Research Data Management workshop @Harvard Libraries

PURDUE UNIVERSITY LIBRARIES

HARVARD LIBRARY

Purdue Libraries

Harvard Libraries

June 18, 2015

RDM Workshop

Welcome

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.

PURDUE UNIVERSITY LIBRARIES

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

Overview of Data Management

Taking a Look at the Landscape

- Paul Bracke: Reflections from past day

Research, data, curation lifecycles

Overview

- 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

PURDUE UNIVERSITY

- **Line Pouchard, PhD**
- **Scott Brandt**
- Purdue Libraries, Research Data

- **06/18/2015**

HARVARD-PURDUE DATA MANAGEMENT WORKSHOP

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)

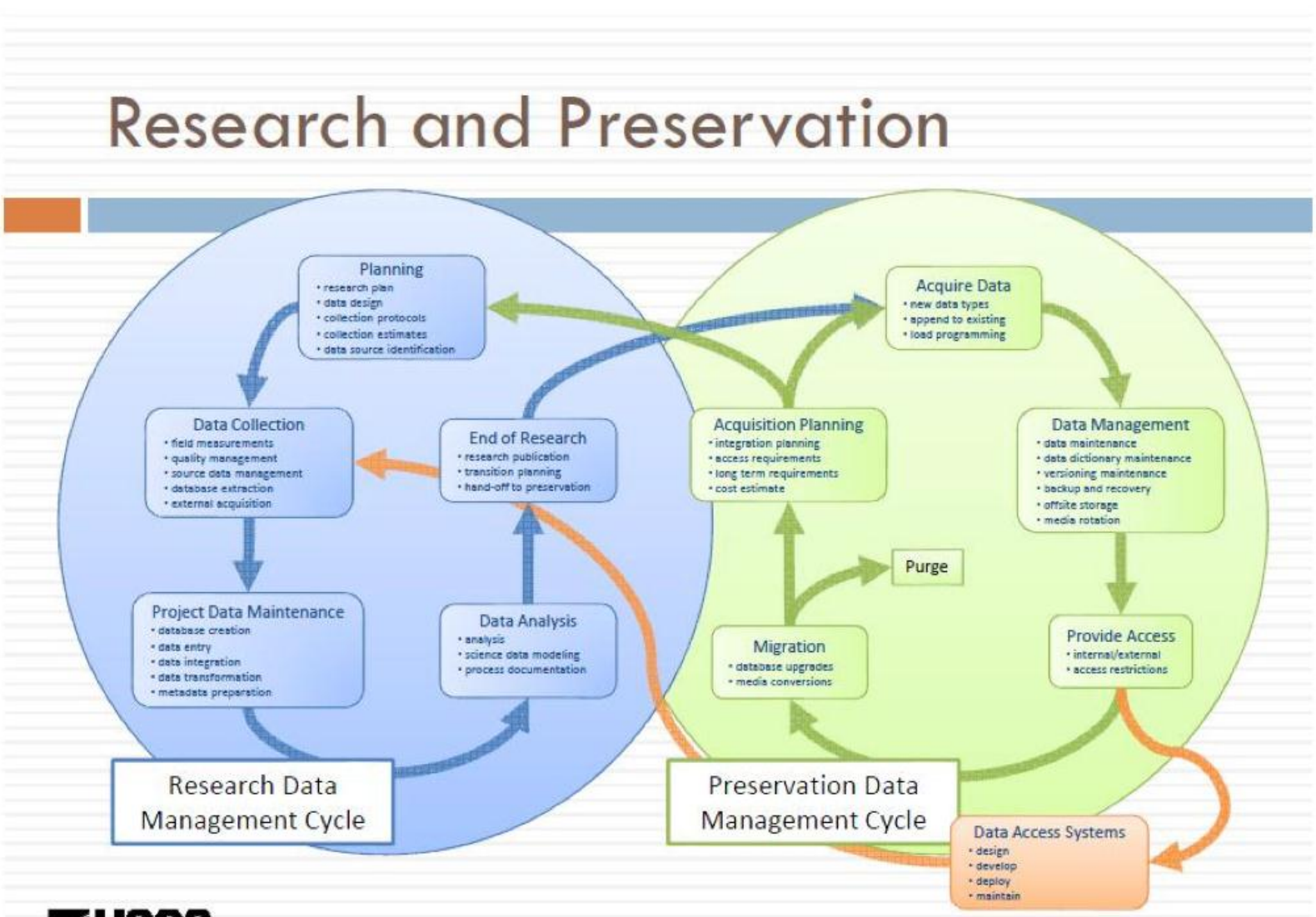
- The active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for reuse over time, and this new field includes authentication,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

