Curation Microservices: Making Data Management Easy
STS Discussion Group
10:30-12:00 Sunday June 26, 2011, Sheraton Hotel, Borgne Room
ALA Annual 2011
Joan Starr
University of California Curation Center
California Digital Library

Presenter Contact information
joan.starr@ucop.edu
Twitter: @joanstarr
Presentation slides available at http://www.slideshare.net/joanstarr/toward-making-data-management-easy

Introductions

- California Digital Library (CDL) – Serves 226,000 students 134,000 faculty and staff, for the entire University of California system
- Curation and Preservation, partnership between the campuses, CDL and some peer institutions-data curation requires a new structure
- Curation- New roles for new times, digital curation for preservation “refers to the actions people take to maintain and add value to digital information over its lifecycle, including the processes of maintaining digital content”
- Digital preservation - focuses on the series of managed assets – ensuring access to digital materials for as long as necessary

The Research Life Cycle

- Scholar has capacity for discovery because curators/libraries have collected
- Gathering because publication has happened, and vice versa- our activities enable the activities of the scholars. Protecting the intellectual investment of the institution and intellectual legacy
- Librarians can jump in at any time
  - Build or acquire tools at any time of the lifecycle
  - Enhance data sharing, collecting and gathering
  - Facilitate data publication
  - Partnerships to facilitate wide access
  - Help researchers comply with new requirements-
- Micro-services and Merritt
- DCXL
- WAS
- Data Paper model
- EZID
  - Partnerships (network level)
    - DataONE and DataCite
    - Data Management Plan Tool
Curation Micro Services

- Individual
  - Can build piece by piece
- Small and self contained
- Components
- Custom combinations
- Can solve complex problems

Building Blocks

- Building block list of critical components of complete curation solutions
- As an institution, you can add the pieces that you want

Merritt: Microservices “Off the Shelf” [www.cdlib.org/services/uc3/merritt](http://www.cdlib.org/services/uc3/merritt)

- Persistent identifiers

Merritt will soon be available to non-CA institutions to access

DCXL: Data Curation Excel

- Why Excel? Many researchers are storing their data in Excel, but not good for preservation
- Creating an open source plug-in to convert, researcher can install the plug-in
  - Permit standardized headers
  - Standard date formats
  - Auto archiving and persistent identifier
  - Discourage macros
  - (Q: Will it create a new format type? – A. not a new type, remains an excel spreadsheet, but has additional features)

WAS – Web Archiving Service

- Service started in 2007
- Capture ephemeral content from the web
- Anything on a website that can be reached by a click can be archived
- Example – archiving the gulf oil spill, captured 946 sites
- Powerful tool for curators
- Can capture the entire website or portions, individual resources, configurable
- If data is on a website, how does it capture? As a website and maintains format

Data Paper Model

- Data publication– particularly in the sciences, movement toward researchers getting “credit” for publication of data sets
- Provide citation support for the data set, some rich text
  - Gives the researcher an additional way to track their data set
  - Next steps would be to have a “data journal”, peer review, discipline grouping
  - Institutional investment
  - Foray into library as publisher
  - Q: Difference between a data paper and a paper about data?
Would a journal host the data or point to the data hosted elsewhere? A: Could work either way, using persistent identifiers so it wouldn’t matter where data was hosted. There are some data journals in the early stages.

EZID

- Problem- An article about data, but no data (paper provides information about data, but has no data sets)
  - When searching, you might eventually find data, but format is unsure (FTP site)
  - EZID provides a link between the data set and the scholarly publication
  - Gives a persistent link to the data set within the tool, adds metadata,
  - Update object location and object metadata
- Way to extend library services by meeting researchers needs
- Can be used early in the research life cycle – use EZID and can create a citable link to the data set before the paper is published. Update the metadata later when published (target URL).
- Can really aid a federated research team – stabilizes all references, citations, links, etc.
- Enables a researcher to “take their data with them” – leaving institution, etc.
- Meets funder requirements- having a data management plan
- Meets library needs to enable a new kind of scholarly output and need to build collections
- Meeting data center needs – new demands for storage, international standards
- http://n2t.net/ezid
- Q: Does EZID have a relationship to Orchid? A: At the present time, no. It could in the future.

DataONE- partnership at the network level

- NSF data project, build a new culture of data access and data sharing
- Building on existing and new infrastructures, and creating new communities of practice
- Librarians are at the table

Data Management Plan Tool

- https://bitbucket.org/dmptool/main/wiki/home
- Collaborative effort
- Funder’s data management sharing policies
- Data sharing requirements of journals
- UC schools, Smithsonian, ICPSR, UVA, and the NSF plus other institutions
  - Goal is to have this tool available for wide use in the Fall of 2011.
- Sharing of data increases citations and boosts discoveries
- Cultivate quantity and efficiency
- Better organization, easier preservation
- Intention is to be available without a fee

Build your Own / Open Source Tools / Off the Shelf Options

CurateCamp- Community “unconference” in August @ Stanford to connect practitioners and technologists interested in digital curation and data management
Discussion/Questions

DOI’s in existence, compatible with EZID? Granted permission to bring in legacy DOIs into the system and will be able to maintain them going forward.

Question to post to STS discussion group. Where will it be hosted? Is it customizable?

Discussion Question:

What are the advantages of automatic assignment of DOIs (to documents, datasets, and objects)?

- What do we do about versioning? Changing of the data sets and derived data?
- DOIs are special, not automatic identifiers,
- Data has to be intended to be taken care of on a perpetual basis and require care
- If data is changed or a new version, a new DOI should be issued.

How can we get involved in the data management planning tool effort?

- Check the wiki and see how to sign up for development and evaluating of the Beta.