Submitted on: 22.07.2017

Supporting Digital Preservation and Access with Fedora

David Wilcox
DuraSpace, Halifax, NS, Canada
Email address: dwilcox@duraspace.org

Evviva Weinraub
Northwestern University, Evanston, IL, USA
Email address: evviva.weinraub@northwestern.edu

Abstract:

Digital preservation is complex, and the vocabulary is not well-defined. A long-term digital preservation and access strategy incorporates many components, and there are levels of preservation to match the risk tolerance and available resources of an institution - there is no "one size fits all" approach. Digital preservation systems with modular components provide the greatest flexibility for organizations to choose an approach that can scale up or down as needed over time. Fedora is an open source, durable repository for digital objects, that is part of a long-term digital preservation and access solution. Fedora is used in a wide variety of institutions including libraries, museums, archives, and government organizations. It is a community-based solution that leverages existing, widely used standards whenever possible to ensure long-term sustainability. Fedora stakeholders from around the world have come together to clearly define how Fedora supports digital preservation, and how it fits into a larger digital preservation solution. This paper will provide an overview of the considerations and complexities of a digital preservation strategy, and describe how Fedora can serve as a key component of a digital preservation and access solution.

Keywords: fedora, open source, repository, preservation

Introduction

Digital Information, and its persistence over time is a looming challenge for cultural heritage and scientific communities. The challenge is not unique to any one institution, and is best solved by working with the people, organizations, and technologies that make up these communities. Active exchange of information is what supports innovation and research, sustains learning, and catalyzes creativity. The challenge of ensuring the long-term viability and sustainability of the academic and cultural output of the human experience is imperative if libraries hope to maintain relevancy in the modern age.

A 2014 study of data growth found that "from 2013 - 2020, the digital universe will grow by a factor of 10--from 4.4 trillion gigabytes to 44 trillion. It more than doubles every two years" (Turner, Gantz, Reinsel, & Minton, 2014). By the authors estimation, only 5% of digital content across the globe was usable in 2013. If the data is not usable, it's not useful and any value that once existed is lost. In 2014, Northwestern University Libraries performed a Research Data Survey where it was found that most academic output is living on personal servers, commercial storage services, or on hard drives sitting on shelves collecting dust. Only 14% of faculty (of 651 respondents) store their work on university servers (Buys & Shaw, 2014).

Digital preservation is more than just saving the bits and bytes that make up a digital object, it's a complex ecosystem of activities which have significant interdependencies and a large number of unresolved issues. While there is general agreement and coalescing around the value of implementing the recommendations of the NDSA Levels of Preservation (Phillips, Bailey, Goethals, & Owens, 2013) there are complexities inherent to building out the various systems necessary for implementation.

Digital Preservation

With varying levels of sophistication, institutions worldwide are in the process of developing local infrastructure and workflows to facilitate curation, management, preservation, and access to digital information resources in their care. Even in the context of a single institutional system where policy and procedures must be implemented to meet the needs and requirements of research data, published and licensed electronic resources, digitized and born-digital collections, and institutional records, among many others, this is no easy task. Constant and continual custodial negotiations are necessary to address these needs making intra-institutional collaborations a complex issue.