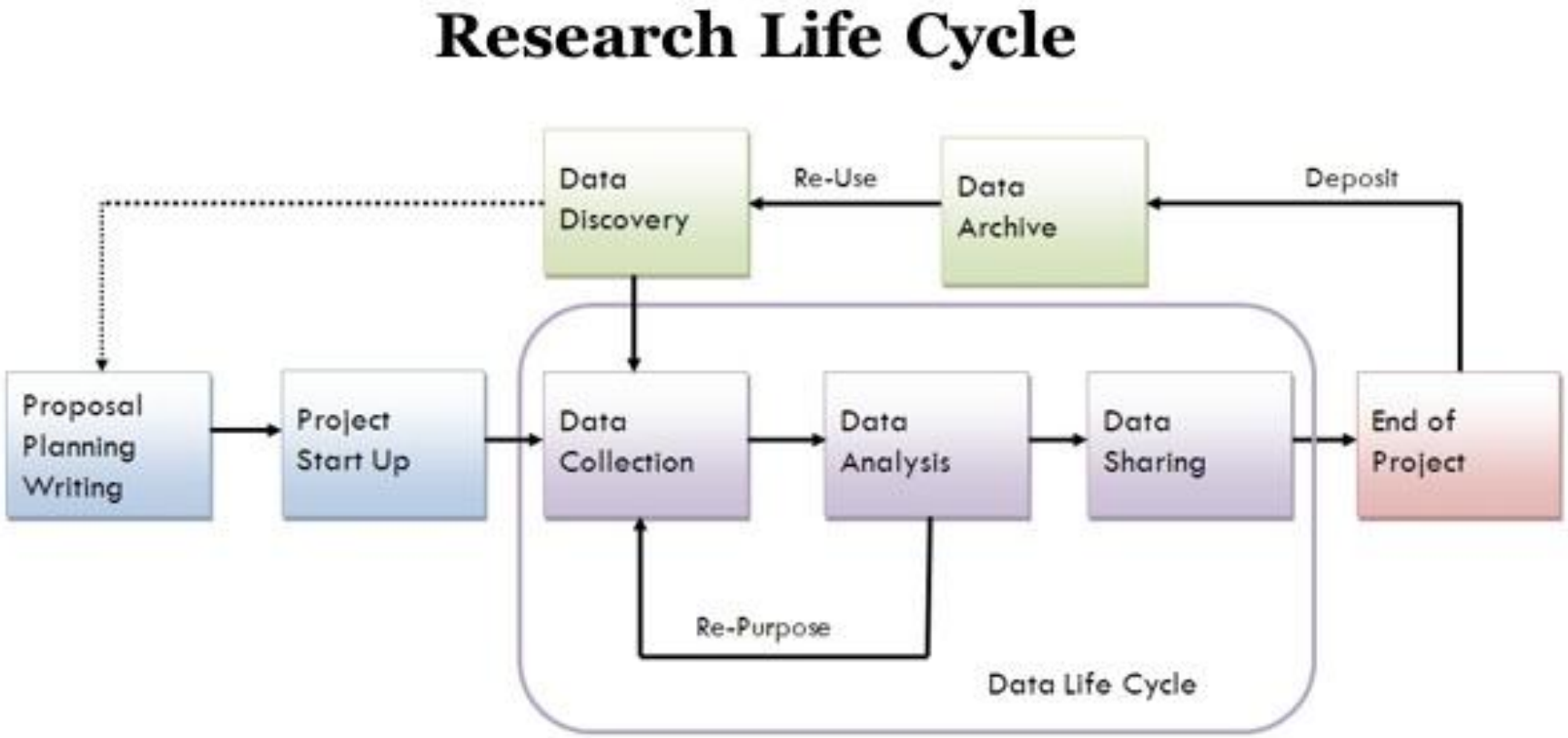
DATA LIFE CYCLES

- 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



U OF VIRGINIA LIBRARY



- 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

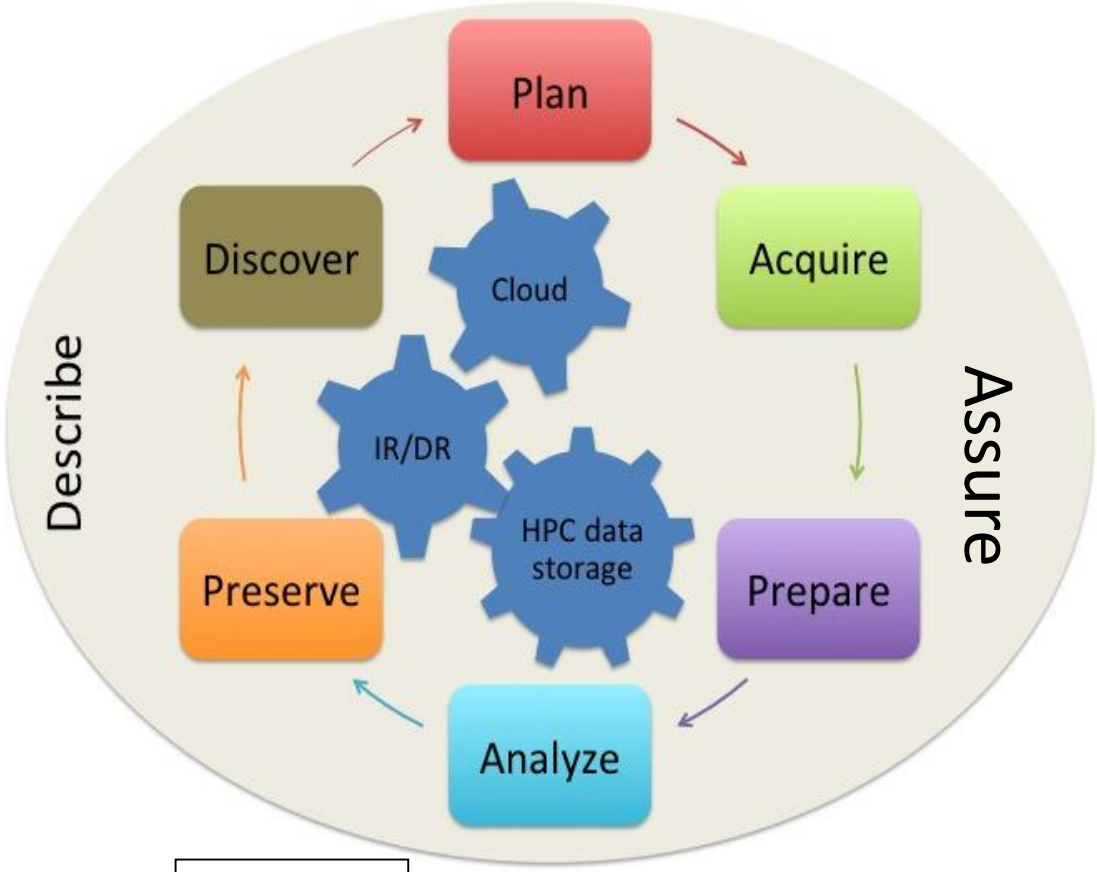
SOURCES

- Data life cycle models and concepts, CEOS V. 1.2

http://ceos.org/ourwork/workinggroups/wgiss/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.
- Digital Curation Center – Digital Curation by Ross Harvey, London and New York, Neal-Schumann Publishers, 2010. http://www.dcc.ac.uk/resources/curation-lifecycle-mode

DATA LIFECYCLE ACTIVITIES



Description of Data

Data Formats

Metadata for Organization

Standards for Interoperability

Documentation of Methodology

Intellectual Property Rights

Software

Description for Discovery

Sharing & Access Policies

Attribution/ citation

Provenance for Preservation

QUESTIONS INFORMING CURATION ACTIVITIES

	Plan	Acquire	Prepare
Volume	What is an estimate of volume & growth rate?	What is the most suited storage (databases, NoSQL, cloud)?	How do we prepare datasets for analysis? (remove blanks, duplicates, splitting columns, adding/removing headers)?
Variety	Are the data sensitive? What provisions are made to accommodate sensitive data?	What are the data formats and steps needed to integrate them?	What transformations are needed to aggregate data? Do we need to create a pipeline?
Velocity	Is bandwidth sufficient to accommodate input rates?	Will datasets be aggregated into series? Will metadata apply to individual datasets or to series?	What type of naming format is needed to keep track of incoming and derived datasets?
Veracity	What are the data sources? What allows us to trust them?	Who collects the data? Do they have the tools and skills to ensure continuity?	Are the wrangling steps sufficiently documented to foster trust in the analysis?

