

Washington D.C. Fedora User Group Meeting - 10 Mar 2014

DC_Fedora_User_Group_2014-03-10.pdf

What

D.C. Fedora User Group Meeting

Where

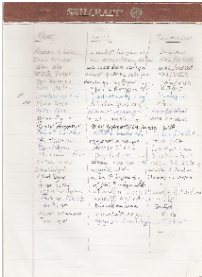
National Museum of the American Indian
Washington, D.C.

Room 4018, which is a space just off the main hall, so no need for special access

When

10 Mar 2014
10:15 am until 5:00 pm

Attendees



Agenda/Presentations

Time	Topic	Presenter
10:15 – 10:30	Welcome and introductions all around	Thorny Staples, Smithsonian
10:30 – 10:50	The Fedora community and DuraSpace	David Wilcox, DuraSpace
10:50 – 11:20	Fedora 4 development	Andrew Woods, DuraSpace
11:20 – 11:50	?mystery?	Bria Parker and Kevin Rice, NASA Goddard Space Flight Center Library
11:50 – 1:00	Lunch on your own	
1:00 – 1:30	Fedora Membership and the DC user community	Jonathan Markow, DuraSpace
1:30 – 1:50	RUcore and their research data portal	Ron Jantz, Rutgers University

1:50 – 2:10	Federal Science Repository Service	Wayne Strickland / Gail Hodge, Information International Associates
2:10 – 2:30	break	
2:30 – 2:50	Sldora, a research support environment	Thorny Staples, Smithsonian
2:50 – 3:30	Short Presentations	<ul style="list-style-type: none"> • Don Gourley, National Agricultural Library • Ben Wallberg, University of Maryland • John Doyle, National Library of Medicine • Anyone else?
3:30 – 4:00	Discussion and wrap-up	

Summary

10:15 – Welcome and introductions all around

- Thorny Staples - Smithsonian
- Andrew Woods - Fedora Tech Lead
- David Wilcox - Fedora Product Manager
- Jonathan Markow - DuraSpace
- Tom Cramer, Stanford
- Ron Jantz, Rutgers
- Don Gorley, National Agricultural Library USDA
 - Ursula Pieper
 - Wei Wu
 - Chuck Schoppet
- Rob Cartolano, Columbia
- Bria Parker - NASA Goddard Library, Space Flight Library
 - Kevin Rice
 - Mitzy Cole
- Stefano - Art Institute of Chicago - new Fedora adopter
- Patrick - Northeastern University
- Fran Stern - Smithsonian
- Ti Amy - NLM, long-time user of Fedora
- Jenny - UMD
 - Ben Wallburg
- Adam Soroka - UVa
 - Mike Durbin - UVa - repository manager
- Josh Wesgard - UMD

10:30 – 10:50 David Wilcox – update on the Fedora community and DuraSpace

- slide show
- revitalized technology
- revitalized pool of contributors
- increase community involvement in project
- use-case driven
- review of steering and advisory group
- Fedora Community
 - 326 Registered Fedora Implementations
 - 21 new instances in 2013
 - 858 members of fedora community mailing list
 - 40 Fedora Sponsors
 - 19 Active Developers
 - 17 Members of Fedora Advisory Group
 - 10 Members of Fedora Steering Group

10:50 – 11:20 Andrew Woods – update on Fedora 4 development

- Fedora 4
- call to engage in generating use cases, testing
- Fedora 4 Features
 - Content Modeling
 - nested or hierarchical structure

- validation - define properties on objects
- Authorization
 - application has pluggable authorization mechanism
- Durable storage
- Versioning
- Scale (large files and many files)
- Linked data / RDF (and external triplestore)
 - although not part of core base
 - important that triplestore be readily available for Fedora
 - pattern for triple store and SOLR index
 - every event that takes place on repository, JMS events come out of repository
 - completely functional message consumer
- Internal and external search
- Transactions
 - any action, largest bottleneck is persistence of action
 - when save takes place, takes significant part of action's time
 - pull together a series of actions into a single transaction
 - one larger save at the end
- Performance
 - recent sprint
 - 30% faster using transactions for update operations
- Clustering
 - consistent, scriptable way to put together Fedora servers

F4 Timeline

- Spring 2014 - 3.7.2 release
 - around code4lib conference
- Spring 2014 - 4.0-beta release
 - feature complete beta in spring Q1/Q2 boundary
- Engagement with the community
- Beta acceptance testing
 - download the alpha and beta, to provide feedback
- Beta pilot projects
- 4.0 Fall 2014
- Mailing Lists
 - fedora-community
 - fedora-tech
 - dc-fedora-users

