Supporting Research Data Management at the University of California

Permalink
https://escholarship.org/uc/item/86p3x8hw

Author
Abrams, Stephen

Publication Date
2017-12-12
Kyoto University, 12 December 2017

Supporting Research Data Management at the University of California

Stephen Abrams
California Digital Library
University of California
Stephen.Abrams@ucop.edu
ORCID 0000-0003-2326-6672
Research data management (RDM)

Effective management of scholarly research data is necessary

- To ensure integrity and accountability
- To avoid needless duplication of effort
- To enable scholarly inquiry, innovation, and advancement
- To promote public awareness and informed discourse
The library’s role in research data management

The continuation of the long-standing mission to provide effective stewardship of the University’s intellectual capital by its libraries, archives, and museums

www.flickr.com/photos/janoma/4754744617
University of California

A large and diverse research University system

- 10 campuses
  - 238,000 students;
  - 190,000 faculty and staff
  - 150 academic disciplines;
  - 900 graduate programs

- 5 medical centers
- 3 national laboratories
- $4.97 billion (¥559 billion) in annual funded research
California Digital Library (CDL)

Providing transformative digital library services, grounded in campus partnerships and extended through external collaborations, that amplify the impact of the libraries, scholarship, and resources of the University of California.

www.cdlib.org
### Four CDL programs

- Bibliographic services
- Open access publishing and special collections
- Licensing and mass digitization
- Curation and preservation (UC Curation Center)

**CDL Services**

<table>
<thead>
<tr>
<th>Collect</th>
<th>Publish</th>
<th>Preserve</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Licensed Resources, Shared Print, Mass Digitization, Digital Special Collections, OAC</td>
<td>eScholarship Open Access Publishing, Publishing Tools &amp; Technologies</td>
<td>UC Curation Center, Microservices, Consulting, Preservation Repository</td>
<td>Melvyl™, UC-eLinks, Resource Sharing, Discovery and Delivery Tools</td>
</tr>
</tbody>
</table>
UC Curation Center (UC3)

Providing innovative solutions for active curation and long-term preservation of the University’s digital resources

uc3.cdlib.org
www.cdlib.org/uc3
What do we mean by curation and preservation?

Curation

“Maintaining, preserving and adding value to digital research data throughout its lifecycle”
www.dcc.ac.uk/digital-curation/what-digital-curation

Preservation

“Policies, strategies and actions that ensure access to digital content over time”
www.ala.org/alcts/resources/preserv/defdigpres0408
UC3’s initiatives in research data management

- Training
- Sharing/(re)use
- Planning
- Research
- Preservation
- Analytics
Planning

- Training
- Planning
- Research
- Analytics
- Preservation
- Sharing/(re)use
Planning

Ideally, data management decisions should be planned before a research investigation starts.

Formal data management plans (DMPs) are now required for funding proposals by all US federal agencies and many private foundations.

- Even if not required, data management planning should be encouraged as a scholarly best practice.
- It is better to be proactive, rather than reactive.
- It is better to be deliberate, rather than ad hoc.
DMPTool

- Create and share plans conforming to funder requirements
- Customized for public and private funders
- Customized with institutional resources and guidance
- Optional institutional review
- Public sample plans

dmptool.org
Best practices for data management planning that will produce FAIR data

Findable, accessible, interoperable, reusable

www.force11.org/group/fairdmp
Collaboration with UK Digital Curation Centre

Common platform consolidating DMPTool and DMPonline

Internationalization

github.com/DMPRoadmap/roadmap

blog.dmptool.org/2017/08/17/dmproadmap-summer-camp-news/
maDMPs – Machine-actionable DMPs

NSF-funded project to explore the transformation of DMPs from static documents into ““information structured in a consistent way so that machines, or computers, can be programmed against the structure”

blog.dmptool.org/2017/09/18/nsf-eager-grant-for-making-dmps-actionable/
Sharing/(re)use

Training

Planning

Research

Analytics

Preservation

Sharing/(re)use
Sharing/(re)use

Data should be published so that they are available for review, re-analysis, and the starting point for new inquiry
What makes data reusable?

