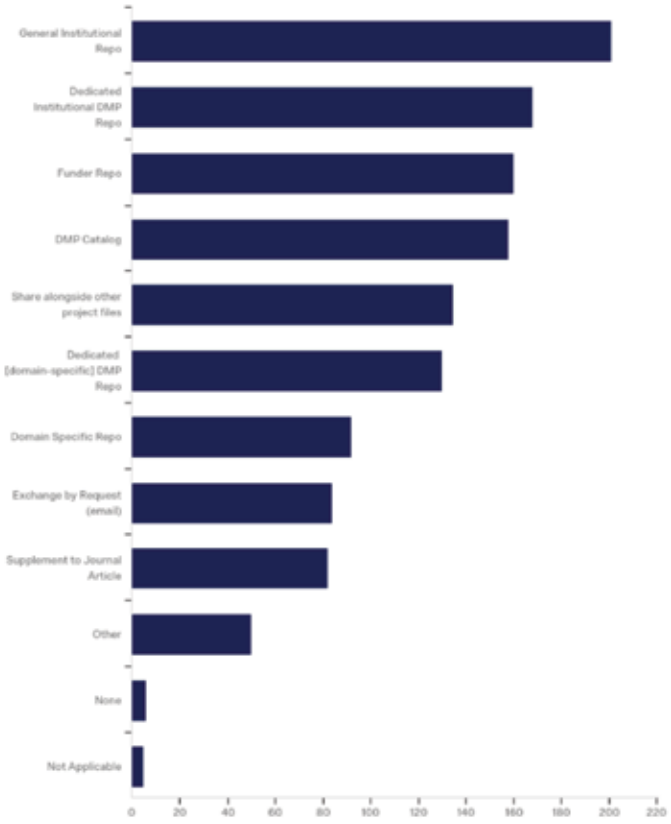


FIGURE 2. RDA Exposing Data Management Plans Needs Assessment Responses to Question: What would be your preferred mechanism/method(s) for publishing DMPs? (Select all that apply).

Researchers' trust in IRs appears to be steady, while at the same time libraries and disciplinary societies are increasingly seeking publisher and vendor provided solutions in an effort to reduce duplication of effort and community-wide expenditures on repository platforms. Data packaged alongside articles badged and promoted as "open data" in reputable journals on a commercially-supported platform appears as "safe as data can be." However, when such data is lost, stakeholders directly experience the difference between sharing and preserving their data. In *Badges for sharing data and code at Biostatistics: an observational study*,[21] **Rowhani-Farid** and **Barnett** observed that 49 out of 76 (64%) badged articles articles at *Biostatistics* had broken links. At *Statistics in Medicine*, 21 out of 53 (40%) had broken links. Upon inquiry, *Biostatistics* indicated that when the publisher (Oxford) switched to a new publishing platform in January 2017, *some of the supplemental material was lost in the transfer*.[22]

FIGURE 3. Explanation of broken links to data on open data badged journal articles (*https://doi.org/10.12688/f1000research.13477.2*)

This story helps researchers and decision-makers alike understand the risk of treating journal platforms as de facto preservation systems.

Registries like OpenAIRE[23] and r*e3data.org*[24] can help identify best-fit preservation solutions. Certifications like CoreTrustSeal[25] and nestorSeal[26] along with standards like ISO 16363[27] provide certainties to stakeholders. Evaluators[28] are emerging which enable tests of a repository's support for the FAIR principles (Findable, Accessible, Interoperable, and Reusable).[29] As repositories become standards-compliant, certified and "FAIR-ify" their systems, the re-usability of preserved data grows. So too does the interoperability potential between our repositories and their ability to connect beyond the repository services.

The Repository Ecosystem

Repository developers and maintainers are responding to changes in scholarly communication norms and intersecting more closely with the researcher toolchain than ever before as we engage with "beyond the repository" services. We also focus now on how researchers will re-use repository content. For example, at Notre Dame we encourage the use of cloud services' connected platforms like the Open Science Framework (OSF.io)[30] for active project and data management to meet researchers where they are. On OSF researchers can access their files on cloud storage like Box,[31] Dropbox,[32] Google Drive,[33] OneDrive[34] or ownCloud[35] and share data in consortial digital projects. As researchers are increasingly expected to share not just their data, but also their scripts and code, we appreciate that on OSF they can version software through code-repository add-ons (e.g., Github,[36] Gitlab,[37] or Bitbucket[38]) and associate it with their projects. Hosted static PDFs and source code alone do not often offer immediate computational or visualization experiences for re-use. So, we turn to other "beyond the repository services" when needed. Platforms like SuAVE, Earthcube's CINERGI Data Discovery Studio,[39] CodeOcean,[40] Papers with Code,[41] and WHOLETALE[42] are all changing the hitherto static expectation of scholarly communication. These services facilitate bundled ways to interactively experience publications, scripts, methods and visualizations. Some leverage Jupyter notebook integrations allowing users to execute code while concurrently supporting reproducibility of scientific methods. Such platforms are increasingly integrated with publishers' journal delivery

