Submitted on: 17.08.2023

A Tale of Two Systems: Research Data Repositories and Digital Preservation

Maria Cotera
Figshare, part of Digital Science, London, United Kingdom
E-mail address: maria@figshare.com

Andrew Mckenna-Foster
Figshare, part of Digital Science, Boston, United States

Adrian-Tudor Panescu
Figshare, part of Digital Science, Iasi, Romania

Abstract:

Research data is being created at an incredible rate and it is near impossible to predict which datasets may become future treasure troves of data. There are many historical examples of this - from weather and sea temperature data recorded in hundred-year-old ship logs from the Southern Weather Discovery project informing current climate models to understand climate change; to the Överkalix study which used historical food harvest records and church logs to make groundbreaking epigenetic discoveries. Nowadays data is born digital; hence it requires digital preservation. Digital preservation is defined as "the method of keeping digital materials alive so that they remain usable as technological advances render original hardware and software specification obsolete". The combined task of making digital records accessible and FAIR while also following best practices in digital preservation is complex, to say the least. Depending on an institution's or collection's needs, this most often requires integrating one or more repositories with a preservation system and managing one or more workflows alongside. Determining how to best preserve research data adds another layer of complexity because research data can be very large in size, have diverse file types, and be described by different metadata schemas. This paper describes why and how research librarians can innovate ways to balance these requirements by focusing attention on interoperability and creative technical solutions. It summarizes how a repository can incorporate some aspects of preservation into the platform itself; and uses case studies to examine what a repository / preservation system integration can look like.

Keywords: Digital preservation. Research data. Research Data Repository. Preservation Systems. Interoperability.

Introduction

Research data is being created at an incredible rate and it is near impossible to predict which datasets may become future treasure troves of data. There are many historical examples of this - from weather and sea temperature data recorded in hundred-year-old ship logs from the Southern Weather Discovery project informing current climate models to understand climate change; to the Överkalix study which used historical food harvest records and church logs to make groundbreaking epigenetic discoveries. Nowadays data is born digital; hence it requires digital preservation.

Digital preservation, defined as "*the method of keeping digital materials alive so that they remain usable as technological advances render original hardware and software specification obsolete*", is a core tenant of the library's role within an institution. Traditionally, a main challenge in digital preservation is finding ways to steward digital files as storage systems change or fail, and file formats become obsolete with technology advancement. This usually involves both institutional workflows and digital infrastructure in the form of software and storage services. Two types of platforms coexist in this space: repositories and preservation systems.

The combined task of making digital records accessible and FAIR while also following best practices in digital preservation is complex, to say the least. Depending on an institution's or collection's needs, this most often requires integrating one or more repositories with a preservation system and managing one or more workflows alongside. Determining how to best preserve research data adds another layer of complexity because research data can be very large in size, have diverse file types, and be described by different metadata schemas.

A tale of two systems

Data repositories and preservation systems typically have different priorities. Most data repositories prioritize maintaining publicly accessible collections and, in doing so, devote resources to make it easier to ingest materials like open access papers, theses and dissertations, and datasets. Repository platforms are also often focused on metadata management, submission workflow, discoverability, and interoperability with other scholarly communication systems and services. It is difficult to balance these functional demands of end user access and discoverability, with the technical needs for preservation. Repositories may have some preservation capabilities and conform to parts of the OAIS Reference Model (CCSDS 2012), but they typically do not meet the strict requirements to be a true preservation system (Rieger et al. 2022).

On the other hand, digital preservation of research data prioritizes minimizing risk for stored files and metadata by protecting them from loss, corruption, and obsolescence and requires merging technical solutions with record management workflows run by knowledgeable staff (Rieger et al. 2022). At a high level, a preservation system typically stores multiple copies of files in multiple places, offers processes to check file integrity, manage file types, tracks provenance at a granular level, and aims to fully conform to the OAIS Reference Model.

While preservation may be included within the scope of repositories, it is difficult to maintain one single system that reliably meets the expectations of both a repository and a preservation system (Weinraub et al. 2018). There is ongoing debate about a repository's role in taking on discovery, access, and preservation in one system (Coalition for Networked Information 2017). Additionally, the library may have digital assets outside of the repository that may also require preservation - therefore the most efficient solution to successfully implement digital preservation of research data within an institution often requires multiple platforms that must be integrated with one dedicated preservation system. Separating the repository and preservation system has the added benefit of not only creating additional copies of files, but also storing them with different services and infrastructures. Adding a

preservation system to a repository is also part of international repository certifications like the CoreTrust Seal (CoreTrustSeal Standards and Certification Board 2022).