- Findable
- Accessible
- Interoperable
- Reusable

www.force11.org/group/fairgroup/fairprinciples
The most significant step to making data reusable

Being managed by an appropriate curatorial program and system

- Minimally-curated data in a managed system can be enhanced over time
- Data that are not actively managed will become unusable sooner or later
Dash data publication service

- Self-service operation by researchers
- DataCite metadata schema.datacite.org/meta/kernel-4.0/
- Overlay layer sitting on top of any standards-compliant repository
- Multi-tenant UI
- Curatorial interface

dash.ucop.edu
Dash data publication service

- Findable
  - DOIs assigned and indexed with DataCite [www.datacite.org](http://www.datacite.org)
  - Authors/contributors, title, abstract, methods, usage notes, funders, geospatial locations
  - Formatted citations conforming to Joint Declaration of Data Citation Principles [www.force11.org/datacitationprinciples](http://www.force11.org/datacitationprinciples)

- Accessible
- Interoperable
- Reusable
Dash data publication service

- Findable
- Accessible
  - DOIs point to permanent dataset landing pages
  - Download data or data paper
  - Data download can be embargoed, while metadata remains accessible
- Interoperable
- Reusable
Dash data publication service

- Findable
- Accessible
- Interoperable
  - DataCite metadata schema.datacite.org
  - Optional references to citing articles and related datasets and data packages
- Reusable
Dash data publication service

- Findable
- Accessible
- Interoperable
- Reusable
  - CC0 or CC-BY licensing
  - Complete version history
Additional Dash features

- Multi-tenant UI with institutional branding
- Login with ORCiD or institutional credentials
- Upload via drag-and-drop or from external hosting sites, e.g., Box, Dropbox, Google Drive, laboratory server, etc.
- Curation interface to add value to managed data
- Usage metrics
Dash overlay architecture

- Ruby/Rails application
- Loosely coupled to an underlying repository
- Communication via standard protocols
- Can be integrated with any standards-compliant repository
Adoption of Dash has been slow

**Problem**

Everyone agrees on the benefit of data publication, but no one wants to do it if it means additional work

**Solution**

Integrate data publication as a side-effect of other activities that researchers are already doing, e.g., article publication

We’re working to integrate Dash into journal publication workflows
Sharing/(re)use

Sharing and reuse occurs not only through open publication, but also between collaborators, and within laboratories
Dat-in-the-Lab

Prototype use of the Dat peer-to-peer data sharing technology (*datproject.org*) in two University of California research laboratories

Streamlining data workflows for research and publication, and afterwards

*[uc3.cdlib.org/2017/09/27/moore-foundation-supports-uc3-research-data-management-project/]*/
Sharing/(re)use

Federal government agencies and laboratories perform about 11% of all research in the US.

The *data.gov* open data portal provides access to over 197,000 (42 TB) government datasets.

The ongoing commitment to sustain this access is unclear.
Mirror of the US federal government open data portal, data.gov

If these data are critical to the University’s mission, then it is our responsibility to help steward it.

datamirror.org
Preservation

- Training
- Planning
- Research
- Analytics
- Sharing/(re)use
- Preservation

UC Curation Center
Preservation

Ensuring that data remain accessible and usable by scholars and researchers now and in the future
Merritt

Preservation and access

- All 10 campuses, 35 curatorial units, 403 collections
- 2.7 million objects, 3.4 million versions, 40.9 million files, 89.3 TB
- Partial cost recovery for preservation storage

merritt.cdlib.org
Currently used primarily for cultural heritage material

But with growing data collections, including all Dash datasets

Integrated with the DataONE network www.dataone.org

CoreTrustSeal self-audit underway www.coretrustseal.org
Preservation at the network level

Digital Preservation Network (DPN) – additional storage replication
dpn.org

HathiTrust – mass digitization of serials and monographs
www.hathitrust.org

International Internet Preservation Consortium (IIPC) – web archives
netpreserve.org

National Digital Stewardship Alliance (NDSA) – advocacy
ndsa.org
Analytics

- Training
- Planning
- Research
- Sharing/(re)use
- Preservation
- Analytics

UC3 Curation Center
Analytics

If data are to be considered first-class research outputs alongside traditional publications, it is important to quantify their impact.

