David M. Tenenholtz. A Complex Web: Upgrading to Linked Data in Digital Repositories. A Master's paper for the M.S. in I.S. degree. April, 2017. 55 pages. Advisor: Ryan Shaw

As libraries, archives, and museums (“LAMs”) adopt linked data for purposes of enhancing their bibliographic and authority metadata, the technologies around digital repositories also are similarly changing course to model digital objects using linked data standards such as RDF. This study explores the digital repository community's engagement and perceptions of linked data modeling. The study is split into two phases consisting of a web survey and semi-structured interviews. Qualitative analysis of the data summarizes key characteristics of the community of practice, and open problems in transitioning to linked data in the redesign of the Fedora storage and preservation architecture commonly used in digital repositories. Other areas of discussion include the perceived concerns in cross-walking MODS to RDF, as well as the community’s recommended implementation of the Portland Common Data Model (PCDM).

Headings:

Data modeling

Digital library software

Digital preservation

Institutional repositories

Linked Data (Semantic Web)

Metadata Object Description Schema (Document type definition)

A COMPLEX WEB: UPGRADING TO LINKED DATA IN DIGITAL REPOSITORIES

by
David M. Tenenholtz

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina
April, 2017

Approved by:

_____
Ryan Shaw

Table of Contents:

List of Figures:

Acknowledgements

I must offer my thanks to many people that helped me in the course of constructing this paper, and the following mentions come in no particular order. To formulate my knowledge of the subject of linked data and digital repositories that was invaluable to conduct this study, I have to thank my master's paper advisor Dr. Ryan Shaw, and Sonoe Nakasone, my supervisor in the Resource Description and Management department at UNC's Davis Library during my first year at SILS. Further thanks go to UNC Libraries staff Ben Pennell, Julie Rudder, and Moira Downey; SILS doctoral students Deborah Maron and Jacob Hill; Code4Lib Data Modeling pre-conference workshop facilitators Christina Harlow, Mark Matienzo, Steve Van Tuyl, and Hector Correa. Instrumental to the survey construction was DuraSpace's David Wilcox, who provided vital feedback during the survey's development. For helping me build the survey itself, I am grateful to The Odum Institute staff members Teresa Edwards, and Paige Ottmar; also at Odum, Paul Mihas, whose course on NVivo was essential for me to conduct the qualitative data analysis presented here. My thanks also go to the survey and interview participants, and the community of practice around the Fedora Project and associated software. Lastly, I am grateful for all of the words of encouragement and support from other students at SILS, from Roey, and my family.

Introduction

The Web of Data and Cultural Heritage Institutions

Since its formal definition in 2006, linked data has moved from a heavily-hyped niche technology within libraries, archives, and museums (LAMs) to a recommended model for digital collections (Stuart, 2014). In 2011, the Andrew W. Mellon Foundation awarded Stanford University $50,000 in order to fund a workshop that explored prototyping linked data for scholarly use. Although these beginnings were humble, in the intervening years more Mellon grants have been issued. The most recent, dated March 10, 2016, for $1,500,000 to Cornell, Stanford, Harvard, and the University of Iowa, was funded to "support library initiatives that develop and advance the use of linked open data."[1] The progression for cultural heritage institutions to adopt linked data as a solution to expose and share data has surpassed the recommendation phase, and now is an accepted practice. Indeed, there is a clear push to advance the implementation of linked data within the LAM community, and the expression of increased funding is only one indication that the tides are now shifting to make this initiative move from aspiration to common practice.

---

[1] Linked Data for Libraries - LD4L Labs : Cornell University | The Andrew W. Mellon Foundation," n.d. Accessed 2016-10-03

The aspiration to transform library data into library *linked* data was perhaps best explained in the 2011 Stanford grant description, which offered this manifesto:

> We in the cultural heritage and knowledge management institutions are discovering better ways of publishing, sharing, and using information by linking data and helping others do the same. Through this work, we have come to value and to promote the following practices:
>
> 1. Publishing data on the web for discovery and use, rather than preserving it in dark, more or less unreachable archives that are often proprietary and profit driven;
> 2. Continuously improving data and Linked Data, rather than waiting to publishing "perfect" data;
> 3. Structuring data semantically, rather than preparing flat, unstructured data;
> 4. Collaborating, rather than working alone;
> 5. Adopting Web standards, rather than domain specific ones;
> 6. Using open, commonly understood licenses, rather than closed and/or local licenses.
>
> While we recognize the need for both approaches in each "couplet," we value the initial ones more. (University, Stanford, & Complaints, n.d.)

To explain how linked data might serve the LAM community, the technologies involved must be defined. To publish structured data on the web, Tim Berners-Lee set out the four principles of linked data:

> 1. Use [Uniform Resource Identifiers] URIs as names for things.
> 2. Use HTTP URIs so that people can look up those names.
> 3. When someone looks up a URI, provide useful information.
> 4. Include links to other URIs so that they can discover more things. (Berners-Lee, 2006)

Tom Heath and Christian Bizer describe how linked data builds on the Web's architecture:

> Just as hyperlinks in the classic Web connect documents into a single global information space, Linked Data enables links to be set between items in different data sources and therefore connect these sources into a single global data space. The use of Web standards and a common data model make it possible to implement generic applications that operate over the complete data space. This is the essence of *Linked Data*. (Heath & Bizer, 2011)

Within the Stanford grant manifesto lie allusions to another vital concept to the changing face of data modeling in LAMs: “open data.” This is a concept similar to “open source” in that the content is freely accessible, published with a non-restrictive license, and is compatible with other pools of materials. Science and government can gain the most from sharing their open data freely through formed relationships with other open data. Discoverability across the web of “open data” is seen as one of the major benefits of this ideological shift, along with transparency (as in the case of government-funded research). Ultimately, this interoperable version of publicly-available science and government data results in a variation of linked data, called Linked Open Data (LOD).

Alongside the recent adoption of LOD for sharing information coming out of LAMs, there also has been a consistent growth in the field of institutional repositories since the early days of pre-print servers for the sciences, such as arXiv.org. In many government-supported locations contributing to research (state-affiliated institutions of higher education being only one such setting) the push to provide “open access” to research data has led to an adoption of an LOD approach for institutional repositories. As a result, digital repository software systems used by institutional repositories have moved from unsophisticated metadata modeling in relational databases to newer structures that support the mandate to expose state-funded research using LOD (Babu, et al., 2012).

Dorothea Salo describes the data challenge that comes with stewardship over research data, and indicates that digital repositories are the ideal location for this purpose. Indeed, particularly *institutional repositories* (IRs) are the most capable setting to handle a variety of digital objects (Salo, 2010). In a 2013 survey of 150 digital repositories, Nicholas segmented out the many content types that IRs can provide preservation and access to. Nicholas suggests that institutions aim to mandate that scholars deposit their research into institutional repositories (Nicholas et al., 2013).

Not everyone in this domain has such a positive outlook for institutional repositories. Former director of the Coalition for Networked Information Clifford Lynch, in discussion with Richard Poynder, expressed skepticism about the outlook for institutional repositories going forward, saying that IRs have not reached their potential since they were first conceptually outlined at the 1999 Santa Fe meeting of the Open Archives Initiative. Reasons for this under-realized potential, Lynch suggested, are due to the barriers to faculty who should self-deposit their research output but choose not to, and a lack of disciplinary archives that are linked together and possibly replicated across repositories (Poynder, 2016).

Fedora

Institutional repositories are typically built on a feature-rich technology stack that incorporates a variety of open-source software tools. These tools provide storage and digital preservation solutions, search interfaces, access

control, and more. In many cases, one tool is based on another, or incorporates some outside features loosely into itself.

Currently, one of the most commonly used digital object storage platforms for digital repositories is FEDORA, which stands for Flexible Extensible Durable Object Repository Architecture.[2] Conceived at Cornell's computer science department by Sandra Payette and Carl Lagoze, and described through their 1998 publication, Fedora has changed shape over time to be an enriched framework with community-driven support (Lagoze, Payette, Shin, & Wilper, 2006; Payette & Lagoze, 2013).

Fedora occupies a unique role within the institutional repository community, because many repositories are commonly built on the Hydra and Islandora digital asset management systems.[3] Hydra is a set of "solution bundles" that at their core include Fedora, the Blacklight discovery layer, Apache Solr search index, and Ruby on Rails web application codebase.[4] Islandora is also a modular stack comprised of the Drupal content management system, Apache Solr, and Fedora.[5] That both Hydra and Islandora use Fedora as the storage and preservation component within their systems is a testament to Fedora's ubiquity in this space.

---

[2] Some members of the Fedora community also consider the "D" to stand for "Digital" rather than "Durable."

[3] In recent practice, the name Fedora is typically spelled only with the first letter capitalized.

[4] http://www.projecthydra.org. Accessed 2017-04-07.

[5] https://islandora.ca/about. Accessed 20170-04-07.

