SW Arch Meeting Minutes – November 10, 2011

Agenda

- Announcements and Updates
- R6.0 (new platform) Status
- R6.0 Test Plan
- R6.1 (SOLR/Lucene, etc) Status
- Discussion - Backend Archival System

Announcements and Updates

Ron briefly reviewed some of the more relevant results from the Digital Library Federation conference (note, many members had already heard the full report in the recent Cyber-Infrastructure meeting). The use of DOIs and the EZID software) was presented by John Kunze (from CDL). They also discussed the DataCite consortium which is using EZID (and DOIs) for citing of data. An annual membership fee is required to join DataCite, however this fee entitles the organization to an unlimited number of DOIs. In CISC, Grace proposed that we investigate both DOIs for use in RUcore and DataCite. Regarding software methodology, the Hydra organization presented their approach, noting that they do continuous integrations and have a fixed release interval (basically every two months). These approaches are worth some further investigation by sw_arch and the software methodology team. Finally, it doesn't look like we will get much help from Fedora in the near term for large file ingests. (However, in a post-meeting email interchange with Chris Wilper – Fedora developer – it looks like we can do local file ingest.) Regarding more immediate changes in our software methodology, we briefly discussed the issues with providing 24 hour turnaround for critical quick fixes – a topic for continued discussion in the next meeting. It was noted that we have now put a hold on all ingests into mss3 since we are at the 90% threshold which would cause thrashing with tape backup. Ingesting of objects will be re-started when we are on the new platform.

Release R6.0 (new platform) Status

Dave indicated that all software and databases were on the production system. However, datastream copying was proceeding slowly, suggesting that we might have to slip our planned final re-synch from Monday (11/14) to Tuesday. (In a later email update, Dave has restarted the copy with the compress flag removed, resulting in dramatically improved transfer speeds. As of this writing, it looks like we are back on schedule). We still need to have Sho (with Chuck Hedrick) install shibboleth on the development server.

Release R6.0 Test Plan

Kalaivani presented the test plan for R6.0. 6.0 test plan. Everyone will help in testing different parts of the system. Testing will start next week on rep-prod once we have everything in place. The testing will be done, initially at least, under rep-prod names, but then there will be a sanity test of public interfaces under mss3 when the name moves.

Release R6.1 (SOLR/Lucene, etc)

All developers reported that there were on target to meet the code complete date of Nov. 15. Jeffery will take the lead and work with Chad to develop the cron that is required for update objects whose "portal" identities have changed.

Backend Archival System

We had a good discussion of different alternatives for ingesting and archiving large files. It is acknowledged that we will need more discussion on what we mean by a “large file”. The principles that we agreed to and which would form part of our requirements are the following:

- All ingests would be done in the background. Users could determine the status of ingest progress by going to a designated website or by a yet to be determined alerting mechanism. This approach eliminates the issues we have had with users waiting for the ingest to complete. This background approach also will enable us to open up OCR to handle larger page counts. In addition, we should consider ingest “complete” when the user can actually go to the portal and find the resource (i.e. the cron re-indexer has been run).
- All archival objects will have explicit datastreams, i.e. no more tar files. This change will significantly improved our ability to manage archival datastreams, to preserve and migrate datastreams, and to provide more precise metadata for each file that comprises the archival master.
- For handling large files, we do not need a second backend Fedora system (given that we will do all ingests in the background). Also, the archival master will be a single object with explicit datastreasm (i.e. no tar files)
- Since all ingests are in the background, we will need to provide a means for the user to determine the status of the ingest. This feature might be provided by returning to WMS to check the status. However, we might also want to provide an alerting system that would not require the user to return to WMS.
- Regarding a dark archive, we may want to consider a second backend Fedora system. Other approaches also need to be investigated.

Agenda Items for Next Meeting

- Progress R6.1
- Fast tracking critical quick fixes
- Discussion of R6.2 content
- Large File Specification – continued discussion

rcj – 12/09/2011