2-5-2014

Defining and Deploying an Institutional Data Repository Service at Purdue (PURR)

Michael Witt

Purdue University, mwitt@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

Witt, Michael, "Defining and Deploying an Institutional Data Repository Service at Purdue (PURR)" (2014). Libraries Faculty and Staff Presentations. Paper 44.
http://docs.lib.purdue.edu/lib_fspres/44

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.
Defining and Deploying an Institutional Data Repository Service at Purdue (PURR)

Michael Witt
Head, Distributed Data Curation Center
Associate Professor of Library Science
http://www.lib.purdue.edu/research/witt
E-mail: mwitt@purdue.edu
OUTLINE

1. Context
2. PURR service definition
   a. Data management planning
   b. Project space for collaboration
   c. Publishing data
   d. Archiving data
3. Platform: HUBzero
4. Roles and collaboration
5. Conclusion
WHAT ARE “DATA”?

[Image: Diagram illustrating the concept of data, information, presentation, and knowledge with associated images of ingredients and a cake]
## FUNDING AGENCY MANDATES

The DMPTool is a comprehensive resource for guidance and resources for your Data Management Plan. It provides a structured way to address the requirements of different funders. Here are some examples:

<table>
<thead>
<tr>
<th>Funder</th>
<th>Funder Link</th>
<th>Sample Plan</th>
<th>Funder Requirements Template</th>
<th>Supported in DMP Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon and Betty Moore Foundation</td>
<td>Guidelines</td>
<td>Template [RTF]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gulf of Mexico Research Initiative</td>
<td>Template [RTF]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institute of Education Sciences (US Dept of Education)</td>
<td>IES Data Sharing Implementation Guide</td>
<td>Template [RTF]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institute of Museum and Library Services</td>
<td>Guidance</td>
<td>Template [RTF]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joint Fire Science Program</td>
<td>Template [RTF]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>National Institutes of Health</td>
<td>Guidance</td>
<td>Template [RTF]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEH-ODH: Office of Digital Humanities</td>
<td>Guidelines</td>
<td>NEH-ODH Sample</td>
<td>Template [RTF]</td>
<td></td>
</tr>
<tr>
<td>NSF - Mathematical Sciences</td>
<td>Advice to PIs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSF-AGS: Atmospheric and Geospace</td>
<td>AGS Advice</td>
<td>Template [RTF]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As new Data Management Plan Guidelines become available, they will be added to this page and eventually to the DMPTool.

**Disclaimer**

The sample plans provided here may or may not be associated with successful grant applications. They may contain details not relevant to your specific project. They are provided only to illustrate representative responses.

**Funder Key**

- NSF = National Science Foundation
- NEH = National Endowment for the Humanities
- IMLS = Institute of Museum and Library Services

**Additional DMP Tool Help**

- DMP Tool Guide
- Video Demo
ECOSYSTEM OF DATA REPOSITORIES

- Publisher, e.g., Dryad
- Sub/Disciplinary, e.g., RKMP
- Consortium, e.g., ICPSR
- Country, e.g., Research Data Australia
- Government, e.g., data.gc.ca
- Research center, e.g., NASA GES DISC
- Instrument, e.g., CHANDRA
- General-purpose, e.g., FigShare
- Roll-your-own, e.g., DataVerse
- University, e.g., PURR
- Many others…
The PURR service is a collaborative effort of the Purdue University Libraries, Office of the Vice President for Research, and Information Technology at Purdue. PURR is a designated university core research facility.

Designated community:
Purdue University faculty, staff, and graduate student researchers; their collaborators; and the current and future consumers of their data.
Data is written into the three pillars of our strategic plan:

- **Learning**
  “…information literacy defined broadly to include digital information literacy, science literacy, data literacy, health literacy, etc…”

- **Scholarly Communication**
  “Lead in data-related scholarship and initiatives”

- **Global Challenges**
  “We will lead in international initiatives in information literacy and e-science and … contribute to international information literacy, learning spaces, data management, and scholarly communication initiatives.”