Questions:

1. Migration path from Fedora 3 to Fedora 4
 - greenfield first
 - options
 - projection of Fedora 3 content in Fedora 4
 - can use immediately
 - copy over time
2. Ron Jantz
 - is there a 3.8?
 - No, we do not expect a 3.8
3. Jonathan
 - acceptance testing using F3 connector?
 - not hardened code yet
 - F4 has one click install, amazingly simple
 - what can we do to lower the bar to get you to test it
4. Transactions
 - we rely on transaction msg that they've actually occurred
 - at end of commit, want JMS msg, that transaction is complete
 - Adam - answers the question

11:20 – 11:50 Bria Parker and Kevin Rice, NASA Goddard Space Flight Center Library – update

- Bria Parker - metadata librarian - since 2010
- Kevin Rice - web programmer since 2012
- Mitzy Cole
 - contractors at Goddard Library
 - they've been downsized
 - originally 2 programming, 1 metadata librarian
- Godard Library Repository
 - <http://gsfcir.gsfc.nasa.gov>
 - Drupal 6 on top of Fedora 3.3
 - harvest, scripting before it gets to Fedora

- two collections
 - JSP scripts load collections from outside Fedora
 - two collections entered right into Fedora using Drupal interface
- what's in it?
 - colloquia, authors and publications
 - haven't ingested PDF
 - use link resolver to get to PDF content
 - separate authority author objects
 - balloon technology
 - scientific documents
 - a list of publications,
 - RDF - all other Goddard authors
 - all happening in Fedora
 - script to check author authority match
 - plans to move to Drupal 7
 - original people that built it are gone
 - some documentation
 - some logic is lost
- Kevin Rice
 - Drupal 6 - piecemeal version of Islandora
 - works in a similar process to Islandora
 - do not intake data through Drupal
 - only display it through Drupal
 - migrating from drupal 6 to 7
 - mysql transactions have changed
 - take a step back
 - fixing Drupal issues
 - will help get to new versions of Fedora
 - change made to core Fedora system
 - by local developer
 - no feature needs for future versions of Fedora
 - harvest publication information from variety of commercial databases
 - since we started, VIVO came out! 8-)

11:50 – 1:00 Lunch on your own

1:20PM - DuraSpace - Jonathan Markow

- overview of chart
 - Sponsorship vs. Membership model
 - "Sponsorships" not part of European approach/budgets
 - many are budgeted for "memberships"
 - libraries have "membership" budget lines
 - membership - more participatory governance
1. Q:Rob Jantz
- challenge with synchronizing our development with Fedora development

1:50 – 2:20 Ron Jantz, Rutgers University – update on RUCore and their research data portal

- DOI and micro-citation
 - RUCore Data Repository
 - Research Data Working Group
 - Established group including library liaisons
 - group provides supports for grants and NSF Data mgmt plan
 - Developed extensive metadata profile for research data
 - Research Data Portal - recent additions
 - ingest of multi-level directories as submitted by the researcher
 - link data to articles and vice-versa
 - Rutgers allows a researcher to give us a full multi-level directory all metadata is typically in the file name
 - using CDL's EZID service
 - 35,000 objects in RUCore
 - Questions
 - Do we want to assign every one of those items a DOI?
 - we think probably not, but then how do we filter out?
 - Not everything in RUCore is research data
 - Example: 3,000 Roman coins
 - we take 7 images per coin
 - we would probably not create a DOI for every image
 - can we do even finer grained linking, subset of a dataset, paragraph in a book, video clip, etc.?
1. Questions:
1. Robin

- ORCID - give them a choice
- Ron
 - fair amount of faculty pushback
 - Universities trying to do University-wide
 - don't want to fill out "yet another profile"
 - some faculty pushback
- 2. Tom
 - upload a whole directory of files, is it one fedora object?
 - Yes
- 3. Any migration to other formats?
 - Ron - we do. We take Word documents, migrate to XML, keep original

2:10 – 2:30 Wayne Strickland, Information International Associates – Update on the Federal Science Repository Service.

