# Fedora and Digital Preservation Survey

Introduction

In the fall of 2016, Fedora Leaders engaged in discussions with the aim to clearly define how Fedora supports digital preservation, and how it fits into a larger digital preservation ecosystem. These discussions led to the publication of a document outlining the functionality provided by Fedora that supports digital preservation practice, as well as a well-attended panel discussion at the Fall CNI member meeting.

As part of the effort to explain the role Fedora plays in digital preservation ecosystems, Fedora Leaders deployed a survey to institutions who have implemented Fedora for this purpose in early 2017. The results in this report provide some context for the preservation functions supported by the software, and utilized by the Fedora community, as well as the other systems used in conjunction with it. The aim of the survey was to better understand the use of Fedora, and to provide data to institutions considering the use of Fedora in a preservation capacity.

Questions

Please tell us who you are - Institution
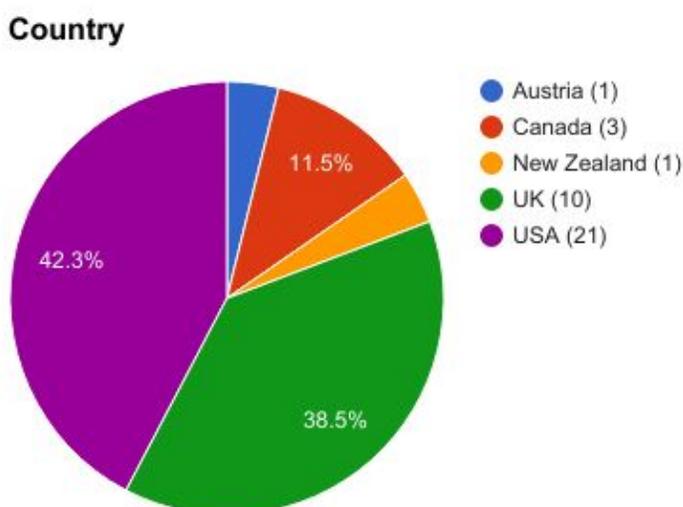
Please describe your repository (not preservation) architecture including ancillary systems.

Answers to this question were difficult to summarize due to the sheer diversity in systems configuration in regards to infrastructure to support ingest, storage, object management, metadata management, and access. Having said that, most indicated their use of a specific framework (Samvera, Islandora, custom) and references to Fedora 2.x vs 3.x vs 4.x. In addition to this, a number of institutions noted the use of multiple installations of Fedora instances used for different purposes (an example- one installation used for ingest and testing and another for public access; some host a separate Fedora for restricted or private content), as well as multiple instances of a single Fedora serving multiple functional applications.

There were many references to apps built in-house (often referred to as 'home-grown') to manage various operations including fixity, metadata administration, and access portals. There were often references to numerous Samvera applications (built on Hyrax, Avalon, etc.) serving up content from a Fedora Repository for different purposes (media streaming, file delivery, geospatial web services, web archive playback, etc.).

Ancillary Systems of Note

Ingest
-      Some respondents spoke to having separate processors for images and AV material
-      Workflow Management System for metadata creation and object ingest
-      Therapy Dog

Preservation/Storage
-      Isilon
-      AWS
-      DuraCloud
-      WayBack
-      Ceph cluster

Access
-      Contentdm for image hosting
-      Video from Vimeo through Fedora/Islandora
-      Blacklight
-      Dedicated image/streaming media server
-      Mirador
-      Avalon

DOM
- OpenETD
- DSpace (w. Symplectic Elements integration)
- ArchivesSpace
- Dataverse
- ArcGIS
- AtoM
- EPrints
- Archon
- Pure
- Archivematica

Describe the size of your repository in GB or TB, and number of objects.

Number of Objects in Repository

257
500
2 200
6600
15000
36000
41639
100000
100000
110000
130000
145000
150000
250000
343108
397000
410635
500000
500000
590000
700000
1100000
1400000
2900000
3000000
4169748
9843363
10043913

Number of items

0   3000000   6000000   9000000   12000000

Space Consumed (TB)

Space Consumed (TB)

0.01
0.01
0.03
0.06
0.1
0.5
1
2.76
3.3
8.4
8.5
10
10
12.5
13.5
17.8
20
20
27
30
30
30
50
58
148
200
200
220
267
279
500

0   150   300   450   600

What is the purpose of Fedora and how are you using it at your institution?

It is clear from the survey results that Fedora is conceptualized in a variety of ways and is being used in a lot of different configurations, including components to address access, preservation, DAMS, IR and primary sources of record.

Access
Fedora was noted for its use in serving up content to platforms such as Samvera/Islandora. Fedora was often referenced for its utility to index and serve derivative access copies, while storing the preservation masters in dark store. Others used the Fedora back-end to create 'special portals' for access to special collections and grant-funded projects.

