The purpose of this workshop is to explore data management, identify activities that can help integrate library services into the research lifecycle, and demonstrate that data management and curation is about building relationships and engaging stakeholders at all levels across the university.
Overview of Data Management

Taking a Look at the Landscape

- Opening: Sarah E. Thomas, Vice President for the Harvard Library, Harvard University
- Remarks: James L. Mullins, Dean of Libraries, Purdue University
- Paul Bracke: Reflections from past day
- Sarah Demb: Local context at Harvard
Paul Bracke: Reflections from past day
• Line Pouchard: Overview of lifecycles: research, data, and curation
• Amy Barton: A model for collaboration
• Chris Erdmann: on education & scale, and using tools
• Scott Brandt: Building relationships and crafting messages
Interconnecting Lifecycles – Research, Data, Curation
DEFINITIONS OF DATA CURATION

• **Data curation** is a term used to indicate management activities required to maintain research data long-term such that it is available for reuse and preservation (Wikipedia)

• The active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for reuse over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation

• Source: Specialization in data curation, Graduate School of Library and Information Science, University of Illinois (https://www.lis.illinois.edu/academics/degrees/specializations/data_curation)
DATA LIFE CYCLES

• There are many life cycle examples:
  - research life cycles, data life cycles, curation & preservation life cycles
• In our practice, we use them both as metaphor and schematic
• Data life cycles help determine our tasks and roles at every step of the research project
• Sometimes it is useful to show them to a researcher, but not always.
EXAMPLE OF LIFE CYCLES WITH THREE PHASES

Research and Preservation

Data Collection
- field measurement
- data design
- data management
- database extraction
- external acquisition

Project Data Maintenance
- database creation
- data entry
- data integration
- data transformation
- metadata preparation

Data Analysis
- analysis
- storage data modeling
- process documentation

Planning
- research plan
- data design
- collection protocols
- collection estimates
- data source identification

End of Research
- research publication
- transition planning
- hand-off to preservation

Acquisition Planning
- data type
- append to existing
- load programming

Acquire Data
- new data types
- append to existing
- load programming

Data Management
- data maintenance
- dictionary maintenance
- versioning maintenance
- backup and recovery
- offline storage
- media rotation

Migration
- database upgrades
- media conversions

Purge

Provide Access
- internal/external
- access restrictions

Data Access Systems
- design
- develop
- deploy
- maintain

Research Data Management Cycle

Preservation Data Management Cycle
Research Life Cycle

- Proposal Planning Writing
- Project Start Up
- Data Collection
- Data Analysis
- Data Sharing
- Data Archive
- Re-Use
- Deposit
- Re-Purpose
- End of Project

http://data.library.virginia.edu/data-management/
ROLES AROUND DATA

Data reference questions (where to find standards)
Reviewing/revising DMPs (providing input/suggestions)
Data management planning (identifying metadata along lifecycle)
Data consultation (may lead to collaborations/grants)
Using repository (local, disciplinary)
Promoting data DOIs
Data information literacy (graduate students/lab)
Finding and using data (e.g., using r3data.org
Developing tools (e.g., Data Curation Profiles)
Developing data resources (LibGuides, tutorials)
Developing local data collections
Promoting open access
• Data life cycle models and concepts, CEOS V. 1.2
http://ceos.org/ourwork/workinggroups/wgisss/documents/

• The Use of Life Cycle Models in Developing and Supporting Data Services by Jake Carlson, Research Data Management: Practical Strategies for Information Professionals, J. Ray ed. Purdue University Press, 2014.

http://www.dcc.ac.uk/resources/curation-lifecycle-mode
DATA LIFECYCLE ACTIVITIES

Plan

Discover

Cloud

Acquire

IR/DR

Preserve

HPC data storage

Prepare

Analyze

Assure

Description of Data

Data Formats

Metadata for Organization

Standards for Interoperability

Documentation of Methodology

Intellectual Property Rights

Software

Sharing & Access Policies

Provenance for Preservation

Attribution/citation

Description for Discovery
# Questions Informing Curation Activities

<table>
<thead>
<tr>
<th>Plan</th>
<th>Acquire</th>
<th>Prepare</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Volume</strong></td>
<td>What is an estimate of volume &amp; growth rate?</td>
<td>What is the most suited storage (databases, NoSQL, cloud)?</td>
</tr>
<tr>
<td><strong>Variety</strong></td>
<td>Are the data sensitive? What provisions are made to accommodate sensitive data?</td>
<td>What are the data formats and steps needed to integrate them?</td>
</tr>
<tr>
<td><strong>Velocity</strong></td>
<td>Is bandwidth sufficient to accommodate input rates?</td>
<td>Will datasets be aggregated into series? Will metadata apply to individual datasets or to series?</td>
</tr>
<tr>
<td><strong>Veracity</strong></td>
<td>What are the data sources? What allows us to trust them?</td>
<td>Who collects the data? Do they have the tools and skills to ensure continuity?</td>
</tr>
<tr>
<td>Questions Informing Curation Activities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td><strong>Analyse</strong></td>
<td><strong>Preserve</strong></td>
<td><strong>Discover</strong></td>
</tr>
<tr>
<td>Volume</td>
<td>Are adequate compute power and analysis methods available?</td>
<td>Should raw data be preserved? What storage space is needed in the long-term?</td>
</tr>
<tr>
<td>Variety</td>
<td>Are the various analytical methods compatible with the different datasets?</td>
<td>Are there different legal considerations for each data source? Are there conflicts with privacy and confidentiality?</td>
</tr>
<tr>
<td>Velocity</td>
<td>At what time point does the analytical feedback need to inform decisions?</td>
<td>When does data become obsolete?</td>
</tr>
<tr>
<td>Veracity</td>
<td>What kind of access to scripts, software, and procedures is needed to ensure transparency and reproducibility?</td>
<td>What are the trade-offs if only derived products and no raw data are preserved?</td>
</tr>
</tbody>
</table>
COLLABORATIONS IN PROJECT

• Collaborations on multi-disciplinary proposals and projects
  • Levels of collaboration
• Developing customized Data Management Plans
• Organizing your data
• Describing your data
• Sharing your data
• Publishing your datasets
• Preserving your data
• Education on best practices
Research Data Management at Harvard

Sarah R Demb, Senior Records Manager/Archivist

THANKS TO: Skip Kendall, Senior Collection Development & Electronic Records Archivist
Harvard’s Research Data Policy

http://vpr.harvard.edu/pages/research-record-retention-and-management
**RESEARCH DATA LANDSCAPE**

- **HU Office of the Vice Provost for Research**
- **Schedules:** RMO, HUA, CHOM, HBS
- **GMAS:** HUA, OSP, OTD, OGC & the University CIO
- **Harvard [CHOM, Loeb Library (GSD), and Baker Library Historical Collections HBS], MIT**
- **Harvard College Library, FAS libraries**
- **HUIT LTS**
- **HLPS**
- **IQSS**
- **H-SCfA/Wolbach**
- **Dataverse (IQSS)**
- **DASH (OSC)**
- **CHOM, Ernst Mayr Library, H-SCfA/Wolbach, Schlesinger Library, IQSS, HL, HCL, Vice Provost for Research, HMS, HUIT**
• Amy Barton: A model for collaboration
A Draft Conceptual Model for Libraries Expertise Conjoining with Domain Expertise to Apply Active Research to Produce Research Data

**Color code:**
- **Light blue** = Research channel (throughput)
- **Dark blue** = Process
- **Lavender** = Domain expertise
- **Purple** = Applied expertise in collaboration
- **Pink** = Results of collaboration of domain expertise, library expertise, and library services...
- **Turquoise** = Scholarship

**Components:**
- Conceptualization
- Data collection
- Data processing
- Collaboration
- Data product
- Data curation
A Draft Conceptual Model for Libraries Expertise/Services Conjoining with Domain Expertise to Apply Active Research to Produce Research Data and Digital Collection

**Domain Expertise:**
- Political Science
- Library Science
- Metadata
- Research Data

**Libraries Services:**
- Archives
- Digital Programs
- Libraries IT

**Model for Collaboration:**
*Amnesty International*
Rewarding the Tool Builders
Emerging technologies and digital initiatives to support research software developers

Christopher Erdmann
Head Librarian, Harvard-Smithsonian Center for Astrophysics
cerdmann@cfa.harvard.edu
@libcce
Software environments and tools are crucial
Make it easy to share and find code
Learn from each others’ code, build on it
Reward the tool builders
The UK Community of Research Software Engineers

The people behind research software

The people behind research software - Research Software Engineers (RSEs) - lack recognition and reward for the incredible contribution they make to research. The RSE Community have come together to raise awareness of this issue, to campaign for change, and to share knowledge and collaborate to improve research software.
Astropy: A Python community package for astronomy