QUESTIONS INFORMING CURATION ACTIVITIES

	Analyse	Preserve	Discover
Volume	Are adequate compute power and analysis methods available?	Should raw data be preserved? What storage space is needed in the long-term?	What part of the data (derived, raw, software code) will be made accessible to searches?
Variety	Are the various analytical methods compatible with the different datasets?	Are there different legal considerations for each data source? Are there conflicts with privacy and confidentiality?	What search methods best suit this data – keyword-based, geo-spatial searches, metadata-based, semantic searches?
Velocity	At what time point does the analytical feedback need to inform decisions?	When does data become obsolete?	What degree of search latency is tolerable?
Veracity	What kind of access to scripts, software, and procedures is needed to ensure transparency and reproducibility?	What are the trade-offs if only derived products and no raw data are preserved?	Providing well-documented data in open access allows scrutiny. How is veracity supported with sensitive and private data?

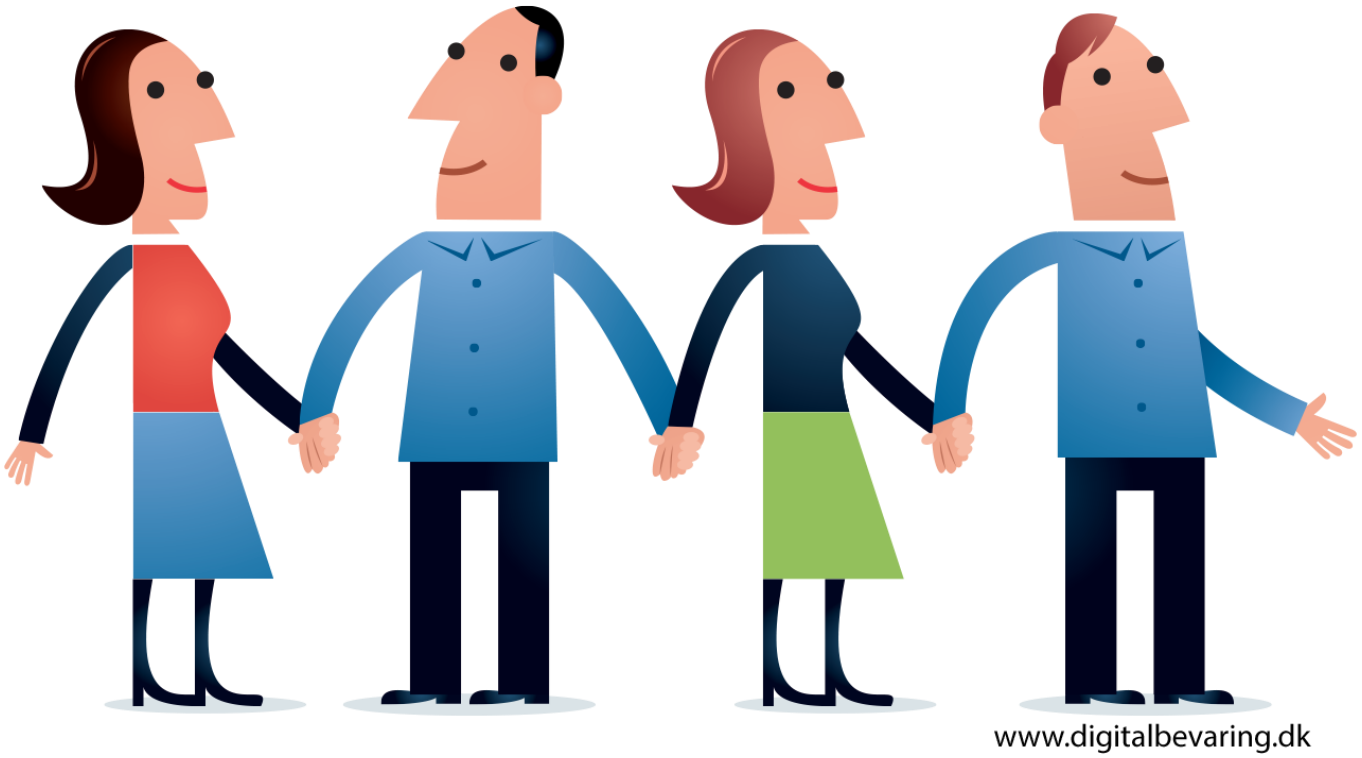
COLLABORATIONS IN PROJECT

- 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, curation lifecycles

Model for Collaboration

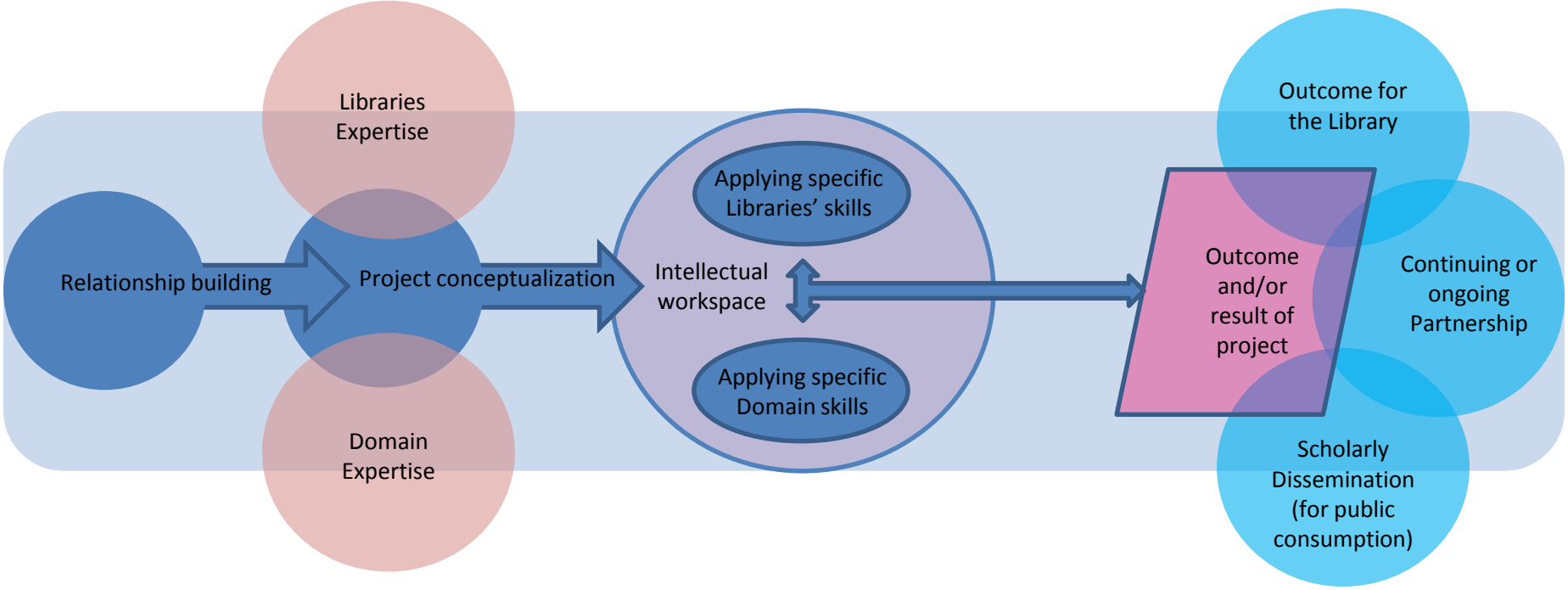
- Amy Barton: A model for collaboration



Research, data, curation lifecycles

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

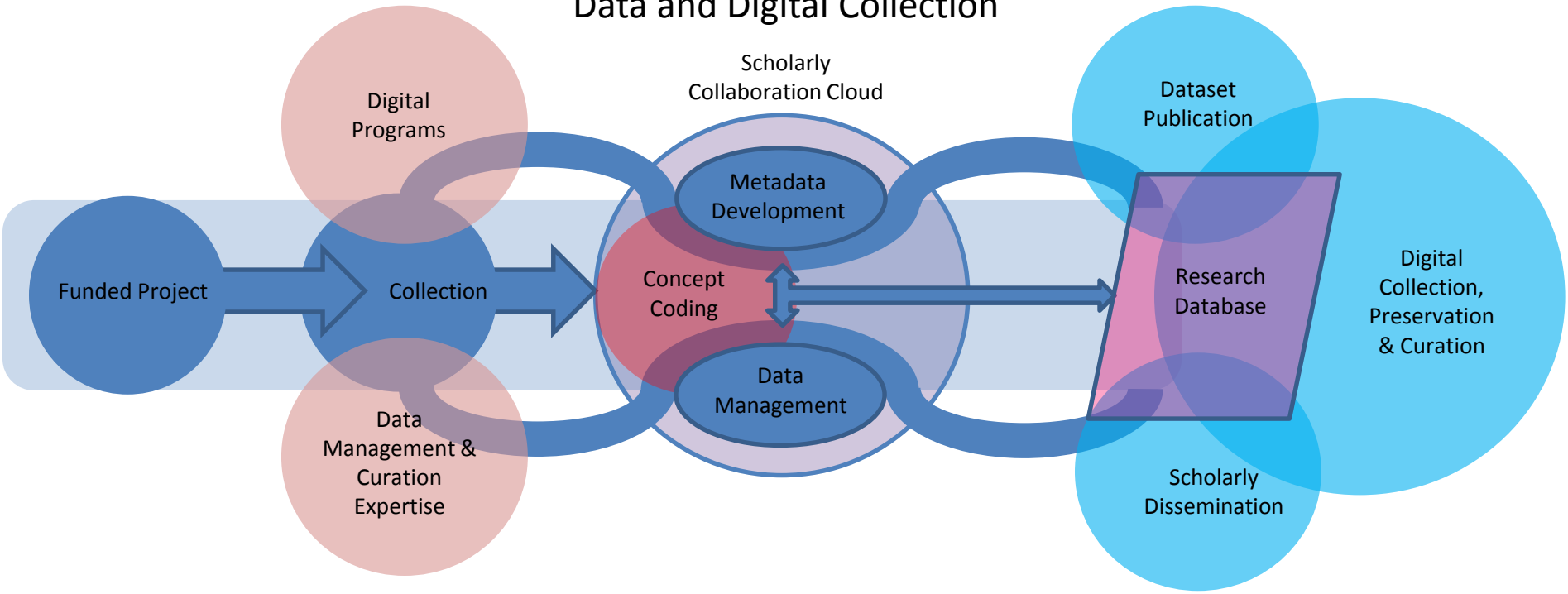
Components:

- Conceptualization
- Data collection
- Data processing
- Collaboration
- Data product
- Data curation

Research, data, curation lifecycles

Model for Collaboration: Amnesty International

A Draft Conceptual Model for Libraries Expertise/Services Conjoining with Domain Expertise to Apply Active Research to Produce Research Data and Digital Collection



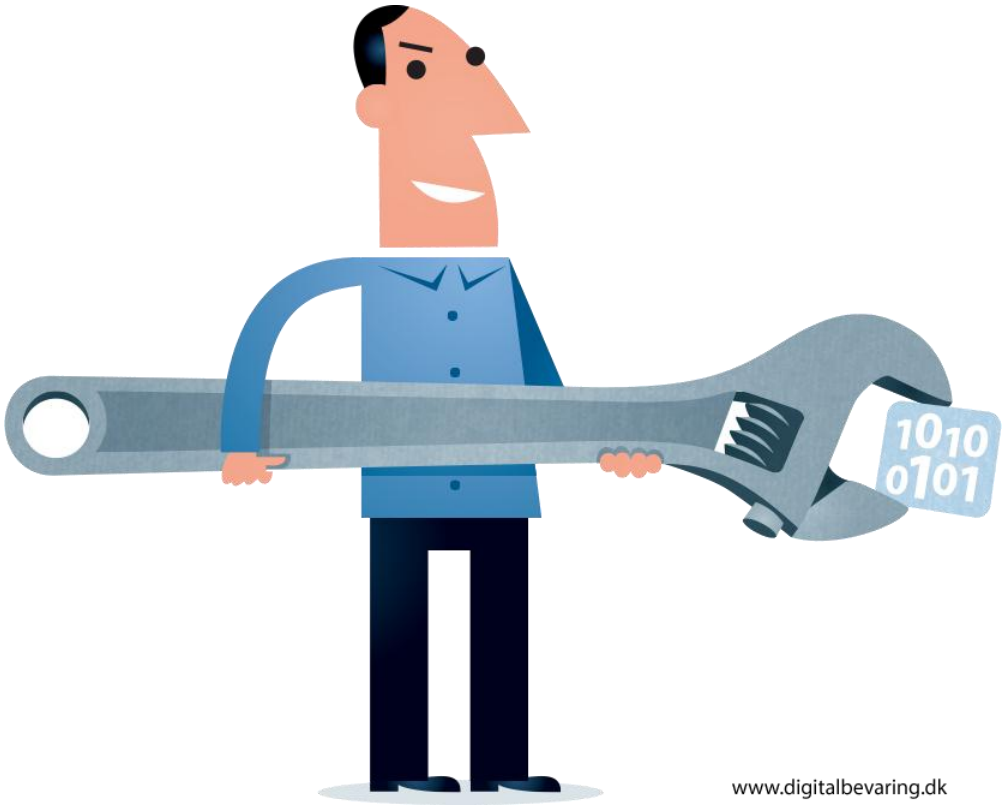
Domain Expertise:

- Political Science
- Library Science
- Metadata
- Research Data

Libraries Services:

