Data Curation as a Form of Collaborative Research

D. Scott Brandt

Purdue University Libraries, techman@purdue.edu

Follow this and additional works at: http://docs.lib.purdue.edu/lib_fspres

Part of the Library and Information Science Commons

Recommended Citation


This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
Data Curation as a Form of Collaborative Research
Research as a Basis for Collaborative Data Curation
<table>
<thead>
<tr>
<th>Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>Exploration of campus research issues (“we need help organizing data!”)</td>
</tr>
<tr>
<td>2005</td>
<td>Strategic Plan → reorganization in Libraries creates position of Associate Deans for Research (ADR) to interface with Research Administration, support Libraries research</td>
</tr>
<tr>
<td>2006</td>
<td>e-Pubs IR (Digital Commons) launched and Distributed Data Curation Center (D2C2) created to leverage multi-disciplinary collaborations on funded research projects</td>
</tr>
<tr>
<td>2007</td>
<td>IMLS funded Data Curation Profile project (“who is willing to share what with whom, and when?”) and Data Research Scientist hired</td>
</tr>
<tr>
<td>2008</td>
<td>e-Data Task Force led to making data services a priority and demonstrated proof-of-concept data ingest into prototype data repository</td>
</tr>
<tr>
<td>2009</td>
<td>D2C2 continued collaborations with HUBzero® to investigate application of emerging library standards and technologies OAI-PMH, Linked Data</td>
</tr>
<tr>
<td>2010</td>
<td>NSF Data Mgmt Plan announcement, Libraries led meeting with IT, VPR, faculty</td>
</tr>
<tr>
<td>2011</td>
<td>Libraries/IT NSF DMP workshops, Purdue University Research Repository WG</td>
</tr>
<tr>
<td>2012</td>
<td>Four year budget to build out PURR, soft launch with data publishing functionality</td>
</tr>
</tbody>
</table>
Librarians provide education on depositing and finding data.

Librarians help shepherd and curate datasets.

Librarians contribute to or collaborate on projects.

Librarians consult on data management plans.

Small science, single PI, small lab

P. Bryan Heidorn, while Program Director at NSF, demonstrated that small science = 80% *
(awards under $350,000)

By small science, we mean usually a single PI (or one who may collaborate with another PI) in a fairly small lab setting, where there is likely on average 2-6 graduate students/post docs and possibly support from the department for a part-time lab manager or secretary.


photo from Science & Technology Review https://www.llnl.gov/str/December01/Orme.html
Small science can encounter unintentional consequences

Small science researchers self report: no specific person for data management/curation; data is likely saved to hard drives in the lab and backed up on CDs, usually by the students. While students have received “research integrity” training (which focuses on making data available upon request by funder, publisher, or FOIA, etc.) it is not likely that anyone could retrieve usable data easily or quickly.*

*D. Scott Brandt. Provost Fellowship, Final Report 2009 (unpublished) all pictures Purdue University
Because single PIs tend not to have mature data curation in place, they are not likely to see where to enter data curation continuum.
Download the Data Curation Profiles Toolkit

The Data Curation Profile Toolkit is available for anyone to download and use. However, we require that you register by creating an account on this site first. The Data Curation Profiles Toolkit is composed of four documents:

1. **The User Guide** - (login to download)
   
   The User Guide provides information about the Data Curation Profiles, including background information, the purpose and use of Data Curation Profiles, and directions on how to construct a Data Curation Profile.

2. **The Interviewer’s Manual** - (login to download)
   
   The Interviewer’s Manual provides the framework for the interview. It contains text and questions to be read to the participating researcher over the course of the interview. Some of the questions to be asked will be in response to the answers given by the researcher in the Interview Worksheet (see below).

3. **The Interview Worksheet** - (login to download)
   
   The Interview Worksheet is to be given to the researcher by the interviewer at the start of the interview. It is the worksheet that the participating researcher will fill out over the course of the interview. In addition to capturing important information, the responses provided by the researcher will serve as the basis for further discussion during the interview.

