psc minutes 20110526

Thanks for everyone who attended today's preservation telephone call! Despite my best efforts, I failed to get everyone's names, for which I apologize. And if I have basic facts wrong in the minutes, please correct me.

I definitely noted Michael Giarlo, Penn State; John Wong, Emory; John Rees and John Doyle, National Library of Medicine; Deborah Kaplan, Tufts University. Who have I missed?

We had a wide-ranging and quite fruitful conversation.

1. Open Repositories

There's no formal get-together for the preservation group at Open Repositories, but several people expressed interest in an informal get-together, either during the day or at dinner or some such. If you would be interested in a get-together, RSVP to the list, and we can put together a Doodle poll for best possible times. John Rees is our resident Austin expatriate and therefore our ringer for knowing where are the good places to go in town for dinner.

People expressed some concern with how long it has taken to see the schedule for Open Repositories, but the schedule is up now.

Preconference events: http://sites.tdl.org/openrepositories/or11preconference/

- Some which might be of interest to this group include "Writing and Deploying Your Own Curation Task in DSpace"

Main conference preliminary program: https://conferences.tdl.org/or/OR2011/OR2011main/schedConf/program

- Some which might be of interest to this group include "The DuraCloud Pilot Partners Share their Experiences: Using the cloud for preservation and access services", " Free Tools for Your Preservation Toolbelt", "User Group Session 1B (Fedora): Media & Preservation".

And I'm sure plenty more.

2. NLM gave a report of where they are with preservation right now. The focus is on the policy side of preservation, with the hope that events metadata can be extracted from fedora. Right now they are limited in what they can do because their dependency on Muradora requires that they use fedora 3.2. If they can replace Muradora with some other access layer (perhaps Blacklight & Vufind, or Islandora), then they can upgrade fedora and then use the APIs to do checksum verification. They want to be a trusted repository, with all that entails.

3. Penn State also gave a report on the state of their repository, which is not fedora-dependent but is instead a suite of applications and microservices. Right now it is primarily access based, using a variety of different tools (such as ContentDM) for delivery of individual content types. The goal is to move towards a more distributed microservices vision, with the idea that even if the design changes, it's easier to rebuild one or two microservices than it is to change the architecture of a large repository.

In Penn State's distributed architecture, ingest will be based on microservices, every object will be a Bag using the BagIt specification, the objects will be stored with version control using Git, using PREMIS as a datastream for tracking preservation events, and so on, so a suite of pre-existing tools and specifications will define the repository. They've had a lot of success getting buy-in from management because they made sure that librarians consider themselves to be stakeholders throughout the process of developing this.

4. Tufts gave a report on the Submission Agreement Builder Tool. We've developed a schema for encoding submission agreements, and the resulting submission agreements are stored in XML into the repository. While the schema was designed with the idea that they can be machine readable for rules-based processing, no project is currently in process to do anything with those machine-readable submission agreements. Right now Tufts is just storing them. However, the essential data elements for future processing are resident in every XML submission agreement which is created by the open source tool, so automated processing of retention dates, reservation plans, etc. could happen in the future.

5. We had a general discussion of what it means to have "Trust" in a repository.

- We talked about how our users usually place (completely unwarranted) trust in cloud-based repositories such as Google docs or YouTube, and how this can bite them when those repositories go away (e.g. geocities).
- We talked about alternating focuses among repository groups: access versus delivery versus preservation and duration versus management services.
- We raised the question of whether it is appropriate to borrow trouble by addressing preservation concerns which have not been raised by our user community.
- Multiple people had things to say about file formats: do we use JHOVE and roll the file formats forward; do we do expectation management with our users; do we convert to a normalized and supportable form do we aim for something with widespread adoption (e.g. PDF) or something which is officially open (e.g. ODF); do we just cross our fingers and pray? And what do we do about older file formats?

6. We talked a little bit about the administrative aspects of this team.

- Several people expressed the desire that we should not be 100% focused on fedora but we should try to produce some kind of product that's useful for the Fedora community.
- Deborah raised the idea of having a rotating chair every year or two, in order to get input from as many different community members as possible. (Interested parties are welcome to raise their hands!)
- We realized there are no Hydra people who are regular meeting attendees, and no DSpace attendees. Deborah will do our reach to the Hydra community, and also to the people who are working on the Dspace duration system, to see if they would be interested in coming in either as regulars or occasional guest speakers to our mailing list and phone calls.