- Federal Science Repository Service
 - Wayne Strickland
 - Gail Hodge - private contract - public/private partnership
 - Don Hagen - Associate Director (Wayne's boss)
- National Technical Reports Library
 - NOAA - deep water horizon project
- FSRS
 - <http://www.ntis.gov/fsrs/>
 - by statute, used to charge for service
 - this is changing
 - use our approach with FSRS
- Technical reports used to be much more important
 - Data needs to accessible
- Object model
 - geospatial data example
- NOAA Repository - Deepwater Horizon Repository
 - 8,000 metadata records
- Two Islandora implementations for programs in D of Commerce, and DoD
- Don Hagen - talking with OSTP
 - open source, open access
 - part of the culture shift

3:00PM - 3:30 Thorny Staples, Smithsonian – update on Sldora, a research support environment

- Putting data that has never been in an organized place is hopeless
 - Rather, create a workspace to better manage research data from the start
 - So that when the research process is done
- Research Project - no formal hierarchy, it's a graph
- Two RDF relationships
- We will collect more and more standards
 - concepts are metadata objects
- Discovery and Collecting Environment - starts collecting data right at that moment. Starts gathering links to things, with links to items already in their project.
- DataSet concept - concept of sets to bound the data
- Analysis Environment:
 - Galaxy - workflow management system, by genomics
 - reflect a Sldora set as a Galaxy set, and all the tools in Galaxy can be used
 - Taverna
 - workflow environment, comes with all the R tools
 - can convert a Sldora set to a Taverna set
- Researchers have their own stuff, own tools on their own desktop/laptop
- Have files look like local filesystem (like dropbox)
- Questions
 1. Tom - access
 - Thorny - wants to use disseminators
 - for tabular data files - the idea is when you upload tabular data, assign to a code book
 - don't upload unless you assign the variables
 - can build disseminators on the dataset concept
 2. Do you expect to assign workflows as objects?
 - Workflow in Taverna is already an XML file
 - Adam Soroka - get connected to Kepler
 3. rtc - enforce requirement to upload codebook?
 - Thorny - yes
 - we will ask for absolute minimum metadata
 - try to fill out the code book and ask them to correct it
 - huge payoff when you upload 1,000 data sets with same codebook
 - can generate spss output for use by R later in workflow
 4. Working on relationship with Oakridge
 - Smithsonian, public-facing front end

- grid ftp in a lab
- lab's work becomes a research project
- each

3:40 - Don Gorley, National Agricultural Library

- OSTP mandated that agencies that meet a threshold need to make results of research publicly accessible
- Saw our area as the place to make materials publicly available
- 2 Fedora instances, and ILS
- be prepared to scale up discovery and access
- Single Fedora instance, with 4 to 5 million objects
 - MODS data streams
- A lot of overlap
- A lot like a PubMed Central system
- Use Islandora as the management platform
- For discovery, it's really working against SOLR index, and not Fedora at all
- staff were reluctant to mix
- separate out discovery interface completely from Fedora
 - Fedora is back-end
- build indexes and content servers
 - java application servers
 - nginx front-end
 - SOLR index
 - filestore of content that we are delivering
 - simple pear tree directory structure
- JBOSS domain controller
- SOLR master server - keep SOLR slaves up-to-date
- Content server - copying
- changes in Fedora repository will change
 - Fedora master
 - content server
 - use Rsync to distribute to front-end

3:45PM Ben Wallberg, UMD

- moved toward management
- building up the development staff
- Fedora - many home-grown applications
 - search service - moving to SOLR
- use plain file system storage
 - research hadoop as a back-end for Fedora
 - hadoop has big growing community
 - if we could use hadoop to store large files
 - share results on code4lib
- four home developed interfaces to Fedora
 - 1 staff facing
 - 3 public facing
- open source - hippo
 - libraries web site
 - use it for other interfaces
- focus on improving loading and backlog of collections
- we are on Fedora 2.2 instance

4:00PM - National Library of Medicine

- John Doyle
- Fedora, Blacklight
- <http://collections.nlm.nih.gov>
- 2 million items
- working with serial content
- Running Fedora 3.6
- hoping to go to Fedora 4
- indexcat - 3.7 million items
 - special Fedora instance to handle ingest of older XML data
 - mainly journal articles cited over decades
 - not full-text content
 - purely for preservation

Thorny - Closing Discussion

- Technical Training for Fedora 4
- Fedora 4 Migration meeting

Other Attendees

- Tom Cramer, Stanford University
- Robin Ruggaber, University of Virginia
- Jennie Levine Knies, University of Maryland Libraries