Preservation
Fedora is mentioned as a 'component' in a preservation platform and as a 'base preservation layer'; a solution for 'at-risk' digitized collections. It is referred to as a 'link manager' maintaining relationships between digital files. One respondent noted it as 'preservation ready', storing checksums, versioning, audit information, etc. to be part of a larger preservation process. Institutions noted that they tried to take advantage of the preservation affordances that Fedora offers, but without indicating it as 'the' preservation solution in their ecosystem. One respondent indicated that 'Fedora provides secure, auditable management of content and metadata.' Some refer to Fedora as their system of record, where others relegate it simply as an object store with relationships to other systems that are considered authoritative.

Institutional Repository
More than one respondent indicated that Fedora was their institution's primary institutional repository, managing a diverse eco-system of collections and materials.

DAMS
One institution noted Fedora's data modeling capabilities and preservation functions as primary reasons for selection as a DAMS to add and manage content for digital books and video. Yet another noted Fedora's purpose in supporting component parts of digital objects and collections operating as a workflow and web services engine. One response pointed to its use as an exhibition space for curating and managing objects in conjunction with an external CMS.

Platform Specific
It is clear from some respondents that Fedora is being used simply because it is a component in the stack of a preferred platform (e.g. Samvera, Islandora). In addition, it
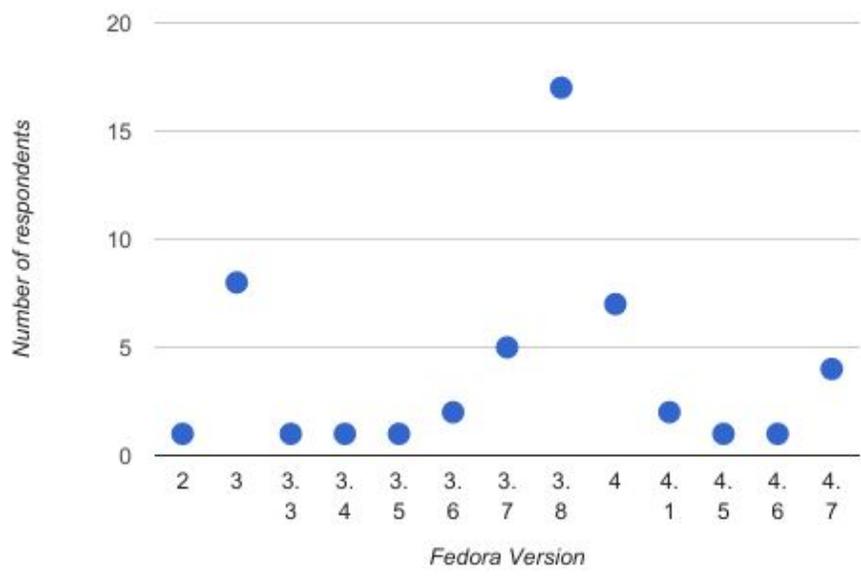
is clear that many institutions use Fedora in a broad eco-system of applications and components, with some integrations in place, but Fedora as a key preservation component in the eco-system/platform.

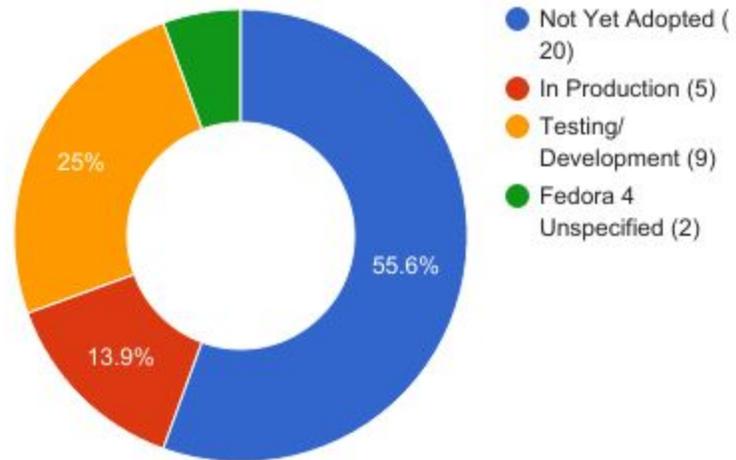# What version(s) of Fedora is/are in use at your institution?

