SW Arch Meeting Minutes – October 27, 2011

Agenda

- Announcements and Updates
- R6.0 (new platform) Status
- R6.1 (SOLR/Lucene) Status
- Release Targets
- Continuation of AV objects with transcripts discussion

Announcements and Updates

Ron reviewed some of the highlights from the recent DC-area Fedora conference. The DuraSpace representative noted that large file handling will be addressed in Fedora release 4.0 – at least one year away. There are approximately 200 Fedora installations and about 1100 Dspace installations. People are beginning to use DuraCloud for backup and restore. It is noteworthy that many government organizations in the DC area are moving to Fedora, including the National Library of Agriculture, the National Library of Medicine, the National Technical Information Service, USGS, and the Smithsonian. Most of these are using the Islandora platform which provides a number of tools and a variety of metadata schemes. There is some promise that this conference can be transformed into a Fedora Mid-Atlantic conference.

Release R6.0 (new platform) Status

It looks like we are on target to cutover to the new platform by mid-November. Accounts have been set up on rep-devel and developers reported that they could support the production release (R5.2.1) on the new server. We re-confirmed that we will use the PHP deprecated code and turn off the error messages in R6.0. We will upgrade (i.e. not use deprecated code) in R6.1. We will use local access rather than RPC in R6.1 for OCR. There are other relatively minor issues to be investigated (e.g. installation of the authentication database for dlr/EDIT, outputds anomalies, etc). We also need to investigate why the ingest process runs slowly on rep-devel. The new server will continue to use "mss3" as the server name. Regarding the transition period, ingests will have to be suspended from November 14 – 17 and Dave will do the last re-synch on Monday, Nov. 14. We should be fully operational by Friday, November 18. It was not clear how to best provide for access to RUcore content during this period (e.g. by a read-only system or other mechanisms). Ron will compose a draft message regarding the transition plan and post to sw_arch for comments, then to CISC, and finally to our external customers.

Release R6.1 (SOLR/Lucene)

Sho reported that PECL has been installed and is ready for use. Shibboleth will also need to be set up. Chad reviewed the updates to the SOLR/Lucene specification. We re-confirmed that access to SOLR/Lucene will be protected in a separate CGI directory by using Apache IP restrictions. Regarding the code complete date, Chad and Jie are targeting November 15. Jeffery and Yang felt that they could finish coding a few days earlier. So, our final target for code complete will be November 15.

AV Objects with Transcripts

Ron reviewed the specification for ingesting audio objects with transcripts. This specification is similar to a case study for a specific project that is to be ingested into the Research portal. For this specific project, we decided the following:

- Since the mp3 audio file cannot be released publicly, we concluded that there was no value in including the mp3 as a presentation format. The only presentation format will be the scrubbed transcript in pdf format. A non-viewable xml datastream will be a available for full text searching.
- Five files will be encapsulated in the tar for the archival master as follows: 1) the .wav for the audio, 2) the unscrubbed transcript in MS-Word format, 3) the scrubbed transcript in MS-Word format, 4) the pdf for the unscrubbed transcript, and 5) the pdf for the scrubbed transcript.
- For format migration (e.g. as new versions of MS-Word are released), a new archival master (tar file) would be created with the new versions of the Word document. Fedora would retain the previous tar file. (Note, we did not discuss the implications of this approach when the versioned files are very large. Ultimately, we may need to move to a policy in which we retain the 1st version and the n-th version the intermediate ones are deleted to save on storage space.)
- We discussed the scenario in which the mp3 could be made publicly available. In this case, there would be two distinctly different presentation formats – the mp3 and the pdf. The consensus was both datastreams should be included in a single object. We would need to have multiple techMD sections (not yet fully implemented) and the two presentation formats would have to be described in the descriptive metadata.

It was noted that some re-structuring of the data objects in the research portal should be undertaken at some point after release of R6.1. In the interim, it would be useful to ingest one of the audio compound objects into lefty64 (or rep-devel) as a test.

Other Items

Relative to AV objects with transcripts, the China Boom project was also discussed. The project has videos, transcripts, and also biographical sketches. The general view was that the sketches should not be ingested into RUcore, but rather should be available on the China Boom website. It was not clear what other types of materials were part of this project. It was also noted that this project is not a candidate for the Research portal. However, given the object complexity, the compound object approach should be considered, possibly with different relationships. Rhonda raised the issue of someone finding a separate object on the Web (e.g. a biographical sketch) and not having the appropriate context. If access is via the persistent ID, the object relationships will be displayed and provide access to the entire compound object. However, direct pdf access will not provide this context – a subject for further discussion. Given the complexity of the China Boom project, the number of videos, and the expected completion date (end of the year), this project should be discussed in CISC.

Agenda Items for Next Meeting

- Progress on R6.0 and R6.1
- China Boom project
- Large File Specification

rcj – 11/07/2011