Software Architecture Working Group
Minutes
May 10, 2011

Present: Chad Mills (chair), Kalaivani Ananthan, Isaiah Beard, Jie Geng, Rhonda Marker (scribe), Sho Nakagana, Jeffery Triggs, Ryan Womack, Yang Yu

1. Updates
   RFP for server status (Mills)
   A decision on purchasing a new server will probably be made this week. The equipment must be delivered, installed, and running (to determine acceptance) by the second week of June, in order to complete the invoicing process by the end of the fiscal year.

2. R5.2 testing update (Ananthan)
   WMS testing is likely to take three weeks. There were additional active bugs last week and this week. Current status is:

WMS	24 active bugs/reports
	7 test status bugs/reports
dlr/EDIT	18 active bugs/reports
dlr/EDIT	13 test status bugs/reports
search	25 test status bugs/reports

   Release 5.3 and Release 5.X have been created for future releases.

   Ananthan explained that the person who files a report is responsible for testing it. If that is not possible, then the originator is responsible for assigning it to someone else to test.

   Mills is working on an option for a new RUcore front page.

   In discussion, it was mentioned that there is a misspelling of the label “qualifier” in the date element in the XML. Triggs will write a script to correct extant records after R5.2 is installed on the production server.

   Server downtime debriefing, redux: Nakagama reported that he and Hoover can fairly quickly replicate the datastream tree on the production server for object metadata. In discussion, there was consensus that we want to have a read-only server available and backed up to a 4-6 hour window, not just 24 hours or several days.

   PDFs being generated by WMS/PDF server are creating errors in various logs. Nakagama and Geng have seen the errors and will create bug reports regarding them.

   Embargo messages differ between RUcore and Showfed displays. They should be consistent. Marker will choose the preferred message. [The preferred message is: Access not allowed for PDF]

   XACML policy is now being applied from the rightsEvent in the WMS. In the WMS when a rightsEvent / type=”Embargo” is entered, a XACML policy is created at the time of ingest. Catalogers need to provide some specific information in this event for the XACML policy to work properly. A function to manage embargo policies in dlr/EDIT is being added in this release.

Schedule of testing for R5.2 (Ananthan/All)
Testing server: Yang will finish outstanding bugs this week. WMS testing continues May 16-20.
Package code by end of business on May 20.
Staging server: Install May 23-24. Testing of staging server tentatively set for May 25-June 1.
Production server: Install on Tuesday, June 7.
Beard will send a downtime notice ("on or about") on May 23. The working group was reminded that the Hill Center will shut down "rci" server services on May 24.

3. Large file support (Marker/All)
   We are encountering large files in many formats.
   Data files: The "cranberry genome" csf file, zipped, is 14GB, 14GB, plus two other files that total 40GB. Upzipped, it is about five times that size.
   Video files: This was our original "large file" category.
   Audio files: We have had audio files so large that they had been carved into several parts. There are 6GB files from Jazz Oral History.
   Text files: We have 10GB files in the test system for the RU yearbooks.
   Image files: We have large image files 5GB and up, especially large scale maps.

   Videos are all processed and uploaded to the repository "offline", that is, not through the WMS. The archival master is an external redirect. We chose this method because of the large size of the files. The workflow in the WMS is dependent on manual intervention. The upload is done manually, and the XML is generated manually: file size, checksum, and original file name. The WMS has been modified to accommodate this: if "video" is chosen as the Content Model, the directory structure is changed with a renamed file name, etc. The workgroup determined that it is possible to configure other file types to go through this process in the WMS if it becomes necessary.

   In answer to a question, the workgroup determined that there are cases where we will have large presentation files, not just large archival files. An example is data sets, for which we are not processing separate presentation files. The original dataset is also the presentation format. Another example is images, especially maps. We have had instances for which the presentation files are so large that it causes DjVu to crash.

   What is "large"?
   Fedora: 2GB is too large
   DjVu/OCR process: 750MB is too large to generate OCR from a DjVu file
   PDF: none known
   DjVu: unknown
   WMS generation of presentation files: 500MB (for the generated file) is too large
   The Working Group determined that there should be a cap of 2GB for Fedora ingest and that size could also apply to other functions.

   Thesis: The process we use to handle files, regardless of their size, should not require manual intervention to evaluate the file size.

   Proposition: Create a utility that evaluates the file size, processing requirements, and rules to examine a file. Make provision to process the file "later" if the file characteristics are outside the bounds of current system capabilities.

There are several possible approaches to large file handling:

Fully automated at the time of request [e.g. WMS file upload]
Fully automated but delayed
Manual process handled offline

Referencing the process of putting video archival files into an external location, it was suggested that we need a decision about what size file would be put into an external location.

For data, we are discussing the possibility of keeping some data on an external server that Fedora points to, so that the file is not only external to Fedora, but also not even on the same server as Fedora.

Triggs has a glossary of the Fedora datastream types. He will forward the glossary to the group. [The Fedora control groups are:

X=XML code = part of the object file itself such as the DC or MODS datastream
M=managed internally = external file associated with the object and managed internally by Fedora, such as our presentation datastreams
E=managed externally = file accessible through the http protocol and retrieved and managed transparently by Fedora upon access
R=redirect = file accessible through the http protocol to which Fedora redirects users upon request]

Fedora cannot directly serve files over 2GB, so anything over 2GB is considered a large file. Any file over 2GB is a "redirect" datastream. The workgroup agreed it might be best to make all ARCH datastreams RARCH which is a "redirected" datastream. Some presentation files over 2GB can be "redirected" datastreams. This is determined by size, not type of datastream. Investigation into browser download resumes and sftp access needs to be done. The workgroup agreed that we need to determine practical presentation file size limits for all datastream types.

Beard/Triggs/Marker will come up with a large file size limit for presentation files. These are the two "bookends" of spec's that we will begin to develop for the next release.