1 Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany
2 Department of Physics and Astronomy, Yale University, PO Box 208102, New Haven, CT 06520, USA
3 Hobby-Eberly
4 Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
5 Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
6 Center for Astrophysics and Space Sciences, University of California, San Diego, CA 92093-0314, USA
7 Department of Astronomy, Columbia University, Pupin Hall, 535 West 120th St, New York, NY 10027, USA
8 Department of Astronomy and Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON M5S3H8, Canada
9 Argonne National Laboratory, High Energy Physics Division, 9800 South Cass Avenue, Argonne, IL 60439, USA
10 Department of Astronomy, Ohio State University, Columbus, OH 43210, USA
11 Kavli Institute of the National Academy of Sciences, 1500 California St NW, Washington, DC 20005, USA
12 University of California, Berkeley, 501 Campbell Hall, Berkeley, CA 94720, USA
13 European Southern Observatory, Karl-Schwarzschild-Straße 2, 85748 Garching bei München, Germany
14 Laboratoire d'Astrophysique de Marseille (LAM) UMR CNRS 7326, 13388 Marseille, France
15 ETH Zurich, Institute for Astronomy, Wolfgang Pauli-Strasse 27, Building HTT, 8093 Zurich, Switzerland
16 Center for Astrophysics and Space Sciences, University of California, San Diego, CA 92093-0314, USA
17 KISO Observatory, Cerro Calán, La Serena, Chile
18 LLOO Laboratory, California Institute of Technology, 120 E. California Blvd, Pasadena, CA 91125, USA
19 Center for Cosmology and Particle Physics, New York University, New York, NY 10003, USA
20 Department of Astronomy and Astrophysics, Pennsylvania State University, 525 Davey Lab, University Park, PA 16802, USA
21 Research School of Astronomy and Astrophysics, Australian National University, Mount Stromlo Observatory, Ginninderra Road, Weston Creek ACT 2611, Australia
22 RAAG, PO Box 3019, Observatory 7935, 7925 Cape Town, South Africa
23 ASTRONational Radio Astronomy Observatory, PO Box 2, 7990 JA, Dwingeloo, The Netherlands
24 Department of Astronomy, Cornell University, Ithaca, NY 14853, USA
25 California Institute of Technology, Pasadena, CA 91125, USA
26 Oxford Astrophysics, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK
27 Department of Physics and Astronomy, University College London, London WC1E 6BT, UK
28 NASA Goddard Space Flight Center, X-ray Astrophysics Lab Code 622, Greenbelt, MD 20771, USA
29 Laboratoire AIM, CEA Saclay, Bât. 307, 91191 Gif-sur-Yvette, France
30 Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany

Received 12 June 2013 / Accepted 23 July 2013

ABSTRACT

We present the first public version (v0.2) of the open-source and community-developed Python package, Astropy. This package provides core astronomy-related functionality to the community, including support for domain specific file formats such as double image transport system (DIT) files, Virtual Observatory (VO) tables, and contains Astronomical Data Explorer (ADEX) support, and physical and quantity conversions, physical constants specific to astronomy, celestial coordinate and time transformations, world coordinate system (WCS) support, and polar coordinates. It is an ecosystem of the following code: Astropy and associated external projects, is a framework for integrating third-party libraries, such as a model fitting framework, VO client, and server tools, spectral and polar methods (PYPY) photon-ray tools. The core development team is actively making additions and enhancements to the current code base, so we encourage anyone interested in participating in the development of future Astropy versions.

Key words. methods: data analysis – methods: miscellaneous – virtual observatory tools

Article published by EDP Sciences

A35, page 1 of 9

$ python setup.py --help install

This will prevent the astropy helpers script from attempting to reach out to PyPI.
<table>
<thead>
<tr>
<th>Citations for</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total citations</td>
<td></td>
<td>83</td>
</tr>
<tr>
<td>Normalized citations</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>Refereed citations</td>
<td>61</td>
<td></td>
</tr>
</tbody>
</table>

VS

Downloads (All Versions):
1338 downloads in the last day
22408 downloads in the last week
34636 downloads in the last month
Citing Code in the Literature

\[\{\omega_b, H_0, A_s, n_s, \tau_{\text{reio}}, \omega_{\text{cdm}+\text{dr}}, \Gamma_{\text{cdm}}, r, \Omega_k\}\],

in addition to the other nuisance parameters for the Planck likelihood, omitted here for brevity. The first five cosmological parameters stand respectively for the baryon density, the

\(^5\text{https://github.com/baudren/montepython-public.}\)
REFERENCES

Merrill P. W., 1936, Contributions from the Mount Wilson Observatory / Carnegie Institution of Washington, 536, 1
Pérez F., Granger B. E., 2007, Computing in Science and Engineering, 9, 21
Sarre P. J., 2006, Journal of Molecular Spectroscopy, 238, 1
Savitzky A., Golay M. J. E., 1964, Analytical Chemistry, 36, 1627
GitHub ≠ Preservation

Danger Zone

Make this repository private
Public forks can't be made private. Please [duplicate the repository](https://help.github.com/en/github/administering-a-repository/duplicating-a-repository) or [contact support](https://help.github.com/en/github/administering-a-repository/contacting-support).

Make private

Transfer Ownership
Transfer this repo to another user or to an organization where you have admin rights.

Transfer

Delete this repository
Once you delete a repository, there is no going back. Please be certain.

Delete this repository
Digital Object Identifiers (DOI) are the backbone of the academic reference and metrics system. If you’re a researcher writing software, this guide will show you how to make the work you share on GitHub citable by archiving one of your GitHub repositories and assigning a DOI with the data archiving tool Zenodo.

**ProTip:** This tutorial is aimed at researchers who want to cite GitHub repositories in academic literature. Provided you’ve already set up a GitHub repository, this tutorial can be completed without installing any special software. If you haven’t yet created a project on GitHub, start first by uploading your work to a repository.
GitHub meets Zenodo
Software Publication Growth

Since beginning of 2014, over 2K software packages released from GitHub and published via Zenodo

(as of November 2014)
“If you allow linking to DOIs for data and software but not citing them, what's the point?”

Benjamin Weiner
**REVISED wigExplorer**, a BioJS component to visualise wig data [v2; ref status: approved with reservations 1, not approved 1, http://f1000r.es/3kq]

Software availability

Zenodo: wigExplorer, a BioJS component to visualise wig data_v2, doi: 10.5281/zenodo.85168


References

wigExplorer, a BioJS component to visualise wig data_v2

Thanki, Anil S ; Jimenez, Rafael C ; Kalthakottil, Cerry K ; Corpas, Manuel ; Davey, Robert P.

wigExplorer is a BioJS component whose main purpose is to provide a platform for visualisation of wig-formatted data.

Note: This is a version 2 of wigExplorer codebase updated version of wigExplorer (DOI: 10.5281/zenodo.7721). In this version, code is optimised to inform user with alert messages in case of error, and code accepts fixedStep and variableStep wig format specified by UCSC (http://genome.ucsc.edu/goldenPath/help/wiggle.html), can work with multiple wigExplorer in same webpage.
DATA SCIENTIST TRAINING FOR LIBRARIANS
@COPENHAGEN

Register for DST4L

The DST4L course is free - but the number of seats are limited - and we will select the applicants best suited for the training program based on your responses to the questions below. The registration is open from the 3rd June 2015 and will close again the 10th June 2015. Shortly thereafter, we will inform applicants whether they have been accepted or not. To register, visit the DST4L @ Copenhagen Application Form.
Thank you. Questions?

cerdemann@cfa.harvard.edu
@libcce
Scott Brandt: relationships and messages

Some extrapolation...

• Data are for use.
• Every researcher his or her data.
• Every data its researcher.
• Save the time of the researcher.
• The world of data is a growing organism.

*with apologies to S. R. Ranganathan*
Research, data, curation lifecycles

Adoption of Technology

The Technology Adoption Lifecycle was developed by Joe M. Bohlen, George M. Beal and Everett M. Rogers at Iowa State University built on earlier research conducted there by Neal C. Gross and Bryce Ryan.

http://blog.ftfnews.com/2012/11/01/transforming-uncertainty-into-opportunity/
Time is on our side...
Research, data, curation lifecycles

Working Together as a Network

Distributed Libraries
Centralized Libraries Services
Campus Services

Support across scholarly communication

Faculty Liaisons
Divisions:
PSET
HLS
HSSEB

Data Management
Data and Metadata Services
Data Education Working Group
Distributed Data Curation Center

OVPR
Research Admin
Sponsored Programs
RCR & IRB

Disciplinary Faculty
Graduate Students
Undergraduate Students

Data and Metadata Services
Data Education Working Group
Distributed Data Curation Center

PURR
Digitization Center

Archives & Spec Collections

Publishing
University Press
e-Pubs IR

Rights Management
University Copyright Office

Provost
IM:PACT

ITAP
Central IT
Ctr for Excellence
Rosen Center for Advanced Computing
Build Relationships with Stakeholders

Once you've identified your key stakeholders, it's time to begin building relationships.