Fedora Versions (Aggregated)

Legend:
- Fedora 3
- Fedora 4
- Fedora 2

29.4%

68.6%
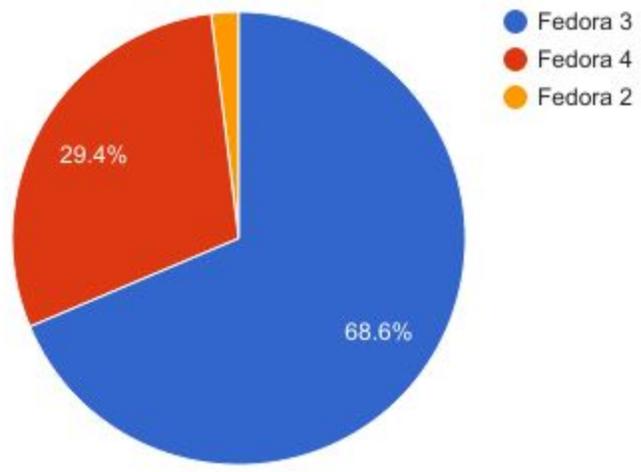
Fedora Versions (Raw responses)

Fedora 4 Adoption & Purpose

- Not Yet Adopted (20)
- In Production (5)
- Testing/ Development (9)
- Fedora 4 Unspecified (2)

Key takeaways from these results:

- Most institutions are running some version of Fedora 3 as their production repository
- The institutions that are running Fedora 4 are largely doing so in non-production capacity, in more of a testing/development capacity
- Fedora 2 is still in use

If you aren't using Fedora 4, do you plan to migrate? If not, why not?
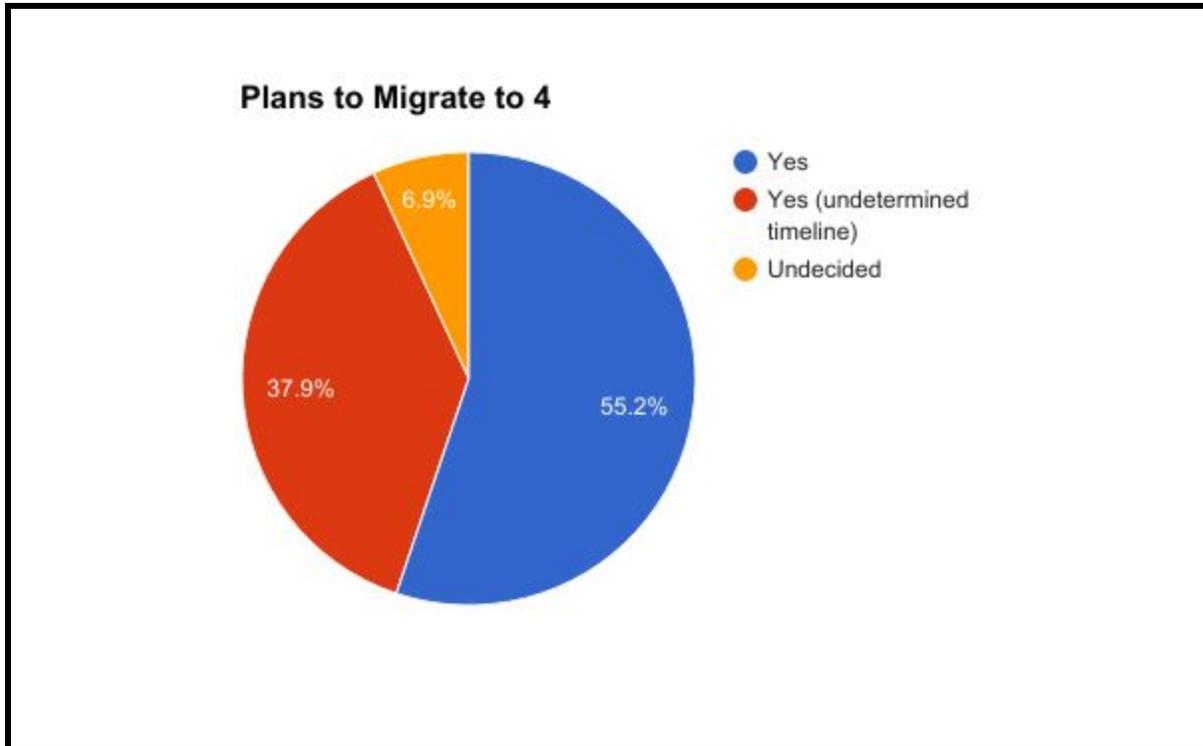
Key takeaways from these results:

- Plans to migrate when management software (Islandora, Hyrax, etc) supports Fedora 4
- Long lead time to migrate due to complexity of migrating software and collections
- Concerns over rapid pace of change and change in direction from Fedora 3 to 4

Please describe your preservation architecture and Fedora's relationship to it.

