Libraries as Partners in Research: the UC Curation Center’s Tools and Services

Patricia Cruse

University of California Curation Center
California Digital Library
data intensive research

• Ever increasing number, size, and diversity of content
• Ever increasing diversity of partners, and stakeholders
• Data are the new oil: “the future belongs to those that turn data into products” Mike Loukides
What keeps users up at night?

- How much will it cost?
- How do I fulfill the data management requirements of my grant?
- Why should I care about preservation? I just need a place to put my data.
- How can I share my work with my colleagues?
- How can I provide access to my work?
- How can I publish the data associated with my publications?
- How can I make sure I get credit?
- Can't my work be included in the Web of Science?
- What are “metadata”?
- Who owns my data?
- Are data included in open access?
Challenge for Researchers
Role of libraries

apply library knowledge and expertise to data challenges: manage, organize, describe, disseminate, preserve information

• Neutral service provider: work across the entire institution
• Intellectual Property experts: dealt with copyright, can translate to data
• Ability to advise on preserving research outputs
• Knowledge to advise on data management and curation
• Knowledge on complying with funder mandates, including open access
Resulting service landscape

- Create, edit, share, and save data management plans
- Create, share, archive, publish data
- Create and manage persistent identifiers
- Curation repository: store, manage, preserve, and share research data
- An infrastructure to publish and get credit for sharing research data
- A service to collect, manage and preserve Web-published content
DMPTool

Meeting funding agencies data management plan requirements

• Connect researchers to resources to create a data management plan
• NSF and directorates, NIH, NEH, IMLS, foundations plus
• Customizable

Primary Functions

1. Step-by-step “wizard”
2. Templates and examples
3. Links to institutional resources and agency information
4. Plan publication and sharing
DMTool usage

Campus activities:
- Campus researchers are active users
- Campus libraries have customized the tool for local use
Create, share, archive, connect, publish data

Excel is the database of choice for many researchers

**Primary Functions**

1. An Excel 1) add-in and 2) cloud application
2. Document data
3. Check for good data practices
4. Obtain identifier and citation
5. Archive and share
EZID: Long term identifiers made easy

- Precise identification of a dataset (DOI or ARK)
- Credit to data producers and data publishers
- A link from the traditional literature to the data (DataCite)
- Exposure and research metrics for datasets (Web of Knowledge, Google)

**Campus library activities:**
- Nine out of ten campus libraries subscribe to the service
- Campus libraries monitor how to expand / index

1. Create persistent identifiers
2. Manage identifiers (and associated metadata) over time
3. Resolve identifiers
Merritt Repository

- Curation repository open to the UC community and beyond
- Discipline / content agnostic
- New kind of cost model: *pay as you go* or *pay once and store for X*

Primary Functions

1. Deposit
2. Manage (metadata, versions, etc)
3. Access (expose)
4. Share (with other researchers)
5. Preserve
Merritt’s Diverse Service Offering: Meeting Campus Needs

• Dark archive for important digital assets

• Bright archive with direct discovery and access

• Preservation back-end for existing or new discovery and content management systems

• Integration with distributed data grids

Campus library activities:
• Use the service to archive library content
• Reach out to researchers
• Help grow the service
Vision for a “data paper”

• Wrap the unfamiliar in a familiar façade
• Minimally, a cover sheet and a set of links to archived artifacts
• Cover sheet contains familiar elements: title, date, authors, abstract, identifiers
• Just enough metadata to permit basic exposure to and discovery
  – Indexing by services such as Web of Science, Google Scholar
  – Instilling confidence in the identifier’s stability

Multi-decade, spatially explicit population studies of canopy dynamics in Michigan old-growth forests


Kerry D. Woods
Natural Sciences, Bennington College, Bennington, Vermont 05201 USA

Abstract
Established in 1933, a regular grid of 256 permanent plots includes about 10% of a 100-ha old-growth forest at the Dake Research Natural Area in northern Michigan, USA. Woody stems have been remeasured 3–7 times providing extensive quantitative records of population and community dynamics over periods of up to 72 years. Woody stems in upland hemlock–northern hardwood stands, about half of the study plots, have been mapped and individually tracked since about 1990. Remaining plots are in swampy stands dominated by Fraxinus nigra and Thuja occidentalis. Detailed, long-term demographic data for late-successional forests are rare in general, this data set is both of exceptional duration and unusual in spatial intensity and detail. Because sample plots are in a regular array over the stand, they can support analyses of spatiotemporal patterns at various scales. A major wind disturbance in 2005 provides a unique opportunity to compare disturbance response to baseline dynamics. Several publications based on this data set have already provided new insights into late-successional processes, but general availability of the data set with metadata should permit a range of further comparative and integrative analyses. The study is ongoing, and new measurements will be added to the archived data set. Several ancillary data sets are available from the author.

Key words: Acer saccharum; Betula alleghaniensis; Fagus grandifolia; Fraxinus nigra, long-term studies; northern hardwood forest; old-growth forest; permanent plots; succession; Thuja occidentalis; tree mapping; Tsuga canadensis.

Data Files
Files are ASCII text, tab-delimited. No compression schemes were used.
- all plots 1933-1948.txt -- data for all stems measured in 1933 and 1948.
- upland plots 09-07.txt -- data for upland plots mapped and measured two or more times, 1989 through 2007.
- species codes.txt -- four-letter codes and full names for all species.
- sampling history.txt -- table summarizing sampling history for all plots.
Data Publishing at the CDL

UC Curation Center
- Merritt Curation repository
- EZID: Persistent id management and resolution (ARFs, DOIs, et al.)

Publishing Services Program
- Online journals, with peer review
- Scholarly communication: grey literature to post-prints
- Search and display tools (XTF)
53 public archives
120+ archives total
58K crawls
7,500 + sites
600 million + URLs
60+ TB
24 institutions
What are people using WAS for?

Archiving at-risk websites and publications
Archiving their own university domains
Building web archives to complement library collections
Documenting web coverage of significant events

Campus library activities:
• Build collections: locally and collaboratively
• Provide service to campus community
New kind of cost model

• Understand costs in order to plan for and implement sustainable preservation services

• Investigate *paid-up* pricing in order to address
  – Boom-or-bust budget cycles
  – Fixed-term, grant funded projects

*Source: www.sharedidiz.com/*
For more information

UC Curation Center

http://www.cdlib.org/uc3
uc3@ucop.edu

Stephen Abrams Mark Reyes
Patricia Cruse Abhishek Salve
Scott Fisher Joan Starr
Erik Hetzner Rosalie Lack
Greg Janée Carly Strasser
John Kunze Marisa Strong
Margaret Low Adrian Turner
David Loy Perry Willett