Start now. Begin introducing yourself to policy makers immediately with the understanding that relationships take time to establish and build.

Be visible. Be visible in the community and at meetings. Re-introduce yourself and say hello, and let policy makers get to know you BEFORE you need them to know you (e.g. in a crisis).

Be attentive. Listen for community needs and think about whether and how the library can meet those needs.

One of the best ways to advance a relationship is to meet face to face. This can be the hardest part, but it's so important for sharing your message. And you need to meet stakeholders in person, because that is how they connect with your library. Here's how to make that connection:

Think first. Before you set up a meeting, think about what the people you want to meet with care about. Then, try to connect your library with their interests. For example, if you are meeting with a city council representative, does that person sit on any committees that affect the library? Have they been an advocate for youth? You can provide them with information about youth in the community they may not know. You can also tell the story about how your library is essential for the youth in your community.

Find a mentor. If you've never participated in this kind of meeting with a stakeholder, ask to attend a meeting with a trusted colleague or two. You can pay attention to what they do, how they speak, and how they describe the importance of your library and its role in the community. Afterwards, ask
50 conversations → 10 discussions → 1 meeting

• Read: news, websites, announcements, etc.

• Go to seminars, talks, presentations, where people hangout for coffee
  – Let people know who you are

• Email people hello, congratulations, etc.
  – See if you get a nibble

• Suggest going out for coffee...
INTRODUCTION
Librarians at Purdue University are beginning to identify the scientific datasets that are being generated by our faculty and researchers as information assets to be collected, preserved, and made accessible as a function of the library’s collection development. These librarians are subject-area specialists, and many have advanced degrees in their respective disciplines in addition to a degree in library science. They have all been trained in collection management; however, much of this training was related to traditional formats such as monographs and serials and not datasets. In our experience, one of the most effective tactics for eliciting datasets for the collection is a simple librarian-researcher interview. In this piece, we share a set of ten questions that a librarian can use as a starting point for such a “data interview.” It is not a comprehensive strategy but instead a practical tool to draw out information that needs to be considered in order to evaluate the suitability of a dataset for the collection and the requirements for the infrastructure and services that will be needed for data curation.

#1 What is the story of the data?
Begin the interview with an open-ended question that allows the researcher to talk freely about his or her research, scientific workflow, and community of practice. This lends some insight into the value of the dataset and how it may fit into your collection and be used, and it provides the context for understanding how and why the dataset was created and how it was processed and analyzed.

#2 What form and format are the data in?
What computing environments (e.g., software) are required to use the data? If the data are in proprietary structures, you may consider reformating them into an agnostic, for many use cases that can be more easily re-versioned. Is there any existing metadata, either external to the data or description that could be extracted from it? Ideally the data could be described to be discoverable by researchers from another discipline.

#3 What is the expected lifespan of the dataset?
In many cases, there are distinctions in the utility of a dataset as it begins in a raw state and then is analyzed and processed into new forms and versions as a result of different steps in the research workflow. Different entities may have custody of the data and use it for different purposes at different times, affecting its provenance. Proving agencies may require that data be archived for a prescribed period of time or you may forecast its future value and the amount of time it should be retained. The data may be described and archived for effective preservation to ensure its accessibility and integrity over time.

#4 How could the data be used, reused, and repurposed?
This is a primary selection criterion that also impacts how the data are accessed and what policies may be needed to govern its use. As data are archived and shared, new and unintended uses for the data may increase in value. For example, a research dataset may be repurposed as a learning object.

#5 How large is the dataset, and what is its rate of growth?
It is important to quantify the size of the data for storage and network provisioning. If you intend to ingest it into your repository, what is its physical (bits) and logical (research) scale? Is the dataset static or dynamic? Ask for a sample of the data to examine.

#6 Who are the potential audiences for the data?
Information regarding potential users of the data and the users’ needs is paramount. Along with potential uses for the data, this is another primary selection criterion. In some cases, the data may need to be embargoed or restricted to a limited group of users who are granted permission to access it.

#7 Who owns the data?
Establishing and maintaining the intellectual property represented by the data should be discussed at the earliest opportunity, and any concerns should be resolved up-front. Many organizations have a submission policy that asks the contributor to verify that they own the data and have the right to submit it.

#8 Does the dataset include any sensitive information?
All data should be reviewed for information that violates confidentiality, such as identification information on human subjects. Data curation activities should be informed by institutional review board requirements.

#9 What publications or discoveries have resulted from the data?
The researchers may have a bias regarding the importance of their data. The purpose of this question is to establish an objective metric for determining the value of the data for the collection. Different metrics may be more appropriate in determining the selection criteria for different kinds of data and data collections.

#10 How should the data be made accessible?
There is value in making data accessible using a conventional web-based user interface, but machine-to-machine interfaces should also be evaluated. These methods of access will be informed by the answers to the previous questions, and this question can be asked in an open-ended manner to fill in any gaps remaining at the conclusion of the interview.

BUILDING RELATIONSHIPS: DATA INTERVIEW

http://docs.lib.purdue.edu/lib_research/81/

SUMMARY
Although building robust collections of datasets present several complexities and challenges to resolve, the process of looking at scientific datasets as information assets and exploring what is needed to develop and manage data collections is similar to the traditional collection development practices that have been successfully employed by librarians for decades. We offer these ten “data interview” questions as a springboard for librarians to explore data curation in greater depth and specialization.

Michael Witt (mwitt@purdue.edu)
Assistant Professor of Library Science

Jake Carlson (jcarlso@purdue.edu)
Data Research Scientist

Purdue University Libraries
Distributed Data Curation Center
http://D2C2.lib.purdue.edu

“Conducting a Data Interview”

Michael Witt & Jake Carlson, Purdue University Libraries, West Lafayette, Indiana, USA
Research Framework

- **Information Collection:**
  Interviews, project participation, consultation, feedbacks, etc.

- **Analysis:**
  Interview transcripts, survey results, email feedbacks, notes -> Nvivo 10

- **Action Plan:**
  Research Support
  Learning
  Outreach
**Results**

**Research Support**

- **Spatial Visualization**
  - Humanities
  - Educational purpose
  - Data sharing, broad impact

- **Data Integration**
  - Data from different sources
  - Data between peers
  - Data from external sources

- **Spatial Analysis**
  - Liberal Arts: getting started
  - STEM: advanced analysis
Results

Research Outcome

Geospatial Expectation (humanities):

- Spatial data integration
- Spatial visualization
  - online
  - user friendly
- search/query function
- scholar communication
• Wow-How-Now
  – Something to get attention, tell how you can help (i.e., what you do), give a current example

• The New Elevator Pitch
  – Conversation, emotional benefit, quantify, “close”

• Conflict:
  – ID a conflict (problem), escalate the conflict (say why it’s problematic), resolve the conflict (discuss ways to address the problem)

http://speakingppt.com/2012/07/26/3-best-elevator-pitches/
• Wow
  – “Data management seems to be so complex.”
• How
  – Describe how it involves steps in a lifecycle, keeping track of what’s going on, and documenting it...
• Now
  – Discuss how the library knows of checklists that can be adopted by graduate students in their work.
Conversation starter
  - “Do you think the people restoring the USS Constitution document every step of their work?”

Emotional benefit
  - “That’s got to be hard to do. But we try to help people do that for research: document the data lifecycle.”

Quantify
  - “We helped people write [X, many, some] document management plans last year.”

“Close”
  - “If you know of anyone we could help, here’s my card.”
In Conclusion

“...the difference between what the most and the least learned people know is *inexpressibly trivial* in relation to that which is unknown.”

http://www.rpmwin.com/user/padagge/einstein.htm
BREAK 10:45 - 11:00 a.m
A conversation between Harvard and Purdue on data repositories

Michael Witt and Mercè Crosas

mwitt@purdue.edu  mcrosas@iq.harvard.edu
Outline

> Introducing Harvard Dataverse & PURR

1. Metadata & digital object management
2. Sensitive data
3. Identifiers and citations
4. Access to data
5. Interoperability
6. Staffing & governance

> Turning the conversation over to you
What Does My Data Management Plan Need to Address?

There are many things that need to be included in your data management plan. Watch our video tutorials, read the step by step instructions, or view a completed data management plan to help you get started. You can also click the link below to start now.