There were some interesting answers here, particularly the references to external preservation tools and services. Other common answers included references to preservation metadata, fixity, and using Fedora as the system of record. Some institutions appeared to have a strategy for preservation (whatever their definition of that may be), while others were in an 'investigatory mode'. Numerous respondents referenced contracted services with vendors including DuraCloud, AWS, Arkivum, DPN, and Iron Mountain.

A limited number of respondents referred only to Fedora as their preservation solution- 'Fedora provides storage, access, and fixity of our assets,' calling it their object of record. Many others referred to it as an object store manager in a larger preservation architecture, or referenced its capabilities in managing technical and descriptive metadata as part of their preservation strategy. Still others noted specifically that they did not store binaries in Fedora, but treated them as external content and used Fedora simply as an object reference manager; one respondent spoke to the future of Fedora being outsourced from their stack as it 'has nothing to do with preservation and only exists on the dissemination side'.

Many had custom functions built to serve needs around indexing, fixity checking, generation of checksums and verification/alerting processes. A number of respondents spoke to external systems in use for preservation activities including Archivespace, Archivematica, JISC Research Data Shared Service, Chronopolis, IRODS, APTrust, HathiTrust, Internet Archive, Preservica, etc. Some of these systems were integrated with Fedora, others not.

Some respondents referenced preservation in regards to their storage infrastructure. As an example, one respondent spoke to their processes with AWS' S3 services for backup and integrity checking, and local replication to grounded storage; another to their process for storing master copies in the cloud with access and metadata records stored in Fedora; others just referred to preservation consisting of file system backups/tape storage on institutional offerings; yet another spoke of using Box. One response spoke of participation in a preservation collective like the MetaArchive Cooperative which utilizes a private LOCKSS network, and finally, one respondent spoke of managing content on external hard drives

Some of the responded referred to their preservation strategy in terms of the processes, actions, and lifecycles that their digital assets piped through on their way to ingest and storage. These discussions did not always include Fedora as a part of that lifecycle; one indicated interest in using Archivematica for preservation and storing resultant PREMIS files in Fedora alongside the asset. Some institutions responding in this vein referred to Fedora simply as an object store for the discovery front end. Fedora was also frequently referred to for its 'tracking' capabilities- in essence, it's features for relationship and version management of digital assets.

In summary, respondents generally had an individualistic 'definition' of preservation or preservation architecture, in various states of maturity, with Fedora manifesting its utility in a variety of ways.

If you do not already have a preservation strategy, when do you expect to have one?

Most of these answers revolved around timeframe (if a strategy didn't already exist). Some were generic, speaking to their preservation efforts as 'it's a work in progress'; others referred to having processes in place that support preservation but operating without a full-scale preservation strategy or plan. Still others referred to definition of workflows and policy as the start of a preservation strategy that would inform on architecture decisions, and emphasized the foundation of such for operational preservation strategies.

Respondents acknowledged that the scope and nature of digital preservation keeps changing and as such a 'strategy' that is fixed is difficult. The word 'evolving' was specifically referenced. Some referenced participation in digital preservation trial efforts, such as with JISC, as the start of their preservation strategy definition.

Finally, some respondents simply stated that they had no plans to develop a preservation strategy, while other institutions emphasized resources earmarked for hiring preservation staff that were dedicated to such activities, noting them as 'growth areas'. Once again, responses indicated a lot of diversity and transience around definition of preservation, and defined operational strategies for implementing at their institutions.

If you do not already have a preservation infrastructure, when do you expect to implement one?

The results of this question were not conclusive. Institutions were all over the map.

In reference to the previous question in the survey, some respondents simply spoke to this work as 'in-progress' and evolving, noting that infrastructure and preservation activities would change as new solutions for content areas were on-boarded. One respondent noted that their institution was in 'R&D mode' on the topic.

One respondent spoke to this activity happening in tandem with infrastructure improvements, not as a proactive activity where policy would come before infrastructure, so in essence, very much evolving and organic.

Conclusion

The responses to this survey are diverse, illustrating the many differences in thinking and approaches to digital preservation, even within the Fedora community. While Samvera and Islandora are commonly used frameworks, respondents also make use of a variety of ancillary systems for ingest, preservation, access, and management. This makes sense given the different purposes for which Fedora is used: preservation and access along with digital asset management. With regard to preservation, many respondents make use of external preservation services, such as DuraCloud, DPN, and Arkivum. In these cases, Fedora plays a role in the overall preservation strategy but it is only one component in a larger system.

Additionally, many respondents indicated that they are currently pursuing or plan to pursue a digital preservation strategy, but this is a moving target and in general not well-understood. This represents an opportunity for Fedora to better position itself as a flexible application that can be used in a variety of different preservation configurations, even as strategies and components change over time. As a middleware system with a focus on integrations, Fedora is ideally suited to playing an important role in the digital preservation ecosystem.