Fedora's development has been gaining speed and at the time of this writing is at release version 4.7.2. Fedora's latest major upgrade from version 3.8.2 to 4.0 implemented a new data model, one that is based on a conceptually different idea for how digital repositories should function and take advantage of emerging web technologies. At its core lies native support for linked data.

Fedora has a growing member-driven community, with stewardship provided by DuraSpace. In 2016, the Fedora community was adding members rapidly, and it will soon have 100 member institutions defining their IRs as "built on Fedora." There are already an estimated 400 digital libraries that leverage Fedora's conceptual architecture in some way.[6]

Fedora 3.x is now a legacy system not under active development. It used a proprietary method of associating various digital objects through an XML set it called RELS-EXT and RELS-INT. In June 2014, Fedora 4.0 Beta was released. The release was a conceptual departure in various ways in that it allowed for more flexibility and scalability, but most predominantly through the adoption of an emerging technology called the Linked Data Platform (LDP), which defines a RESTful web services API mechanism to interact with digital assets.[7] In a conference report from May 2016, Carol Minton Morris defined what Fedora 4 would mean to the broader digital library community:

---

6 This metric was provided by David Wilcox during his workshop on Fedora at the 2016 Digital Library Federation Forum meeting in Milwaukee, WI, November 8, 2016.

7 "Fedora 4 implements the LDP specification for create, read, update and delete (CRUD) operations, allowing HTTP, REST, and linked data clients to make requests to Fedora 4." https://wiki.duraspace.org/display/FEDORA47/Features Accessed 2016-10-15.

> The nature of Fedora 4, a modular repository framework that scales, has linked data capabilities, provides research data support, is easier to develop with, and more. Recent community work towards developing a stable, independently-versioned Fedora RESTful API will make it possible for future users to adopt alternative implementations that meet the Fedora specification. Attendees discussed the idea of a bundled trademark as a way to brand the product as part of a technology stack to ensure that the community understands the value of having 'Fedora inside'. (Minton Morris, 2014)

Open-Source IRs: Development and Community Steering

At the core of the early philosophy for open-source software development lies the concept of "simplicity."[8] For large open access repositories that preserve research data at universities, this ideal is upheld at least as an aim, but the need to manage large amounts of heterogeneous data at an enterprise scale creates complex data management plans and codebases. How do the sweeping changes in the Fedora codebase, and the decisions within the community to adopt varied data models and services, support this ideal?

Coping with profound codebase changes is a difficulty within the interconnected world of software development. Even in open-source codebase development, there is centralization of what becomes a part of the software and what does not. Typically this is called the "Cathedral" model in open-source development (Raymond, 2001). Providing a definition of the general mission of the next version of a piece of software, and helping developers not be surprised by foundational changes, goes a long way to continued support

[8] http://openpreservation.org/technology/principles/simplicity/ Accessed 2016-11-29.

and involvement from the community for which the software is intended.[9] Assessing how the change is accepted, and how it is disruptive, can help achieve greater adoption. Providing an implementation path towards one drastically changing feature of Fedora 4 is one aspect of this development work that must be completely understood.

With the move from Fedora 3.x to 4, repository development teams must work to rebuild their technology stack to accommodate the new data model. Not having an easy upgrade path may cause some disruption, and some frustration for this growing community.[10] The data model changes also spark some more general questions about where LAMs stand in the adoption of linked data to share their data across the web.

Purpose of This Study

The purpose of this study is to evaluate the work that is done in digital repositories to implement linked data, and understand how to migrate to Fedora 4.[11] The Linked Data Platform used with Fedora 4 supports interaction with collections of linked data, but this technology's new place within the

---

[9] One example of an open-source technology that did not successfully broadcast the major changes from a version 1 to 2 upgrade was Google's AngularJS framework. After the version 2 announcement in 2014, there was profound backlash from the developer community. https://jaxenter.com/angular-2-0-announcement-backfires-112127.html Accessed 2016-10-15.

[10] One research presentation in January 2016 suggests that confusion starts with conceptualizing linked data (Deng, 2016). Is this confusion easily overcome for developers working to support repositories?

[11] Numerous research institutions around the world are now at Fedora version 4. At the time of this writing, Lafayette College is the one academic institution in the United States that reports a full commitment to the Hydra Community's efforts to move to linked data. They are listed on the Fedora Commons registry as having implemented Fedora 4 (in this case, 4.4) http://registry.duraspace.org/registry/fedora?f%5b0%5d=field_institution_type%3Aacademic Accessed 2016-10-15.

development community bears exploration. While there are discernable aspirations to use linked open data for collections of scholarly research data, there is a fundamental need to scan the work being done within this community of practice. This study starts with an evaluation of what factors into the migration to Fedora 4.

Literature Review

The LOD Community of Practice

As discussed in Chapter 1, a community of practice within cultural heritage institutions has developed in recent years centering on linked open data. Now organizations like Linked Open Data in Libraries, Museums, and Archives (LODLAM)[12] and Semantic Web in Libraries (SWIB)[13] are well-established. A newer addition to this community is the Andrew W. Mellon Foundation grant-funded project called Linked Data for Libraries (LD4L)[14], which is at the time of this writing the most well-funded initiative, and encompasses work at Cornell, Stanford, Harvard, the University of Iowa, and other institutions. These communities of practice offer guidance for the unique challenges the "web of data" poses in their domain, and helps to extend the work done by the organizations involved.

---

12 http://lodlam.net/ Accessed 2016-09-10.
13 http://swib.org/swib16/ Accessed 2016-09-10.
14 https://www.ld4l.org/ Accessed 2016-09-10.

In the past three years, there has been an international survey of LAM organizations by OCLC researcher Karen Smith-Yoshimura, as well as a lengthy assessment through multiple American Library Association (ALA) *Library Technology Report*s by Erik Mitchell. The resulting analyses from Smith-Yoshimura and Mitchell provide more insight into how the LOD community within cultural heritage institutions can be better served and continue to develop. Looking at the outcomes of her study, Smith-Yoshimura posed the following most common barriers that speak to the needs of the projects implemented within this community:

1. Steep learning curve for staff
2. Inconsistency of legacy data
3. Selecting appropriate ontologies to represent our data
4. Establishing the links
5. Little documentation or advice on how to build the systems
6. Lack of tools
7. Immature software
8. Ascertaining who owns the data (Smith-Yoshimura, 2016)

In her analysis, Smith-Yoshimura categorized some of the predominant use cases within LAMs. There are modern bibliographic cataloguing efforts through BIBFRAME[15], published datasets and exploratory "proof-of-concept" implementations, as well as some successful pilot projects, such as the work done at North Carolina State University and at the University of Nevada-Las

---

15 The Bibliographic Framework Initiative (BIBFRAME) is "an initiative to evolve bibliographic description standards to a linked data model, in order to make bibliographic information more useful both within and outside the library community." https://www.loc.gov/bibframe/docs/bibframe2-model.html Accessed 2016-11-22.

Vegas (UNLV). Sylvia Southwick described UNLV's work with linked data, and how they worked from a point of confusion about how to structure data in triples to building a digital collection represented with linked data functionality and properties (Southwick, 2015).

The vast majority of the OCLC survey respondents reported that they had set forth only one project that used linked data in some way. Many of those respondents had yet to complete their projects at the time they responded to the survey. Of the ninety total respondents, there were six that reported that their project had to do with Fedora and/or Hydra. These institutions were: (1) Dartmouth College, (2) Villanova University, (3) New York University, (4) University of Tennessee-Knoxville, (5) The Chemical Heritage Foundation, and (6) The Big Data Institute. For institutional repositories using Fedora 4, where there is a mandate to do development using the Linked Data Platform, the eight barriers as described by Smith-Yoshimura not only apply, but they must also be overcome in order for the repository to function.[16]

In writing his *Library Technology Reports* for ALA in 2013 and Jan 2016, Erik Mitchell gave a nod to the emerging standards within the community of practice around Fedora, but focused more on assessing the larger LAM community. In his most recent survey of linked data, he argued that there is an element of "hype" around library linked data implementations, saying:

---

[16] The sixth point of Stanford's 2011 grant "manifesto" concerned using open licensing. Some data are intentionally silo-ed due to license restrictions on external consumption. Although this use case is not the focus of this paper, it is an important concern for IR development, and for LAM adoption of linked data.

> A related policy question surfaced in this survey is how LAM institutions should approach LD production or adoption. It appears that despite the transition to Linked Data for large-scale and core services such as the transformation of library MARC platforms and the migration of EAD finding aids, the community has not yet distilled a set of activities or systems into an 'easy-to-implement' platform or adoption approach. Indeed, LD efforts might still be categorized as existing in the startup phase of a technology adoption hype cycle given the variation in standards, tools, approaches, and perceived benefits documented in survey results and published literature. (Mitchell, 2016, pg. 6)

Implementing a small linked data project for a cultural heritage institution, let alone a large-scale restructuring of an institutional repository's data to drive new functionality, is seemingly dependent on factors unique to the organization, as well as the data. In order to go further, this established community of practice related to linked open data must grapple with each of these outlined barriers.