Learn More
Choose the metadata fields to use in dataset templates and when adding a dataset to this dataverse.

- Use metadata fields from Demo Dataverse
- Citation Metadata (Required)  
  View fields + set as hidden, required, or optional
- Geospatial Metadata  
  View fields + set as hidden, required, or optional
- Social Science and Humanities Metadata  
  View fields + set as hidden, required, or optional
- Astronomy and Astrophysics Metadata  
  View fields + set as hidden, required, or optional
- Life Sciences Metadata  
  View fields + set as hidden, required, or optional
- Journal Metadata  
  View fields + set as hidden, required, or optional
- MRA Metadata  
  View fields
- Graduate School of Design Metadata  
  View fields
- Alliance for Research on Corporate Sustainability Metadata  
  View fields
- Political Science Replication Initiative Metadata  
  View fields
- PSI Metadata  
  View fields
- CHIA Metadata  
  View fields
- Digitaal Metadata  
  View fields

Metadata & digital object management
This video supplements Purdue University’s Steel Bridge Research, Inspection, Training, and Engineering (S-BRITE) Center report "Fatigue and Fracture Library for the Inspection, Evaluation, and Repair of Vehicular Steel Bridges."
<table>
<thead>
<tr>
<th>Color</th>
<th>Type</th>
<th>Protection</th>
<th>Additional Security Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>Non-confidential</td>
<td>Clear</td>
<td>Open</td>
</tr>
<tr>
<td>Green</td>
<td>Not harmful</td>
<td>Clear</td>
<td>Email, OAuth verified registration</td>
</tr>
<tr>
<td>Yellow</td>
<td>Potentially harmful</td>
<td>Encrypted</td>
<td>Password, Registered, Approval click-through DUA</td>
</tr>
<tr>
<td>Orange</td>
<td>Sensitive</td>
<td>Encrypted</td>
<td>Password, Registered, Approval, signed DUA</td>
</tr>
<tr>
<td>Red</td>
<td>Very sensitive</td>
<td>Encrypted</td>
<td>Two-factor Auth, Registered, Approval, signed DUA</td>
</tr>
<tr>
<td>Crimson</td>
<td>Maximum sensitive</td>
<td>Double Encrypted</td>
<td>Two-factor Auth, Registered, Approval, signed DUA</td>
</tr>
</tbody>
</table>
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.

Are you planning to upload datasets containing any sensitive or restricted data? 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.

- 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.

Note: Your project information has been saved. You can exit your browser now. You will be contacted by e-mail within 2-3 business days.
Identifiers and citation

Principle 2: Credit and Attribution
Principle 4, 5, 6: Unique Id Access Persistence
Principle 7: Specificity and Verifiability
Principle 8: Interoperability and flexibility:
Repository exports citation metadata in XML, JSON formats

Resolves to landing page with access to metadata, docs, and data

Authors, Year, Dataset Title, DOI, Data Repository, UNF, version
Virginia Avenue Bridge Hits

By Robert Connor¹, Jason L. Loyd¹, Teresa Washeleski², Teresa Morris¹, Nathan Riggs³

1. Purdue University 2. Michael Baker, International 3. Indiana Department of Transportation

The Virginia Avenue Bridge over I-65/I-70 in Indianapolis had been struck by many oversize trucks. In fall 2012, INDOT and Purdue University installed a camera to monitor the strikes.

Listed in Datasets

Citations

Non-affiliated (0) | Affiliated (2)
Affiliated authors

http://dx.doi.org/10.5703/1288284315507

http://dx.doi.org/10.5703/1288284315508
Access to data

Terms of Use

Waiver
Our Community Norms as well as good scientific practices expect that proper credit is given via citation. Please use the data citation above, generated by the Dataverse.

CC0 - "Public Domain Dedication"

Restricted Files + Terms of Access

Restricted Files
There is 1 restricted file in this dataset.

Terms of Access
Must request access in order to be granted access to this dataset.

Request Access
Users may request access to files.

Guestbook

Guestbook
There are no guestbooks available in Harvard Purdue Data Management Symposium Dataverse to assign to this dataset.
Terms of Use

Please read these Terms of Use carefully before accessing PURR. The PURR site is operated by a community of researchers and hosted by Purdue University. Your access to PURR is subject to your agreement to these Terms of Use. By accessing PURR, you accept, without limitation or qualification, these Terms of Use.

Modification of the Agreement

We maintain the right to modify these Terms of Use and may do so by posting modifications on this page. Any modification is effective immediately upon posting the modification unless otherwise stated. Your continued use of PURR following the posting of any modification signifies your acceptance of that modification. You should regularly visit this page to review the current Terms of Use.

Eligibility, Registration, Access, and Termination

The PURR site is limited to non-commercial educational and research activities. Posting of spam messages or other commercial advertisements is a violation of these terms and may result in termination of your account.

Certain portions of PURR are limited to registered users and/or allow a user to participate in online services by entering personal information. You agree that any information provided to PURR in these areas will be complete and accurate, and that you will neither register under the name of nor attempt to enter PURR under the name of another person or entity.

You are responsible for maintaining the confidentiality of your user ID and password, if any, and for restricting access to your computer, and you agree to accept responsibility for all activities that occur under your account or password. PURR does not authorize use of your User ID by third-parties.

We may, in our sole discretion, terminate or suspend your access to and use of PURR without notice and for any reason, including for...
API Guide

We encourage anyone interested in building tools to interoperate with the Dataverse to utilize our APIs. In 4.0, we require to get a token, by simply registering for a Dataverse account, before using our APIs (We are considering making some of the APIs completely public in the future - no token required - if you use it only a few times).

Rather than using a production installation of Dataverse, API users should use http://apitest.dataverse.org for testing.

Contents:

- **SWORD API**
  - Backward incompatible changes
  - New features as of v1.1
  - curl examples
    - Retrieve SWORD service document
    - Create a dataset with an Atom entry
    - Dublin Core Terms (DC Terms) Qualified Mapping - Dataverse DB Element Crosswalk
    - List datasets in a dataverse
    - Add files to a dataset with a zip file
    - Display a dataset atom entry
    - Display a dataset statement
    - Delete a file by database id
    - Replacing metadata for a dataset
    - Delete a dataset
    - Determine if a dataverse has been published
Interoperability
Turning the Conversation to You

- Metadata and Digital Object Management
- Sensitive Data
- Identifiers and Citations
- Access, Terms and Licenses
- Interoperability (APIs)
- Staffing and Governance
LUNCH 12:00 - 1:30 p.m
• Marianne Stowell Bracke: Overview of DMPs, examples
• Scott Brandt: Supporting DMPs
• Natasha Johnson: Approaches in Science & Engineering
• Jud Harward: Approaches in the Humanities
• Alex Caracuzzo: Approaches in Business
• Marianne Stowell Bracke: Overview of DMPs, examples
DATA MANAGEMENT PLANS

ELEMENTS, OPTIONS, & WHERE TO GET HELP...

PURDUE UNIVERSITY LIBRARIES
What is a DMP?

• Managing, disseminating and sharing research results.

• Describes
  – what data will be produced
  – how data will be standardized, organized, tagged, archived, shared, distributed, secured, preserved, etc.
  – identifies who will take responsibility for these actions.
  – Generally brief
Data Management Plan
Arthropod responses to grassland nutrient limitation.

1. Types of Data Produced
We will collect insects annually from the 30 experimental plots at each of the eight sites (see body of proposal for sampling details). Samples will be immediately deposited in sealable containers labeled with the data, site code (already existing), block, plot, and subsample. An associated record of any observations or notes will be entered in a field tablet computer and labeled with the same information. We will also record environmental information including temperature and general observations. Labeled samples will be transported back to the laboratory, where they will be sorted and identified using a dissecting microscope. We will identify and count the arthropods to the classification of order, with the exception of members of the order Auchenorrhyncha, which will be identified to species or morphospecies. Identifications will be reviewed by multiple researchers associated with the project and verified with the assistance of Stuart McKeay of the Systematic Entomology Laboratory of the USDA Agricultural Research Service. Representatives of the identified species and morphospecies will be vouchered to the Bell Museum of Natural History at the University of Minnesota (U of M).

Abundance for each group will be recorded by hand in a laboratory notebook during sorting. These data will be transcribed into an Excel spreadsheet as each sample is completed. The spreadsheets will be stored on a controlled-access U of M server directory that is backed up offsite nightly. Files will be named according to the format site_year_plot.csv using existing unique site codes. Lind will be responsible for the data during and after data collection until publication.