Repository integrations with digital preservation systems come with their logistical challenges and associated cost implications. It costs money to store multiple copies of files, and one must then also track files across different storage locations and different services. The growth of digital research data adds complexity because research data files can be extremely large in size, making storage and transfer quite expensive. Also, as funders gradually expand their expectations and requirements around data sharing, institutions will face growing pressure to responsibly store, share, and preserve data for required time periods. Therefore, finding ways to properly preserve digital files, especially large ones, while controlling storage costs is essential for the sustainability of the library's research support and digital collection efforts.

It is difficult to assess the success of repository and preservation system integrations as the descriptions of actual implementations are rarely published (Barrueco and Termens 2022, Weinraub et al. 2018). And it is noted that many times organizations do not have the resources or technical capabilities to even carry out preservation within best practices in the first place (Rieger et al. 2022).

Thus, libraries that do have the resources for comprehensive digital preservation of research data must navigate both repository and preservation platform options, along with integration options for their choices. Universities around the world use different combinations of repository platforms and preservations systems. Some institutions develop their own integrations for their repositories and preservations systems and may make them available for others to use (e.g., Fryson and McNicholl 2023) while others may rely on manual processes or completely custom technical integrations. The authors of this paper present their experience from the perspective of a repository platform assisting its users with digital preservation of research data via preservation system integrations.

Figshare is an open access repository platform used by institutions globally for all types of academic outputs, especially research data, including large datasets, images and videos. It can handle very large files and file sets, up to 5TB in size. Figshare focuses on building and supporting tools that help researchers and libraries create FAIR records that meet or exceed best practices around access and reuse. This means Figshare is very focused on discovery and active reuse, rather than archiving. While the platform includes many preservation features, it is not a preservation system. It can work with most storage options, whether cloud or local, and it is set up to store multiple copies of files in multiple locations. It calculates a checksum for every file, maintains access and change logs for every item, and has a business model for long term sustainability. Figshare has endorsed the TRUST principles and already conforms to parts of the OAIS reference model. Figshare has no built-in functionality around file obsolescence and specific preservation metadata. However, its openly documented API enables the development of additional automated preservation processes, such as a file type tracker or a periodic checksum checking application. Despite these digital preservation functionalities, Figshare's primary use is a data capable institutional repository that aims to integrate with dedicated preservation systems. There is the acknowledgement that most libraries have multiple digital content platforms that require preservation, and the most efficient solution is to have all of them feed one dedicated preservation system. In that vein, Figshare supports an API to facilitate these integrations and helps institutions with customized integrations.

Figshare has helped institutions integrate with a large variety of preservation systems (these include Arkivum, Ex Libris's Rosetta, DuraCloud's Chronopolis, DANS EASY, and Archivematica and Preservica via the Jisc RDSS) and in this paper we briefly summarize two case studies. The first case study covers the integration with Arkivum - to date Figshare has assisted six institutions in its implementation. The second case study describes efforts to integrate with the Data Archiving and Networked Services (DANS) EASY service, presented here as an example of how large datasets can create preservation challenges.

Case Study 1 - Arkivum

Arkivum (https://arkivum.com/) is a Software-as-a-Service platform that works with all types of digital outputs, including scholarly works like datasets and papers. It can accept very large files (e.g over 1TB) by offering a storage solution that can receive multipart, or chunked, uploads. Figshare's first implementation of an Arkivum integration was in 2015. The implementation is relatively straightforward because Arkivum accepts the files individually and can handle extremely large files. An institution must manage its own Arkivum appliance but, once the integration is complete, the two systems do not require manual intervention. To date, four institutions have a Figshare supported integration that uses Arkivum just to store files and two send files and metadata. Figshare is working with two of these institutions to further support large files.

The integration process begins by working with the institution to understand its particular integration needs around what repository content will be preserved and what metadata should be included. All metadata is stored by Figshare in a central database. The file(s) associated with the metadata go through the following process within the Figshare system (Fig. 1): 1) user uploads files, 2) file is immediately stored on Figshare's temporary storage, being available for download 3) An MD5 checksum of the file is computed and stored for future integrity checks 4) two operations are being performed in parallel: the preview of the file is generated and stored on a separate Figshare storage instance, and the file is copied from the temporary storage to its final storage location, which can be either supplied by Figshare (e.g. Amazon Simple Storage System) or the institution 5) the file is mirrored on one or more 3rd-party storage solutions (e.g. a preservation system).