- Archives
- Digital Programs
- Libraries IT

- Chris Erdmann: Rewarding the Tool Builders



Scott Brandt: relationships and messages

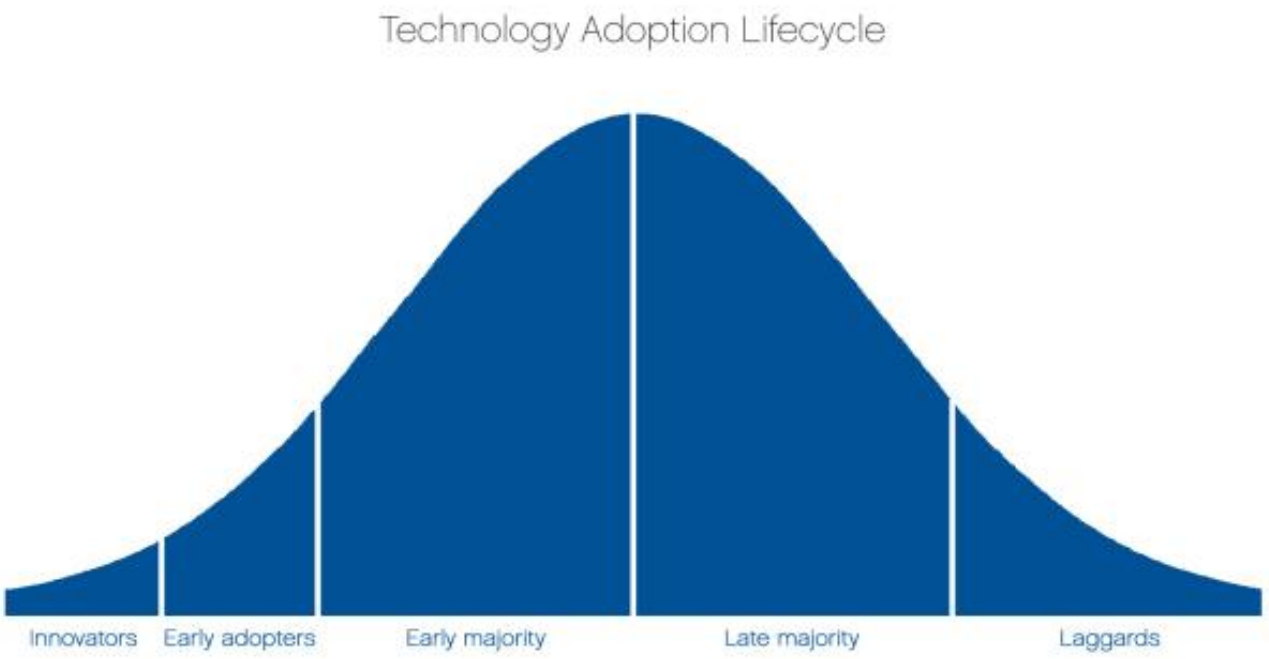
Some extrapolation…

- 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

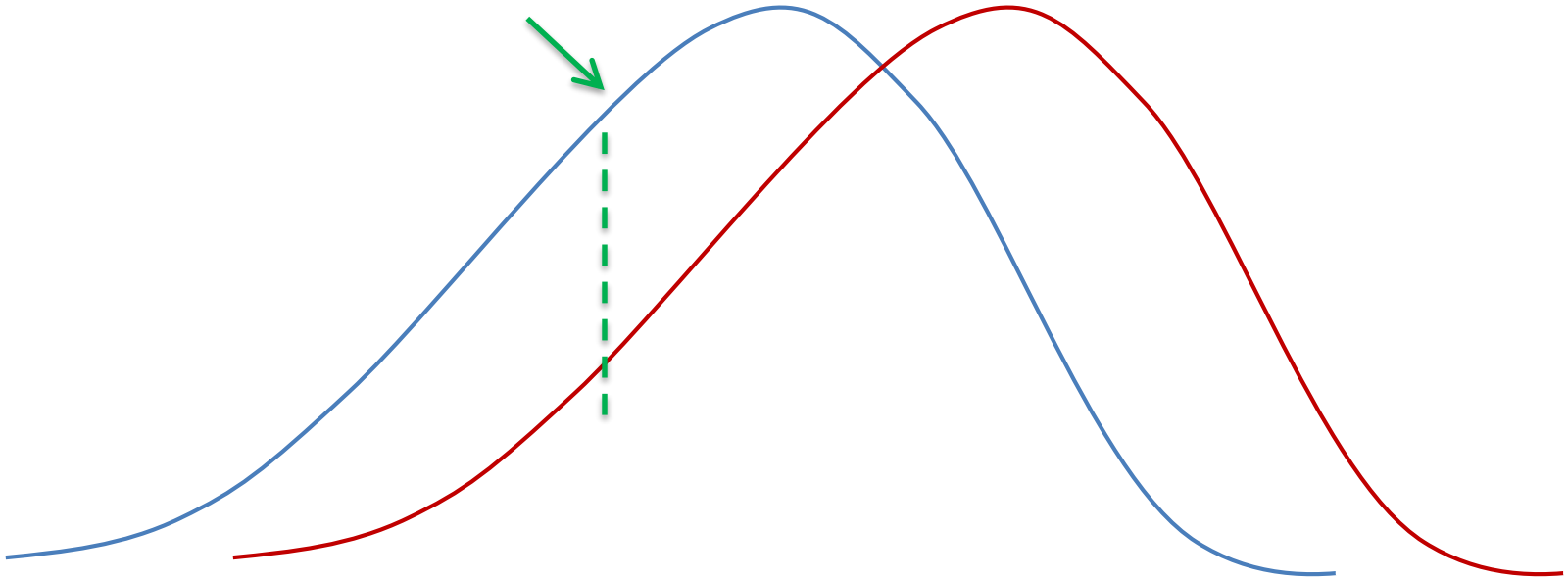
Adoption of Technology



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…

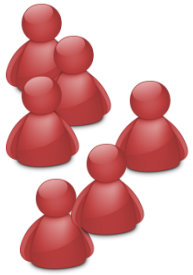


Research, data, curation lifecycles

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

PURR

Digitization Center

Archives & Spec Collections

Publishing
University Press
e-Pubs IR

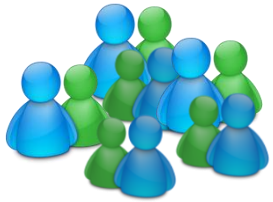
Rights Management
University Copyright Office

OVPR
Research Admin
Sponsored Programs
RCR & IRB

Provost
IM:PACT

ITAP
Central IT
Ctr for Excellence
Rosen Center for Advanced Computing

Disciplinary Faculty
Graduate Students
Undergraduate Students



Research, data, curation lifecycles

Building Relationships

everydayadvocacy
LEARN, SHARE, MAKE A DIFFERENCE

ALA Co

AN INITIATIVE OF THE ASSN. FOR LIBRARY SERVICE TO CHILDREN, A DIVISION OF THE AMERICAN LIBRARY ASSN.