Towards a Common Data Model

In contrast to earlier data modeling in LAMs that emphasized strict authority control, and that could be constructed automatically, the flexibility and "latitude" offered in linked data provides many breakdowns for authority control. Myntti and Cothran argue that the abilities of linked data to bypass constructing matches on authority files can either help or hinder data sharing (Corbett, 2016; Myntti & Cothran, 2013). However, if the data are weakly represented, the result is a loss of semantic precision (Isaac & Baker, 2015).

In terms of metadata standards implementations, IRs typically describe data using Dublin Core. Salo explains how the IR reliance on Dublin Core suits a particular interoperability need:

> Most repositories rely on Dublin Core metadata, largely because the OAI-PMH metadata exchange standard asserts unqualified Dublin Core as a minimum interoperability layer. Few repositories venture beyond qualified Dublin Core. Those who do, or wish to, find that much repository software can only manage key-value pairs. Now that many, if not most, metadata and exchange standards for research data use XML or RDF as a base, this limitation seriously vitiates repositories' ability to manage datasets. (Salo, 2010)

DuraSpace, the not-for-profit governing body over the Fedora project, recommends that IRs moving to Fedora 4 implement a metadata application profile using the emerging Portland Common Data Model. In contrast, Fedora 3.x used Dublin Core, along with complex uses of many metadata application profiles truncated together, and constructed in a proprietary schema called FOXML. The now-outmoded FOXML representation incorporates various models of information objects into one complex file (Dublin Core, MODS, RDF, thumbnails, and more). Noticeably, RDF takes up a portion of the FOXML specification. This is because relationships between objects stored in Fedora had to be made. Using the data streams "REL-EXT" and "REL-INT," the Fedora Digital Object Model can form these associations.

A major bottleneck in moving towards RDF is due to a silo-ed digital object description resulting from the Metadata Object Description Schema (MODS). Hardesty describes the nuanced conflicts in the conversion of MODS records to RDF that is mandated in an upgrade from Fedora 3 to 4 (Hardesty, 2016). A community has formed to resolve this issue, led in part by Steve Anderson at the Boston Public Library and others. While the problem of transforming MODS to RDF is not fully resolved yet, the Fedora 4 community must invent a solution.

The digital asset management community, and particularly the Hydra Community, is now building towards a more formalized integration of data models. The emerging standard that will drive an off-the-shelf solution called "Hydra-in-a-Box" is the Portland Common Data Model (PDCM).

The choice to implement the Linked Data Platform with the upgrade from Fedora 3 to Fedora 4, along with the more recent data modeling effort of PCDM, leaves steering Fedora development, and the larger Hydra Community, in an uncomfortable and unresolved area. Is the community successfully managing the changes? How does the PCDM implementation for Hydra-in-a-Box help or hurt the institutions that look to build on this platform? Assessment of these system integration and vendor options provide increasing challenges. How did the DuraSpace/Fedora team resolve to move to PCDM, and how does this impact institutional repositories, and also the broader LAM community?

Research Questions

With the forthcoming move to Fedora 4, and the varied understanding of how to implement linked data in LAMs, this study aims to ask research questions as follows:

RQ1. What do digital repositories moving to Fedora 4 see as positives and negatives?

RQ2. Which data model does a given digital repository setting choose to implement? Why reject the data models that weren't chosen? Why reject the Portland Common Data Model in particular, as it is a recommendation for Fedora 4?

RQ3. At a given digital repository setting, how do the developers evaluate their institutional repository technology stack functioning as a whole with the Linked Data Platform implementation through Fedora? What sort of benchmarking can be done to evaluate if the potential benefit(s) were achieved?

RQ4. What sort of training has been implemented at the given digital repository setting to describe resources using linked data, or in working to support and interact with Fedora?

Research Design

Phase 1: Web Survey

The aim of the study is to examine the Fedora community of practice and what some perceptions within the community might be around implementing Fedora 4 and linked data for digital repositories. To this end, I conducted a web survey, distributed through three Google mailing lists—Fedora-Community, Hydra-Community, and Islandora. To develop the survey questions (see Appendix B), I consulted with the Odum Institute at UNC, and built the survey using Qualtrics. In order to establish more contextually appropriate questions on the nature of the Fedora data model, and evaluation over the adoption of Fedora 4, I solicited the aid of DuraSpace's Fedora Product Manager, David Wilcox. In order to incentivize responses, I provided an option to users to include an email address to opt in for inclusion into a random drawing for a $25 Amazon Digital Gift Card. Using a random-number-generator,

I established the four winners, and then their reward was disseminated electronically forty-eight hours after survey closed.

Phase 2: Case Study

To conduct this study with deeper understanding of Fedora, I formulated a set of ten questions that could be asked of a library department that is currently involved in the "upgration" process of moving from Fedora 3 to Fedora 4. The ten questions (see Appendix C) were then asked in semi-structured interviews of two staff members at a university that implements a large institutional repository using Fedora 3. A third staff member received the questions via email, and due to six of the questions being out of scope of their knowledge-base, they answered the four questions that they were able to. The two staff members at this location that were interviewed were the Repository Program Librarian and the Lead Repository Developer. The aim of these interviews was to learn about the planning to implement Fedora 4, and what unique challenges these staff members had experienced in the course of their work to implement Fedora 4. Additionally, the topic of linked data, and perceived usefulness to repositories, was discussed. For these interviews, I recorded the in-person sessions, and generated full transcriptions that were used during the analysis phase. To analyze the transcripts using the NVivo software, I applied a coding scheme to the transcripts, which provided qualitative data in the form of groupings of themes for further analysis (see Results section).

Results

Phase 1 of this study comprised the web survey, which was built using Qualtrics, and then distributed through three Google groups “fedora-community”, “hydra-community”, and “islandora.” Being that there may have been a sizeable number of duplicate users across two or more of these Google groups, the actually total of potential unique responders is unknown, but can be reasonably determined to be under 2,500.

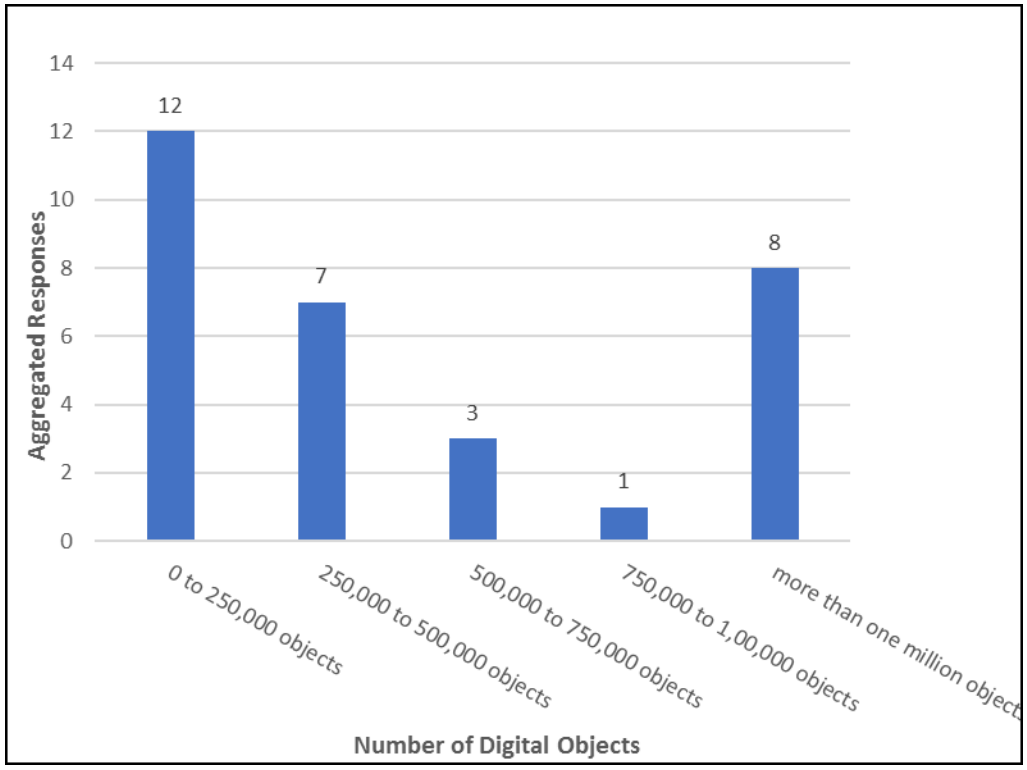
The survey remained open for responses for the duration of three weeks. At the close of the survey, there were thirty-three total responses, with six responses being entirely incomplete (other than the required first question), and another two responses being largely incomplete. The reason for this abandonment is a conundrum, but may be due in part to first question of the survey being designed to enforce that the user provided a response. None of the other questions were built with this limitation. As a result, a portion of the responders did not answer the questions, despite clicking through to the end of the survey. In total, there were twenty-five completed responses, where the responders provided answers to all of the questions presented to them.

