DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2011

Metadata for a Micro-services-based Digital Curation System

Kevin Clair
Pennsylvania State University Libraries, USA
kmc35@psu.edu

Abstract

The Libraries and Information Technology Services at the Pennsylvania State University are in the process of developing a service architecture for supporting digital curation and preservation activity at the university. This system, called Curation Architecture Prototype Services (CAPS), is built on the micro-services approach to digital curation pioneered by the California Digital Library. The paper details methods and philosophies related to metadata development for this system, and how those methods align with the general approach of the micro-services model. The current state of production of this architecture is detailed, along with future metadata services to be embedded in the system, and how those services will be deployed in collaboration with its stakeholders.

Keywords: metadata; digital preservation; digital curation; electronic records management

1. Introduction

In 2011, the Pennsylvania State University Libraries and Information Technology Services began work on the Curation Architecture Prototype Services (CAPS) initiative for digital curation services. CAPS uses a micro-services approach to digital curation in order to enable agile development of curation services that can be deployed quickly, in a decentralized way, in a number of different operating environments. A key component of this work is the development of extensible metadata guidelines to enable certain micro-services to function, for the development of agile information architecture, and for the metadata contained in the CAPS environment to be compatible with emerging linked data and Semantic Web applications.

This paper reports on the first phase of development of the CAPS platform at Penn State, with a particular emphasis on how its baseline metadata schema was designed. Metadata development for CAPS aligns generally with the principles of agile development underlying the micro-services model of digital curation. The paper details the ways in which direct engagement with prospective users and early adopters of the system led to an early baseline for descriptive metadata, and the specific vocabularies CAPS supports. It concludes with a discussion of the transition from CAPS to the Open Curatorial and Archival Services Architecture (OpenCASA), a planned institutional repository, and how future metadata development will evolve to support its needs.

2. About Curation Micro-services

The curation micro-services model was initially developed by the California Digital Library in order to re-conceptualize digital preservation as a set of inter-connected preservation services, developed and deployed in a decentralized manner, rather than as a centralized, monolithic *place* at which curation occurs (Abrams, Cruse, & Kunze, 2008). The advantages of such an approach are that individual services are easier to maintain than the same service embedded within a larger repository framework, as well as being easier to optimize for specialized environments outside of the central research library or IT unit.

There are four main strategic goals addressed by the micro-services approach to digital curation (Abrams, Kunze, & Loy, 2010):

https://doi.org/10.23106/dcmi.952135737

- Safety through redundancy ("lots of copies keep stuff safe")
- Maintaining meaning through description ("lots of description keeps stuff meaningful")
- Utility through services ("lots of services keep stuff useful")
- Adding value through use ("lots of uses keep stuff valuable")

In effect, by encouraging continued use of the curation platform, both in quantity and diversity of use cases, the strategic goals of the micro-services approach are meant to ensure the sustainability of both the system and the digital objects contained within it.

Embodied within the CDL development process of a micro-services curation framework are a number of agile development principles, such as a continuous flow of development through rapid prototyping and testing of services, and early and persistent engagement with the intended user community to develop functional requirements. Combined with an iterative approach to building, releasing, and refining new and existing services within the framework, this allows for sustainable development of the platform.

As previously mentioned, the distinction of the micro-services model is its emphasis of the services associated with digital curation, as opposed to the location in which those activities take place. To this end, metadata services are implicated in a variety of curation activities within a micro-services curation system, to a variety of degrees. This is most apparent in the second goal of micro-services (maintaining meaning through description), and in the services of inventory and annotation. Metadata makes the digital objects discoverable within the system; relatable to other objects within and outside of the local environment; and able to be enriched by curators and end-users for continuous improvement of discoverability and usability within the curation system. Because these metadata services are embedded throughout the stack of micro-services, a plan for metadata that encompasses the whole stack of applications and the entire lifecycle of a digital object within it is essential to such a system's long-term success.

Micro-services with direct applications for metadata services include identity (minting persistent, unique identifiers for a digital object), inventory (associating metadata with digital objects), and annotation (adding metadata to an object). Once a collection of metadata to describe digital objects is in place, other services that act upon the metadata index can be utilized to a greater extent. This includes notification (alerting users to the presence of new digital objects in the system), transformation (creating derivatives of existing digital objects), as well as basic searching and indexing services. Expressing this metadata using agreed-upon standards such as JSON and RDF, and creating it using widely-used controlled vocabularies such as DBPedia or Library of Congress Subject Headings, ensures its interoperability across system boundaries.