4. **The Template** - (login to download)
   
   The Data Curation Profile Template describes the structure of the Data Curation Profile. Each section or subsection within the Data Curation Profile template contains a brief definition of the information that is needed to populate an individual Data Curation Profile for the participating researcher.
What data sets do they have?
How are they managing these data sets?
What would they like to do with them?
2. How long would your dataset be useful or have value for you or others if it were to be preserved?

___ My dataset does not need to be preserved.
___ Less than 3 years.
___ 3 years or more but less than 5 years.
___ 5 years or more but less than 10 years.
___ 10 years or more but less than 20 years.
___ 20 years or more but less than 50 years.
___ 50 years or more but less than 100 years.
___ Indefinitely.

Have the interviewee answer question 2 on the worksheet relating to the length of time the data should be preserved. Then ask the interviewee to talk a little about his/her response:

• Why did you select this length of time?

Section 13 - Preservation

This section contains information about the needs / desires of the data client regarding the preservation of the data set under discussion.

If needed, a general statement about the researcher’s preservation needs can be inserted here.

13.1 - Duration of preservation

A statement about the length of time the data is to be preserved. The duration may be event based rather than time based, though estimation for the length of time related to the event should be noted where possible.
13.1 - Duration of preservation
The scientist believes that her data should be preserved for 10 years or more but less than 20 years. The scientist believes that the mathematical model they are working on will be perfected within this timeframe. When the model is considered complete the data will lose much if not all of its value. In addition, the device used in generating the data may become obsolete within this time period and replaced, again diminishing the value of the data.
Module 10: Data Sharing

This module presents information on data sharing and includes some helpful exercises for determining a researcher's conditions for sharing their data. Before beginning this module, it is recommended that you have an understanding of the data lifecycle as presented in module 9.

Source: http://www.urisa.org/publications/journal/articles/when_data_sharing
Distributed Liaisons + Centralized Libraries Services + Campus Services = Support across the research lifecycle

OVPR → Experts

IT → Faculty
What about my current proposal?

- Option 1: use the Purdue Data Management Plan
  - Libraries/OVPR will provide a citable overview of basic, long term issues
  - You append details relevant to your proposal guided by questions provided by Librarians
  - Libraries are available to provide review of your proposal
  - Nelson, M. Stowell Bracke, and J. Carlsson

- Option 2: exploit data management in place for your discipline
  - Example: data management resources of the Consortium for Political and Social Research
  - http://www.icpsr.umich.edu/icpsrweb/ICPSR/

- Option 3: personalized solution based provide and control

What about my current proposal?

- If you elect to use the Purdue University Research Repository (PURR)
  - The quotable overview is here:
    - http://vpr.hubzero.org/about/usehub
  - Pre-award staff also have a short paragraph suitable for inclusion in the “Facilities, Equipment and Other Resources” section of your proposal

  It explains the current prospect of a centrally supported PURR resource available at no direct cost to the project.
7a-d are related to **Export Control**.

<table>
<thead>
<tr>
<th>Does this project require sponsor approval of publications?</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the announcement restrict participation based on citizenship?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Does this proposal contain technical data within the statement of work that are restricted for purposes specific to military or space applications?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Does this proposal contain technical data within the statement of work that are confidential or proprietary to a company or Purdue (i.e. requires a confidentiality agreement)? If yes, list the relevant pages.</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

8  Are you receiving private health information that is protected by HIPAA?  

<table>
<thead>
<tr>
<th>Project Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

5  Please identify all ITaP units which have or will contribute to this project:  

- Envision Center for Data Personalization  
- Rosen Center for Advanced Computing  
- Other ITaP Units  
- None

6  Is this proposal a resubmission of a previously submitted application?  

7  Are you an NSF Beginning Investigator?  

8  Does the funding agency require a management plan for data or digital products produced using the grant?  

- Yes, and we plan to use the [Purdue University Research Repository (PURR)](https://researchhub.purdue.edu)  
- Yes, and we plan to use another option  
- No, funding agency does not require

9  Will historical sites be affected?  

| If Yes, please provide an explanation: |

10  If this project has an actual or potential impact on the environment, has an exemption been
Will you be uploading any data that may be considered by the University to be sensitive or restricted? REQUIRED

☐ Yes, this project may involve uploading sensitive or restricted data to PURR.