The twenty-five complete responses were tabulated and analyzed in Microsoft Excel. Initial questions were constructed to profile the responders’ job roles, years of experience, and the repository development team size where the responder worked.

Phase 1 Results

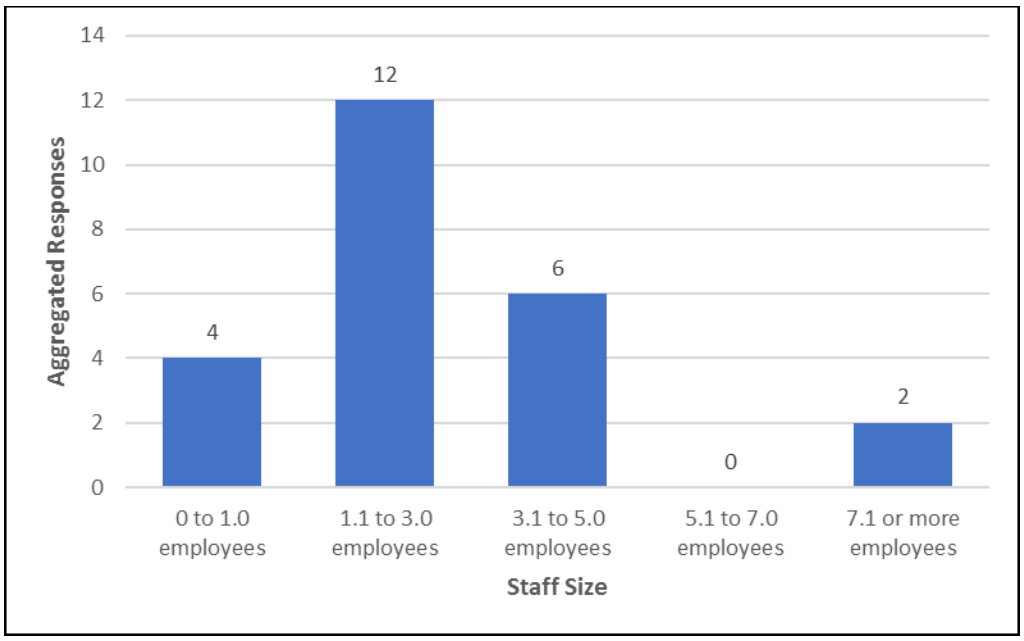
The survey delivered results that related to RQ2 and RQ4. The majority of responders were experienced with development for digital repositories, with an average of 6.25 years of experience, and a maximum value reported of 15 years. The responders either worked at locations developing smaller-scale repositories, choosing "0 to 250,000 digital objects," or very large-scale repositories, responding with the choice of "more than one million objects" (Figure 1).

FIGURE 1: SIZE OF REPOSITORIES (IN DIGITAL OBJECTS)



In response to question Q2.4, in most cases, the department that directly developed the digital repository consisted of three full-time employees. The data for this survey metric was nuanced, however, since one responder provided a response of “50” full-time employees (Figure 2).

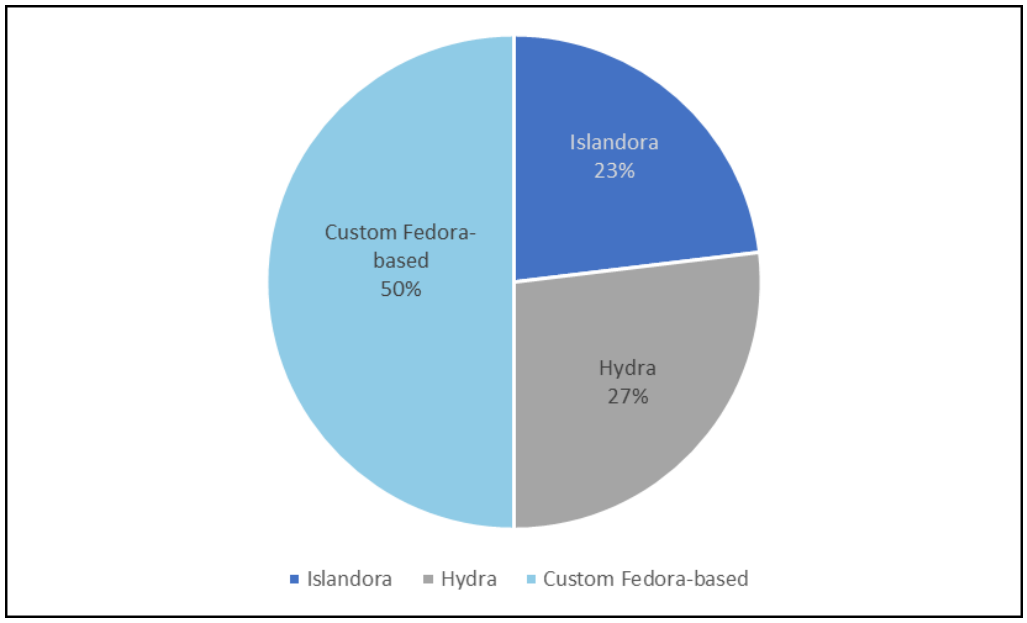
FIGURE 2: STAFF SIZE, N=24



Within the group of twenty-five responses that completed the survey, and accounting for the fact that question Q3.2 allowed responders to select more than one answer, the result points to a strong selection of customized Fedora-based repositories, as opposed to “off-the-shelf” bundled solutions as provided by Islandora or Hydra. With 50% of responders indicating their work location builds a customized Fedora-based repository, and may also include an “off-the-shelf” instance, the results suggest that development teams in these locations

would need to understand the underpinning Fedora architecture for their repository to support all use cases.

FIGURE 3: CHOSEN REPOSITORY PLATFORM



Regarding the move from Fedora version 3 to version 4, an often-cited pain point for many Fedora-based repositories, the responses to questions Q3.3 through Q3.5 all suggest that the vast majority of repositories remain on version 3 without feasibility to complete the move to version 4 within an eighteen-month window. The survey provided the following aggregated results for Q3.3, Q3.4, and Q3.5:

Q3.3: What version of Fedora do you currently use?

Responses:

LEFT BLANK	VERSION 3	VERSION 4
11	19	3

Q3.4: Have you started the upgrade/migrate process (what DuraSpace calls "upgration") from version 3 to version 4 of Fedora?

Responses:

LEFT BLANK	YES	NO
14	9	10

Q3.5: Will your repository complete the "upgration" process in the next 18 months?

Responses:

LEFT BLANK	YES	NO	I'M NOT SURE
14	6	9	4

In terms of staffing resources devoted to the “upgration,” there was an operationalized metric of “man-years” (i.e. 2,000 work hours). This metric was used in questions Q3.6 and Q3.7 to ask responders about their perceived time and effort up to present, as well as going forward. Results to question Q3.6 had marked outliers (Figure 4A) which made averages and other aggregate data meaningless. There was some indication, based on the range of responses, that the survey question may have been too difficult to provide an adequate response. Excluding the outliers of “5,000” and “700” man-years, other responses to Q3.6 resulted in a tighter set of values: [0.5, 1, 2, 4, 6]. This adjusted set has an average of 1.25 man-years (Figure 4B).

