SW Arch Meeting Minutes – April 19, 2012

Agenda

- Announcements and Updates
- Status Update – R6.1.2
- Progress on Non-release projects
- Checksum specification – final review
- Complex objects- presentation
- File Format Tools
- Multiple techMDs

Announcements and Updates

Ron indicated that we will review the release summary in our next meeting with the purpose of reducing the size of the proposed R7.0. A proposed release structure for discussion is as follows: 1) R6.5 – RUcore R6.1.2 on Fedora 3.5 (May/June), 2) R7.0 – Large files and complex objects (August/September), 3) R7.x – Faculty deposit and video (late Fall), and 4) R8.0 – EADs in RUcore (Spring, 2013).

Release R6.1.2

Kalaivani will discuss one remaining issue with Chad. We should be able to deploy to production on Monday (April 23). For these dot releases, we do not need an announcement.

Non-Release Related Updates

The status of each project is summarized below:

- JPE. We are still having problems with directly linking to JPE videos. Dave, Chad, and Jeffery will continue to work this issue.
- XML-1 Datastreams and Jpeg Thumbnails. The scripts have been updated to handle the case of an embargoed resource. Jeffery will work with Dave and Anu to run the scripts on staging and to evaluate the results. We will plan to do this on Wednesday (April 25). We should be able to run the scripts on production shortly thereafter.

Checksums

Jeffery reviewed his specification on checksums. We agreed to move forward with sha256 and using the Fedora compare API. WMS will populate the datastream contentDigest attribute and will no longer insert the checksum in the techMD. Jeffery's compare capability will continue to support sha1 in old objects and sha256 for objects ingested with the updated WMS capability. The specification can be considered final with two updates: 1) it should be noted that replacing an archival datastream would require the manual deletion of the sha1 checksum and 2) as a continued investigation, we will explore the possibility of using a time-base cron rather than number of objects (e.g. 5 hours rather than x number of

objects) and the possibility of running the nightly compare on staging where there are lots of extra cpu cycles.

Complex Objects – Presentation

Chad reviewed the approach for a presentation structural map and related scenarios. Regarding filenames, we agreed that the default would be to preserve the original filename in all cases, i.e. we will not ask the user. A major issue related to delivering the original filenames when a user downloads a file. We agreed that this is desirable and is particularly important for data in which a user might want a single file in a complex directory and can only select the file based on documentation that references the file by the original filename. Given the complexity of implementation and the consideration that we don't have very many of these complex objects as yet in RUcore, Ron proposed a phasing approach in which the user could either a) request a zip of the entire directory or b) be presented with a directory structure where the user selects the files to be downloaded (as suggested by Ryan and Aletia). These files would be in the original master format, i.e. we wouldn't as yet deal with automatically generating derivatives from files within a complex structure. There were no issues regarding the faculty deposit scenario in which no new questions are asked and the archival and presentation structural maps are not required. For the next meeting, Chad will present the 3rd part of this specification dealing with the user interface.

File Format Tools

Isaiah reviewed two file format tools to be used in creating technical metadata and validating file formats. The ExifTool provides support for still images and born digital documents and the Mediainfo tool works best for audio and video. Everyone agreed that we should proceed with the integration of these tools into RUcore and WMS. However, there are several questions that should be addressed in an MDWG review: 1) will there be extensive manual editing of the resulting metadata, 2) how do we resolve differences in conventions, e.g. the use of GB or GiB rather than displaying the number of bytes for the size of the archival master, and 3) in some cases the complete set of metadata might not be appropriate for inclusion in techMD. We will need to decide which fields from these tools will be used. Isaiah suggested we might want to consider encapsulating the complete XML file in the archival master. At some point, we might also want to update all of the techMD of all previously ingested objects.

Multiple techMDs and File Policy

Yang reviewed how digital file information would be recorded in METS and FOXML. There were only a few issues with the basic approach (e.g. how do we deal with different extensions for the same file type – jpg and jpeg?). For the FOXML approach, we decided that we would not need the digital file – digital file relationship (part B under the FOXML section). Most of the discussion focused on

representing the relationship between the techMD and the archival master. We will use rels-int to represent the relationship, however there were at least two major issues. First, for a book with 400 tiff images, there would be four hundred archival masters files. As a result, there would be four hundred lines of xml in the resulting rels-int section. Should we include all of this information in the inline xml that is part of the Fedora object. The concern is the size of the resulting object and the impact on performance. The second question related to the suggestion that the rels-int could be a managed datastream and thus would not be included as inline xml. However, we suspect that Fedora would not deal with any updates or edits that impact the resulting rels-int. For example, if someone deletes one of the archival master files, then the related reference in the rels-int must also be deleted. It appears that we would have to provide some mechanism to keep the rels-int up to date. It appears that we may have this problem if the rels-int is inline as well. We will continue discussing these issues in the next meeting.

Agenda Items for Next Meeting

- Complex objects/structural map specification – user interface (continuation) – Chad
- Multiple techMDs (editing rels-int and using managed datastreams) – Yang
- Fedora 3.5 on staging - Dave
- Status of Non-release projects
- Quick check on approach for file policies (response from developers who will use the xml file)
- Background ingest – Kalaivani and MDWG
- Review of release targets
- Pending
    - Enhanced UI for the landing page (a possible framework) – Chad and Jeffery
    - WMS – creation of relationships (rels-ext) for data projects

rcj – 05/04/2012