We need an infrastructure for tracking and reporting usage similar to that in place for the published literature.

  dx.doi.org/10.1038/sdata.2015.39
Collaboration between UC3, DataONE, and DataCite

COUNTER code of practice for data-level metrics (DLM)
www.projectcounter.org

Extending DataCite/Crossref EventData to support DLM
www.datacite.org/eventdata.html

makedatacount.org
Data literacy training

Training

Sharing/(re)use

Planning

Research

Preservation

Analytics
Data literacy training

Most scholars and researchers have never received any data literacy training.

They do not view the library as the natural place to turn for advice and guidance.
**Self-assessment maturity guide**

- Intended to assess individual researchers, not institutions
- Informative, not prescriptive

---

### Practice Maturity

<table>
<thead>
<tr>
<th>Planning for data</th>
<th>When it comes to my data, I have an informal “way of doing things” but not a formal plan.</th>
<th>I have a formal DMP document that outlines how I plan collect, manage, and save my data.</th>
<th>I have an active data management plan that is revisited throughout my project’s lifecycle.</th>
<th>My data management plan is the hub of all of my research activity.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saving data</td>
<td>I save my data only on my local machine(s).</td>
<td>I save my data on an external hard drive, server, or in the cloud.</td>
<td>I save my data in multiple locations.</td>
<td></td>
</tr>
<tr>
<td>Organizing data</td>
<td>I have identified the data I need to keep organized, including data types, sources, and file formats.</td>
<td>I apply consistent naming and structuring schemes to all of the files associated with my data.</td>
<td>I apply community or discipline-specific schemes when naming and structuring my files.</td>
<td>I work to ensure my data exists in a form that is interoperable and suitable for long term preservation.</td>
</tr>
<tr>
<td>Preparing data</td>
<td>I format my data consistently. I use the same units and formats across variables.</td>
<td>I document the format and structure of my data in a data dictionary, codebook, or readme.</td>
<td>I describe my data using a discipline-appropriate metadata schema.</td>
<td>Because of my well-defined organizational practices, my data is already prepared for analysis.</td>
</tr>
<tr>
<td>Analyzing data</td>
<td>I keep notes on the parameters, procedures, and protocols applied throughout my data analysis workflow.</td>
<td>I maintain a lab notebook that documents the specifics of my analysis workflow as well as my decision making process.</td>
<td>My protocol, lab notebook, or analysis workflow is collocated with the results of my analyses.</td>
<td></td>
</tr>
<tr>
<td>Sharing data</td>
<td>I communicate my data via tables and figures in a poster, presentation, or paper.</td>
<td>Any description of my data includes either a data availability statement or my data as a supplementary material.</td>
<td>I deposit my data in a database, repository, or system that provides a persistent identifier.</td>
<td></td>
</tr>
</tbody>
</table>

---

Library carpentry

What is Library Carpentry?

Library Carpentry is made by librarians, for librarians to help you:

- automate repetitive, boring, error-prone tasks
- create, maintain and analyse sustainable and reusable data
- work effectively with IT and systems colleagues
- better understand the use of software in research
- and much more...

Library Carpentry introduces you to the fundamentals of computing and provides you with a platform for further self-directed learning.

librarycarpentry.github.io
uc3.cdlib.org/2017/11/06/skills-training-for-librarians-expanding-library-carpentry/

- Increase awareness, develop new training modules, and coordinate national activities
Supporting research data management

- Training
- Planning
- Research
- Sharing/(re)use
- Preservation
- Analytics
Supporting research data management

at the University of California

UC Curation Center
California Digital Library
University of California

www.cdlib.org/uc3
uc3.cdlib.org
uc3@ucop.edu
@uc3cdl

Stephen.Abrams@ucop.edu

ORCID 0000-0003-2326-6672