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 you 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

Research, data, curation lifecycles

Elevator Pitches: also Takes Time…

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…

Research, data, curation lifecycles

Building Relationships: Data interview

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

PURDUE UNIVERSITY Libraries *Access. Knowledge. Success.*

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 poster, 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 reformatting them into agnostic formats or ones 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***. Funding 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 its 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 (records) ***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 conflicts 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.

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 (jrcarlso@purdue.edu)
Data Research Scientist

Purdue University Libraries
Distributed Data Curation Center
http://d2c2.lib.purdue.edu

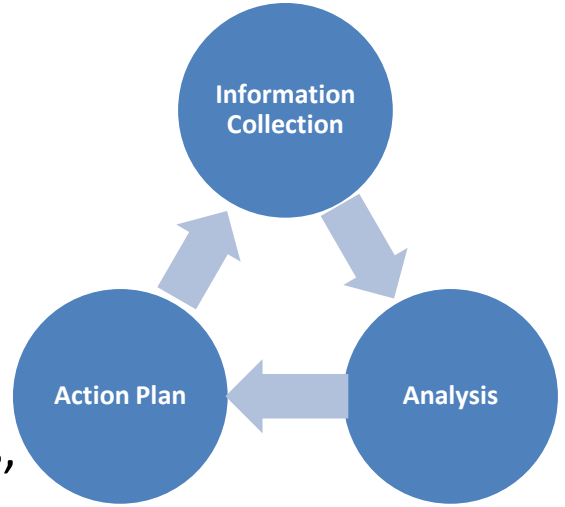
D2C2 Distributed Data Curation Center

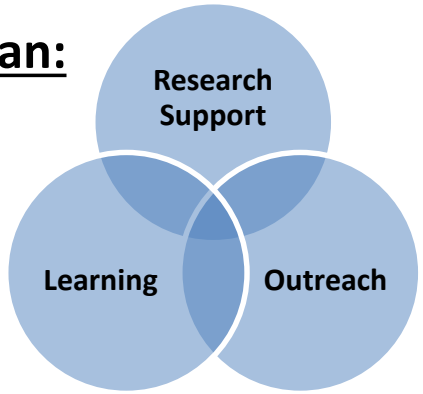
"Conducting a Data Interview"

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

Research Framework

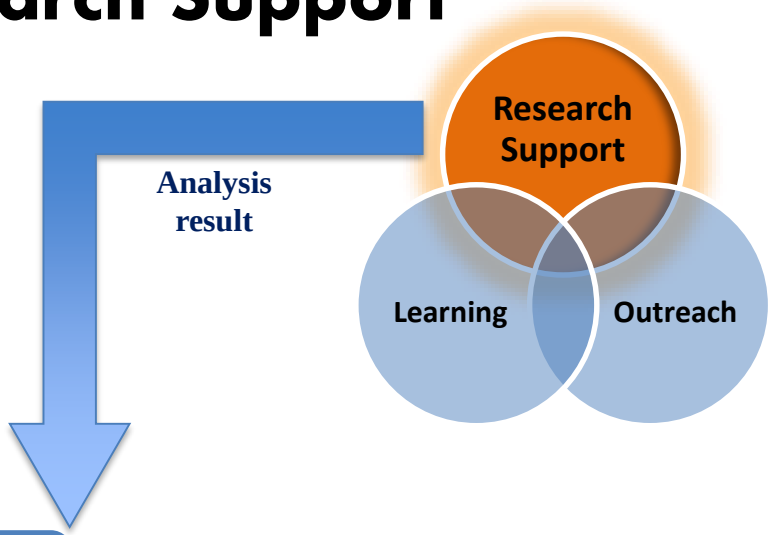
- **Information Collection:**
  Interviews, project participation, consultation, feedbacks, etc.
- **Analysis:**
  Interview transcripts, survey results, email feedbacks, notes -> Nvivo 10
- **Action Plan:**





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):

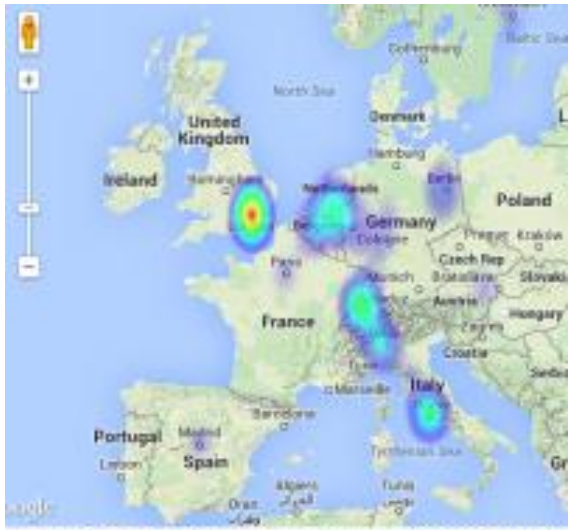
- ❑ Spatial data integration
- ❑ Spatial visualization

  online

  user friendly

  search/query function

  scholar communication



kson Pollock in Western Europe 1958-1962 Ce
ffusion

period between 1958 and 1962 marks the highest point in the visibility of Amer
ressionism and Jackson Pollock in Western Europe. The American artist could
ver Europe, while *New American Painting, Jackson Pollock 1912-1956*, and *A*
ok: *Paintings, Drawings and Watercolours from the Collection of Lee Kram*
led from city to city.

Elevator Pitches: Overview

- 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/

Research, data, curation lifecycles

Elevator Pitches: Wow-How-Now

- 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.

