IFLA WLIC 2017 WROCŁAW

Submitted on: 02.06.2017

Stewarding Research Data with Fedora

David Wilcox
DuraSpace, Halifax, Canada.
E-mail address: dwilcox@duraspace.org

Abstract:

Beyond the complexities faced by typical asset management or institutional repository systems, research data presents a number of complications, including complex hierarchies of related objects that must be modelled and displayed, a wider array of data formats that must be supported, and domain-specific metadata that is necessary to make data intelligible. Managing these complications often leads to software that is tailored to particular data, making it difficult to maintain or share.

Fedora is a flexible, extensible, open source repository platform for storing, managing, and preserving digital content, including research data. Fedora is used in a wide variety of institutions including libraries, museums, archives, and government organizations. Fedora 4, the latest version of Fedora, supports research data management by providing key repository features such as support for millions of resources and files of any size, native linked data functionality, advanced data modeling, and preservation services. Fedora is also extremely well-suited to integrations with existing researcher workflows via a well-documented REST-API and event-based messaging service. Fedora's interoperable design paved the way for integration with the Open Science Framework; a free, open source tool that connects with the applications and services researchers already use to support the entire research lifecycle. This paper will provide an overview of how Fedora support supports research data management, including integration with the OSF and the roadmap for future development and integrations.

Keywords: fedora, repository, open source, research data, linked data.

Introduction

Research data management is complicated. Research data are heterogeneous both within and across domains, and present many unique requirements compared to typical digital asset management or institutional repository use cases. Data modelling is one such requirement; some datasets are relatively simple in terms of structure, while others require complex hierarchies or graphs to represent fully and accurately. Additionally, the diversity of file formats found in research datasets presents a major problem for both preservation and access; simply preserving the bits does not address the need to ensure that files can be accessed and used in the future. Metadata can also be an issue, as different standards are used within different domains, leading to difficulties with sharing and understanding information. Managing all of these difficulties often leads to highly customized software that is difficult to maintain over time. Fedora, the flexible, extensible, open source repository platform, seeks to remedy this situation by providing common back-end infrastructure for research data management that can be customized with different interfaces and integrations. This paper will provide an overview of Fedora with a focus on its research data management capabilities.

Fedora

The Fedora project began as a concept articulated in a research paper published in 1998 by Sandy Payette and Carl Lagoze entitled, "Flexible and Extensible Digital Object and Repository Architecture (FEDORA)". In this paper, Payette and Lagoze describe a model for a durable, digital object and repository architecture 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). These basic design principles, articulated twenty years ago, have continued to guide and inform Fedora development efforts through to the current iteration (Fedora 4.x). Extensibility and interoperability, while not features in and of themselves, are critical to success of Fedora as a component of larger infrastructure and systems, which is particularly relevant in the case of research data management.

Research Data Challenges

Research data management presents unique challenges beyond those presented by typical digital asset management use cases. Data modelling is a prime example; many repositories deal with a limited set of resource types that are relatively simple to model. Some typical examples have been described as part of the Portland Common Data Model effort ("Portland Common Data Model," 2016); these include things like basic and large images, newspapers, journals, and theses. All of these resources types can be modelled in similar ways, with a limited number of components and relationships. But research data are very heterogeneous; they can take a myriad of forms that differ not only across but within disciplines.

In addition to complex data modelling needs, research datasets often include a variety of file formats, with no guarantee of standardization. Simpler resource types, such as theses, newspapers, and photographs, are much more limited in terms of the file types that are typically present, and there is a greater degree of standardization. High resolution images, for example, are typically stored as TIFF files, with JPEG or PNG used as a lower resolution formats for things like thumbnails. Research datasets, by contrast, may contain a variety of file types, some of which may correspond to different versions of the software used to create and access them.

A third complicating factor is metadata; there are a wide variety of metadata standards used within and across scientific disciplines, and in many cases custom metadata fields are added when existing standards are perceived to be insufficient to describe a particular resource.

Features to Support Research Data

Fedora is a flexible, extensible repository platform that is agnostic regarding the types of resources it manages and the ways in which those resources are modelled. This flexible approach, while initially daunting due to its openness, is well suited to address the needs of research data management. Fedora is an implementation of the Linked Data Platform ("Linked Data Platform 1.0," 2015), and in that context acts as a linked data server. Resources created in Fedora are assigned HTTP URIs and are described using RDF triples. Resources are also modelled using RDF triples, allowing for both simple and complex use cases. RDF facilitates flexible, highly expressive relationships, thereby allowing datasets to be represented as accurately as possible within the context of the repository.

Fedora places no restrictions on the file formats that it manages, so even the most esoteric files within datasets can be stored, managed and preserved. Files can be uploaded via the Fedora REST-API, which has been tested with files over a terabyte in size ("Large File Ingest and Retrieval," 2017). Fedora is equally agnostic regarding the metadata used to describe the resources it manages. Metadata can either be represented as RDF triples associated with a resource, in which case any number of metadata schemas may be used, or as a binary XML file associated with the resource it describes. In this way, Fedora does not impose any particular limits on the metadata schemas or formats being used, though representing metadata as RDF does confer certain advantages in terms of participating more fully in the semantic web.

Fedora and the Open Science Framework

More recently, members of the Fedora community have pursued an integration between Fedora and the Open Science Framework (OSF). The OSF is a free, open source project management platform for researchers that, "enables connections to the many services researchers already use to streamline their process and increase efficiency." ("Open Science Framework," n.d.). The OSF already connects to common storage providers such as Dropbox, Google Drive, and Box. With the integration of Fedora as a storage provider add-on, researchers will be able to archive and preserve their OSF projects and files in their Fedora-based institutional repositories without leaving the OSF interface. This addresses a key problem of depositing research data during the research process rather than trying to obtain it afterwards once the project is complete.

Conclusion

The complexities of research data management are well known. Most repository platforms are designed to work with relatively common digital asset management use cases, which do not fully address the needs of research data. Data modelling requirements are often complex and difficult to represent in traditional folder hierarchies, while diverse and sometimes proprietary file formats are not always supported. Metadata presents a major obstacle; the diversity of standards and practices used across disciplines are difficult to consolidate in a single system. Fedora was designed with this complexity in mind, and provides a set of features that support research data management in a flexible, extensible way. With native support for linked data, Fedora allows administrators to use RDF to model datasets in a rich, accurate way. This flexibility is enhanced with support for any file format, as well as support for large files. Finally, support for any metadata standard in both RDF and binary XML formats ensures a broad range of research data can be described. Taken together, these features provide a strong

basis for research data management using Fedora as a repository platform, but Fedora can also be plugged into existing systems and services to provide a similar level of support within existing researcher workflows. The Open Science Framework is one example of such an integration, allowing researchers to use OSF as a workbench while also storing and preserving their data in Fedora throughout the research lifecycle.

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
Large File Ingest and Retrieval. (2017, January 25). Retrieved May 31, 2017 from the Fedora 4.7.1 Documentation wiki: https://wiki.duraspace.org/display/FEDORA471/Large+File+Ingest+and+Retrieval

Linked Data Platform 1.0. (2015, February 26). Retrieved from https://www.w3.org/TR/ldp/

Open Science Framework. (n.d.). Retrieved from https://cos.io/our-products/open-science-framework/

Payette, S., & Lagoze, C. (1998). Flexible and Extensible Digital Object and Repository Architecture (FEDORA). Lecture Notes in Computer Science, 1513, 41-59. doi:10.1007/3-540-49653-x_4

Portland Common Data Model. (2016, August 9). Retrieved March 29, 2017, from https://github.com/duraspace/pcdm/wiki