After identification, arthropod samples from each experimental plot will be subsampled and sent to the University of St. Thomas Kay lab for stoichiometric analysis. We will receive a spreadsheet of data after processing is complete. This spreadsheet will include the insect identification (including site code, date, year, plot, and arthropod identification) and percent by mass of carbon, phosphorus, and nitrogen. These files will be saved as .csv files in the previously described server directory.

Our data set will be used in combination with the existing Nutrient Network (nunet.unr.edu) data on plant responses to nutrient manipulation. The NutNet data is currently stored and managed in a MySQL relational database housed at the Minnesota Supercomputing Institute and accessed through a secure internet connection. We will add our data and metadata to the NutNet relational database. The existing .csv files will be read into temporary tables in the MySQL database, and then inserted into permanent data tables using insert query statements. The existing database schema links tables of data observations to a “plot” table describing the experimental unit. New tables will be created for each of the arthropod data types (abundance and stoichiometry) containing the unique plot identifier. Multiple tables may be necessary for efficient data storage and management; for example, an “Arthropod” table holding scientific names for use can be used to constrain the labels of abundance records to acceptable possibilities.

2. Data and Metadata Standards
The project will leverage existing metadata standards currently stored in Ecological Metadata Language (EML) format for the NutNet project. We will add additional metadata entries for the arthropod community composition and arthropod stoichiometry; field notes taken during the time of collection will be recorded. Morpho software will be used to generate the metadata file in EML. We chose EML format for our metadata since it allows integration with existing NutNet data housed in the Knowledge Network for Bioconsmplexity (KNET) data repository.

Example DMP - NutNet
© DataONE 2011
Data management and quality assurance plan for Deer Creek watershed, 56 pages

Fortunately, this document and the procedures can be re-used
Why do I need a DMP?

• Funding agency requirements
  – part of their Data Sharing Policy
  – ensure ready availability
  – use of research data beyond life of a project.

A DMP fulfills part of a central ethical principle and the responsible conduct of research.
Basic Elements of an NSF DMP

- **Types of data**, samples, physical collections, software, curriculum materials, other materials to be produced
Basic Elements of an NSF DMP

✓ Standards to be used for data and metadata format and content (if existing standards are absent/inadequate, this should be documented with proposed solutions);

Disciplinary Metadata

While data curators, and increasingly researchers, know that good metadata is key for research data access and reuse, figuring out precisely what metadata to capture and how to capture it is a complex task. Fortunately, many academic disciplines have supported initiatives to formalise the metadata specifications the community deems to be required for data re-use. This page provides links to information about these disciplinary metadata standards, including profiles, tools to implement the standards, and use cases of data repositories currently implementing them.

For those disciplines that have not yet settled on a metadata standard, and for those repositories that work with data across disciplines, the General Research Data section links to information about broader metadata standards that have been adapted to suit the needs of research data.

Search by Discipline

- Biology
- Earth Science
- General Research Data

http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp
Basic Elements of an NSF DMP

✓ Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights;

Ease of Access / Persistence of Data

- When Researchers contact PI directly
- Deposit raw research data with journal (e.g., J-Share)
- Website or file sharing service (e.g., ICPSR, JHU Data Archive)
- Data Archive (e.g., ICPSR, JHU Data Archive)

or PURR...

http://dmp.data.jhu.edu/preserve-share-research-data/sharing-your-research-data/
# Basic Elements of an NSF DMP

- Policies/provisions for re-use, and derivatives

## LICENCES

<table>
<thead>
<tr>
<th>License</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://ontheline.trincoll.edu/oral-history/creative-commons/cc-licenses-terms/">CC BY</a></td>
<td><strong>Attribution</strong>&lt;br&gt;Others can copy, distribute, display, perform and remix your work if they credit your name as requested by you</td>
</tr>
<tr>
<td><a href="http://ontheline.trincoll.edu/oral-history/creative-commons/cc-licenses-terms/">CC BY SA</a></td>
<td><strong>No Derivative Works</strong>&lt;br&gt;Others can only copy, distribute, display or perform verbatim copies of your work</td>
</tr>
<tr>
<td><a href="http://ontheline.trincoll.edu/oral-history/creative-commons/cc-licenses-terms/">CC BY ND</a></td>
<td><strong>Share Alike</strong>&lt;br&gt;Others can distribute your work only under a license identical to the one you have chosen for your work</td>
</tr>
<tr>
<td><a href="http://ontheline.trincoll.edu/oral-history/creative-commons/cc-licenses-terms/">CC BY NC SA</a></td>
<td><strong>Non-Commercial</strong>&lt;br&gt;Others can copy, distribute, display, perform or remix your work but for non-commercial purposes only.</td>
</tr>
</tbody>
</table>

---

For more information, visit [http://ontheline.trincoll.edu/oral-history/creative-commons/cc-licenses-terms/](http://ontheline.trincoll.edu/oral-history/creative-commons/cc-licenses-terms/)
Basic Elements of an NSF DMP

- Plans for archiving data, and other research products, and for long term preservation of them
Where can I find official information on the NSF Data Management Plan requirement?

– **Policy statement:** NSF grant Awards and Administration Guide (AAG), §VI.D.4, Dissemination and Sharing of Research Results


– **Implementation as a basic requirement of all proposals:** §II.C.2.j of the Grant Proposal Guide (GPG).


– Some Directorates and Divisions provide additional program specific instructions.

Other Agency Requirements

– NIH Data Sharing Policy and Implementation Guidance

– Data Management Plans for NEH Office of Digital Humanities

– EPA Funding Opportunities, G2010-STAR-N1 &N2 (Research & Data Plan: “describe plans to communicate and share data”)

– Department of Energy Statement on Digital Data Management
  http://science.energy.gov/funding-opportunities/digital-data-management/

With OSTP memo, others are coming on board
• Scott Brandt: Supporting DMPs

Is there a guide to the data anywhere? Yes, of course!

Data Sharing and Management Snafu in 3 Short Acts
By Karen Hanson, Alisa Surkis & Karen Yacobucci
NYU Health Sciences Libraries
August 3, 2012 (Last Update: December 12, 2012)
http://www.youtube.com/watch?v=N2zK3sAtr-4
Research Data Mgmt & Curation

- **Why manage & share your data?**
  - MIT Libraries

- **Managing Your Data**
  - University of Minnesota Libraries

- **Data Management Planning Support**
  - University of Virginia Libraries

- **Research Data Management**
  - University of Edinburgh
Data Management Planning Tool
Create, review, and share data management plans that meet institutional and funder requirements.

PUBLIC DMPS
List of sample data management plans provided by DMPTool users.
- CAREER: Parietal Cortex and the Transformation of Spatial Cognition into Action
- Biosignature Suites: Using Connections between Microbes and Minerals to understand Biogenic Carbonates
- A unified approach to preserving cultural software objects and their development histories

DMPTOOL NEWS
Latest information about data management and the DMPTool.
- US Dept of Energy data management requirements
- We need API use cases!
- DMPTool downtime this Saturday 7/12/14
- Misc. Stats for DMPTool2
- DMPTool workshop at DataONE Users Group...

DMPTOOL HELP
Overview of how to use the tool, plus resources and guidance on data management.
- Frequently Asked Questions
- Create a DMP
- Administer the DMPTool
- Data management guidance
- Community resources

View All
More News
View All
DMP Tool

Creating an account

1. You’ll be directed to login via your career account...

2. We suggest registering for a unique author ORCID, which helps publishers keep track of your work.

3. As it implies, DMP Tool is a tool you can use to create data mgmt plans.

4. Then create your account.

5. To register at ORCID site click Look up then click on the “here” link.

btw: when ORCID asks you to create an account this is not your career account (and you can use any password you want).
Creating a DMP

1. Click to get started—you can select a template for guidance, or copy an existing DMP...

2. Use this space to respond to the Template outline of questions & guidance, then save and print as needed.

3. Output may need to be reformatted to fit two pages…

EXAMPLE: PURDUE UNIVERSITY DIGITAL DANCE COMMONS (PurdueDDC)

1 Roles and responsibilities
Data management plan introduction:
Dance as an art-form is an inherently ephemeral, multi-faceted, and temporal-spatial form of expression that is difficult to recreate identically over time. Traditional methods of documentation such as notation, motion-capture, and film/video are unable to preserve all the components that construct a dance work. As such, dance scholars do not always have the reference materials they need for their research and dance heritage, traditions, and history are in danger of being lost. The Purdue University Digital Dance Commons (PurdueDDC) is a digital repository prototype for preserving the dance works of faculty and students in the Dance Department at Purdue University. The main intent is to identify key components essential to preserving dance works digitally for a specific audience. This Data Management Plan addresses the curation of two kinds of data:

- Data sets generated from work performed during the course of project implementation. This includes documentation, survey results, and observations ABOUT the use and implementation of PurdueDDC.
- Data sets consisting of dance works generated primarily content deposited by creators INTO PurdueDDC.