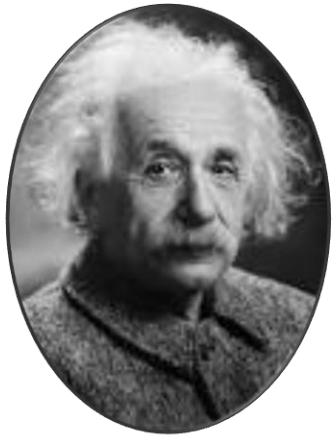
Research, data, curation lifecycles

Elevator Pitches: The New Elevator Pitch

- 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

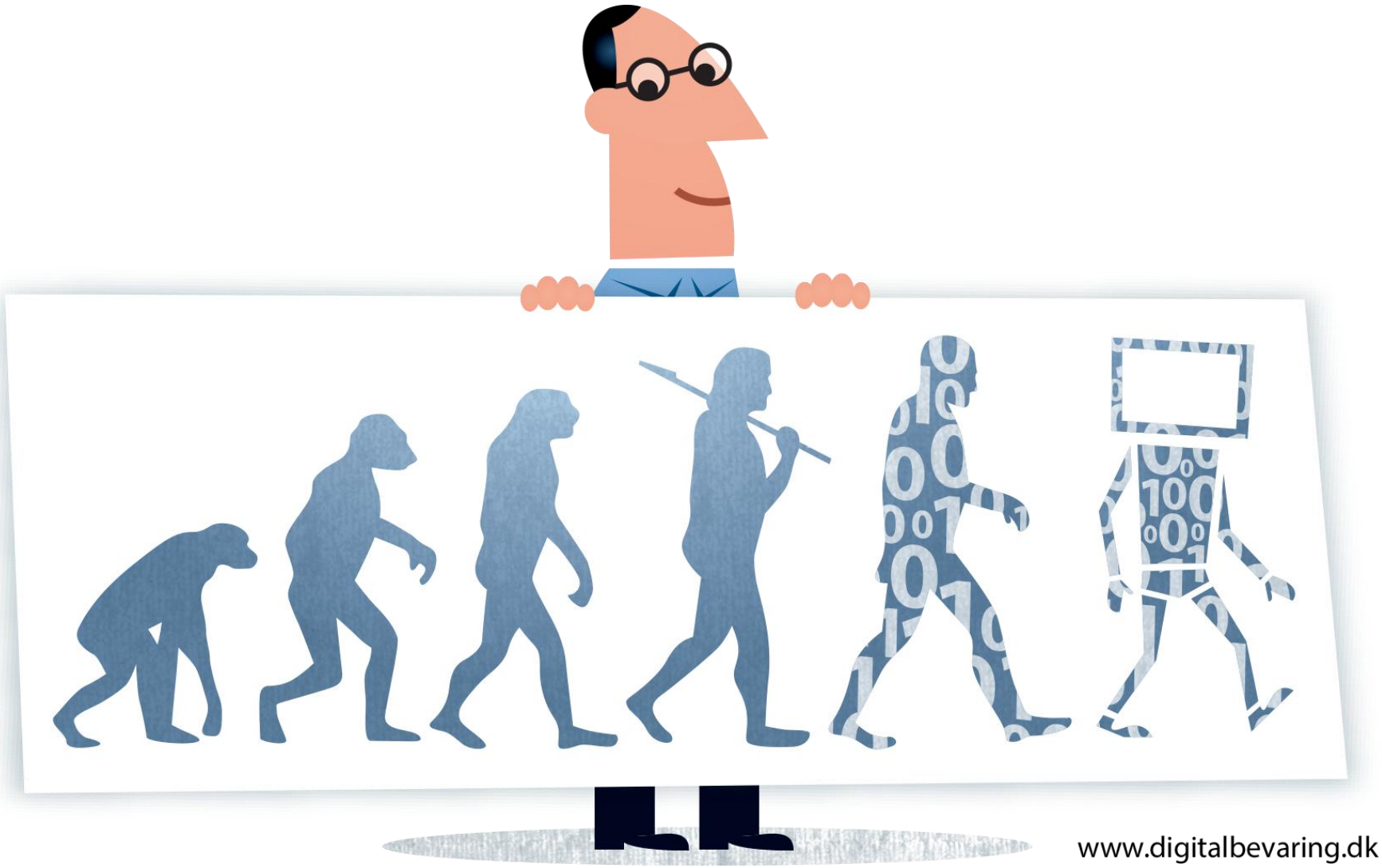
BREAK 10:45 - 11:00 a.m

Institutional Repositories

A Discussion

- Mercè Crosas: Dataverse
- Michael Witt: PURR

LUNCH 12:00 - 1:30 p.m



www.digitalbevaring.dk

Line-Up

- Marianne Stowell Bracke: Overview of DMPs, examples
- Scott Brandt: Supporting DMPs
- Nastasha Johnson: Approaches in Science & Engineering
- Jud Harward: Approaches in the Humanities
- Alex Caracuzzo: Approaches in Business

Data Management Plans

- 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

PURDUE UNIVERSITY LIBRARIES

DMP vs. dmp

DataONE www.dataone.org

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 date, 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 Auchenorrhynca, 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 McKamey 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_mmddyyyy_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 (nutnet.unm.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 Biocomplexity (KNB) data repository.

3. Policies for Access and Sharing
After publication of manuscripts based on the data we collect, we will share our data and metadata with the NutNet community via data updates sent annually as .csv files from the existing central relational database. Other NutNet users will need to contact Lind for access to the data.

We will also submit both of our datasets (abundance and stoichiometry) to the U of M Digital Conservancy, an archive for digital preservation. Borer has access to this resource as a faculty member. This will occur within a year of publication. The data will be publicly available via the Digital Conservancy, which provides a permanent URL for digital documents.

4. Policies for Re-use, Distribution
Access to databases and associated software tools generated under the project will be available for educational, research and non-profit purposes. Such access will be provided using web-based applications, as appropriate.

Materials generated under the project will be disseminated in accordance with University/Participating institutional and NSF policies. Depending on such policies, materials may be transferred to others under the terms of a material transfer agreement.

Those that use the data (as opposed to any resulting manuscripts) should cite it as follows:

> Lind, E, E Borer and A Kay. yyyy. Grassland Arthropod abundance and stoichiometry associated with nutrient manipulation. [URL]; accessed on ddmmyyyy.