*Please indicate the type of data that may be involved. Check the box if you're not sure.*

☐ This project may involve government-restricted, export-controlled, or proprietary company information (without permission).

☐ This project may involve data that is governed by an Institutional Review Board (IRB) Approval Protocol.

☐ This project may involve HIPAA data or Protected Health Information.

☐ This project may involve FERPA data or student records.
Office of the University Chief Information Officer

The Office of the University Chief Information Officer strives to develop and deliver world-class computing and digital resources to Purdue faculty, staff, and students. The goals of the office are to enable research in science and engineering, provide classroom technologies to support and improve learning, and present innovative tools for engagement.

Purdue is a partner in XSEDE, the most advanced, powerful and robust collection of integrated digital resources and services.
HOW HUBS WORK

HUBZero is a new way for scientists and engineers to publish and share information. The latest hub will be used to study the causes and effects of earthquakes for the Network for Earthquake Engineering Simulation, or NEES. This hub joins others focused on topics such as nanotechnology, pharmaceutical products, advanced manufacturing, cancer care, assistive technologies for people with disabilities, heat-transfer issues in engineering, and several others. New hubs are being created at a rate of about one per month.

1) A scientist named Hugh creates software for scientific modeling and uploads it to the hub to share with colleagues, much as he would share a video on YouTube. Other scientists and engineers can use the tool and rate or comment on it.

2) Another researcher, Sue, runs her data on the new software tool using a simple Web interface. The demanding computations are done using cloud computing, and are automatically farmed out to available computers across the nation. Sue's data is sent to supercomputers connected to the TeraGrid, Open Science Grid, or the new DiaGrid, and she receives her results within minutes.

3) Sue has questions about her results, so she asks the community of scientists on the hub what they think about the finding.

4) A third researcher, Drew, posts a possible answer to Sue’s question. Then, other scientists vote “thumbs up” or “thumbs down” on whether they agree with Drew's answer. Drew and the other scientists participate by voting on the best answer, all receiving hub points for their efforts.

5) Sue also has an idea for a new feature for the software tool, so she adds her idea to a wish list for the software on the hub. Because this is a feature she really wants, she offers a bounty of 200 of her hub points for the feature.

And Sue decide to collaborate on their next project, which is successful, so they write a research paper and give a presentation at a conference. They post the presentation slides on the hub and then track how many users view the materials. Drew uses the presentation and uses some of the slides in his class to explain this new area of research to his students. He creates a homework assignment which his students use Hugh’s simulation tool to investigate the phenomena, and then Drew posts the assignment on the hub for other educators.
### Storage space for PURR projects currently covered by central resources

<table>
<thead>
<tr>
<th>Option</th>
<th>For what</th>
<th>Space</th>
<th>For how long</th>
<th>For who</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default/trial projects</td>
<td>Just trying things out, or don't need much</td>
<td>500 MB</td>
<td>3 years</td>
<td>Anyone with a Purdue Career Account</td>
</tr>
<tr>
<td>Default data publications</td>
<td>Great for small publications</td>
<td>50 MB</td>
<td>10 years</td>
<td>Anyone with a Purdue Career Account</td>
</tr>
<tr>
<td>Supported projects</td>
<td>Funded projects with PIs from Purdue</td>
<td>100 GB</td>
<td>Life of project</td>
<td>Purdue faculty with a verifiable grant or account number</td>
</tr>
<tr>
<td>Supported data publications</td>
<td>Publishing work done on a funded project</td>
<td>10 GB</td>
<td>10 years</td>
<td>Purdue faculty with a verifiable grant or account number</td>
</tr>
</tbody>
</table>