continued on page 72

systems, allowing readers to follow the steps in a publication, recreate results, or even use the provided methods or code to produce or analyze new data.

The Scaling Emulation and Software Preservation Infrastructure program (EaaSI)[43] is another "beyond the repository service." Led by the Digital Preservation Services team at **Yale University Library**, EaaSI is focused on the development of technology and services to expand and scale the capabilities of Emulation-as-a-Service software. Through EaaSI, thousands of computing environments are being configured and shared to the EaaSI Network service. As a node within this larger network of software emulation servers, **Notre Dame** will be able to offer access to preserved software running in emulation anywhere within the network. We recognize our IR will likewise share our users' attention with EaaSI, OSF, CodeOcean, WholeTale (as well as many other commercial and social network tools) interoperating with and re-using data from these platforms in a broad scholarly information ecosystem.

The preservation needs of diverse content and formats may demand different migration and retention schedules. Is there a need to preserve software, for example, for decades? Use cases served by EaaSI show that there may be dependencies that demand immersion in a running application to render particular data formats. However, for some other scientific software, audiences may be just as interested in referring back to source code[44] or software designs. For these users of preserved source code, preservation value lies more in the study of software development design patterns/conventions rather than actually running the preserved software. The Research Data Alliance Software Source Code Identification Working Group[45] recognizes that because most research datasets are created and/or transformed using software, a common standard for software identification will enable better traceability and reproducibility of research data. The group aims to author concrete recommendations for the academic community to ensure that software identification solutions for sharing and preservation adopted by the academic players are not only mutually compatible but also aligned with software development practice. Because of the diversity of software preservation user needs and use cases, we acknowledge that software preservationists must assess anticipated use cases, digital content, and file formats independently to determine how and which technologies, systems, and policies best facilitate future reference, use, and reuse.

The Future of IRs

Given all this, what's the role for Institutional Repositories going forward? A preprint server? The funded researcher's go-to for in-house data preservation and sharing compliance to meet funder mandates? Organizational memory? Data/Software Repository? Dissertation/Thesis repository? Conference Presentation/Poster repository for organizationally-hosted events? The answer can be "Yes" to all. As funders, disciplinary societies, and publishers generate mandates for code and data sharing, our repositories are evolving right alongside the scholarship, accommodating the need for scholars to share re-usable data, re-runnable code, workflows and more.

We're continually testing the integration of active data management platforms like OSF with our Fedora-based preservation repository (*Curate.ND.edu*). We recognize that working with content in IRs via APIs improves interoperability and usability with the computational environments of users, so we are improving our IR's APIs and web services. We've participated in data transfer pilot projects, through a National Data Service Dashboard integration project[46] and our current PresQT project (IMLS LG-70-18-0082-18)[47,48] aims to develop repository-agnostic tools which function as middleware between systems for improving data and software preservation quality. At the same time we are developing a Unified Preservation and Exhibition Platform[49] with support from the **Mellon Foundation** which will unite previously independent efforts to build digital infrastructures, building on existing platforms and tools.

The IR of the future may be an abstraction known to its users under one organizationally-branded name that leverages federated search to expose content and metadata from multiple storage locations, each with its own deposit, access and retention policies interoperating with "beyond the repository" extensions. Perhaps the best question facing the IRs of the future is whether and how they can leverage repository extensions, content discovery features and content curation options. In other words, accepting that content will be deposited in many places, what are the links and partnerships needed to realize common interoperability through adoption of FAIR standards that help our scholars seamlessly move beyond the limitations of single repository solutions?

endnotes on page 73

ATG Special Report – Endnotes

1. **Arlitsch, K.**, and **Grant, C.** (2018). Why So Many Repositories? Examining the Limitations and Possibilities of the Institutional Repositories Landscape. *Journal of Library Administration*, *58*(3), 264–281. *doi: 10.1080/01930826.2018.1436778*

2. The Information Maintainers, **Olson, D.**, **Meyerson, J.**, **Parsons, M. A.**, **Castro, J.**, **Lassere, M.**, … **Acker, A.** (2019, June 17). Information Maintenance as a Practice of Care. Zenodo. *doi:10.5281/zenodo.3251131*

3. COAR Next Generation Repositories Working Group. (2017). *Next Generation Repositories: Behaviours and Technical Recommendations*. Retrieved from COAR website: *https://www.coar-repositories.org/files/NGR-Final-Formatted-Report-cc.pdf*.