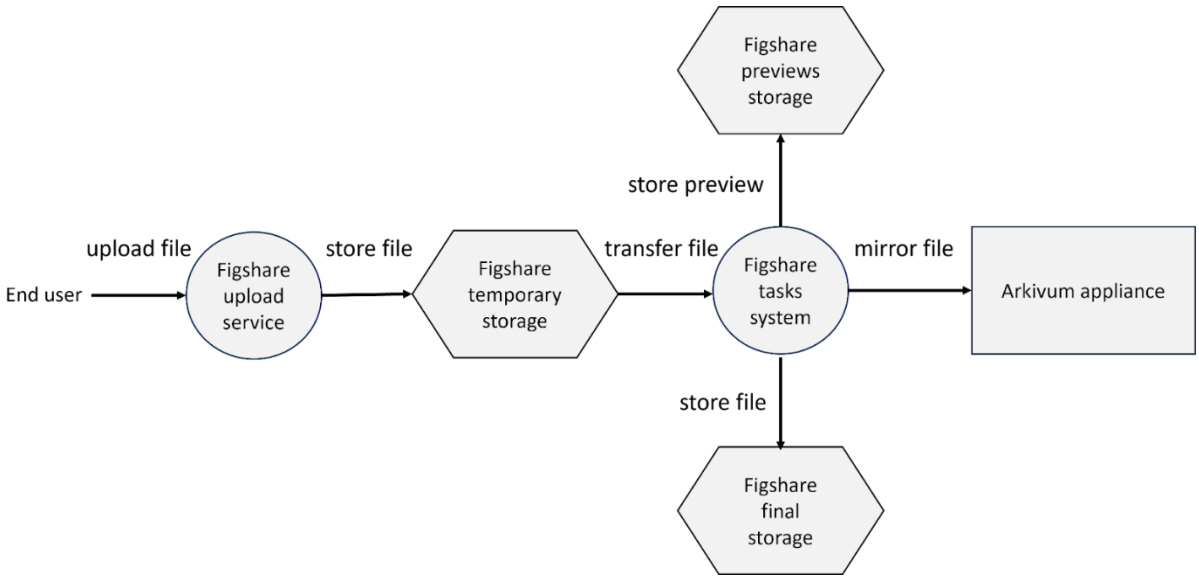

Figure 1. The file upload process and mirroring to an Arkivum appliance.

Figshare uses Arkivum's REST API to mirror the file and send it to the Arkivum data center. Once this is done, the files stored in Arkivum are available for the institution to access outside the Figshare system.

Figshare can also perform a third API call to retrieve the status of the file ingestion in the Arkivum system and this becomes very important for storage management. By confirming the successful transfer, one can decide to evict rarely accessed files from repository storage and retrieve them as necessary from archival storage. This provides a way to efficiently use storage space, especially as the size of digital files and the overall volume of stored files is increasing and can be very expensive. The eviction process can be set to start when a percentage of the repository storage is used (e.g. 90%) and it can stop using several different halting conditions: 1) a certain percentage of the storage is free (e.g. 60% free), 2) no more files exist that have not been accessed within a given timeframe, or 3) there are

no more files that have not been uploaded within a given timeframe. Files that go through this eviction process are downloadable but not immediately. A user who tries to download the file is notified that the file is being retrieved from archival storage and that they should check back later to access the download. The file is moved from the Arkivum datacenter to the institution's Arkivum system and then it is moved to Figshare storage. It must be noted that the extra steps for file retrieval reduce the overall FAIRness (Findable, Accessible, Interoperable, Reusable) of the records because accessing files is slightly more difficult. Despite this, an institution may view it as necessary for managing storage costs.