The reason for having two types of expected data comes from fundamentally a development initiative supporting arts research by user needs.

Roles & Responsibilities:
Project Data Personnel
The following lists the project personnel who may generate some form of data and the actions or items that would result in a data set.

- Project Director: Documentation of project progress, system design, instructional materials.
- Information Technology Department: Bundling virtual images, software back-ups.
- Development Student Assistant: Documentation of system configuration, any customized code.
- Ingest Student Assistant: Documentation and feedback of the ingest process.

Dance Data Sets Personnel
The following lists the project personnel who may generate some form of data and the actions or items that would result in a data set.

- Dance Collaborators: Choreographic materials, documents about the choreographic work, any derived or relate documents.
- Ingest Student Assistant: Digitized files of analog choreographic materials.

Generated by the DMPTool on August 27, 2014 13:04
"Hi, I know it’s short notice [Friday, 9 am] but I have to submit a data mgmt plan with my grant, and the deadline is this afternoon. Could you look this over and give me any advice?"
• Under #2. Expected data... One questions to ask is _who will be likely to use this data?_ which relates to impact. Although you allude to it, you could explicitly identify likely users of data (you may already do this elsewhere).

• Under #5. Data dissemination (etc)... By using PURR you will be able to: publish the data, which assigns metadata for discovery and access, a Digital Object Identifier for persistence, and a standardized format for citation. The DOI is the big thing because published data are easier to track, especially the citations of your data by others. But you can embargo the data until the paper comes out.

• Under #6. Data storage and preservation... You might organize this section explicitly into two groups— that protected by IRB, and that which can made publicly available—and state what happens to each.
ASSISTANCE WITH DATA

Funding Agency

Note: various agencies have data sharing/plan requirements

Data management plan requirement

Proposal with data management plan

University OVPR

There must be a plan to be able to submit proposal

SPS Pre-awards

Pre-Awards may recommend contacting Libraries for help

Libraries

Researchers

PURR
• Natasha Johnson:
  Approaches in Science & Engineering
In single Chemistry Lab:
Experiments/Reactions → Processes → Lab Habits → Transferability → Naming Conventions → Backups

Organic: (size?, formats)
• Publishable data??
(preservation, naming & file conventions, metadata, institutional repository, sharing)

Theoretical/Physical Chemistry:
Publishable data
(planning, managing, storing, preservation, et. al)

Opportunities:
pre-award, award, post-award, lab

Image Source: https://https://www.chemistry.nus.edu.sg/_images/research/org.PNG
Physics & Astronomy

- **national/international projects**
  - Dedicated data managers
  - Mediating between interdisciplinary and/or transdisciplinary researchers
  - Standards developed in practice
  - **Opportunities:** few

- **institutional projects**
  - Lab-specific practices that may not translate outside of that particular lab
  - **Opportunities:** pre-award, award, post-award, lab

Earth Atmospheric & Planetary Sciences

- Long history of longitudinal, systematic data curation and practices
- Not reactive practices, but proactive practices
- **Opportunities:** few

Image Source: Chad Evans, WLFI, Feb 2, 2014, https://www.youtube.com/watch?v=4gbnNrVDcbM
• Every person for themselves!
• Very little to no consistent practices within the field
• Project or lab level only
• **Opportunities:** pre-award, award, post-award, lab

At first glance
- Individual sport but...
- Contributes to the development and research of many other science disciplines

Opportunities
- Metadata
- File formats
- Naming conventions
- pre-award, award, post-award, lab

Image Source: https://mathemagicalsite.files.wordpress.com/2015/04/446-pure-math-vs-applied-math.png?w=474&h=226
• Varies by project
• “map and organize data so they can better understand relationships, patterns and trends”
  — Nicole Kong, GIS Specialist

• **Opportunities:** pre-award, award, post-award, lab, end-users

70000+ cameras all over the world, tons of images

*Image Source: https://engineering.purdue.edu/HELPs/Images/cam2screenshot.jpg*
• Conversations/Interviews
• One Size Does Not Fit All
• Mediating & Negotiating
• Standards developed in practice for projects and labs
• **Opportunities**: pre-award, award, post-award, project planning, anywhere and everywhere

Image Source: [https://ccph.memberclicks.net/assets/Images/ccph%20partnership%20graphic.png](https://ccph.memberclicks.net/assets/Images/ccph%20partnership%20graphic.png)
Jud Harward: Approaches in the Humanities
Research Data in the Humanities

Dr. Judson Harward
Harvard-Purdue Data Workshop
June 18, 2015
eScience and Big Data Research Computing
The Importance of Models in Science Today

Numerical models and “Big Data” analysis are compute intensive!
Do Humanists Use Computationally Intensive Models?

- Most Arts and Humanities computing involves searches, design, interactive analysis of datasets or writing.
- The scholar or student is always in the loop.
So What about Data Needs in the Humanities?  
The Growth of Harvard’s DRS

138 (current) terabytes in the DRS is a respectable amount of data. For comparison, the Sloan Digital Sky Survey contains 140 TB.

Chart courtesy of Andrea Goethals and Wendy Gogel, Harvard Library
Conclusion

Arts and Humanities computing requires low latency (fast) access to well indexed data sets, but rarely requires compute intensive processing.

Science, engineering, and bioinformatics computing often require very compute intensive processing that can proceed for hours or days without human intervention.

High performance compute facilities like the MGHPCC are optimized for the second, not the first.
A Recent Example: Visualizing Broadway

• Conceived and led by Professor Derek Miller, English Department
  – (http://visualizingbroadway.com/Visualizations/Broadway/)

• The Broadway theatre scene is Darwinian. It “is a field of battle. Productions fight for resources, tangible and intangible: money, artists, space, critical attention, audiences.”

• To study that field, Derek Miller has gathered data from two public domain web databases and started to visualize it:
The Growing Dominance of Musicals Over Plays

Cumulative Run Length of Plays and Musicals (1900-2014)

- Plays Run Longer than Musicals
- Musicals Run Longer than Plays


Run Length (Days):
- 0
- 5,000
- 10,000
- 15,000
- 20,000
- 25,000
- 30,000
- 35,000

The chart shows the cumulative run length of plays and musicals from 1900 to 2014. The orange area indicates that musicals have run longer than plays since around 1970, with a significant increase in recent years.
Who Are the Most Important (Connected) Figures on Broadway?
Who Are the Most Important (Connected) Figures on Broadway, 2
Two Further Examples: Women’s Worlds in Qajar Iran (WWQI)

• Virtual archive of documents about 19th and 20th c. Iranian women led by Prof. Afsaneh Najmabadi: www.qajarwomen.org

• Documents and objects digitized from private and public collections.

• Current browser interface allows navigation by type of object, collection, geography, period, subject, etc.

• Here we have the page that introduces the photograph collection.

• Note the “filters” to the right and keyword search above.
The Giza Archives, Tomb of Hetepheres

In Prof. Peter der Manuelian’s Giza Archives project, 3D navigation over the web, on a connected workstation, or in a special 3D theater allows users to navigate to reconstructed objects in their original contexts.
Giza Archives, Tomb of Hetepheres, Databases

Many visuals in the 3D reconstruction are hot linked to records on the archive web site.

**Eastern Cemetery: G 7000 X**

- Shafts only (no superstructure)
- **PorterMoss Date:** Khufu
- **Tomb Owner:** Hetepheres I (G 7000 X)
- **Excavator:** George Andrew Reisner
  - American, 1867–1942
- **Attested:** Snefru
- **Attested:** Khufu
- **Attested:** Huni
- **Attested:** Kawab (G 7110-7120)
- **Attested:** Meresankh III (G 7530-7540)
- **Attested:** Ankhnesmeryre

**Selected bibliographical references:**
- Dunham, Dows. *The Egyptian Department and its...*
Conclusion, 2

• To curate humanities data, we not only need the data (and the metadata), but we also need to preserve the functionality to interactively search, visualize, and summarize the data.

• What kinds of data? What kinds of functionality?
  – Database data?
  – Economic and other numeric data?
  – Social network data?
  – Annotations?
Alex Caracuzzo and Rachel Wise: Approaches in Business
Research Data Program at HBS: A Data Management & Curation Approach for Business

Alex Caracuzzo, Research Data & Collections Librarian
Rachel Wise, HBS Archivist
Community

234 Faculty & 150 Doctoral Students

KLS: Baker Research Services & HBS Archives

Support from: Research Computing Services, Division of Research & Faculty Development (DRFD), Doctoral Programs
KLS’ vision for RDP at HBS

Formalize processes to *Share* and *Archive* researchers’ data

An extension of established, successful services:

- *Find* data
- *Buy* data
- *Gather* & *Create* data
- *Analyze* data
HBS Dataverse brings together HBS datasets (15) deposited in HU Dataverse since 2010

Full-service deposit service for faculty
(1 dataset published in 1st year; 2 more in process)

Self-service deposit for doctoral students
(1 dataset in process)

Researchers can choose terms for sharing data
Why Share & Archive data?

