The Functions of (Meta)Data: Lessons Learned with a Fedora Digital Repository

Jennifer M. Eustis
University of Connecticut - Storrs, jennifer.eustis@lib.uconn.edu

Follow this and additional works at: <https://docs.lib.purdue.edu/charleston>

Part of the [Library and Information Science Commons](#)

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

Jennifer M. Eustis, "The Functions of (Meta)Data: Lessons Learned with a Fedora Digital Repository" (2012). *Proceedings of the Charleston Library Conference*.
<http://dx.doi.org/10.5703/1288284315139>

The Functions of (Meta)Data: Lessons Learned with a Fedora Digital Repository

Jennifer M. Eustis, Catalog/Metadata Librarian, University of Connecticut Libraries

Abstract

The University of Connecticut Libraries began building a Fedora digital repository last year. Because of the differences between Fedora and relational databases, it was necessary to understand how Fedora works with objects and data streams. The repository team realized that with Fedora, there existed several options on how to store data. This realization encouraged looking at metadata differently. For starters, we began to emphasize functions over types of metadata. Secondly, we saw the advantages of striping meta from the word metadata. This change allowed us to conceptualize a broader application of functional data within the repository. My presentation would like to explore our emphasis on the functions of data rather than types of metadata and how this is helping to create a better digital repository.

Fedora and the University of Connecticut Libraries

The University of Connecticut Libraries, like many of its peer institutions, provide solutions for the institutional output of the university. Currently, these solutions are both vendor specific. The first is provided by bepress and called Digital Commons. Digital Commons is essentially the University of Connecticut Libraries' institutional repository. The second vendor solution is ContentDM from OCLC. Used primarily for images, ContentDM is used in conjunction with other institutions throughout Connecticut for materials housed in archives and special collections. These platforms work and have been able to meet much of our digital initiative needs until recently.

The recent change that prompted reevaluating our current platforms for digital collections revolves around scientific research, digital humanities, data from the state, digital preservation, and TRAC compliancy. Scientific research, sometimes referred to eScience or eResearch, is currently undergoing a huge transformation in terms of how data is discovered, accessed, and shared (Gabridge, 2009). It is even a question of the amount of data produced, especially in intensely computational scientific research areas. Researchers are searching for a number of solutions that include huge amounts of storage, file management systems, and training on a number of issues ranging from data security to how to write a data management plan. The

second phenomenon affecting the University of Connecticut Libraries is the recent creation of a committee to investigate solutions for the digital humanities. Such solutions include looking into the creation of a digital humanities center or how to promote and support digital research and teaching in the humanities. A third recent development is that the University of Connecticut Libraries Map and Geographic Information Center is digitizing its collections in the public domain in addition to its being part of the Connecticut State Center, whose mission is to assist the state in meeting its obligations to provide reliable census data.

The last two concerns to affect how we store and manage digital collections touch on more internal questions posed by the library. These are digital preservation and TRAC compliancy. Digital preservation is an upmost concern in regards to the fragile nature of digital materials. A good example is the number of web sites that are ephemeral in nature or bit rot that can make content unusable. TRAC, or Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC, 2007), is a document that sets out to guide users on how to ensure their repositories are trustworthy in the way that data is stored and managed. According to TRAC, "the definition of a trusted digital repository must start with 'a mission to provide reliable, long-term access to managed digital resources to its designated community, now and into the future'" (TRAC, 2007).

The conflict at the University of Connecticut Libraries arose when these initiatives and internal questions posed by the library met with the current two platforms used for digital collections. Though both Digital Commons and ContentDM work and fit the needs of the library at the time they were implemented, both solutions are unable to deal with scientific research, digital humanities on a large scale, digital preservation, and especially TRAC compliancy. With our instance of Digital Commons, large scientific data sets that are in different formats from a Word document, spreadsheet, or even PDF are not able to be added to Digital Commons. There is also an issue with the size of the document, which we currently limit to 5kb, which is hardly enough storage space for many of the scientific data sets produced. The same is true with ContentDM. Both solutions also don't offer a trusted and reliable repository according to the criteria of the TRAC document.