### Estimated cost of additional space *

<table>
<thead>
<tr>
<th>Option</th>
<th>Space Available</th>
<th>For how long</th>
<th>For who</th>
<th>Cost:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extra project space</td>
<td>As needed</td>
<td>Life of project</td>
<td>Anyone with a Purdue Career Account</td>
<td>$2,102.57 per TB per year **</td>
</tr>
<tr>
<td>Extra publication space</td>
<td>As needed</td>
<td>10 years</td>
<td>Anyone with a Purdue Career Account</td>
<td>$14,297.48 per TB</td>
</tr>
</tbody>
</table>
Applying library science to information problems

“You know thirty years ago, good laboratory practice was... you took good notes, you took fifteen or twenty data points... and you had a nice little lab book. But we’ve scaled now to getting this mega amount of information and we haven’t scaled our laboratory management practices.”

“In an ideal situation we would somehow have some sort of standard under which we named things and stored things and kept track of things and we would, you know, have a way to get this information to our students.”
Identify

Utilize Data Curation Profiles to collect information about current data gathering, workflow and documentation

Assess

Review applicable data &metadata standards, as well as best practice to document workflow (e.g., USDA, DFataONE, AgMES, AGROVOC)

Analyze

Perform requirements analysis to determine specifications for new user demands which include enhanced data management & sharing

Model

Develop a data model that accommodates collection, recording and workflow of data as well as application of thesaurus/ontology

Demonstrate

Demonstrate through a use case that data can be ingested and used by a modeler; and compliant with DataONE API

Problem: Water quality research is conducted by several groups which results in various amounts and sets of data which need to be organized
Problem: How are digital objects of dance identified, accessioned, and prepared for distribution? And what are the best practices and policies required to produce a multimedia-based dance history website?
Data curation is a collaborative process...

*How this is manifested in PURR*

http://research.hub.purdue.edu
Researchers are guided to PURR for help with data mgmt plans by Pre-Awards, workshops and promotion, and by word-of-mouth

Librarians consult on data management plans.
Researchers can create projects at any time, invite others to join... the goal is to help facilitate research development.

Librarians can contribute to or collaborate on projects.
Once a grant is awarded, researchers get an increase in space allocation and length of time for project and data.

Project space increases from 50 MB to 100 GB.
To make data sets publicly discoverable and available, there is a submission and “publishing” process.

Librarians help shepherd and curate datasets.
To make data sets publicly discoverable and available, there is a submission and “publishing” process.
PURRR functionality & service

PURRR policy allows for a specified time for discovery, and then decisions are made regarding long-term preservation.

Librarians provide education on depositing and finding data.
General Library Services

Reference: assessing complex information needs
- Meet researchers in the lab, office, classroom, Starbucks
- Discuss issues/problems of all aspects of scholarly communication
- Identify resources and ways to meet those needs

Instruction: enhance finding, evaluating, using research
- Explore activities related to creating and using information
- Partner with faculty to teach various illiteracies (info, data, etc.)

Collection mgmt: appraising local collections, including data
- Identify content for different dissemination modes
- Determine selection practices for new collections

Liaison: engaging researchers in new ways, in their environments
- Discuss research initiatives, projects, outputs...
- Collaborate on funded research projects— apply library science
Specific Data Services

- Data reference
- Data mgmt planning
- Data consultation (may lead to collaborations/grants)
- Using PURR
- Promoting data DOIs
- Data mgmt education and information literacy
- Finding and using data
- Developing tools (DCP 2.0, DataBib, DMP-SAQ)
- Data visualization/GIS

- Developing data resources (LibGuides, tutorials)
- Linking data to articles and dissertations
- Promoting open access (Authors rights, IR deposit)*
- Leveraging publishing opportunities*
- Developing local collections*
- Collection mgmt of “e” (journals, data, archives)*
- Integrating systems *(i.e., finding data in Primo)
  * As relates to data
Data curation is a process based on partnerships & collaborations