Interoperability of Metadata for Thematic Research Collections: A Model Based on the Walt Whitman Archive

Brown University Report

Brown University Library's Center for Digital Initiatives (CDI), directed by Patrick Yott, focused on two areas of operations: first, in participating in the development and review of protocols for encoding metadata across three standards (METS, TEI, and EAD); and second, in testing the resulting metadata by ingesting it into the CDI's digital repository.

This work was conducted as part of a long-term research effort at Brown focusing on effective methods of bringing encoded materials into library collections. With the exception of this project, the emphasis so far has been on metadata and bulk encoding of printed materials. Researchers at the CDI are now focusing attention more carefully on how to accommodate full-text digital objects, and how to integrate them usefully with other types of digital objects such as finding aids, METS records, and digital images. As part of this work, the CDI has developed a METS-based digital repository service that uses the structural map elements in a METS record as the core organizational information for digital objects to be ingested.

For Brown, the CDI and the WWP, this grant presented an opportunity to take a close look at metadata standards from the practical viewpoint of large-scale ingestion, rather than from the usual perspective of expressiveness with respect to an individual project of collection. The metadata records in this case were being created outside Brown, based on criteria arising from the Walt Whitman Archive and its internal metadata needs, while the ingestion process on Brown's side was designed to take information from a comparatively broad range of possible formats and target it to specific metadata fields used internally by the digital repository system. This project thus enabled us to test how well the Brown ingestion process anticipates the metadata being provided, and also how well METS serves as a metadata interchange format.

In Brown's repository system, the METS record serves both as a way of organizing and managing digital objects, and also as a way to determine and guide the user's experience with any digital object. The METS record associates each type of digital object with a set of disseminators (stylesheets, viewing programs, user input, and manipulation options) that permit different kinds of user interaction appropriate both to the digital object type and to the collection of which it is a part. Disseminators are typically written in XSLT and XQuery and constructed using open source tools such as SOLR and eXist. Sample disseminators might include:

- Metadata only
- Image only
- Text transcription only (either in the absence of an image, or to foreground a full-text transcription; usefully with very difficult-to-read texts)
- Text and image displayed side by side
- Image only with embedded text search
- User-controlled text display (permitting control over features such as the use of colors to display textual features such as revision and deletion, or the choice of which texts to juxtapose in a side-by-side view)
- Temporal views of a document's history, showing successive revisions

From a reader's point of view, these disseminators are task-specific, but they generalize well based on digital object type (page image, TEI transcriptions of various kinds, EAD finding aid, METS record, audio file, etc.) and together they constitute a layered approach to a scalable interface. The options presented to the reader reflect the presence or absence of different kinds of digital objects; if only a page image and metadata are available, those options are offered, but if a TEI transcription with revision information is available, then a disseminator that can show revision history (e.g. through a specialized stylesheet) will be available as well.

While this approach can accommodate variable presentation of specific project materials (by associating a specific project's materials with project-specific stylesheets and behaviors as disseminators), its structure emphasizes uniformity of access and the opportunity for cross-repository searching, comparison, and even project development. Materials for any given project (such as the Walt Whitman Archive), once ingested, are potentially available as a set of digital objects, metadata, and disseminators that can be aggregated in ways that cut across the project's original boundaries. Thus a new project on 19th Century journalism might access a subset of the Whitman Archive's materials and combine them with materials from the repository's Lincoln Broadsides collection and temperance pamphlets.

The results of our test ingestion yielded several insights. First, the ingestion of the Whitman Archive's METS was successful: the metadata required by the Brown digital repository was readily extracted from the records provided, and brought into the repository in a functional state, without hand modification. Second, the success of this process suggests that there are significant benefits to this model of metadata ingestion. The accompanying diagram illustrates the planned workflow for the digital deposit process for the CDI repository, much of which is devoted to metadata creation and the management of digital object files. The METS record used in the repository is created fairly late in the process, but in the case being tested here, that record can be derived directly as part of the ingestion process, thus providing something of a short cut. Finally, the specifications developed by CDI to guide the ingestion process will scale well to other METS-encoded metadata, making future ingestion of METS data from other projects straightforward.

This project also revealed some problems with both the METS metadata structure and the METS Profile schema. First, we discovered that there is no efficient way to associate an XSLT document with an XML file in a single METS instance, particularly when there are multiple XSLT files relating to the XML source. This is particularly problematic since the combination of the source tree with the XSLT represents the scholars' intent in creating the object that is being shared. This issue has been raised with the METS Editorial Board by member Patrick Yott, and may be addressed in a future revision. It is possible that the work of the Digital Library Federation is doing with asset actions will provide a solution. Second, the METS profiles can become too narrative and therefore not "machine actionable." Having a more actionable profile would make the transformation and validation tasks required for successful metadata interoperability easier to design and execute. This problem is one that the METS board is currently attempting to tackle.