July 2010 Partners Meeting Breakout Session

Preservation Tools and Services: JHOVE2 and PREMIS, Session #4
NDIIPP Annual Meeting
July 21, 2010
2:45 p.m. – 4:00 p.m.

Presenters: Stephen Abrams, California Digital Library
Rebecca Guenther, Library of Congress

Attendees: 22

Overview:
Stephen Abrams discussed features of JHOVE2. It covers four main aspects in the characterization of digital objects: format identification (the purported format from looking at internal and external signatures); feature extraction (deriving properties specific to that format); validation (what is it really? Does it follow the rules of that format?); and assessment (new to JHOVE2; what do those properties mean?).

He discussed the distinction between validation and assessment: validation is an objective determination based on community consensus on requirements; assessment is a subjective determination based on local policy rules. JHOVE2 provides a tangible syntax to make policy regimes actionable. The JHOVE2 feature set includes a multi-stage processing model, with signature-based format identification using the Droid tool. Everything is configurable; users can control the stages, change the order, or turn off certain features. They have extended the underlying data model to understand logical objects, whether spanning multiple files, or subsets of single files. It features recursive processing of objects, and granular modularization with generic plug-ins to create customized workflows. It includes clean APIs (not backward-compatible with the original JHOVE tool), common module design patterns, buffered I/O for performance issues, and standard Java mechanisms for internationalization. There is extensive documentation, including a users' guide, architectural overview, module specifications, and a programmers guide.

JHOVE2 can presumptively identify over 500 formats from the PRONOM registry for identification; the validation stage includes a smaller set of formats, including ICC color profiles, JP2, PDF, SGML, shapefile, and zip.

Unsupported formats (at this time) include AIFF, GIF, HTML, and JPEG, however HTML can be expressed as SGML or XML (which are supported). They are investigating funding for GIF and JPEG modules.

All code was written in Java 1.6 under a BSD open source license; code is managed on an external hosting site (bitbucket), it will soon be open for public access.

The final production release will be September 2010. Results of a user survey indicate that 2/3 of respondents assume they'll use JHOVE2 in the first six months of release, and 3/4 within the first year.

Three project partners have committed to provide self-funded maintenance (bug fixes, not development). Possible future development efforts include additional format modules, training and tutorials, and identifying a permanent organizational home.

Rebecca Guenther discussed the PREMIS in METS Toolbox (PiM: http://pim.fcla.edu/), and the Authorities and Vocabularies web service (http://id.loc.gov/).

PiM was developed by the Florida Center for Library Automation. It is all open source (source code will be made available on sourceforge.net), and consists of three components: validate, convert, and describe.

- "Validate" takes a PREMIS in METS document and uses schematron to validate against the <premis> and <mets> schema and the PREMIS in METS best practice guidelines.
- "Convert" converts data (either uploaded or referenced via a uri) from <premis> to PREMIS in METS (putting it into the <digiprov> section of the METS file) using XSLT.
- "Describe" uses DAITSS and droid/jhove to generate an xml file (a PREMIS file object) for files that are uploaded or referenced via a url. She demonstrated each of the features on the PiM website. Not all file formats are supported, but audio, video, text, image and html are. When creating <mets>, users can point to a directory of files and get a basic <structmap>.

The Authorites and Vocabularies web service (id.loc.gov) makes LC-owned and maintained authorities available as linked data on the web, and allows for search and download. LCSH went online about a year ago; other vocabularies were added in May. SKOS is an RDF application for knowledge organization systems. It has a defined element set with rich tagging available, and can have a dereferenceable uri for concepts. Rebecca discussed linked data concepts and the benefits of making controlled vocabularies available as linked data. The technical infrastructure behind the id.loc.gov service includes Django (Python); SKOS RDF is generated at the time of the request.

Action items:

- Mark up to identify facets
- Enhance vocabularies to show relationships between terms
- Add new vocabularies (Name authorities; Premis controlled vocabularies; MARC country, geographic area, and language codes; ISO 639-2 and 639-5)
- Enhance the PiM tool to validate <premis> vocabulary tools