**Compliance** with HU Provost’s policies on *Research Data Security* and *Research Record Retention & Management*  
[No external mandates]

**Visibility**
- Help create a wider audience for research
- Share data with other interested researchers

**Preservation**
- Persistent data identifiers
- Access to the data and documentation for years later  
[Licensed data cannot be archived]
Depositing data is a collaborative process

- Faculty research must be completed

- KLS consults with faculty researcher regarding:
  - security issues (researcher must bring data into compliance)
  - file format (BRS will re-format if needed)

- **HBS Archivist** creates a metadata record in DV

- Researcher signs **Deposit Agreement/Release Form**

- **HBS Archivist** deposits the data on researcher’s behalf
Metadata record

- Intake process helps us gather information to use in the catalog record
- Faculty member approves record and either writes or enhances the description field
- Challenge to come up with local practice for cataloging so we are consistent across records and can give self deposit guidelines

<table>
<thead>
<tr>
<th>Dateverse Field label</th>
<th>Dataverse Obligation</th>
<th>KLS Obligation</th>
<th>USC metadata values</th>
<th>Justification</th>
<th>Use case assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Required</td>
<td>Required</td>
<td>HBS Survey on U.S. Competitiveness</td>
<td>Identifies the work</td>
<td>Share data with HBS faculty</td>
</tr>
<tr>
<td>Publication Citation/Related Publications</td>
<td>Recommended</td>
<td>Strongly recommended</td>
<td>Porter, Michael E., Jan W. Rivkin. Prosperity at Risk: Findings of Harvard Business School's Survey on U.S. Competitiveness. Boston, MA: Harvard Business School, 2012.</td>
<td>Not &quot;required,&quot; because we are assuming that faculty may want to share data that has not been written up or published anywhere, but strongly recommended for data that has been</td>
<td>Share data with HBS faculty</td>
</tr>
<tr>
<td>Publication ID type</td>
<td>Recommended</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publication ID #</td>
<td>Recommended</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publication URL</td>
<td>Recommended</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study Global ID</td>
<td>Required</td>
<td>(automatically generated by system)</td>
<td></td>
<td>Identifies the work</td>
<td>Secure and preserve</td>
</tr>
<tr>
<td>Author Name</td>
<td>Recommended</td>
<td>Required</td>
<td>Porter, Michael E.; Rivkin, Jan W.</td>
<td>Identifies main responsibility for the work. Part of citation</td>
<td>Share data with HBS faculty</td>
</tr>
<tr>
<td>Author Affiliation</td>
<td>Recommended</td>
<td>Required</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Producer</td>
<td>Recommended</td>
<td>Required</td>
<td>Harvard Business School</td>
<td>Org or person that sponsored research (financial or admin responsibility). Part of citation</td>
<td></td>
</tr>
</tbody>
</table>
HBS Survey on U.S. Competitiveness

Porter, Michael E.; Rivkin, Jan W., 2013, "HBS Survey on U.S. Competitiveness", http://dx.doi.org/10.7910/DVN1/22665, Harvard Dataverse, V1

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Description
As part of the U.S. Competitiveness Project, Harvard Business School asked its alumni to complete an in-depth survey on U.S. competitiveness. Nearly 10,000 business leaders responded worldwide, resulting in a first-of-its-kind analysis of data from a broad group of central actors in the global economy. The survey results provide strong evidence that America faces a deepening competitiveness problem and help pinpoint where the roots of the problem lie. The survey findings inform the March 2012 issue of Harvard Business Review, which presents analyses of critical areas that drive U.S. competitiveness as well as action agendas for restoring America’s economic vitality.

Keyword
HBS Alumni, United States, Competitiveness

Related Publication
Research Data Program

Baker Library’s expert data specialists have deep experience with business, economic and social science data and databases, and we can help you with everything from finding and vetting sources, to analyzing your data, to archiving and storage. We will make sure to get you the right data, add value to the data, and save you time.

Do you need to:

- **Find** data? We can help you find the best source for your needs, whether it is something freely available, available at Harvard, needs to be purchased, or needs to be created from scratch. Please contact Alex Caracuzzo, HBS Research Data and Collections Librarian, for more information.

- **Buy** data? We will vet the vendor for reliability and negotiate the appropriate contract and best price for you. We will help you unpack the data when it arrives and put it into the optimum format for your project. Please contact Alex Caracuzzo, HBS Research Data and Collections Librarian, for more information.

- **Gather/Create** data? Our experts know the ins and outs of the various licensed and freely available business and economic databases. We can download or scrape data for you, merge it, clean it, and prepare it for your analysis. Please contact Sarah Friksen, Director of Baker Research Services, for more information.

- **Analyze** your data? We can help you with pilot studies, methodological and statistical consultations, and provide services for analyzing and visualizing your data. Please contact Sarah Friksen, Director of Baker Research Services, for more information.

- **Share** your data? We can help you with metadata creation and data deposit in a secure repository, like Harvard Dataverse Network or ICPSR. We will make your data discoverable by other academics. Please contact Alex Caracuzzo, HBS Research Data and Collections Librarian, for more information.

- **Archive** your data and surrounding research records? The HBS Archives is the repository for research and teaching records of HBS faculty and has been collecting research data since the founding of the school. Please contact Rachel Wise, HBS Archivist, for more information.

SERVICES

Data Acquisition
- Buying & Licensing Data
- Permission / Use

Collecting & Managing Data
- Database Creation & Data Collection
- Data Management & Analysis

Depositing & Storing Data
- Full Service Data Program
- Self Service Data Program

Archiving Data
- HBS Archives

QUESTIONS? Contact Us
Future steps and ideas

- Grow the program
- Deposit datasets created from Baker Library’s collections
- Improve Dataverse and website
- Explore possible collaborations with others who have a stake in HBS’ research data
BREAK 2:30 - 2:45 p.m.
A Dozen Things We Do...

12 Things We Are Already Doing That Can Translate to Research Data Management
• Marianne Stowell Bracke: Intro & Things we do: reference and disciplinary expertise
• Line Pouchard: consultation & interventions
• Natasha Johnson: Things we do: instruction
• Sarah Demb: Things we do: archives
• Amy Barton: Things we do: description and metadata
• Diane Sredl and Nancy Quinn: Things we do: purchasing data
• Marianne Stowell Bracke: Intro & Things we do: reference and disciplinary expertise
Reference Interview Skills

How long do you keep your data?

Okay. When you say forever...
Data curation profile toolkit

Ask in-depth or follow-up questions:

Current plans for storage?
Would you be willing to share your data? If so, at what stage?
What do you need to do to meet funder mandates? Is there a gap between current practice and how you would like to manage data?
Disciplinary information expertise...

...Disciplinary cultures of data practice

The response of the Contrarian Curmudgeon:

“You can have my data when you pry it out of my cold, dead hard drive.”
Disciplinary cultures

Where are they in relationship to data sharing? (ex. genomics vs. applied agriculture)

Do they routinely deal with sensitive data, such as human subjects or IP-rich research?

Are they interested in just better data management for better science/research vs. an interest in sharing?

Are they open to librarians assisting in developing standards or best practices?
Connect researchers to articles...

...connect researchers to data sets

[Logos for DRYAD, DATA.GOV, DatanONE, ICPSR, re3data.org]
• Natasha Johnson: Things we do: instruction
A Dozen Things We Do...

Traditional Publications

- Orientations
- Invited instructional sessions
- Library-disciplinary faculty co-teaching partnerships
- Librarian-led for-credit courses
- Librarian-led elective courses
- Instructional modules

Data Sets

- Lab consultations
- Departmental consultations
- Faculty consultations
- Graduate student consultations
- Departmental office hours
- Seminars
- Pre-award DMP consultations
- Instructional modules
- Marketing/outreach
A Dozen Things We Do...

**Distinctions of Quality**

### Traditional Publications
- Library-disciplinary faculty co-teaching partnerships
- Librarian-led for-credit courses
- Librarian-led elective courses
- Instructional modules

### Data Sets
- Lab consultations
- Departmental consultations
- Faculty consultations
- Graduate student consultations
- Departmental office hours
- Seminars
- Pre-award DMP consultations
A Dozen Things We Do...

