SW Arch Meeting Minutes – February 9, 2012

Agenda

- Announcements and Updates
- R6.1 – discussion of what we can learn to improve the release process (postponed)
- Non-release related Projects
- RUcore Release R6.5 and Fedora 3.5
- Large Archival Masters – specification proposal

Announcements and Updates

Ron distributed charts that show RUcore download activity. We had 349,350 downloads for calendar year 2011. This is a 450% increase over 2010 and represents a download about once every 1.5 minutes. ETDs accounted for about 60% of the total downloads. Chad's subsequent tracking in 2012 indicates that the download rate continues to increase and we will probably achieve a download per minute in 2012. The early 2012 download snapshot also indicates significant spread over the whole collection, showing that 30% of the ETDs had been downloaded. Kalaivani noted that empty search now is almost 10 times faster than in R5.2.1. This is all good news and is a result of much better performance on the new platform and significantly improved Google Scholar searching. Regarding DOIs, we are still moving ahead however a cabinet discussion probably won't occur until March.

We discussed the need to put out an R6.1.l dot release to handle special indexing for the data portal and a database issue regarding Faculty Deposit. Everyone agreed that we should process and code change have already been made. Jeffery will update the SOLR schema to accommodate the indexing change. The plan is to do testing on the development and staging servers next week and push out the release to the public in the early part of the week beginning February 20.

Non-Release Related Updates

The status of each project is summarized below:

- Cranberry. The project has been moved to production. Ryan will do a few metadata updates. We need to deal with research provided file names in R7.0, perhaps by enabling editable datastream labels.
- Research data. After we release R6.1.1, we should be able to proceed with ingesting new projects.
- JPE. We are ready to move the objects to production. Dave is still trying to get Chuck to make the appropriate changes so we turn on shibboleth.
- Analytic. Chad is making some bug fixes and we also need shibboleth on development in order to run the Analytic software.

- XML-1 Datastreams. Jeffery's script is ready to go subject to modifications to insert the embargo xacml for the XML-1 datastream. We agreed to move ahead with the current script which should handle a large percentage of the objects that do not have XML-1 datatreams. Jeffery will need to rewrite the diagnostic routine to identify the few remaining objects that do not have the XML-1 datastream.
- Jpeg Thumbnails. Jeffery will need to identify objects that need thumbnails. Work has not yet started on this item.

Kalaivani also brought up the issue of replacing PDFs when there is an embargo policy in place. We acknowledged that a replacement process will allow a download to take place during a very small interval (probably less than a minute). Kalaivani will work with Jeffery and Rhonda to do the manual replacement. Jeffery will need to devise a more automatic procedure as part of our management software.

Fedora 3.5 and RUcore Release 6.5

There has been ongoing informal discussion as to how we can get Fedora 3.5 installed without disturbing our current release process. Fedora 3.5 is critical for our next release, not only to pick up many bug fixes, but to also move ahead with new Fedora features including security layer (FESL), file upload capability, messaging, and checksum generation. Dave and Sho presented an approach in which we could use the staging server during intervals when it was not being used for release testing. The critical breakthrough here is that Fedora 3.5 with the current release software can be quickly installed as necessary and the backed out when we need the staging server for testing. A typical scenario would look something like the following:

- Proceed with the release of R6.1.1.
- After R6.1.1 is released, install Fedora 3.5 with the current application software (i.e. R6.1.1)
- Proceed on the staging server to verify or fix any problems with the current software on the new Fedora release.
- If, at any time, the staging server is needed for release testing, Dave and Sho can make it available in short order (a matter of hours).
- After the sanity check of the current release on Fedora 3.5 has been completed, we will install Fedora 3.5 on the development server.
- The current release (e.g. R6.1.1) will then be pushed out to the public on Fedora 3.5.
- Upgrades to the development server will then be made for PHP (3.6 to 3.9) and for SOLR/Lucene.

- Development of R7.0 will then proceed with Fedora 3.5.

Sho noted that SUSE updates may be installed at various points in this process as needed. Everyone agreed that the above process should work and that we will proceed, noting that no changes should be made during the next week or so until we R6.1.1 is released.

Decisions for Large Archival Masters

Ron reviewed the decisions we have made about how to proceed with the development of the large archival master capability. These decisions are briefly summarized below:

Simple Objects. These objects may have multiple archival master datastreams, however there is no need to capture directory hierarchy and filenames. The archival master will have explicit file datastreams (no tars). Naming conventions for the datastream IDs are ARCH-PDF1, ARCH-WAV1.

Complex Objects. For complex objects, we will need to capture directory relationships, original filenames, and be able to map these filenames to datastream IDs. The structure map will be used to capture the relationships and filenames. For example, in the case where the user requests a download of the complete project, the structure map can be used to recreate the directory relationships and original filenames. It should be noted that many of the archival master datastreams will be the same as the presentation datastream. In the above situation, we felt that security would not be compromised by using the archival master datastreams to create the gzip package for downloading. A proposal under consideration is to use rels-int in the presentation datastream to point to the corresponding archival master datastream. So, for a primate tooth surface file of file type .sur, we would have a presentation datastream of ID=SUR1 which would point to the archival master datastream with ID=ARCH-SUR1.

Technical Metadata. Given explicit archival master datastreams, users will, in certain cases, want to provide unique technical metadata for one or more of these datastreams. As a result, we will need to provide links between the techMD and the associated datastream. We should investigate rels-int for this purpose.

Checksums. With explicit datastreams the archival master, we are likely to be computing many more checksums. We should investigate the Fedora capability to both compute checksums and do comparisons. The Fedora API for comparing checksums and also using the messaging service to indicate whether there was a pass or fail. As a result, we also need to explore Fedora messaging.

Specification Structure for Large Archival Masters

Ron then proposed an implementation specification structure as follows where each item represents a document to be written and reviewed in sw_arch.

File Configurability. We have noted previously that there will be many new file types that will be part of the archival master and will also be used for presentation formats. Therefore, all of our software must configure appropriately without having to make software changes. We decided that the locus for this change is the WMS and that Yang will provide a web service to other developers. Chad, Jeffery, and Jie will provide their requirements to Yang.

Technical Metadata. As indicated above, users may want to create unique metadata for some or all of the archival metadata streams. As a result, we should build on the capability of WMS to create multiple techMDs. These techMDs will need to be related to the respective datastream. The Fedora rels-int capability should be investigated for this linking purpose. This is Yang's area with metadata input from Kalaivani, Rhonda, and Isaiah.

Structure Map. We have proposed that the structure map can be used to represent the directory hierarchy and the mapping between Fedora datastream IDs and the original file names. Chad will develop the specification with the objective of adhering to METS standards if possible. Chad will also investigate presentation scenarios in which the original directory structure with original file names must be recreated for download to the user. He will investigate the possibility of presentation datastreams using rels-int to point to the archival master datastreams. This approach can eliminate the redundant data storage in the case where the archival master and presentation files are the same.

User Interface. Ryan, working the Data Working Group, will develop the user interface and related scenarios that demonstrate the various user interface functions required for complex datasets. These scenarios should include the download of a single object with multiple datastreasm (e.g. one primate tooth sample), the download of all data objects in a project (e.g. all the samples for a single primate species), and selectively extracting all files of a single type (e.g. a surface file) from all samples of a species. In addition to the scenarios, a mockup of the user interface should be included that illustrates how the user would make the appropriate selections.

Checksums. Fedora can compute checksums where the user can select the appropriate algorithm (md-5, sha1, etc). We should investigate the use of this capability. As mentioned earlier, the compare checksum capability uses the messaging service so this specification should include the use of Fedora messaging. Jeffery and Jie should develop this specification.

We did not discuss specific target dates for the above specifications, however we should move ahead expeditiously if higher priorities items have been completed (e.g. the non-release projects). Target dates for R6.5 and R7.0 will be discussed in the next meeting. There was some concern that we are again in jeopardy of having too many features in the next release.

Agenda Items for Next Meeting

- Status of R6.1.1 and non-release projects
- Target dates for R6.5 and R7.0
- Streaming server specification
- Jpeg 2000 specification

rcj – 02/20/2012