In light of this, the University of Connecticut Libraries reevaluated how they store and manage digital collections and decided upon Fedora as a solution. Fedora, or Flexible Extensible Digital Object Repository Architecture, is an open source repository software to manage and preserve digital content. Fedora does not dictate the type of content model architecture needed to design the system. In other words, Fedora does not limit you to one pattern of content models and how these different types of content (books, journals, video, etc.) relate to one another. This freedom allows institutions to develop a content model architecture that fits their needs and the needs of their users. On the other hand, this means that Fedora requires a substantial effort in its design and implementation.

Part of the design and implementation necessitates thinking about the type of data required and options that will be a part of the content. These data include the data that will be submitted to users in addition to metadata. Fedora only requires one type of metadata,

namely simple Dublin Core, formatted to the OAI-PMH standard for harvesting; actually, it is even possible to side step this requirement. Beyond that, Fedora does not require any other metadata standards, a particular set of simple Dublin Core elements, or any data content prescriptions. These questions are open to development on the part of the team that builds Fedora for their institution. The result is that there are a number of issues to think through when it comes to developing best practices and implementing them with metadata in Fedora. Before considering an approach on how to do this, I want to take another detour into some common approaches to metadata.

A Common Approach to Metadata

A common approach to metadata can be coined as the singular standard type. In short, metadata is referred to in the singular. The focus is on the idea of metadata or metadata as a global concept. This concept is then organized into types. These types act as broad categories. Examples are descriptive, technical, rights, administrative, preservation, and/or structural. These types are then further categorized by standards. For instance, the metadata standard METSRights falls into the type of rights metadata, the standard PREMIS is preservation metadata, etc. Another type of approach is to list the categories by the type of data as in data content standards, data value standards, or data structure standards. Unlike the first example, this melds the discussion on types and standards together.

There are three good examples of this approach. The first is presented by Steven Miller and his metadata resource web page (Miller, 2011). Figure 1 is a section of Miller's metadata resource page that illustrates the singular standard type approach to metadata.

Miller begins with a singular definition of metadata and then passes quickly onto the broad categories of metadata. Notice that Miller uses

Metadata

- **Meta + Data = Data about data: structured statements (properties and values) that** gh
Metadata is logically separate from the resource it is about, even when embedded within the

Introductions to Metadata for Cultural Heritage Applications

- Miller, Steven J. *Metadata for Digital Collections: A How-To-Do-It Manual*. New York: N
- Coyle, Karen. *Understanding the Semantic Web: Bibliographic Data and Metadata*. Vo
- Zeng, Marcia Lei & Jian Qin. *Metadata*. New York: Neal-Schuman, 2008. ISBN: 97815557
- Foulonneau, Muriel, and Riley, Jenn. *Metadata for Digital Resources: Implementatio*
- Baca, Murtha (editor). *Introduction to Metadata*. Second Edition. Los Angeles: Getty R
- NISO. *Understanding Metadata*. National Information Standards Organization, 2004. Online
- Smiraglia, Richard P., editor. *Metadata: A Cataloger's Primer*. New York: Haworth, 200
- Haynes, David. *Metadata for Information Management and Retrieval*. Facet, 2004.
- Hillmann, Diane I., and Elaine L. Westbrooks. *Metadata in Practice*. American Library Asso

Typology of Metadata Standards

1. **Data Structure Standards** (element sets; schemes, schemas, or schemata) (examples: D
2. **Data Content Standards** (cataloging rules, input standards, best practice guides) (example
3. **Data Value Standards** (controlled vocabularies, encoding schemes) (examples: LCSH, AA'
4. **Data Format / Technical Interchange Standards** (encoding standards for machine proces
5. **Data Presentation Standards** (examples: ISBD punctuation, CSS and/or XSLT for display, OI

Figure 1. Section of Miller's Metadata Resource Page

the second approach where metadata is broadly categorized by type of data standard.