4. **Skinner, Katherine**. (2019). *Mapping the Scholarly Communication Landscape 2019 Census*. Retrieved from Educopia Institute website: *https://educopia.org/2019-census/*.

5. **Quinn, B.**, **Schaefer, S.**, **Weinraub, E.**, **Alagna, L.**, and **Caizzi, C.** (2018, January 4). Beyond the Repository: Integrating Local Preservation Systems with National Distribution Services. *https://doi.org/10.21985/N28M2Z*.

6. **Lynch, C. A.** (2003). Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age. *Portal: Libraries and the Academy*, *3*(2), 327–336. *doi: 10.1353/pla.2003.0039*

7. **Chodacki, J.** (2019, February). *CDL's Path Towards Data Publishing Adoption: Community Infrastructure*. Presented at the 14th International Digital Curation Conference, University of Melbourne, Melbourne. Retrieved from *http://www.dcc.ac.uk/sites/default/files/documents/IDCC19/Slides/CDLPathTowardsDataPublishing_JohnChodacki.pdf*.

8. **Scherer, D.**, and **Valen, D.** (2019). Balancing Multiple Roles of Repositories: Developing a Comprehensive Repository at Carnegie Mellon University. *Publications*, *7*(2). *doi: 10.3390/publications7020030*

9. ICPSR. Retrieved July 31, 2019, from *https://www.icpsr.umich.edu/icpsrweb/*.

10. NeuroVault: An open data repository for brain maps. (n.d.). Retrieved July 31, 2019, from *https://neurovault.org/*.

11. The Dataverse Project. (n.d.). Retrieved July 31, 2019, from *https://dataverse.org/*.

12. Dryad Digital Repository. (n.d.). Retrieved from *https://www.datadryad.org/*.

13. Figshare. (n.d.). Retrieved July 31, 2019, from *https://figshare.com/*.

14. Mendeley Data. (n.d.). Retrieved July 31, 2019, from *https://data.mendeley.com/*.

15. ZenodoResearch. Shared. (n.d.). Retrieved July 31, 2019, from *https://zenodo.org/*.

16. SuAVE | Survey Analysis via Visual Exploration. (n.d.). Retrieved July 31, 2019, from *http://suave.sdsc.edu/*.

17. **Lynch, C.** (2017). *Rethinking Institutional Repository Strategies: Report of a CNI Executive Roundtable Held April 2 & 3, 2017*. Retrieved from Coalition for Networked Information website: *https://www.cni.org/wp-content/uploads/2017/05/CNI-rethinking-irs-exec-rndtbl.report.S17.v1.pdf*.

18. **Johnson, R.**, **Meyers, N.**, and **Wang, J.** (2018, April). *PRESQT: Assessing Researcher and Library Needs for Research Data & Software Preservation Quality Tools*. Presented at the CNI Spring 2018 Membership Meeting, San Diego, CA. Retrieved from *https://www.cni.org/topics/digital-preservation/presqt-assessing-researcher-and-library-needs-for-research-data-software-preservation-quality-tools*.

19. **Brower, D.**, **Gesing, S.**, **Greenawalt, B.**, **Johnson, R.**, **Meyers, N.**, **Spies, J.**, and **Wang, J.** (2017, September 18). PRESQT Needs Assessment. Retrieved from PresQT Data and Software Preservation Quality Tool Project website: *https://doi.org/10.17605/OSF.IO/XFWS6*.

20. Research Data Alliance Exposing Data Management Plans Working Group. (n.d.). Exposing DMPs Needs Assessment. Retrieved July 31, 2019, from Research Data Alliance website: *https://www.rd-alliance.org/exposing-dmps-needs-assessment*.

21. **Rowhani-Farid, A.**, and **Barnett, A.** (2018). Badges for sharing data and code at Biostatistics: An observational study [version 2; peer review: 2 approved]. F1000Research, 7(90). *doi: 10.12688/f1000research.13477.2*

22. **Leek, J.** (2017, November 8). *Personal Communication*.