FIGURE 4A: MAN-YEARS DEVOTED TO UPGRATION, N=16

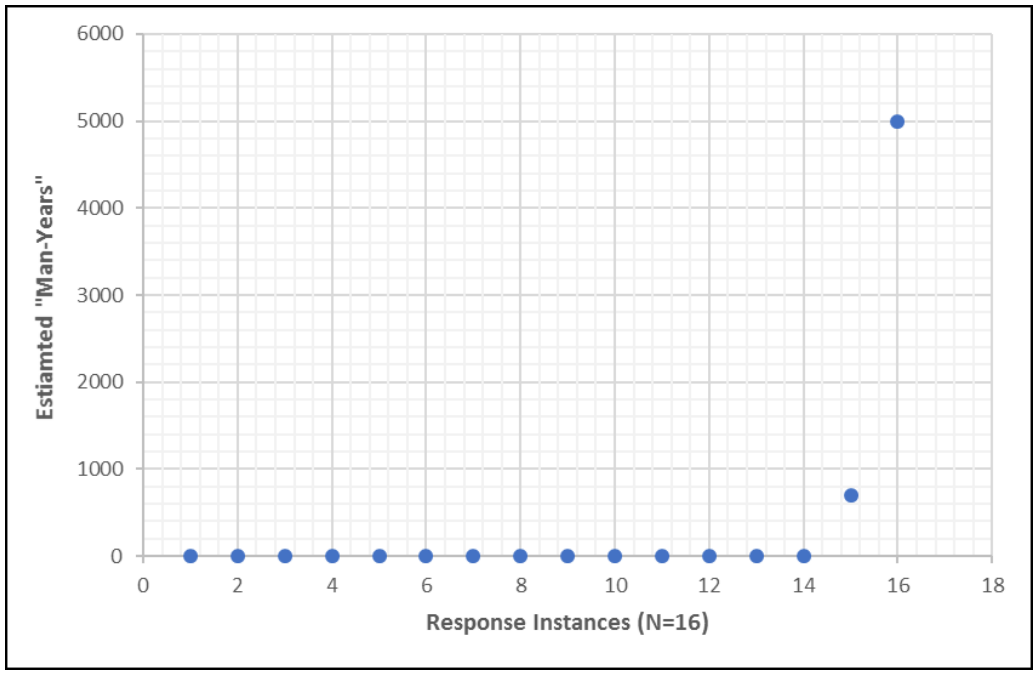
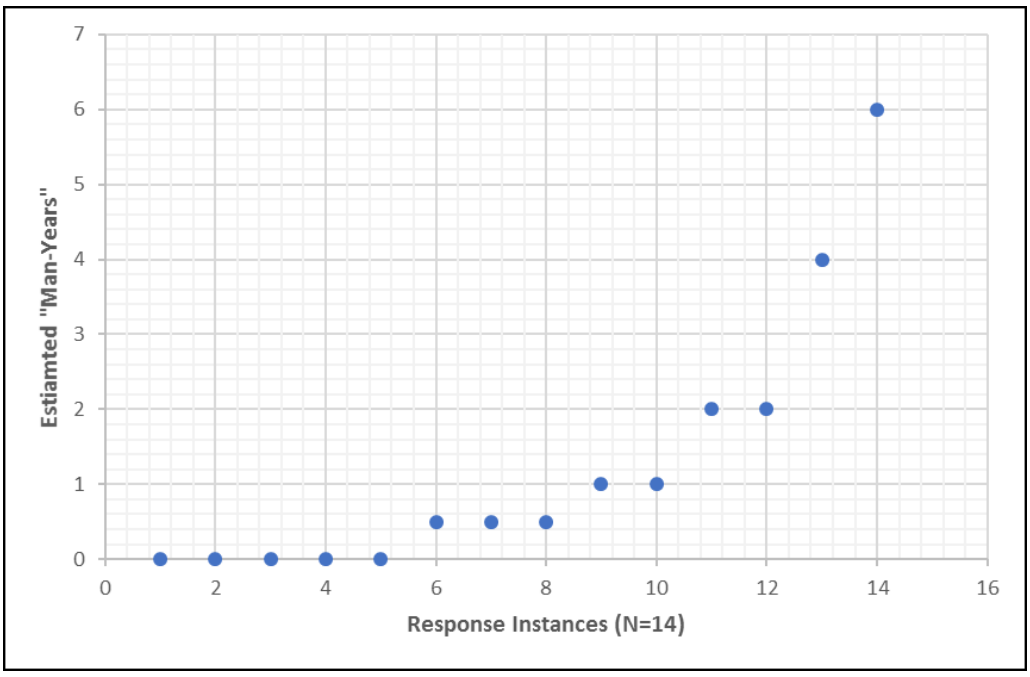


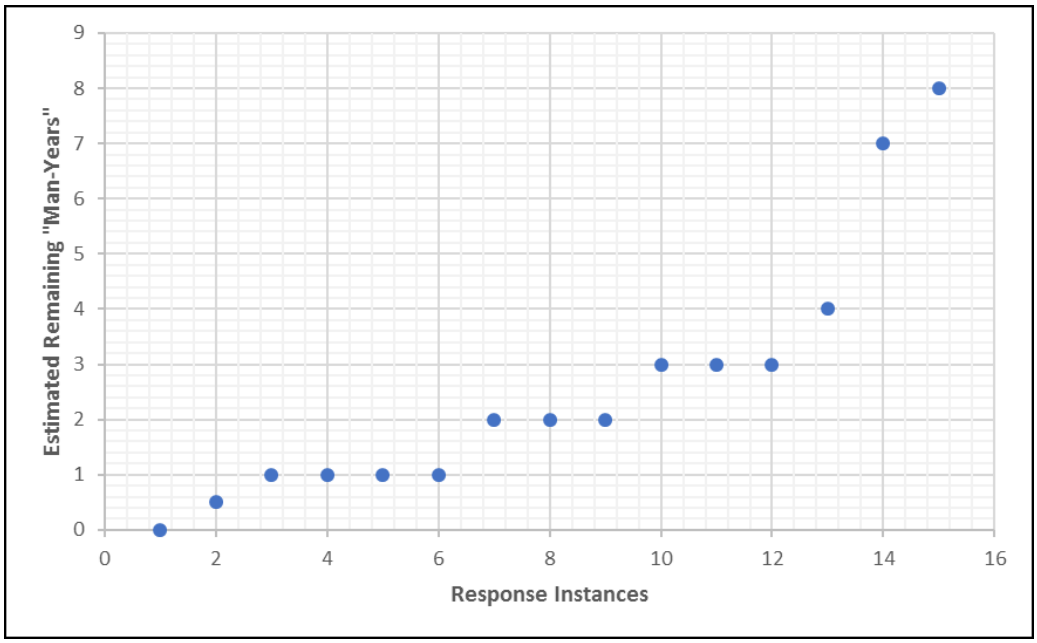
FIGURE 4B: MAN-YEARS DEVOTED TO UPGRATION, REMOVING TWO OUTLIERS, N=14



Within the responses to question Q3.7, again there was an outlier that skewed data. The single outlier value of “3,000” was removed, resulting in a range of

values between zero (0) and eight (8) remaining man-years. The resulting set of responses with exclusion of the single outlier had an average of 2.5667 remaining man-years to complete the upgration (Figure 5). The result from question Q3.7 provided some support to the results of question Q3.5, which suggested that many Fedora-based repositories would remain at version 3 over the next eighteen months, or longer, going forward.

FIGURE 5: MAN-YEARS REMAINING TO COMPLETE UPGRATION[17]



The perceived positive and negatives aspects of moving to linked data within digital repositories was explored within the survey. The aggregated responses might provide some illustration of the perceived benefits of linked data across this community, but the small sample size of twenty-five completed responses to the survey was not large enough to point to any obvious conclusion. Question

[17] Excludes one outlier entry of "3,000" remaining man-years.

Q4.2 was constructed to be an open attempt to gather responses to the notion of perceived benefits of implementing linked data. The results presented some strong indication that the digital repository community was not skeptical of linked data, in theory or practice. The majority of responders indicated there was “moderate benefit.” Notably, no responders chose to answer “no benefit” to question Q4.2.

Q4.2 How much benefit, if any, do you find from implementing linked data in digital repositories?

Responses:

LEFT BLANK	NO BENEFIT	LITTLE BENEFIT	MODERATE BENEFIT	GREAT BENEFIT
10	0	7	10	6

Within the resulting set of twenty-three responses, there is one clear indication. Although the majority of Fedora-based repositories have not yet completed the upgration to a linked data architecture for their data (i.e. version 4), there is a strong sense that the movement in that direction is perceived as positive, and that Fedora 4 offers a discernable benefit to the community.

Survey responses concerning data modeling choices and the adoption of PCDM, were varied and problematic. The question logic itself made an assumption that if you were not adopting PCDM, you inherently had decided to reject it. At least three responses to the questioning in this portion of the

survey suggested that if PCDM was not chosen at that time, it was still a consideration for the future.

There were, however, numerous free-text responses that suggested that PCDM does not support the use cases of many digital repositories. One responder said the following:

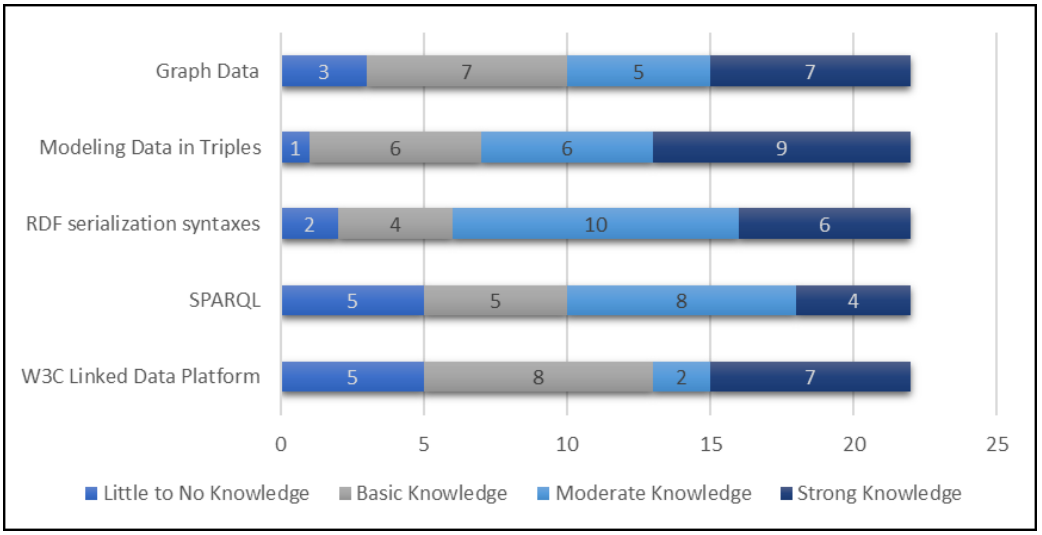
> “There is already a great ontology for structural metadata: ORE. It is an established, well-known standard, and there is no need, from our perspective to re-invent the wheel with PCDM. PCDM divides the world into three categories: Collections, Objects and Files. The notion of a pcdm:Collection is fraught, since it makes particular assumptions about the "identity" of the objects it contains. We needed a "looser" notion of collection: hence dctype:Collection, which we are using. The notion of pcdm:File is unnecessary, since any pcdm:File is already an ldp:NonRDFSource. In our opinion, restating the obvious is not necessary. What is left is a pcdm:Object, which at this point doesn't really serve any purpose beyond being an ore:Aggregation. Furthermore, we are not using explicit typing (rdf:type) in our repository, relying instead on rdfs entailment, which is already a fundamental part of the semantic web architecture and which allows us to infer types. Hence, there is no need for us to use PCDM.”