Publish Datasets with DOIs

Use PURR to publish datasets with Digital Object Identifiers (DOI) that make it easier for people to cite your data and give you credit. Purdue is a founding member of DataCite, the international agency that registers DOIs for data.

http://purr.purdue.edu

Digital Curation Centre’s Curation Lifecycle Model: [http://www.dcc.ac.uk/resources/curation-lifecycle-model](http://www.dcc.ac.uk/resources/curation-lifecycle-model)
PURR Service – Internal Model

Draft 7/14/2011
PURR Working Group

PURR Workflow mapped to OAIS Reference Model
(required for TRAC certification)
PURR SERVICE – EXTERNAL MODEL

Plan

Collaborate

Archive

Publish
INTRO TO PURR VIDEO

http://www.youtube.com/watch?v=Yw0IJj7FqA8
FACT:

MANY FUNDING AGENCIES REQUIRE DATA MANAGEMENT PLANS WITH GRANT PROPOSALS.

purr.purdue.edu
THE PROPOSAL PROCESS AT PURDUE

When you're developing a grant proposal, a number of entities at Purdue are here to assist you. But whom do you contact and where? Individual requirements and circumstances will vary, but this infographic will help point you in the right direction.

Research CORES provides access to state-of-the-art cost-effective technology and data analysis https://engineering.purdue.edu/VPR/CORE

Research Regulatory Compliance will walk you through compliance issues vprregulatory@purdue.edu and/or, for PI's, vpr@purdue.edu

DATA MANAGEMENT PLANS

• Boilerplate text
• Example DMPs
• DMP Self-Assessment
• DMPTool
• Workshops
• Tutorials
• Reference and consultation with subject-specialist librarian and/or data services specialist

https://purr.purdue.edu/dmp
CREATE PROJECT AND COLLABORATE

Create:
- any Purdue faculty, staff, or graduate student researcher can create projects
- describe the project
- disclaim use of sensitive or restricted data
- receive a default allocation of storage
- register a grant award to increase allocation
- invite collaborators to join project

Collaborate:
- git repository to share and version files (Google Drive integration)
- wiki
- blog
- to-do list management and project notes
- newsfeed
- stage data publications
Sensitive data: Information whose access must be guarded due to proprietary, ethical, or privacy considerations. This classification applies even though there may not be a civil statute requiring this protection.

Restricted data Information protected because of protective statutes, policies or regulations. This level also represents information that isn't by default protected by legal statue, but for which the Information Owner has exercised their right to restrict access.

http://www.purdue.edu/securepurdue/policies/dataConfident/restrictions.cfm

- FERPA ➔ Registrar
- HIPAA ➔ Health Center
- IRB ➔ Human Research Protection Program
- Export Control ➔ Vice President for Research
Just one more thing before you get started... (our lawyers made us do it)

Because your project may involve data sharing, we need to make sure you understand and acknowledge some important Privacy terms. This is particularly vital if you work with sensitive data such as health or student records.

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.
  
  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).**
  
  **STOP! This site cannot be used to store government restricted, export-controlled, or proprietary company data (without permission).**

- **This project may involve data that is governed by an Institutional Review Board (IRB) Approval Protocol.**
  
  **You must read and acknowledge to continue:**
  
  Institutional Review Board (IRB) approvals have been received, and I take responsibility for ensuring compliance with legal requirements for all project members.

  - Yes, I read, understand and acknowledge the IRB Statement. **REQUIRED**

- **This project may involve HIPAA data or Protected Health Information.**
  
  **STOP! This site cannot be used to store HIPAA protected health information (PHI).**

- **This project may involve FERPA data or student records.**
  
  **You must read and acknowledge to continue:**
  
  Projects that may include data that are restricted by FERPA regulations must be reviewed by the appropriate University staff before proceeding. Your project will be temporarily suspended while it is being reviewed. You will be contacted by e-mail within 2-3 business days, and you will be able to continue with your project after it has been reviewed for FERPA compliance and approved.

  - Yes, I read, understand and acknowledge my project will be suspended until it has been reviewed. **REQUIRED**

- No, this project will not involve any forms of sensitive or restricted data.