Institutional Level

• Library-disciplinary faculty co-teaching partnerships
• Librarian-led for-credit courses
• Librarian-led elective courses
• Administrative partnerships
• Orientations
• Marketing/outreach
• Instructional modules

Personal or Lab Level

• Lab consultations
• Departmental consultations
• Faculty consultations
• Graduate student consultations
• Departmental office hours
• Seminars
• Pre-award DMP consultations
A Dozen Things We Do...

Depositing of Thesis/Dissertations

- Orientations
- Invited instructional sessions
- Training Workshops
- Library Workshops
- Knowledge Base

Depositing of Data

- Invited instructional sessions
- One-on-one consultations
- DMP follow-ups
- Lab consultations
- Knowledge Base
• Line Pouchard: Things We Do: consultation & interventions
Research Data management @ Purdue university Libraries
Welcome to CAM²

CAM², the Continuous Analysis of Many CAMeras, is a system for analyzing streaming data built by a team of Purdue University researchers.

Dr. Yung-Hsiang Lu, PI
CURATION ISSUES IN CAM2 PROJECT

• Data access and re-use
  - policies of video streams and CCTV
  - Sparse or piecemeal legal framework US
  - Policies are mostly ad hoc

• Data ownership

• Data storage

• Data organization
  - naming scheme
  - metadata

• Protect metadata storage – where the intellectual property lies

• Data information literacy skills for Big Data
WHAT WE DO AT THE ARCHIVES – documenting the data context

- Dataverse (contextualized datasets)
- General Records Schedule and specific schedules (eg OSP)
- PROVOST FOR RESEARCH PRINCIPLES
- HUA DATA SETS and PROJECT RECORDS
- DASH (published articles)
- Business systems (GMAS, FSS)
• Amy Barton: Things we do: repository intermediaries, metadata & data identification
A Dozen Things We Do...

E-Pubs Repository Intermediaries: Roles

Scholarly Repository Specialist

David Scherer
Tel: (765) 494-8511
E-mail: dscherer@purdue.edu

Dave is a trained Librarian and Archivist who is in charge of the Purdue e-Pubs online publishing platform and Purdue research collection. He is an advocate around the University for Scholarly Publishing Services, and provides support for faculty, staff, and student projects ranging from online conference proceedings to electronic theses and dissertations.
## A Dozen Things We Do...

### NEAL A. HARMeyer
- **University Title:** Digital Archivist
- **Division/Unit(s):** Archives & Special Collections
- **Location(s):** Stewart Center 434

### CARLY C. DEARBORN
- **University Title:** Digital Preservation and Electronic Records Archivist
- **Division/Unit(s):** Archives & Special Collections
- **Location(s):** Stewart Center 475

---

**Purdue e-Pubs** - A digital document repository including e-books, papers, reports, and more by Purdue Authors

**e-Archives** - The digitized archives and special collections of Purdue University

**PURR** - A platform for research collaboration and data management for Purdue researchers
A Dozen Things We Do...

PURR & Research Data
Intermediaries: Roles

Purdue e-Pubs - A digital document repository including e-books, papers, reports, and more by Purdue Authors

Publishing

Archives

Research Data

e-Archives - The digitized archives and special collections of Purdue University

PURR - A platform for research collaboration and data management for Purdue researchers
Feed the Future Food Processing and... (usaid)

Private project by Amy Barton

Files

- [ ] Branding-Marketing Materials
- [ ] Forms
- [ ] FPL-Docs
- [ ] land_pcp_paperegressions_cleaned.dla
- [ ] Policy Documents
- [ ] Templates
A Dozen Things We Do…

Traditional Cataloging
Metadata

<table>
<thead>
<tr>
<th>MARC 21 Format for Bibliographic Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library of Congress</td>
</tr>
<tr>
<td>Network Development and MARC Standards Office</td>
</tr>
</tbody>
</table>

**File interchange handbook for images, audio, and metadata**
Brad Gilmer; Society of Motion Picture and Television Engineers.; European Broadcasting
2004


<table>
<thead>
<tr>
<th>Print in Print</th>
<th>Details</th>
<th>Virtual Bookshelf</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title:</strong> File interchange handbook for images, audio, and metadata</td>
<td><strong>Author:</strong> Brad Gilmer; Society of Motion Picture and Television Engineers</td>
<td><strong>Subjects:</strong> Digital video – Handbooks, manuals, etc.; Digital television – Handbooks, manuals, etc.; Image processing – Digital techniques – Handbooks, manuals, etc.; Digital compression – Handbooks, manuals, etc.</td>
</tr>
<tr>
<td><strong>Description:</strong> Convergence of information technology and traditional television organization of metadata and the SMPTE metadata dictionary -- The digital 3D/DV general exchange format -- The material exchange format -- Advanced systems format -- QuickTime file format.</td>
<td><strong>Publisher:</strong> Burlington, MA : Focal Press</td>
<td><strong>Edition:</strong> 2004</td>
</tr>
<tr>
<td><strong>Format:</strong> xi, 307 p. : ill.; 24 cm.</td>
<td><strong>Language:</strong> English</td>
<td><strong>Identifier:</strong> ISBN 0240806050</td>
</tr>
<tr>
<td><strong>Creation Date:</strong> 2004</td>
<td><strong>Citation Source:</strong> PURDUE ALMA</td>
<td><strong>Type:</strong> Book</td>
</tr>
<tr>
<td><strong>MMS ID:</strong> PURDUE_ALMA21496265720001081</td>
<td><strong>Includes bibliographical references and index.</strong></td>
<td>**505</td>
</tr>
<tr>
<td></td>
<td>**650</td>
<td>Digital video</td>
</tr>
<tr>
<td></td>
<td>**650</td>
<td>Video compression</td>
</tr>
</tbody>
</table>
A Dozen Things We Do...

Digital Object Identifier (DOI)

Scholarly Publishing Services
CrossRef DOIs

Archives + MetaArchive
DataCite DOIs

PURR
DataCite DOIs
A Dozen Things We Do…

Dataset Persistent Identifier

DATA ➔ PUBLICATION ➔ DISCOVERABILITY ➔ ATTRIBUTION ➔ IMPACT FACTOR!
• Connie Rinaldo: Wrap-up
Wrap-up for the Workshop & Next Steps for Harvard?

Connie Rinaldo
Things we do, gathered from presentations

- Reference interview; how long is forever?
- Disciplinary expertise; understanding how researchers/labs feel about sharing, for instance
- Connect researchers to datasets & let them see how important metadata is by searching
- Instruction: teach about data: orientations, naming conventions, instructional sessions, co-teaching with faculty; modules, office hours, outreach, consultations, institutional partnerships
- Thesis/dissertation data deposit
- Consultation & intervention
- Curation: Access & re-use policies; ownership; storage; organization (metadata); data info literacy
- Archives: records management; documenting data context
- Repositories
- Purchasing data: archives, e-resources, license negotiation, access
Rick Luce Summary in one Slide

• Key Nodal Connector; Relationships
• Aggressively assert role as authoritative, trusted repository of curated, preserved, persistent knowledge: radical change
• Live with what we create
• Enable connections in vertical domain (research) because library enables horizontal integration (multi-institutional, cross-disciplinary; multi national)
• Data wranglers: IP policies & rights; integration with grants; DMP; metadata creation; collection tools, annotation/documentation; software code documentation; Standards: QA, open, procedures, methods, tools; Preservation; assessment
• Infrastructure from decentralized to micro-cloud
• Open science: reproducibility requires access but also code description/software development; standards & $$
• Value-added workflow; metadata (auto-generate); documentation; time stamp, tools
Harvard Next Steps: Practical & beyond

• Planning/consultation within the library
  – Reimagining
  – Some further along

• Coordination/consultation across divisions & offices
  – OSR
  – DASH/Dataverse
  – Connect whole research cycle & coordinate with libraries

• Talking to faculty, students, staff
  – Elevator speeches/consistent message: make opportunities: start building those relationships into nodal connectors
  – What kinds of data & what should be preserved &/or saved?
  – Records schedule.

• Populate DMPTool (multiple options)
• Parse out records management/special collections
• Promote/assist with dataverse (HBS models) & other tools & find tools to fit appropriate sizes
• Why?: Provost policies; visibility & preservation (marketing opportunity)
Next Steps: Audience Discussion Summary

- Stewardship Committee will work on integrating the Schools’ data management programs into a “Harvard Blueprint” for data (next meeting week of Symposium):
- DMPTool will be populated, promoted & used (Susan Gomes/Kristen Bolt/Stewardship Committee)
- Data management will be part of conversations at all levels
  - liaison discussions with faculty & students
  - Undergraduate RTL
- Data management will be inserted into RTL classes/events. Emphasis depends on class but should be part of all teaching opportunities
- Critical need for an informed “Data Questions Answered” team for triage & referral