All this is to say that when we talk about Digital Preservation, the landscape is complex, messy, and has a variety of moving parts associated with its workflow. Often, Digital Preservation is synonymous with backups, but they aren't the same thing at all. True digital preservation is more than identifying objects for digitization or providing online access to born-digital materials; it requires curation and identification of whether the object is worthy of long term preservation activities. It requires us to think of digital things in much the same way we view physical objects - curatorial activities need to occur because the digital preservation is costly. If we think of digital storage like a storage unit, you pay for a space to put the things you own. Those things can sit in the storage unit for years without being touched. Perhaps you've labeled your boxes with detailed information, or perhaps you've labeled your box "photos" and assumed you'd get back to it later. You ignore the storage unit, perhaps dumping more stuff into it from time to time. One day, you find out that there has been a leak in your storage unit and some of your stuff is ruined. You have no real idea of what was there and no real ability to retrieve or restore those objects. You have no idea what was valuable or useful and you then need to make a decision about what needs to be taken care of and how. Digital Preservation, however, is a much more time consuming and complex procedure. To extend the metaphor, your storage unit has a manifest of things you've put into it and you've likely put your objects inside some sort of packaging that protects it from the elements. You make regular visits to the storage unit to touch each one of your packages to make sure nothing has happened to it while you've been doing other things. If, when you poke a package, it feels weird, you can check it against the manifest you created when you deposited the content. You can then talk to another storage unit with the exact same objects and transfer an exact replica from a storage unit far far away back into your storage unit. You spend a lot of time making sure the climate is right, the roof doesn't leak, and the way you've described the objects is still usable. Your local systems are keeping track of what you've ingested, where it is, and when last it was checked to make sure it's still usable.

The cost of true digital preservation shouldn't be underestimated. The cost of curation, the cost of maintaining and touching copies of your objects on disparate architecture and in multiple environments is difficult enough to consider in a physical environment; when you're dealing with technologies that change quickly, it gets even more complex. From a non-technical perspective, the preservation of digital objects carries with it issues around intellectual property, concerns about breach of privacy, and the reality that you can't really get rid of anything from the digital landscape. This brings up a host of ethical issues around the acquisition of objects that put people and institutions at risk, the right to be forgotten, and other policy issues which professionally we've been grappling with for years, but which become more pressing in the digital world. Our approaches to Digital Preservation have, until recently, been around the idea of just keeping whatever is given to us, because storage is cheap. But we need to move into a mindset of thinking of our digital collections in the same way we view any other collection development activity. If we decide we need to keep it, we need to make preservation decisions; otherwise, we'll end up with a storage unit full of broken and unusable things because the costs and technologies around doing this right are significant. Fedora is just one piece of this incredibly intricate landscape.

Fedora and Digital Preservation

Fedora, the Flexible, Extensible, Durable Object Repository Architecture, was first conceived in 1997 by Sandy Payette and Carl Lagoze as a conceptual model based on the principle of openness: "A fundamental requirement of an open architecture for digital libraries is a reliable and secure means to store and access digital content. FEDORA is a digital object and repository architecture designed to achieve these requirements, while at the same time providing extensibility and interoperability" (Payette & Lagoze, 1998, p. 1). The concept of durability has always been a key component of the Fedora architecture; this includes not just the preservation of the bits as they reside in a storage layer, but the accessibility of digital objects over time. While it is not a complete digital preservation solution, as no one piece of software can be, Fedora has a set of characteristics and provides a number of features that support an overall digital preservation strategy.

As a community-led project, Fedora is developed and maintained by and for the global cultural heritage and scientific community. Libraries, archives, museums, research centres, and government organizations use Fedora for disparate use cases and data types. Fedora, as previously indicated, is open source, covered under an Apache 2.0 license. In addition to the code being more visible to more people, Fedora is openly governed by representatives from stakeholder institutions; the project is not dependent on any one organization. Fedora is standards-based; the project adopts widely used standards rather than developing custom solutions, and prefers standards adopted by the broader Web rather than less widely used standards. Finally, Fedora is interoperable; by providing a robust RESTful API framework and common data models, Fedora positions itself to support the long-term viability of the digital objects it manages.

These characteristics are important aspects of Fedora's support for long term digital preservation, but Fedora also includes a number of specific features that target digital preservation use cases. One such feature involves persistence; files are stored on the filesystem in predictable locations based on their checksums, and metadata describing files and digital objects are stored in a database and can be exported on demand. This means that the content of a Fedora repository can be accessed and retrieved independent from the software. Fedora also supports fixity in a number of scenarios; on ingest Fedora can calculate a file's checksum using several algorithms and optionally compare the calculated value with a value provided alongside the file. Assuming there is no mismatch, in which case the upload will be rejected, the checksum is stored with the file and can be recalculated on demand via the REST-API.