According to the survey results, there is some impasse between the data modeling recommendation for Fedora, and what is needed among responders. Responses to the question asking whether or not the given repository had adopted PCDM as their chosen primary data model (Q4.3, see Appendix B), responses were ten (10) “Yes” and fourteen (14) “No.”

The research question RQ3, regarding evaluation of the Linked Data Platform (LDP) and benchmarking, was not substantially covered through the survey questions. The final research question RQ4 that focused on training and the knowledge of linked data and associated tools, was covered in full. There were numerous useful short answers, and the multiple-choice matrix question

provided constructive results (Figure 6). It could be concluded that the majority of responders had some strong-to-moderate knowledge of graph data, modeling data in triples, RDF serialization syntaxes, and SPARQL. The noticeable limitation, however, is that the responders did not report the same strong-to-moderate knowledge-base majority regarding the Linked Data Platform. A total of thirteen responders had no-little-basic knowledge, while twelve reported that they had moderate or strong knowledge. While familiarity with linked data itself is not a struggle for most responders, there is still not enough comfort with the W3C Linked Data Platform for the purposes of interacting with Fedora 4.

FIGURE 6: SURVEY RESPONSES FOR FAMILIARITY WITH LINKED DATA



Phase 2 Results

The semi-structured interviews were conducted in-person with participants, and the sessions were recorded and later transcribed. Final drafts of the transcriptions were provided to the participants to proofread and

provide clarifications, and one participant provided edits to the transcription to this end. Resulting texts were imported into NVivo Pro 11, and subsequent themes were created as "nodes," which aided in inductive work to understand the contextual implications of the interviews. Ultimately, the NVivo coding work resulted in conceptual themes grouped to illustrate thoughts around the research questions RQ1 (perceived benefits of Fedora and linked data), and RQ3 (evaluation and benchmarking of the Linked Data Platform).

Separating out themes into nodes within NVivo, numerous areas of concern re-appeared: 1) Fedora's performance concerning large ingest operations and other transactions, 2) descriptive metadata cross-walking from MODS to RDF, and 3) a generalized perception that the Fedora community was un-settled on solutions with regard to areas #1 and #2. Although the sample size for Phase 2 of this study was extremely limited (N=2), some of the resulting transcript data provided greater elucidation of the survey responses, which also expressed similar concerns.

The interviewees did convey that the design of Fedora version 3 allowed for consistent preservation of digital objects. This positive sentiment tended to focus on the software's ability of "keeping track of things well" and that "it definitely gives us stuff in a structured way. For us, it's been really reliable as far as storing our data in a way that's understandable and that we can work with."

The interviewees also conveyed that linked data and RDF have appreciable benefits, but that the promise of what is offered as an increased ability to share data is not yet seen, with one participant saying:

> “It allows us to integrate vocabularies and standardizations, to be able to link out to those things and not define them all locally. It also gives us the promise of interacting with the web in a standard way that is known outside of library-land. So, we like that, because that helps our materials be discoverable, which is right along the lines of why we do things. I would say that we haven't seen the benefits of that stuff yet. It's still very much at the beginning stages.”

With regard to evaluative performance measures and benchmarking of linked data (RQ3) within the given repository at the interviewee’s location, some details came to the foreground during the sessions, but remain open as the location does not have the current infrastructure to do this work. Over time, as the repository at that location completes the Fedora 3 to 4 upgration, they will stress-test their system using a similar amount of data to what is currently in their repository, and will make an effort to do a needs assessment one year after the upgration is complete. This needs assessment will be designed to make performance improvements, begin taking advantage of the aspects of linked data within Fedora 4 that they identify as beneficial, and identify further areas of enhancement.

Discussion

Research Question 1

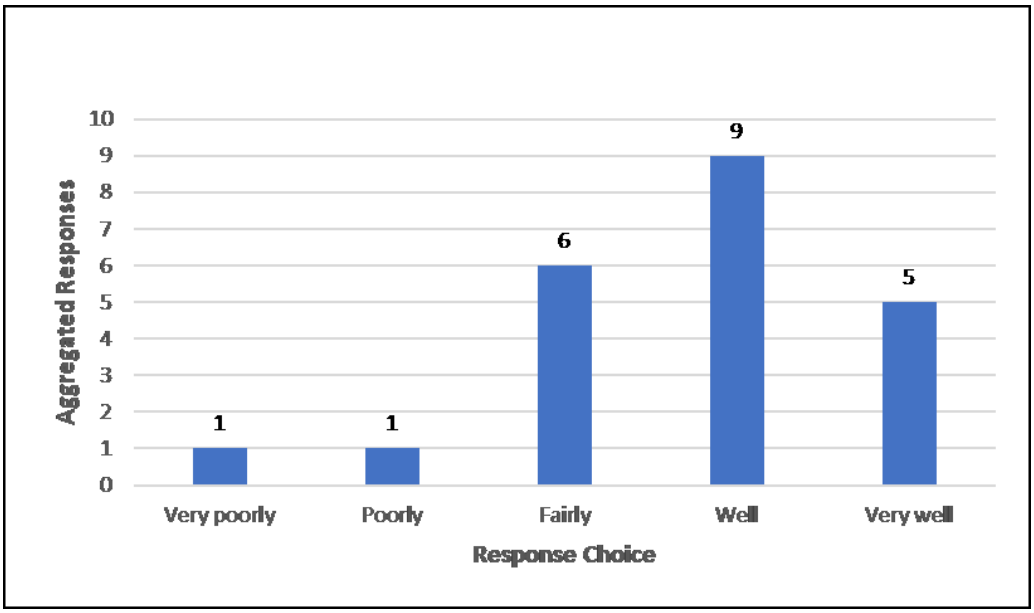
The perceived benefits and ease-of-use of RDF and LDP within Fedora 4 remain limited in implementation, as evidenced by the survey responses. Once response regarding this challenge summarized the problem with clarity:

> "Figuring out how to manage descriptive metadata in Fedora 4 has been challenging. We use MODS and there is no direct simple RDF mapping. Fedora 4 has a hard time managing complex hierarchical RDF so figuring out how we'll be handling descriptive information moving forward has been a big blocker."

Research Question 2

Responses to questions regarding the choice of a data model had at least four re-occurring themes: 1) The majority of responders were not yet committed adopting PCDM, despite the recommendation for that ontology, 2) the Open Archives Initiative's ontology called Object Reuse and Exchange (ORE) had some design benefits over PCDM, 3) there is a development issue in taking descriptive metadata in the form of nested XML (MODS) and translating that metadata at large scale into RDF types, and 4) the majority of responders felt their location implemented their chosen data model well (see Figure 7).

FIGURE 7: HOW WELL DOES YOUR REPOSITORY IMPLEMENT YOUR CHOSEN DATA MODEL?



Going forward, PCDM will continue to be updated, versioned, and enhanced, but at the time of this writing, the data model itself has well-articulated limitations for many digital repositories. One survey responder provided a more high-level reasoning as to why PCDM had adoption in some areas, and was to be avoided in others:

> "PCDM, while a community effort, is probably overly complex for most uses. Community members bend to using PCDM, perhaps because they use Sufia, such as we do, but absent something that 'hides' the complexity of PCDM, such as Sufia does, one would likely implement a simpler data model. Sufia's use of ACLs [Access Control Lists] complicates things, mostly by adding many, many ACLs to the repository unnecessarily."

The MODS-to-RDF open problem requires added attention in order to have more repositories successfully move from Fedora 3 to 4. This problem is compounded by the heterogeneity and inconsistency in data modeling across collections and repositories. One survey response provided a thorough picture of this open problem:

> "Complex XML-based descriptive metadata (such as MODS) is extremely difficult to model as RDF without using blank nodes. Doing this properly requires minting hundreds of thousands of additional objects representing titles, creators, subjects, geographic entities, collections, etc. Fedora is not well-suited to this type of entity creation, meaning an external triplestore is likely to be needed. RDF is also less human-readable, making spot-checking and quality assurance more difficult."