Did you read and agree to the Privacy Terms? **REQUIRED**
PURR project tutorial video: http://www.youtube.com/watch?v=q5xGO_oF9uQ
# STORAGE MENU

Storage space for PURR projects currently covered by central resources

<table>
<thead>
<tr>
<th>Option:</th>
<th>For what:</th>
<th>Space Available:</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 space</td>
<td>10 GB</td>
<td>3 years</td>
<td>Any Purdue faculty, staff, and graduate student</td>
</tr>
<tr>
<td>Default data publications</td>
<td>Great for small publications</td>
<td>1 GB</td>
<td>10 years</td>
<td>Any Purdue faculty, staff, and graduate student</td>
</tr>
<tr>
<td>Supported projects</td>
<td>Funded projects with PIs from Purdue</td>
<td>100 GB</td>
<td>Life of grant</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>Per year</td>
<td>Any Purdue faculty, staff, and graduate student</td>
<td>$2.10 per GB per year **</td>
</tr>
<tr>
<td>Extra publication space</td>
<td>As needed</td>
<td>10 years</td>
<td>Any Purdue faculty, staff, and graduate student</td>
<td>$14.30 per GB</td>
</tr>
</tbody>
</table>

[https://purr.purdue.edu/about/pricing](https://purr.purdue.edu/about/pricing)
Graph of Flickr Photo-Sharing Social Network Crawled in May 2006

By David F. Gleich
Purdue University

Crawl of the Flickr photo-sharing social network from May 2006 returning a graph with 820,878 nodes and 9,837,214 edges. Dataset is distributed as a SMAT file with README file with code to read file in Python and MATLAB.

PURR publication tutorial video: http://www.youtube.com/watch?v=jYBcsfiRhio
Initial commitment of 10 years

- data producer or dept can fund for longer
- otherwise remanded to library collection

Design guided by ISO 16363 / TRAC

- Organization infrastructure
- Digital object management
- Technical infrastructure & Security Risk Management
Bagit “bag” contains:

- bag declaration file, manifest file, data files

Metadata file (XML):

- METS wrapper
- Dublin Core and MODS (descriptive metadata)
- PREMIS (preservation metadata)

MetaArchive: LOCKSS replication network (7 copies)
SUPPORTING POLICIES

- Terms of Deposit
- Collection Development Policy
- Preservation Policy
- Preservation Strategies
- File Format Recommendations
- Preservation Support Policy

https://purr.purdue.edu/legal/terms
REPOSITORY SOFTWARE: HUBZERO

- HUBzero, open source software: [http://hubzero.org](http://hubzero.org)
- Maintained by HUBzero Foundation, originally funded by NSF
- Over 50 hubs online, supporting different virtual scientific communities, hundreds of thousands of users
- [http://nanoHUB.org](http://nanoHUB.org) - grandfather of the hubs, exemplar
- Built to facilitate virtual communities and online, scientific collaboration, research/teaching
- Collaborate, develop, publish, access, execute, and manage content using a web browser
- Software tools, documents, multimedia, learning objects, datasets, etc.
- Social network functionality and collaboration features
- LAMP stack, Joomla framework, OpenVZ and Rappture, git, etc.
- EZID interface to mint DataCite DOIs
- Some extensions customized for PURR not in core distribution
• **Executive Committee**: Dean of Libraries, Vice President for Research, Chief Information Officer

• **Steering Committee**: 2 from libraries, 2 from IT, 2 from research office and sponsored programs, 3 domain faculty researchers

• **Personnel**: Project Director (.50), Technologists (3.85), HUBzero Liaison (.35), Metadata Specialist (.20), Digital Archivist (.25), Digital Data Repository Specialist (1.0)
PURR Project Director (50%)

Michael Witt

Three examples of responsibilities:

• resourcing (personnel, budget, coffee, etc.)
• oversees development roadmap, service definition and design
• communicates across constituencies
Digital Data Repository Specialist

Courtney Matthews

Three examples of responsibilities:

• primary point of contact for helping users and librarians utilize PURR
• coordinates outreach, support, and development (tons of community engagement)
• helps to acquire, organize, and ingest data collections
Digital Library Software Developer

Mark Fisher

Three examples of responsibilities:

• developing a module to create archival information packages from datasets published in PURR
• integrating PURR with MetaArchive, an LOCKSS preservation network
• web and graphics design to keep the PURR website current and dynamic
Digital Archivist (25%)

Carly Dearborn

Three examples of responsibilities:

- define and implement AIP as well as long-term digital object management and supporting practices
- lead policy development and documentation such as PURR’s preservation policy, preservation strategies, file format recommendations, and preservation support policy
- consult with data producers and librarians on file formats, appraisal of data collections, and data management planning
Metadata Specialist (20%)