Case Study 2 - DANS

The DANS EASY service is a national archiving system for research data based in the Netherlands. Figshare developed an integration for a client using Figshare for an institutional data repository. EASY service includes a Digital Preservation Plan (https://dans.knaw.nl/en/preservationplan/) and includes dedicated preservation features such as checking file formats and storing detailed provenance metadata. Unlike Arkivum, EASY requires both files and metadata to be sent together in a bag using the DANS Bagit profile. While Figshare files and JSON formatted metadata are available through Figshare's API, EASY requires a specific XML format that Figshare builds when generating the archival package. Significant time was spent mapping metadata fields and setting up the application on the Figshare side that would format the metadata and package it with the files. On the one hand, creating bags is a nice way to bundle files and metadata, and perhaps easier for a preservation system to manage. However, on the other hand, it requires a much more complex processing application to be developed between the two systems. Creating a common Bagit profile for preservation systems is one recommendation from a survey of repository and preservation system integrations (Weinraub et al. 2018).

The biggest challenge, however, was working with large files. At the time of the integration efforts, the teams could not find a way to reliably move extremely large files. Figshare handles large files by chunking the files and uploading them in parts. This is reliant on the storage solution's ability to accept chunked files. Chunking offers efficiency in case there is a disruption; file transfer can continue with the disrupted chunk rather than transferring the entire large file again. There was no way to chunk files during the transfer to EASY. Ultimately, the teams opted to limit the file size to 100GB. The bags are transferred using the SWORD protocol and the Figshare system can check that the transfer was successful. Because many institutions use Figshare precisely because it accepts large files, the EASY integration made it clear that integrating with preservation systems with chunked upload capabilities would be important in the future.

Conclusion

Repository managers looking to follow best practices for digital preservation of research data need to plan what platforms they will use, considering that research data can be very large in size, have diverse file types, and be described by different metadata schemas. While it would be ideal if a repository could be 'plug and play' for most preservation systems, this is not feasible due to the needs around metadata formatting and file transfer. An existing preservation system integrated with a repository may not be able to accept larger datasets that the repository accepts. Some institutions can afford to build custom solutions for preserving larger files (e.g., Rice and Sutherland 2023) but many may struggle just to host large files in their repository. It is also important that the chosen data repository platform supports robust integrations with the library preservation platform.

References

Barrueco, J.M. and Termens, M. (2022), "Digital preservation in institutional repositories: a systematic literature review", Digital Library Perspectives, Vol. 38 No. 2, pp. 161-174. https://doi.org/10.1108/DLP-02-2021-0011

Becker, R. (2019) “Why century-old ship logs are key to today’s climate research”, The Verge. Retrieved July 20, 2023, from https://www.theverge.com/2019/5/3/18528638/southern-weather-discovery-ship-logs-climate-change

CCSDS Secretariat Space Communications and Navigation Office. (2012). Reference Model for an Open Archival Information System (OAIS): Recommended Practice CCSDS 650.0-M2. Retrieved July 19, 2023, from https://public.ccsds.org/pubs/650x0m2.pdf

Coalition for Networked Information. (2017, May). Rethinking Institutional Repository Strategies: Report of a CNI Executive Roundtable Held April 2 & 3, 2017. Retrieved July 19, 2023, from https://www.cni.org/wp-content/uploads/2017/05/CNI-rethinking-irs-execrndtbl.report.S17.v1.pdf

CoreTrustSeal Standards and Certification Board. (2022). CoreTrustSeal Requirements 2023-2025 (V01.00). Zenodo. https://doi.org/10.5281/zenodo.7051012

Fyson, W., and McNicholl, R. (2023, June 14). Preservation and the repository: Practical interoperability between EPrints and Arkivum. Open Repositories 2023 (OR2023), Stellenbosch, South Africa. Zenodo. https://doi.org/10.5281/zenodo.8091663

Överkalix study. Wikipedia. Retrieved July 20, 2023, from https://en.wikipedia.org/wiki/%C3%96verkalix_study

Rice, R., and Sutherland, I. (2023, June 13). DataVault: One institution's answer to Big Data & confidential data. Open Repositories 2023 (OR2023), Stellenbosch, South Africa. Zenodo. https://doi.org/10.5281/zenodo.8091511

Rieger, O. Y., Schonfeld, R. C., and Sweeney, L. (2022, July 19). The Effectiveness and Durability of Digital Preservation and Curation Systems. https://doi.org/10.18665/sr.316990

Weinraub, E., Alagna, L., Caizzi, C., Quinn, B., & Schaefer, S. (2018). Beyond the repository: Integrating local preservation systems with national distribution services [Report]. Institute of Museum and Library Services. https://apo.org.au/node/127411