Research Question 3

Neither the survey nor the interview sessions provided much useful data on evaluation of the Linked Data Platform (LDP) within Fedora, but given the results from RQ4 that determined that most responders did not have moderate or strong understanding of LDP, there could be a future study that explores this question further.

Research Question 4

In analyzing the varied approaches that responders took to learning about Fedora, linked data, and associated technologies, there were re-occurring themes that could be summarized as follows.

Textual sources of information included:

- Google, Internet sources, articles.
- W3C documentation
- Reading *Semantic Web for the Working Ontologist* (Allemang & Hendler, 2011)
- Reading *Linked Data for Libraries, Archives, and Museums* (Hooland & Verborgh, 2014)

Audio-visual sources of information included:

- Webinars

- Courses on RDF by Library Juice Academy

In-person consultative sources of information included:

- Engagement with the Fedora 4 community
- Slack channels like #pcdm
- Conferences and workshops (20 responders out of 22 replied "Yes" to having attended Fedora workshops; see Appendix B-Q5.5)

Some responses to the survey also showed that many responders were active practitioners with these technologies, and reported that they are "learning by doing" and experimenting. At least three survey responders were contributors to the Fedora project. This subset of Fedora developers committed code and participated in development sprints.

Study Limitations

The four research questions as designed could be approached through a combination of the two phases, but not completely through any one phase. Centering on only the two main themes (the perceived benefits of linked data, and the state of the Fedora software) through survey methods may have limited the survey's scope to more quantitative questions on those topics. However, the data collected through the survey method was useful for coverage of the broad state of the community, as it relates to adoption of Fedora 4, PCDM, and experience and knowledge of staff.

The interviews did provide a more atomic view of the challenge of upgration, which was not as directly approachable through the survey. Conversely, RQ2 (broad choices around data modeling, and adoption of PCDM)

also was not as approachable via the conducted interviews, being that the goal of RQ2 was to scan the larger community to understand their decisions.

The interviews, had they been increased in scope to include possibly up to ten distinct repository instances at various institutions, could have been used to advance the research questions of adoption of PCDM and data modeling choices, but was not feasible in this study. Additionally, since the "PCDM" Google mailing list was not included as a channel to distribute the survey, there may have been some additional users that could have participated, and possibly offered a more specialized set of responses regarding PCDM.

Conclusion

While the transition to linked data within digital repositories remains ongoing, this study may provide some reporting on the current state of the effort. Responses to the survey pointed to the slow adoption of Fedora 4, which in most cases is still out of reach. Although the upgration process is an open problem, the level of activity within the community provides a bolster to digital repositories going through this process. In terms of data modeling, there are established best practices, and there is no need to start from scratch. However, the slow adoption of PCDM points to an awareness that PCDM itself does not cover several common use cases. Increasing adoption of PCDM may entail a revision based on further input from the community.

Although the majority of responses indicated that the principles of linked data have promise within the context of digital repositories, and the

knowledge-base of linked data technologies among the community is strong, there are also known challenges. Inconsistent data modeling is one such challenge, which is made worse by the need to translate from nested XML structures like MODS to the graph-based structure of RDF. Although the performance at large-scale of Fedora 4 is in some ways related to a limitation of the RDF modeling, it is also considered a separate issue from other metadata efforts focused around overcoming the large heterogeneity of content types within digital repositories.

Another nuance of Fedora is that there are numerous "off-the-shelf" digital asset management solutions such as Hydra, Hydra-in-a-Box, Sufia, and Islandora, which serve as alternatives to a customized Fedora repository (see Figure 3). Approaches to building these "off-the-shelf" systems are more volatile within the community of practice and at the time of this writing, there is substantial "churn" of designs and attitudes: Sufia is now merging with another Fedora-based solution called Curation Concerns to become "HyRax"; Hydra development practice has numerous "bundled solutions" that each support distinct use cases (images versus electronic theses and dissertations, for example); the Hydra Project itself is undergoing a larger rebranding with a forthcoming name-change. The linked data implementations and unique data modeling decisions within these unique platforms could be an area for further exploration in smaller-scale studies.

Purpose-built open-source software has firm standing within LAMs, and Fedora's status as a leading storage and preservation system is firmly cemented

within the LAM digital repository community. As Fedora 4 implementations increase in numbers, it will be necessary for the community to continue shaping the features that are core to the application, and how to ensure that these core features are modeled consistently and effectively going forward. While the Linked Data Platform (LDP) serves as an architectural component to this end, the data collected in this study suggests that LDP's mechanisms and features are not fully understood yet. Increased training on LDP structures and processes within Fedora 4 from DuraSpace, and across the entire LAM community, could shore up broader initiatives to accomplish the task of migrating from Fedora 3 to 4.

Bibliography

Arlitsch, K., Obrien, P., Clark, J. A., Young, S. W. H., & Rossmann, D. (2014). Demonstrating library value at network scale: Leveraging the semantic web with new knowledge work. Journal of Library Administration, 54(5), 413-425. doi:http://dx.doi.org/10.1080/01930826.2014.946778

Berners-Lee, T. (2006). “Linked Data - Design Issues”. Web.

Borst, T. (2014). Repositories on their way into the semantic web: Semantically driven interoperability as perspective for repositories. Bibliothek, 38(2), 257-265. doi:https://dx.doi.org/10.1515/bfp-2014-0034

Corbett, L. E. (2016). Linked Data Advice Anyone? (Who Uses Google?). Technicalities, 36(1), 1–7.

Deng, S. (2016). Preparing for linked data in digital repositories. Accessed 2016-10-13

Hallo, M., Luján-Mora, S., Maté, A., & Trujillo, J. (2016). Current state of linked data in digital libraries. Journal of Information Science, 42(2), 117.

Hardesty, J. L. (2016). Transitioning from XML to RDF: Considerations for an effective move towards Linked Data and the Semantic Web. Information Technology & Libraries, 35(1), 51–64. doi:https://doi.org/10.6017/ital.v35i1.9182

Heath, Tom, and Christian Bizer. “Linked Data: Evolving the Web into a Global Data Space.” Synthesis Lectures on the Semantic Web: Theory and Technology 1.1 (2011): 1–136. CrossRef. Web.

Introduction to communities of practice | Wenger-Trayner. (n.d.). Retrieved from http://wenger-trayner.com/introduction-to-communities-of-practice/

Isaac, A., & Baker, T. (2015). Linked Data Practice at Different Levels of Semantic Precision: The Perspective of Libraries, Archives and Museums. Bulletin of the Association for Information Science & Technology, 41(4), 34–39.

Jones, E. (2016). Linked data for cultural heritage (UK ed.). London: Facet Publishing.

Konstantinou, N., Kouis, D., & Mitrou, N. (2014, June). Incremental export of relational database contents into RDF graphs. In Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14) (p. 33). ACM. Accessed 2016-10-13.

Konstantinou, N., Spanos, D. E., Kouis, D., & Mitrou, N. (2015). An approach for the incremental export of relational databases into RDF graphs. International Journal on Artificial Intelligence Tools, 24(2), 1540013. Accessed 2016-10-13.

Konstantinou, N., Spanos, D.E. and Mitrou, N. (2013). "Transient and persistent RDF views over relational databases in the context of digital

repositories". In Metadata and Semantics Research (MTSR 2013), Thessaloniki, Greece, pp. 342-354. Accessed 2016-10-13.

Lagoze, C., Payette, S., Shin, E., & Wilper, C. (2006). Fedora: an architecture for complex objects and their relationships. International Journal on Digital Libraries, 6(2), 124–138. https://doi.org/10.1007/s00799-005-0130-3

Latif, A., Borst, T., & Tochtermann, K. (2014). Exposing data from an open access repository for economics as linked data. D-Lib Magazine, 20(9), 2. doi:http://dx.doi.org/10.1045/september2014-latif Accessed 2016-10-13.

“Linked Data Platform 1.0 Primer.” N.p., n.d. Web. 2 Apr. 2017.

Minton Morris, C. c. (2014). Unpacking Fedora 4. D-Lib Magazine, 20(7/8), 11-12. Accessed 2016-10-13.

Mitchell, E. T. (2015). The Current State of Linked Data in Libraries, Archives, and Museums. Library Technology Reports, 52(1), 5-13.

Myntti, J., & Cothran, N. (2013). Authority control in a digital repository: Preparing for linked data. Journal of Library Metadata, 13(2-3), 95-113. doi:http://dx.doi.org/10.1080/19386389.2013.826061 Accessed 2016-10-13.

Nicholas, D., Rowlands, I., Watkinson, A., Brown, D., Russell, B., & Jamali, H. R. (2013). Have digital repositories come of age? The views of library directors. Webology, 10(2), 1.