23. OpenAIRE. (n.d.). Retrieved July 31, 2019, from *https://explore.openaire.eu/*.

24. Re3data.org. (n.d.). Retrieved July 31, 2019, from *https://www.re3data.org/*.

25. CoreTrustSeal – Core Trustworthy Data Repositories. (n.d.). Retrieved July 31, 2019, from *https://www.coretrustseal.org/*.

26. **Nestor**. Nestor-Siegel. (n.d.). Retrieved July 31, 2019, from *https://www.langzeitarchivierung.de/Webs/nestor/EN/Services/nestor_Siegel/nestor_siegel_node.html*.

27. International Organization for Standardization. (2012). *Space data and information transfer systems — Audit and certification of trustworthy digital repositories* (ISO 16363:2012). Retrieved from *https://www.iso.org/obp/ui/#iso:std:iso:16363:ed-1:v1:en*.

28. RDA FAIR Data Maturity Working Group. (2019). Results of an Analysis of Existing FAIR Assessment Tools. *Research Data Alliance*. *doi: 10.15497/RDA00035*

29. **Wilkinson, M. D.**, **Dumontier, M.**, **Aalbersberg, I. J. J.**, **Appleton, G.**, **Axton, M.**, **Baak, A.**, … **Mons, B.** (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018–160018. *doi: 10.1038/sdata.2016.18*

30. Center for Open Science. (n.d.). Open Science Framework. Retrieved July 31, 2019, from *https://osf.io*.

31. Box: Secure File Sharing, Storage, and Collaboration. (n.d.). Retrieved July 31, 2019, from *https://www.box.com/*.

32. Dropbox. (n.d.). Retrieved July 31, 2019, from *https://www.dropbox.com/*.

33. Google Drive. (n.d.). Retrieved July 31, 2019, from *https://drive.google.com*.

34. Microsoft OneDrive. (n.d.). Retrieved July 31, 2019, from *https://onedrive.live.com/*.

35. ownCloud: Secure Enterprise File Sharing (EFSS). (n.d.). Retrieved July 31, 2019, from OwnCloud website: *https://owncloud.com/*.

36. GitHub. (n.d.). Retrieved July 31, 2019, from *https://github.com/*.

37. GitLab. (n.d.). Retrieved July 31, 2019, from *https://about.gitlab.com/*.

38. Bitbucket. (n.d.). Retrieved July 31, 2019, from *https://bitbucket.org/*.

39. Community Inventory of EarthCube Resources for Geosciences Interoperability: CINERGI. (n.d.). Retrieved July 31, 2019, from *https://www.earthcube.org/group/cinergi*.

40. Code Ocean. (n.d.). Retrieved July 31, 2019, from *https://codeocean.com/*.

41. Papers With Code. (n.d.). Retrieved July 31, 2019, from *https://paperswithcode.com/*.

42. The Whole Tale. (n.d.). Retrieved July 31, 2019, from *https://wholetale.org/*.

43. Scaling Emulation and Software Preservation Infrastructure Program. (n.d.). Retrieved July 31, 2019, from *https://www.softwarepreservationnetwork.org/eaasi/*.

44. Software Heritage. (n.d.). Retrieved July 31, 2019, from *https://www.softwareheritage.org/*.

45. Research Data Alliance Software Source Code Identification WG. (2018, June 14). Retrieved July 31, 2019, from RDA website: *https://www.rd-alliance.org/groups/software-source-code-identification-wg*.

46. **Meyers, N.** (2016, October). *Scaling the Open Science Framework: National Data Service Dashboard, Cloud Storage Add-ons, and Sharing Science Data on the Decentralized Web*. Presented at the International Workshop on Science Gateways, Melbourne. Retrieved from *https://drive.google.com/file/d/0BwK61gB7NIcielRLM1ZGamhZdTQ/view*.

47. PresQT project. (n.d.). Retrieved July 31, 2019, from *https://presqt.crc.nd.edu/*.

48. LG-70-18-0082-18: (n.d.). Retrieved July 31, 2019, from Institute of Museum and Library Services website: *https://www.imls.gov/grants/awarded/lg-70-18-0082-18*.

49. University of Notre Dame. (2018, March 13). Notre Dame receives Mellon Foundation grant to develop software platform to help universities access library and museum holdings. Retrieved July 31, 2019, from Notre Dame News website: *https://ntrda.me/2DmE75K*.