Another example (Figure 2) is the visual representation of metadata standards by Jenn Riley (Riley, 2010).

In this example, the focus is on metadata standards. In terms of function, the goal is to see how a standard meets the needs of users. What is interesting about this approach is that metadata is conceived as a collection of different standards used according to domain, community, purpose, and function.

A last example is the document, “Understanding metadata” (NISO, 2004). The first question asked by this document published by NISO is: what is metadata? It then goes on to explain the types and standards of metadata. This approach echoes that of Steven Miller and Jenn Riley as presented in Figures 1 and 2.

There is absolutely nothing wrong with this approach. Indeed, this is how I learned about metadata, in addition to many of my colleagues. However, there are other ways to conceptually think of metadata. Working with Fedora has helped to think differently about metadata. More

Figure 2. Seeing Standards by Jenn Riley

specifically, Fedora can help go beyond the singular standard type approach by adding a more flexible perspective on metadata. Instead of metadata is, metadata are. Instead of types of metadata, what functions do metadata need fulfill? Instead of metadata standards, metadata are data, which brings into play how these data will be stored, managed, preserved, presented, etc. In brief, the flexible architecture of Fedora almost begs to adopt a flexible process with metadata.

Meeting of Fedora and this Common Approach to Metadata

In the previous section, I mentioned the importance of flexibility with metadata. Remember that Fedora in and of itself is flexible. It requires only a simple Dublin Core as part of the metadata that corresponds to the OAI-PMH standard. Beyond that, there are no prescriptions. It should be known that this simple OAI Dublin Core data is in every Fedora object where data is stored. Because of Fedora's flexibility, the number and type of objects that constitute the content model architecture will vary by institution. For example, at the University of Connecticut Libraries, we have a granular content model architecture. Metadata will be found on three levels which we call grouping, container, and media objects. The grouping object refers to a "grouping" of similar digital resources. The container consists of a specific resource. The media object is where the actual digital resource is stored and managed. For example, for the Eric Reeves Papers, the grouping object would refer to this general, overall group of digital resources that constitute the Eric Reeves Papers. One container object would be the images; another could be letters, etc. The media object directly related to the images would refer to one media object per image. Metadata is found in each level and every object. By default, each Fedora object has a simple OAI Dublin Core record. In this way, users can add metadata to provide information about the Eric Reeves Papers as a collection of resources, the images in that collection or individual digital resources. Each data can be related to one another through a Fedora RDA

Relationship Ontology. After a fashion, this is similar to FRBR.

One issue is that there is no predetermined set of implementation guidelines for metadata. Even with the knowledge that the simple OAI Dublin Core record is required in every Fedora object, there are no Fedora requirements for what this record must contain in terms of data. This leaves the question of how to implement data very openly. It is certainly possible and a good idea to take the approach of singular standard type data. In this case, you would look into what metadata types you need to have in your repository and then the appropriate standards that fit. It could be the case that you will just use the simple OAI Dublin Core since it is already required by Fedora. However, this might not be the case. This either/or case begs the question of just what data you need to store and manage. This leads to the question of why do you need metadata or what is the purpose of the digital repository. Remember that in the case of the University of Connecticut Libraries, we would like a Fedora digital repository to help researchers store and manage data sets. Further, we would like our repository to be TRAC compliant, provide indexing and discovery tools, data visualization and manipulation, metadata management, permanent citable links, and among other things adherence to community standards to satisfy grant funder requirements. It became clear to us here at the University of Connecticut Libraries that relying solely on simple OAI Dublin Core records was not enough. Moreover, thinking of just types of metadata and related standards was not enough. This was one of the many lessons learned with Fedora.