Payette, S., & Lagoze, C. (1998, September). Flexible and extensible digital object and repository architecture (FEDORA). In International Conference on Theory and Practice of Digital Libraries (pp. 41-59). Springer Berlin Heidelberg. Retrieved from http://arxiv.org/abs/1312.1258

Preedip Balaji Babu, Kadari Santosh Kumar, Nilesh A. Shewale, & Abhinav K. Singh. (2012). Rationale of institutional repository categories and IR development challenges in India. Library Review, 61(6), 394–417. doi:https://dx.doi.org/10.1108/00242531211284320

Radio, E., & Hanrath, S. (2016). Measuring the impact and effectiveness of transitioning to a linked data vocabulary. Journal of Library Metadata, 16(2), 80-94. doi:http://dx.doi.org/10.1080/19386389.2016.1215734

Raymond, E. S. (2001). The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. Sebastopol, CA, USA: O'Reilly & Associates, Inc.

Salo, D. (2010). Retooling libraries for the data challenge. Ariadne, (64). Accessed 2016-10-13.

Smith-Yoshimura, K. (2016). Analysis of international linked data survey for implementers. D - Lib Magazine, 22(7), 1. Retrieved from doi:https://doi.org/10.1045/july2016-smith-yoshimura

Solodovnik, I. (2013). Development of a metadata schema describing institutional repository content objects enhanced by 'LODE-BD'

strategies. JLIS.it: Italian Journal of Library and Information Science, 4(2), 109-144. doi:http://dx.doi.org/10.4403/jlis.it-8792

Southwick, S. B. (2015). A guide for transforming digital collections metadata into linked data using open source technologies. Journal of Library Metadata,15(1), 1-35. doi:https://doi.org/10.1080/19386389.2015.1007009

Staples, T., Wayland, R., & Payette, S. (2003). The Fedora Project. D-Lib Magazine, 9(4), 1082-9873. http://dlib.org/dlib/april03/staples/04staples.html Accessed 2016-11-21.

Stuart, D. (2010). Linked data and government data: More than mere semantics. Online, 34(3), 36.

Stuart, D. (2014). Librarians Should Embrace Linked Data. Research information, 71, 14.

Stanford University & Complaints, C. 94305 C. (n.d.). Linked Data. Retrieved April 1, 2017, from https://library.stanford.edu/projects/linked-data

Appendices

Appendix A: Glossary

- Community of Practice: “Communities of practice are groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly.” (“Introduction to communities of practice | Wenger-Trayner,” n.d.)
- FEDORA: Originally published by Sandra Payette and Carl Lagoze as a “Flexible and Extensible Durable Object Repository Architecture,” the FEDORA concept has evolved to become an open-source storage and digital preservation platform suited for digital repositories.
- Linked Data: According to the W3C’s Linked Data Platform 1.0 Primer, Linked Data “refers to an approach to publishing data that puts linking at the heart of the notion of data, and uses the linking technologies provided by the Web to enable the weaving of a global distributed database. By naming real world entities - be they web resources, physical objects such as the Eiffel Tower, or even more abstract things such as relations or concepts - with http(s) URLs, whose meaning can be determined by dereferencing the document at that URL, and by using the relational framework provided by RDF, data can be published and linked in the same way web pages can.” (“Linked Data Platform 1.0 Primer,” n.d.) A layperson’s definition might be that rather than a “web of documents” as originally implemented through

- the World Wide Web, and which has become the prevailing web architecture that links information, the “web of data” connects individual data points via unique identifiers (Uniform Resource Identifiers, or URIs). This makes it possible to form bridges between data points that otherwise would exist in silo-ed HTML structures and documents.

Appendix B: Qualtrics web survey for Phase 1 of study

Survey Title: “Upgrading to Linked Data in Digital Repositories”

Q1.1 Does your current job entail responsibilities or involvement in any aspect of a digital repository?

- Yes
- No

Q2.1 In terms of digital objects, what is the estimated size of the repository you work on?

- 0 to 250,000 objects
- 250,000 to 500,000 objects
- 500,000 to 750,000 objects
- 750,000 to 1,00,000 objects
- more than one million objects

Q2.2 How many years of experience do you have in developing or supporting a digital repository? (Please enter a number.)

Q2.3 What duties are you tasked with for your digital repository?

Q2.4 Considering all staff, what is the total Full-Time Equivalent (FTE) labor effort spent directly on the digital repository? One FTE represents 40 hours of labor per week. For example, if there are three employees that work 40 hours per week for the digital repository, and one that works 20 hours per week, the total is 3.5 FTE. Please record the total number of FTE.

Q3.1 Does your repository currently use Fedora? (If not, what storage and preservation architecture do you use?)

- Yes
- No _____

Condition: No Is Selected. Skip To: End of Block.

Q3.2 Do you currently use a Fedora-based implementation such as Hydra or Islandora? (You may select more than one response.)

- Hydra
- Islandora
- customized Fedora implementation not using Hydra or Islandora

Q3.3 What version of Fedora do you currently use?

- Version 3
- Version 4

Condition: Version 4 Is Selected. Skip To: End of Block.

Q3.4 Have you started the upgrade/migrate process (what DuraSpace calls "upgration") from version 3 to version 4 of Fedora?

- Yes
- No
- I'm not sure

Display This Question: If What version of Fedora do you currently use? Version 3 Is Selected

Q3.5 Will your repository complete the "upgration" process in the next 18 months?

- Yes
- No
- I'm not sure

Q3.6 Considering the work done so far to accomplish the upgration process, how many "man-years" has it taken to get to the point you are at now with this process? (Please enter a number.)

One man-year is equivalent to the number of hours a full-time employee works in a given 52-week period. We can say one man-year amounts to 2,080 hours (40 hours x 52 weeks). To calculate man-years, you can consider the total hours worked on the "upgration" from all staff that are involved, and divide by 2,080. For example, if total staff for a given institutional repository do "upgration"-related work for a combined total of 16,000 hours in a given year, the resulting man-years are 16,000/2,080 = 7.7 man-years.

Q3.7 Based on the point your digital repository is at currently, how many more man-years do you estimate it might take to complete the upgration? (Please enter a number.)

Q4.1 What data modeling difficulties, if any, have you experienced with the digital objects in your current repository architecture?

Q4.2 How much benefit, if any, do you find from implementing linked data in digital repositories?

- no benefit
- little benefit

- moderate benefit
- great benefit

Q4.3 Have you adopted the Portland Common Data Model (PCDM) in your repository?

- Yes
- No

Condition: No Is Selected. Skip To: What factored into your decision to reject…

Condition: Yes Is Selected. Skip To: What factored into your decision to adopt...

Q4.4 What factored into your decision to adopt PCDM?

Condition: What factored into your dec... Is Displayed. Skip To: How well do you think your digital re....

Q4.5 What factored into your decision to reject PCDM?

Q4.6 How well do you think your digital repository implements your chosen data model?

- Very poorly
- Poorly
- Fairly
- Well
- Very well

Q5.1 How would you describe your own level of knowledge with each of the following linked data concepts?

	little to no knowledge	basic knowledge	moderate knowledge	strong knowledge
graph data				
modeling data in triples				
RDF serialization syntaxes (Turtle, RDF/XML, JSON-LD, etc)				
SPARQL				
The W3C Linked Data Platform				

Q5.2 How did you gain the knowledge you currently have of linked data? Here you can provide details on the courses, workshops, or webinars you've taken part in, useful reference sites, and any other details you like.

Q5.3 Please describe your experience working with RESTful web APIs.

This could include interacting with web services via HTTP methods, writing API documentation, and/or building RESTful web APIs.

Q5.4 In which ways do you interact with the Fedora community of practice?

- Through Google mailing lists like "Fedora Tech", "Fedora Community", "Hydra Tech", "Hydra Community", "Islandora", etc.
- Through IRC channels like #code4lib, #pcdm, etc
- DuraSpace scheduled tech meetings
- in-person consultations

- other _____
- I don't have any interaction with the Fedora community

Q5.5 Have you attended a workshop on Fedora by DuraSpace or another workshop that teaches the concepts and features of Fedora?

- Yes
- No

Appendix C: Interview questions for Phase 2 of study

Introduction:

1. Please describe your title and responsibilities within your digital repository.

2. What is the mission of your digital repository?

3. Please describe the current data modeling for digital objects in your repository.

Benefits, Ease-Of-Use, and Usefulness of Linked Data:

4. In general, how does implementing linked data support the mission of the repository?

5. Through implementing Fedora 4, how does the repository achieve its goals?

6. In your current implementation, what feature(s) of the Fedora software are a pain point for you and/or your staff?

7. What feature(s) of the Fedora software work well in your current implementation?

“Upgration" to Fedora 4:

8. What parts of your future implementation of Fedora might require the most staff time and effort?

9. Since you have been working on the “upgration” in going from Fedora 3 to Fedora 4, what would you say is the current prognosis?

10. What factors would you say make your repository’s "upgration" challenging? What do you feel is the most important thing to focus on right now?

Other Questions:

(If we discuss scale or performance issues like caching, etc.)

11. Have you implemented a strategy to deal with problems related to look-up times, i.e. caching of URIs and external linked data endpoints?