Fedora also supports versioning; a new version of a digital object or file can be created on demand when any action is performed using the REST-API. Versions can be restored on demand, or deleted if they are no longer required. Details about changes to objects and files can be captured using the audit trail; preservation metadata, modeled using the RDF-based PREMIS and PROV-O ontologies, can be created, stored, and indexed for search and retrieval. This provides a mechanism for tracking changes in the repository and maintaining a complete audit history. Finally, the entire contents of the repository, or a selection of resources, can be exported as a standardized serialization of RDF. This both allows resources in Fedora to be exported to external digital preservation systems and prevents platform lock-in by providing a mechanism for extracting a complete representation of the repository, including binary files, in a standardized format.

How Fedora Fits into a Digital Preservation Solution

Fedora has a number of characteristics and features that support long term digital preservation, but no single application can support a complete digital preservation workflow. With this in mind, Fedora aims to be a key component of an overall digital preservation strategy. To discover how this functions in practice, the Fedora Leadership Group distributed a survey to the Fedora community that focused on the ways in which Fedora was being used to support digital preservation across organizations and sectors. The survey received 36 responses from a variety of institutions, and provides a wealth of information on the use of Fedora in digital preservation environments.

Survey respondents use a variety of systems and services in conjunction with Fedora to support digital preservation. Applications like Archivematica are used to process digital objects in compliance with the OAIS functional model that can then be deposited into Fedora for storage and preservation. Frameworks like Islandora and Samvera (formerly Hydra) provide management, discovery, and access functionality to resources in Fedora. External preservation systems such as DuraCloud, AWS, Arkivum, DPN, and Iron Mountain provide additional long term digital preservation functionality. These systems do not compete with Fedora; rather, they provide the ability to maintain multiple copies, as outlined in the NDSA Levels of Preservation, of resources that are also stored within Fedora, thereby increasing the likelihood of recovery in case of a disaster. Fedora's import/export capability supports these kinds of services by providing a standard mechanism for exporting relevant resources and packaging them in the required formats.

Conclusion

Because of the flexible nature of Fedora, different institutions have configured Fedora to meet their local needs including implementing issues around access, preservation, digital asset management, and institutional/research repository activities. Additionally, its strong relationships with the Samvera, Islandora, and ICPSR communities, as well as its ability to work with external systems like ArchiveSpace, Archivematica, JISC Research Data Shared Service, Chronopolis, IRODS, APTrust, Hathitrust, Internet Archive, and Preservica, amongst others, to serve needs around indexing, fixity checking, checksum generation, and verification/alerting processes, there are over 400 implementations of Fedora worldwide.

Fedora has the capability to be a key component of any Digital Preservation strategy, whether as an access repository, an asset management platform, a preservation tool, or some combination thereof. Digital preservation plans and their implementation are moving targets, but because Fedora is a community based project, it is actively being worked on and enhanced by the people, organizations, and technologies that will best serve the needs of our communities.

Acknowledgments

This paper would not have been possible without the support of the Fedora community, particularly those institutions that have joined DuraSpace as members in support of the Fedora project.

References

Buys, C., & Shaw, P. (2014, May 22). Report on Data Management Survey, Northwestern University (Rep.).

Payette, S., & Lagoze, C. (1998). Flexible and Extensible Digital Object and Repository Architecture (FEDORA). Research and Advanced Technology for Digital Libraries Lecture Notes in Computer Science, 41-59. doi:10.1007/3-540-49653-x_4

Phillips, M., Bailey, J., Goethals, A., & Owens, T. (2013). The NDSA Levels of Digital Preservation: An Explanation of Uses. Proceedings of the Archiving (IS&T) Conference. Retrieved July 20, 2017, from http://ndsa.org/documents/NDSA_Levels_Archiving_2013.pdf

Turner, V., Gantz, J. F., Reinsel, D., & Minton, S. (2014, April). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Retrieved July 20, 2017, from https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm