Software Architecture Working Group

May 5, 2015

Present: Beard (notes), Mills, Jantz, Triggs, Yang, Marker, Hoover

Announcements

FEDORA Conference and discussion of other institution's moving to FEDORA 4, possible options and move to other platforms (Hyrda, Islandora, DRUPAL module).

Yang attended the DC Fedora meeting. "All about Hydra". People talked about Fedora 4 migration, especially Mike Durbin. He discussed a bunch of issues with his attempt at migration from 3 to 4. Many people are planning to migrate, but not there yet. They are interested in Hydra for Fedora 4. University of Maryland is using Hydra and Blacklight. They needed Ruby on Rails consulting. UMD are trying to get a working Fedora 4 by this summer. They lost their Ruby programmer, and are short of money right now. Ron thinks we might meet with Mike Durbin at some point. PCDM (Portland Common Data Model) was lightly mentioned. Islandora was not mentioned, and it appears it is being completely rewritten for Fedora 4. Yang thinks Islandora (a Drupal module in PHP) might be a better fit for us. All of these options might be too expensive, however. If we get the grant, we would have an opportunity to learn about 4.0.

We need to think about what 8.2 might be.

TestCAS issue: TestCAS does not work off-campus, nor on university WiFi. Testing on rep-test, and rep-staging versions of WMS must be done on-campus at a workstation (Ethernet, not campus wifi). Local accounts however, will continue to work regardless of location. Usability on Rutgers VPN has not been tested at this time.

RULMAIL transition and distribution lists

Mailing lists from existing Zimbra system will continue working for up to a year. We should analyze what lists on Zimbra are currently being used by our architecture and have them migrated to the Office 365 platform.

R8.0: Checksum Issue and related items

An issue identified with FEDORA 3.8 (and 3.8.1) results in a failure to generate SHA-256 hashes, preventing effective digital signature generation and file integrity checking. It is unknown for certain if this is an intended feature of 3.8.x or a bug, but the latter is believed to be the case.

Updated code is now in place on staging to get around the issue and generate signatures. The code has been tested and works, but a new issue with LibTIFF was uncovered, preventing the upload and processing of TIFF files. Dave will be compiling a new versions for LibTIFF to rectify this new bug, in order to complete testing and deploy in production.

With regards to the 71+ objects that have been ingested without signatures: the most viable option is to use WMS to re-ingest these objects once the code fix is in place, which should induce the generation of SHA-256 hashes for these datastreams, and thus restore effective integrity checking.

New Dataset project, approximately 1.3TB of research data

The Research Data Exploratory Team has been charged with ingesting a number of research data sets to explore the feasibility, effort and expense of expanding RUcore to handle Research data and server that community within the university. 5 datasets of significantly smaller size have been ingested. We have, as a last test, a large 1.3TB dataset consisting of 8,453 separate files, compressed, of various types including NetCDF, TSV and CSV. Grace Agnew and Melissa Just are being consulted regarding next steps; we have a copy of the data in staging. Concerns include choke points in software, time required to ingest and disk space consumption (we currently store two identical copies of research data: one for archiving, one for presentation). WMS also keeps a copy in ./workarea, which will require quick cleanups there and in ./temp_upload. This data cannot sit on production temp space for a lengthy period of time.

More information is needed from the researcher. Can this data be partitioned, and if so, how? Should users be allowed to download individual portions of the dataset? We need information on expectations for how other researchers will utilize this data. This will dictate how we organize and ingest this dataset, and will also dictate our workflow in handling the data. We also need documentation from the researcher about the nature and use of this dataset to better understand it. These issues will be brought back to RDET for resolution.

RUcore 8.1 Progress Report and impressions

WMS testing continues. This is a compressed testing window to accommodate schedules; all persons with testing items assigned should expeditiously test and report any issues we find.

One issue encountered was a MySQL “too many connections” error. This may be related to the use of persistent connections in MySQL; Yang will investigate and consider reducing or eliminating its use.

Some good news: This was an opportunity for concurrent ABBYY testing, multiple users and multiple instances of OCR generation. Testing suggests that ABBYY is handling multiple users and heavy workloads better than AdLib software has in the past.
More on ABBYY: MS Office document conversion works, but not through OpenAPI web services as expected; functionality appears to be console-only. Our technical contact with the vendor has been notified, and we are waiting a response.

For 8.1: AdLib will continue to be used for MS office document conversion and URL HTML conversion. Once overarching issues are resolved with ABBYY, we anticipate switching over in a future release.

Next testing session is Tuesday morning, SOAR testing, with a further testing session (search, discovery layer) on Thursday.

8.1 Thumbnail and High-res image creation/export

Server implementation: The IIIF API base URI and IIPImage server was discussed for facilitating PTIFF delivery to end users via the web client. Some issues involve the usage of specific characters in file naming conventions by FEDORA, compared to what IIPImage will accept (e.g. FEDORA's use of the "+" character is something IIPImage rejects). Chad developed a workaround involving the use of symlinks that appears to work well.

PTIFF: In looking at legacy still images, one stumbling block is that a large number of still image objects have tar files as their ARCH datastreams, complicating the ability to create PTIFFs out of the original TIFF files. Rep-test has unbundled objects, so testing can commence there.

THUMBJPEG-1 update: A job was run on RUcore-dev and RUcore-test to generate new THUMBJPEGS if the current version does not meet current thumbnail requirements.

Additionally: Chad was able to use logic from the analytic tool to generate THUMBJPEGS for moving image objects via FFMPEG. These have not yet been committed to FEDORA, but is useful and upon checking to ensure all look well, could be committed to these objects.

Approximate run time to normalize THUMBJPEGS on production would be about 6 hours.

Jobs and Reports

The initial batch of RARCH datastreams from the first 9 days of scripts has been cleared from the ./RARCH directory, and corresponding RARCH datastreams purged from the RUcore records, clearing 1.9TB of disk space made redundant by the conversion of these objects to managed archival datastreams.

After analyzing the amount of space that will be used by the remaining scripts, they will subsequently be run to finish out the conversion of RARCH datastreams. Following this, the unbundling of tar files in archival datastreams can proceed.