2017-08-17 - Fedora Tech Meeting

Time/Place

This meeting is a hybrid teleconference and IRC chat. Anyone is welcome to join...here's the info:

- Time: 11:00am Eastern Daylight Time US (UTC-4)
- Dial-in Number: (712) 775-7035
  - Participant Code: 479307#
  - International numbers: Conference Call Information
  - Web Access: https://www.freeconferencecallhd.com/wp-content/themes/responsive/flashphone/flash-phone.php
- IRC:
  - Join the #fcrepo chat room via Freenode Web IRC (enter a unique nick)
  - Or point your IRC client to #fcrepo on irc.freenode.net

Attendees

- Danny Bernstein
- Carrick Rogers
- Ben Pennell
- Esmé Cowles
- Andrew Woods
- Kevin S. Clarke ⭐
- Nick Ruest
- Ilya Kreymer
- Yinlin Chen
- Daniel Lamb
- Aaron Birkland

Agenda

1. Any pending issues from the last two weeks?
2. Webrecorder Integration with Fedora:

   - Current proof-of-concept: http://fedora.webrecorder.net/
   - Webrecorder writing WARCs, reading from Fedora (no data model, just flat list so far)
   - Using Fedora's HTTP range request support

   Goals:

   - Create PCDM data model for web archives
   - Store WARCs, as well as other web archiving objects created by Webrecorder
3. 4.7.4 release

   a. 4.7 LTS?
   b. Next release: 5.0.0
   c. Migration approaches from 4.7 to 5.0
4. Performance lessons from PREMIS events (Ben Pennell)
5. Volunteer for next week's tech meeting (8/24)?
6. ...
7. Status of "in-flight" tickets

Ticket Summaries

1. Please squash a bug!

key	summary	type	created	updated	due	assignee	reporter	priority	status	resolution

Unable to locate Jira server for this macro. It may be due to Application Link configuration.

2. Tickets resolved this week:

key	summary	type	created	updated	due	assignee	reporter	priority	status	resolution

Unable to locate Jira server for this macro. It may be due to Application Link configuration.

3. Tickets created this week:

key	summary	type	created	updated	due	assignee	reporter	priority	status	resolution

Unable to locate Jira server for this macro. It may be due to Application Link configuration.

Minutes

1. Any pending issues from the last two weeks?

    There is a PR for the import/export tool.

2. Webrecorder Integration with Fedora

    Webrecorder project (Ilya and others) have been looking into having Fedora as a backend. Webrecorder is an interactive web archiving tool. Anyone can use it to record sites. They'd like to add a preservation backend. One of the things they'd like to do in the future is to have a standardized way to preserve web recordings. This is an area that's currently lacking as far as web archives are concerned. The current prototype, linked in the schedule, was a very quick weekend project, intended as a proof of concept (but, as Andrew noted, works beautifully). There is currently no tool that provides preservation and access. This integration that they're proposing addresses that.

    The Webrecorder folks are interested in a discussion on the data model for web archives. The PCDM discussion group (pcdm@googlegroups.com) would be the best place to ask these questions. They're interested in these sorts of discussions and that would be the best way to move the conversation forward. The Webrecorder folks have started brainstorming about the data modeling in a Google doc: https://docs.google.com/document/d/1RiZnX4g3u1ydwX9odu5Y1s2ajquhIkqYema5Tzp5UOQ/edit?ts=596fe0a4.

    Web archives are large objects so they're interested in learning how well Fedora handles this type of material. Andrew reports Fedora (Esme) has tested up to a TB file and that the tests have been successful.

    State of S3 backend storage and clustering? Clustering, as far as Andrew is aware, has not been exercised very much (or at all). It's sort of a feature Fedora gets from Modeshape for free. There have been issues at the Modeshape level that have driven their recent work. So there is clustering in Fedora, but there are sprinkled caveats all around it. As for S3, with the most recent release, there is official support for S3 as a backend. Danny Bernstein has done some testing of the performance of S3 as a backend. Its performance is more or less in line with a local installation. S3 may not be deployed anywhere in production since it's really new. It is going into the Hyku deployment though so it will be pushed on more. Several Samvera people report that there is also S3 integration at the Samvera level (though this is different from Fedora's integration).

    S3 support would be a higher priority for the Webrecorder groups because currently they store everything on S3. Fedora's S3 support is really undocumented at this point so maybe the Webrecorder folks working through this might be a good way to get some documentation. Or maybe some back and forth between Fedora and Webrecorder folks would be a good way to generate some documentation.

    Related: there is work going on around specifying the formal API of Fedora, which will probably be slightly different from the one Webrecorder is currently using. Just as a note.

3. 4.7.4 release

    4.7.4 release is out now. Fedora will be targeting a 5.0 release next that will have some breaking changes as the Fedora specification is finalized. The idea of Fedora having a long term support (LTS) version that the Fedora community would support for a period of years was discussed. This would mean that patches would continue to be applied to 4.7.x. Fedora 5 is only notional thing at this point and it will be quite some time before people migrate to it. Discussion agreed that having an LTS release is a good idea. What types of fixes can folks expect? Security fixes are definitely in. The group ought to articulate what will and won't be done. Another example is the project's dependencies (Java, itself, and otherwise); should underlying versions be upgraded over time? Java, definitely. Maven dependencies might be upgraded on a case by case decision(?)

4. Performance lessons from PREMIS events (Ben Pennell)

    Ben has tested different ways of storing PREMIS events (objects vs. RDF logs (serialized RDF in a binary)) There are some graphs in the Google Groups message. Storing events as objects, of course, results in more objects in the repository and Ben did find performance implications for this. At around 50k objects, the performance was getting significantly slower for creating events as resources/objects. Ben and his group's conclusion was that it didn't seem like it would be a good idea to keep events as objects in Fedora. As an alternative, for any object in the repository, there would be an RDF log where PREMIS events would be stored.

5. Volunteer for next week's tech meeting (8/24)?

    Someone willing to host next week's call? Andrew will be at a Fedora users group meeting in Texas. Aaron volunteered.