Amy Barton

Three examples of responsibilities:

• consult with data producers and librarians identify and apply appropriate metadata schemas and vocabularies to describe datasets
• design and implement metadata for preservation, findability, and citability (i.e., DataCite DOIs)
• enhance and provide quality assurance for metadata for acquired data collections
<table>
<thead>
<tr>
<th>SUBJECT</th>
<th>NAME</th>
<th>PHONE</th>
<th>EMAIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accounting</td>
<td>Iliana Barnes</td>
<td>46243</td>
<td><a href="mailto:ibarnes@purdue.edu">ibarnes@purdue.edu</a></td>
</tr>
<tr>
<td>Aeronautical and Astronautical Engineering</td>
<td>Charlotte Erdmann</td>
<td>42872</td>
<td><a href="mailto:erdmann@purdue.edu">erdmann@purdue.edu</a></td>
</tr>
<tr>
<td>African American Studies</td>
<td>Jamilah Gabriel</td>
<td>43093</td>
<td><a href="mailto:jgabriel@purdue.edu">jgabriel@purdue.edu</a></td>
</tr>
<tr>
<td>Agricultural &amp; Biological Engineering</td>
<td>Marianne Stowell Bracke</td>
<td>69620</td>
<td><a href="mailto:mbracke@purdue.edu">mbracke@purdue.edu</a></td>
</tr>
<tr>
<td>Agricultural Economics</td>
<td>Mary Dugan</td>
<td>49825</td>
<td><a href="mailto:mdugan@purdue.edu">mdugan@purdue.edu</a></td>
</tr>
<tr>
<td>Agricultural Education</td>
<td>Marianne Stowell Bracke</td>
<td>69620</td>
<td><a href="mailto:mbracke@purdue.edu">mbracke@purdue.edu</a></td>
</tr>
<tr>
<td>Agricultural Extension</td>
<td>Marianne Stowell Bracke</td>
<td>69620</td>
<td><a href="mailto:mbracke@purdue.edu">mbracke@purdue.edu</a></td>
</tr>
<tr>
<td>Agriculture</td>
<td>Marianne Stowell Bracke</td>
<td>69620</td>
<td><a href="mailto:mbracke@purdue.edu">mbracke@purdue.edu</a></td>
</tr>
<tr>
<td>Agronomy</td>
<td>Marianne Stowell Bracke</td>
<td>69620</td>
<td><a href="mailto:mbracke@purdue.edu">mbracke@purdue.edu</a></td>
</tr>
<tr>
<td>American Studies</td>
<td>Lawrence Mykytiuk</td>
<td>43665</td>
<td><a href="mailto:lmykytiuk@purdue.edu">lmykytiuk@purdue.edu</a></td>
</tr>
</tbody>
</table>
Librarians consult on data management plans in their subject areas.

Creating opportunities for librarians to interact with researchers about data
Librarian is notified by e-mail when a new project is created or a grant is awarded, based on department affiliation of Purdue project owner.

Creating opportunities for librarians to interact with researchers about data
Librarian may consult or collaborate on project if needed.

Creating opportunities for librarians to interact with researchers about data
Librarians review and post submitted datasets.

Creating opportunities for librarians to interact with researchers about data
At the end of initial commitment (10 years), archived and published datasets are remanded to the Libraries’ collection. A librarian working with the digital archivist selects (or not) the dataset for the collection.

Creating opportunities for librarians to interact with researchers about data
• Soft launch in 2012; 2013 was our first full year
• PURR included in 911 data management plans with proposals
• 77 grants awarded
• 1,277 registered researchers
• 239 active research projects
• Average project team size: 4 people
• Average files per project: 67 files

DMP analysis (n=111 NSF proposals from Purdue, Jan-Jun 2013)
• 49% PURR
• 29% Local computer or server
• 14% Disciplinary repository (e.g., ICPSR, Protein Data Bank, nanoHUB, NEES)
• 8% No data or not applicable
THANK YOU

PURR: http://purr.purdue.edu

Michael Witt
Head, Distributed Data Curation Center
Associate Professor of Library Science
http://www.lib.purdue.edu/research/witt
E-mail: mwitt@purdue.edu