3. Project Plan

CAPS began with a three-month development cycle. The goal of this cycle was to determine whether the release of a micro-services curation system within the Penn State University Libraries was feasible. The project team consisted of a digital library architect, a digital collections curator, a software developer, an archivist, a metadata librarian, and a project manager. Additionally, a team of early-adopting stakeholders from within the Libraries, including subject librarians, archivists, and other users familiar with the Libraries' current efforts in digitization and preservation, was assembled to consult on needs assessment and functional requirements.

CAPS required the design of a basic metadata model, suitable for basic description of all types of resources, and around which applications for specific content types could be developed. Preliminary design was done through discussions with the stakeholder team, to determine how they currently manage metadata for digital objects and their specific needs for elements, vocabularies, and so forth. Based on this information, a baseline metadata model was developed to support these needs, as well as allowing for future development based on the principles of linked data and agile development. The model consists of a basic set of elements meant to cover

https://doi.org/10.23106/dcmi.952135737

core descriptive and administrative metadata needs, as determined by CAPS stakeholders in subject libraries and in the digital preservation unit, as well as a set of vocabularies where appropriate to enable CAPS metadata to link with other metadata sets at similar repositories.

The system currently supports the Dublin Core Metadata Element Set for descriptive metadata. Dublin Core was chosen for a number of reasons. It is widely adopted in both the linked open data and library metadata communities, making it an ideal candidate for sharing across both of them—a primary goal of the project. DCMES also benefits from its simplicity and a general familiarity with it among the stakeholder team, through its use as the metadata *lingua franca* for Penn State digital collections published in the CONTENTdm digital asset manager. Extensibility of this schema to cover future applications of the system, i.e. for electronic theses and dissertations or university electronic records, was essential to this process.

CAPS metadata is stored via two different methods. In the first, they are indexed as Resource Description Framework (RDF) N-triples stored in plain-text alongside their associated digital objects in the file system. They are also indexed in an *ad hoc* triple-store in the form of a MySQL database interfacing with a Django application for quick searching. The CAPS implementation of this idea mirrors a traditional triple-store, in which the subject is an ARK identifier minted by the CAPS system at the point of ingest of a digital object; the predicate is a URI conforming to the element in question, and the value is whatever is entered by the user to populate the element. CAPS supports versioning of both objects and their associated metadata, so that provenance of object or metadata edits is assured from point of ingest throughout the lifecycle of the object. In the future, this MySQL database will be migrated to a graph database with native support for the rdflib library in Python.

A data dictionary for the CAPS prototype was developed, in order to allow current and prospective stakeholders to understand the metadata decisions the team made. The dictionary consists of elements currently supported in the system, and the linked open data vocabularies from which the elements are (or will be) populated. Beyond the basic Dublin Core element set, there were two primary challenges in articulating a data dictionary. In addition to an overall shortage of published linked open data vocabularies for the preservation and technical metadata elements requested by the project's stakeholders, some functionalities written into the data dictionary (i.e. implementation of the dcterms:isPartOf element to indicate relationships between objects in the system) are not fully developed at this time.

TABLE 1: Data dictionary for OpenCASA Phase I.

Field name	Vocabulary	RDF property
Title	Literal string	dc:title
Creator	VIAF, LCNAF, DBPedia	dc:creator dcterms:creator
Description	Literal string	dc:description
Subject	LCSH, DBPedia, etc.	dc:subject dcterms:subject
Coverage	GeoNames (if geographic)	dc:coverage dcterms:coverage
Date	n/a	dc:date
Publisher	Institutional identifiers	dc:publisher dcterms:publisher
Type	DCMI Type, MARC genre terms	dc:type dcterms:type
Format	MIME types	dc:format
Language	ISO 639-1 codes, lingvoj, lexvo	dc:language dcterms:language
Rights	None specified	dc:rights
Collection	None specified	dcterms:isPartOf
Capture	None specified	mix:scannerCapture

https://doi.org/10.23106/dcmi.952135737

Capture Details	None specified	mix:scannerCapture
Compression	None specified	mix:Compression
Color	None specified	mix:imageColorEncoding
Color Management	None specified	mix:imageColorEncoding
Color/Greyscale Bar	None specified	mix:imageColorEncoding
Resolution	None specified	mix:resolutionValues
Modification	Preservation event vocabulary	

Many of these issues are related to the extremely aggressive development timeline for the CAPS proof-of-concept service; work on the system began in January 2011 with an expected completion date of March 31. To that end, many of the intended services meant to complement the metadata, such as Turtle serialization and metadata export via JSON, XML, and other formats, have been postponed to future phases of the project. However, as a basic demonstration of ingest and annotation services, the first phase of CAPS has provided a valuable foundation upon which to build future services, and into which interoperabilities with linked open data services may be built.