Lessons Learned

The overarching lesson learned was to change perspective and view metadata beyond that of a singular standard type. This change of perspective involved three ideas. The first was to switch from defining metadata in the singular to the plural. Metadata are. The second was to focus on data and understand that metadata are data. The third was to switch the focus when initially thinking of metadata best practices and implementation from standards to data functions. This change did not entail leaving the singular standard approach

behind. Rather, this perspective acted as an additional method to thinking about metadata that was particularly helpful with the case of Fedora.

The perspective of metadata as data and in the plural helped on a number of fronts. Metadata are data. Just like any data, it is necessary to take into consideration how to preserve metadata. Data also can be expressed in a number of formats beyond text or character strings. At the University of Connecticut Libraries, one issue that we encountered was that metadata was often associated with the MARC record as found in the catalog. To help people move beyond this view, metadata as data helped us promote that metadata are expressed in a number of different formats such as text, date, numbers, or even linked data. In this way, we were able to better conceptualize how (meta)data could be automatically generated and managed for consistency and accuracy. In this case, we even began to envision using calendar widgets to input dates instead of relying on textual strings. Another advantage of this perspective was the consideration that metadata librarians at the University of Connecticut Libraries could participate in the creation, maintenance, and policy making of metadata beyond pure description. In other words, this was an opportunity to enhance the role played by metadata librarians at the University of Connecticut Libraries. Another interesting result of talking about metadata in the plural and as data was the ability to go beyond metadata librarian jargon. This was instrumental in better communicating with other units involved in the Fedora project such as application developers and digital preservation librarians.

Metadata in relation to functions was also instrumental in helping people understand the

role of metadata in Fedora. Before talking about types and standards of metadata, at the University of Connecticut Libraries, we had to figure out why we needed Fedora and what it was going to be used for. Further, we had to develop use case scenarios. This included having discussions on the scope of the project, requirements and limits, people involved, the systems needed, used or wanted, logistics, and data models. Further, it required seeing that metadata has a lifecycle with different needs at each point in the cycle. As a result, we began to see metadata as dynamic and flexible in the different roles these metadata needed to fulfill.

This change of perspective might seem superficial at first glance or simply an obvious change. In the singular standard approach, it is necessary to ask many of these questions on what functions the metadata standards or types of metadata need to fulfill. However, it is sometimes the simpler ideas that have the greatest impact. At the University of Connecticut Libraries, metadata was associated with the standard MARC21 for encoding. Metadata was also seen as a static product that hardly changed in the catalog. In regards to digital collections, metadata even took a second seat since full-text searching was seen as the hopeful solution for complete discovery and access. Thinking of metadata as data, in the plural, and as dynamic, was instrumental in allowing staff to pass beyond MARC21. This also helped staff see how metadata could help discovery and access in addition to full-text searching. As a result, metadata are beginning to be seen as instrumental in fulfilling functions such as timelines, faceted searching, data visualization, mapping, or preservation. Metadata are also seen to be dynamic, having a life cycle, and also needing the attention of preservation to help have a TRAC compliant repository.

References

Fedora Commons. (n.d.). <http://fedora-commons.org/>

Gabridge, T. (2009). The last mile: The liaison role in curating science and engineering research data. *Research Library Issues*, 265. Retrieved from <http://www.arl.org/bm~doc/rli-265-gabridge.pdf>

- Miller, S. (2011). Metadata resources: Selected reference documents, web sites, and readings. Retrieved from <https://pantherfile.uwm.edu/mlt/www/resource.html>
- NISO. (2004). Understanding metadata. <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>
- Riley, J. (2010). Seeing standards: A visualization of the metadata universe. Becker, D. (Illus.). Retrieved from <http://www.dlib.indiana.edu/~jenlrile/metadatamap/>
- Trustworthy repositories audit and certification: Criteria and checklist (TRAC). (2007, February). Center For Research Libraries. Chicago, IL. Retrieved from http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf