SW Arch Meeting Minutes – January 12, 2012

Agenda

- Announcements and Updates
- Status of R6.1
- Large Files
- Handling Masters for China Boom
- Meeting schedule for 2012

Announcements and Updates

Ron mentioned that the Roman Coins project has been re-activated and that an MLS student (Rick Hale) will be working with SPCOL on a number of related tasks – helping organize the 1300 coins, digitizing and starting the work on developing an RUcore collection. On another issue, we will likely be implementing DOIs in our next major release. We are hoping to get a Cabinet decision in February; annual costs for unlimited DOIs will be $2500.

We briefly addressed the issue of date ordering in presentation results. Basically we need to be able to handle many different date formats. Jeffery will investigate the SOLR module that might help with this issue and will also talk with Kalaivani, Linda, and others to understand what date formats in R6.1 are not being properly sorted. If the SOLR module proves out, we may want to consider a dot release (i.e. R6.1.1).

Release R6.1 Testing Status

Testing is complete on the development server. Dave has the software and should be able to move ahead with installation on testing. It appears that we are on target for a public release by January 23.

Large Archival Masters

Ron distributed discussion notes for how we might handle large archival masters. We made good progress which much discussion of the various architectural issues, however we did not reach conclusions on some of the key issues so more discussion is warranted. The discussion is summarized below with areas where we reached agreement.

Datastream ID Naming Convention. In a previous meeting, we had decided to use managed datastreams for all archival master files rather than encapsulating these files in a tar file. We proposed a datastream naming convention for these archival files as follows: ARCH-PDF1, ARCH-TIFF1, etc. The general convention is as follows: ARCH -|| {file type}{n} where n is a sequence number. It should be noted that the proposal to abandon the tar approach appears to work well for simple archival masters, however the issue becomes more complex when the archival master is a directory with embedded relationships (structures that we are seeing with the China Boom project and which will likely develop with complex datasets) – see discussion below.

Technical Metadata for Many Files of the Same Type. The classic example is the book content model that would typically have many tiff files for each page that comprise the archival master. We concluded that this type of content model can best be represented by one master techMD section, rather than having a techMD section for each tiff file. In this situation we would rely on Fedora to provide the checksum, file size, and mime type as attributes for the file datastream. For uniformity across all content models, we should consider using the Fedora checksum, file size, and mime type all archival master files. Jeffery will investigate the Fedora checksum capability.

File Type Configurability. Given that we will have many new file types as we move ahead with different types of datasets, it becomes important to easily configure these file types without software changes. We concluded that WMS would serve as the base for this configuration capability. Yang will provide a web service to be used by other software developers, based on requirements from Chad, Jeffery, and Jie.

Complex Archival Masters. We had an extended discussion on how to handle complex archival masters. We are beginning to see these masters with commercial software such as Final Cut Pro and will encounter more projects with research data and GIS projects. As an example, consider a multilevel directory in which directory relationships and names should be preserved. The entire dataset would be delivered to us for ingest into Fedora. Assume for this discussion, that this directory would form one object in RUcore. How will the archival master be represented? Although there were many different views presented as to how best to handle this type of archival master, we seemed to agree on two points: 1) the best way to preserve the structure and relationships is to encapsulate this directory and all the files in a tar file, thus we would be using a tar file in contrast to what we decided earlier for simple archival masters, and 2) in order to support various presentation and preservation services, the directory structure could be extracted and represented as a structure map. For example, a data user might request to download only a few files from the total structure or we might have to migrate one of the file types forward as a preservation action.

We did not reach a conclusion on whether we would need to represent each file in the directory structure as a separate archival master file. One reason we might do this is to be able to prepare separate techMDs for these files, one of the original reasons for moving away from the tar file. We also did not discuss the logic of how we might extract a few files from the overall directory structure. So, more discussion is needed. Ron will work with Ryan to try to model the “primate tooth” project, considering both presentation and preservation issues.

Fedora Messaging. We did not have time to discuss the use of Fedora messaging for alerting and other functions. For example, the Fedora checksum compare API return a failure and will also log a message in the message queue. Sho indicated that the apache mq library is already installed on devel and Jie will begin experimenting with this capability.

Release R6.5 and Fedora 3.5.x

It would benefit our development process considerably if we had a server where we could experiment and test with various new features. For example, Fedora checksums, use of the file URI for ingest, the security layer (FESL), and Fedora messaging are all new capabilities that have been referenced above. In general we need to have the next major Fedora release installed somewhere. Some of us have thought that lefty64 could be configured for this purpose, however there are also plans for using lefty64 as a drupal development sever. Dave and Sho will discuss this issue next week with Tibor taking into account the following: a) reusing an existing machine or buying a new server, b) costs of commercial SUSE, c) the possibility of drupal and Fedora coexisting on the same server, and d) how a VM system might also be used. Regarding a VM system, periodically we need both a development and a production server for presentation or compute purposes (e.g. PDF, video streaming, OCR). However, there is no need to have physical separate servers for each function, especially for development purposes. Allocating these servers to VM partitions appears to be a cost-effective direction for the future.

Agenda Items for Next Meeting

- Any Further Issues with R6.1
- Continue Large Archival Master Discussion
- Release R6.5 (Fedora 3.5.x)
- DOI update

rcj – 01/23/2012