4. Future Plans

CAPS represents the first phase of a larger initiative toward a comprehensive solution for observed digital curation needs at Penn State. In the next several months, it will transition into the Open Curatorial and Archival Services Architecture (OpenCASA), a planned institutional repository for storing the electronic records and resources generated by the University. These will include electronic theses and dissertations, data sets generated by faculty in the course of their research, and university records. As of July 2011, the OpenCASA team has not yet been formally charged for Phase II of the project; however, future development plans have already been sketched out based on potential use cases for the platform and discussions with stakeholders. Where metadata is concerned, these plans are a combination of initial, established development goals and new feature requests made by stakeholders and team members.

There are various metadata development goals for the OpenCASA platform in future phases. The first is the ability to automatically extract technical metadata from files and from the system itself, and include it alongside the extant descriptive metadata for each object. At the moment, technical metadata is generated as a matter of course by the system for basic POSIX information, such as the creation and last-modification date for objects and the audit trail for file validation. A separate, but related, focus for metadata development is a local data dictionary for preservation events, with the ability to be expressed using existing linked data vocabularies for preservation metadata. Such a dictionary would need to take into account specific preservation actions taken in the course of a digital object's lifecycle, the dates and agents responsible for those actions, and any pertinent information about an object, such as file format or software requirements for using it, making it a candidate for future preservation actions. It is expected that this metadata will be made available in some form, either as locally-defined elements or conforming to existing technical metadata specifications such as PREMIS or MIX. One shortcoming in this area is the lack of a linked data vocabulary for expressing preservation information about an object; a possible breakthrough in this area is the emergence of an OWL ontology for the PREMIS standard (Coppens et al., 2011).

The proof-of-concept nature of this phase of development prevented the team from implementing support for linked data vocabularies within the system. In future phases, integration with existing vocabularies such as DBPedia, id.loc.gov, and GeoNames, for subject analysis, geo-location of digital objects, etc., will be supported in OpenCASA. It is hoped that by integrating with APIs for these vocabularies, as well as periodically synchronizing their RDF expressions with the OpenCASA environment, that services for curators such as auto-completion of subject headings within the system's Web interface will be supported, greatly increasing the overall

https://doi.org/10.23106/dcmi.952135737

usability and ease of annotation of the system. In addition, the use of Python's standard rdflib library for interfacing with RDF data is not currently supported by the OpenCASA system. Alignment with the broader RDF community, and use of tools such as rdflib for managing RDF data in the OpenCASA environment, is expected as the platform matures.

Finally, there is a demand on the user-facing side of the OpenCASA system for the ability to bind objects together in virtual "collections," i.e. groups of objects that can be acted upon as a collective, whether through simple viewing by an end-user or more active engagement such as performing a preservation event. These relationships, among others, will be expressed using the dcterms:isPartOf element. It is not clear at this time what the nature of these "collections" will be, and who will be allowed to create them; for example, if they will be curated exhibits managed by curators within the OpenCASA environment, or if any user will be allowed to manage their own personal collections of digital objects. The technical challenges associated with binding objects together in local collections within OpenCASA will present a number of interesting challenges for future metadata services.

Acknowledgements

I would like to thank Michael Giarlo, Dan Coughlin, and Patricia Hswe for their assistance in clarifying various aspects of the underlying development philosophies of the micro-services model, as well as technical issues associated with the underlying storage model for metadata in the CAPS system.

References

Abrams, Stephen, Patricia Cruse, and John Kunze. (2009). Preservation Is Not A Place. International Journal of Digital Curation, 4(1), 8-21.

Abrams, Stephen, John Kunze, and David Loy. (2010). An Emergent Micro-Services Approach to Digital Curation Infrastructure. International Journal of Digital Curation, 5(1), 172-186.

Coppens, Sam, Erik Mannens, Davy van Deursen, Patrick Hochstenbach, Bart Jansens, and Rik Van de Walle. (2011). Publishing provenance information on the Web using the Memento Datetime Content Negotiation. Proceedings of the WWW2011 Workshop on Linked Data on the Web (LDOW 2011), Hyderabad, India, March 29, 2011. Retrieved April 26, 2011, from http://events.linkeddata.org/ldow2011/#programme/.

https://doi.org/10.23106/dcmi.952135737