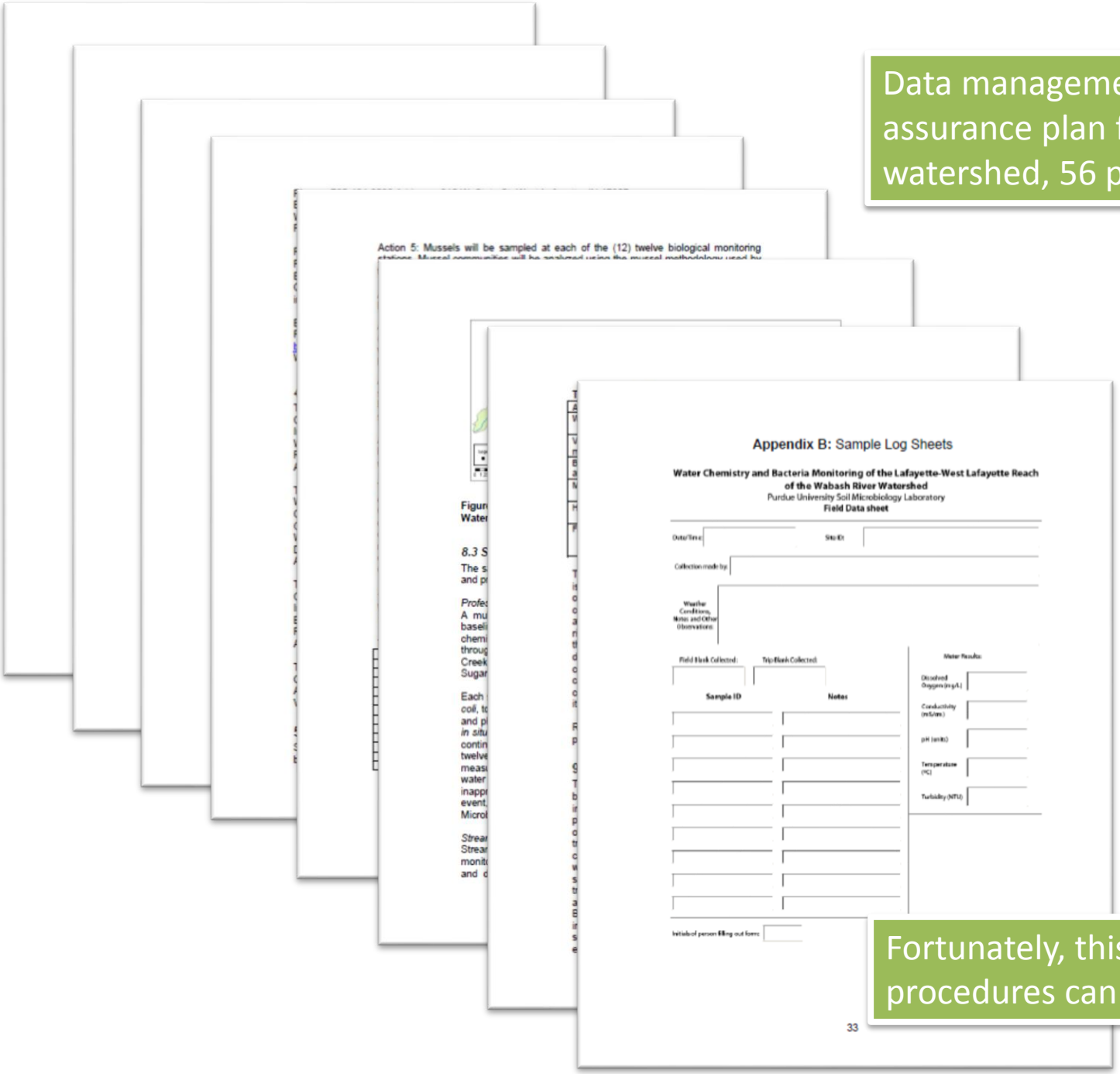
This information will be described in the metadata.

Intended and foreseeable users of the data are NutNet collaborators and participants, as well as other scientists interested in arthropod-plant relationships. This data set could be used in combination with similar data sets from other NutNet sites or for meta-analysis.

5. Plans for Archiving and Preservation
We will preserve both arthropod datasets generated during this project (abundance and stoichiometry) for the long term in the Digital Conservancy at the U of M. We will include the .csv files, along with the associated metadata files. We will also submit an abstract with the datasets that describe their original context and any potentially relevant project information. Borer will be responsible for preparing data for long-term preservation and for updating contact information for investigators.

LIBRARIES

Data management and quality assurance plan for Deer Creek watershed, 56 pages


Action 5: Mussels will be sampled at each of the (12) twelve biological monitoring

Appendix B: Sample Log Sheets

Water Chemistry and Bacteria Monitoring of the Lafayette-West Lafayette Reach of the Wabash River Watershed

Purdue University Soil Microbiology Laboratory

Field Data sheet

Date/Time:

Site ID:

Collection made by:

Weather Conditions, Notes and Other Observations:

Field Blank Collected:

Trip Blank Collected:

Sample ID

Notes

Water Results:

Dissolved Oxygen (mg/L):

Conductivity (mS/m):

pH (units):

Temperature (°C):

Turbidity (NTU):

Initials of person filling out form:

33

Fortunately, this document and the procedures can be re-used

Why do I need a DMP?

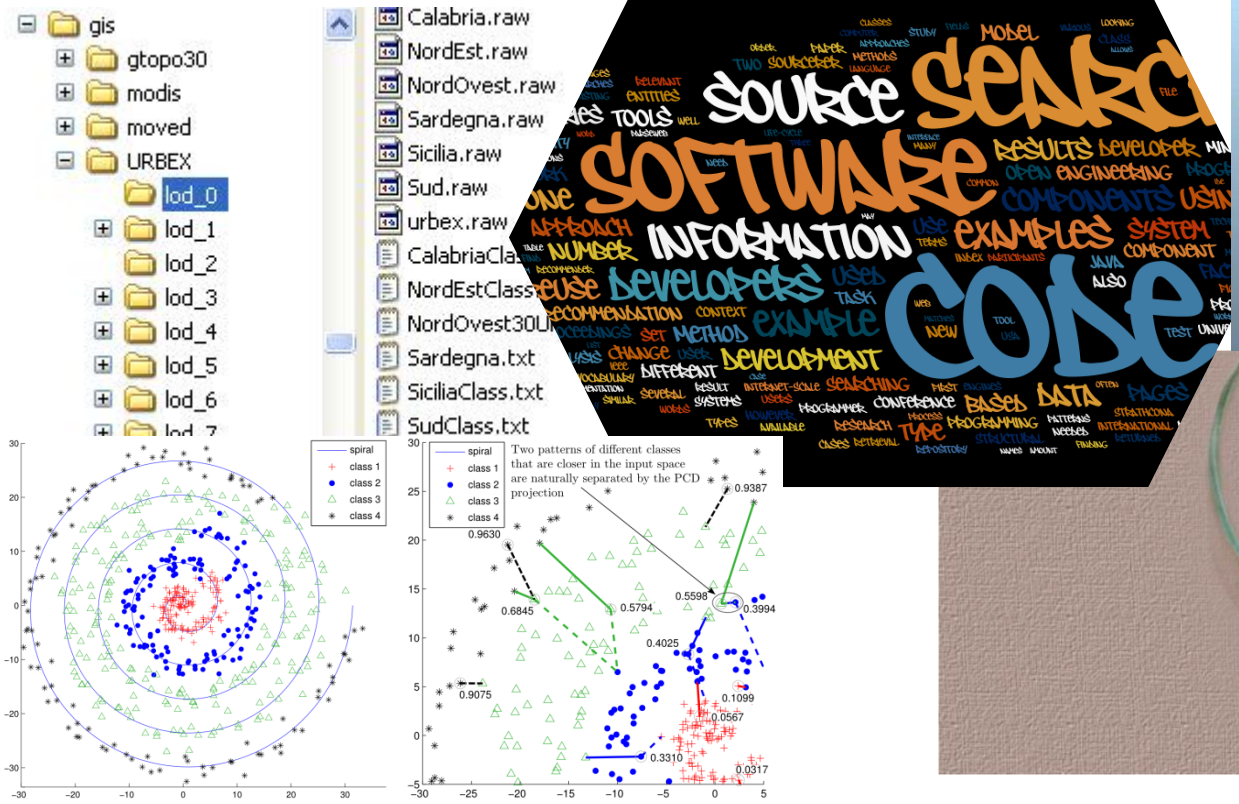
- Funding agency requirements
  - part of their Data Sharing Policy
  - ensure ready availability
  - use of research data beyond life of a project.

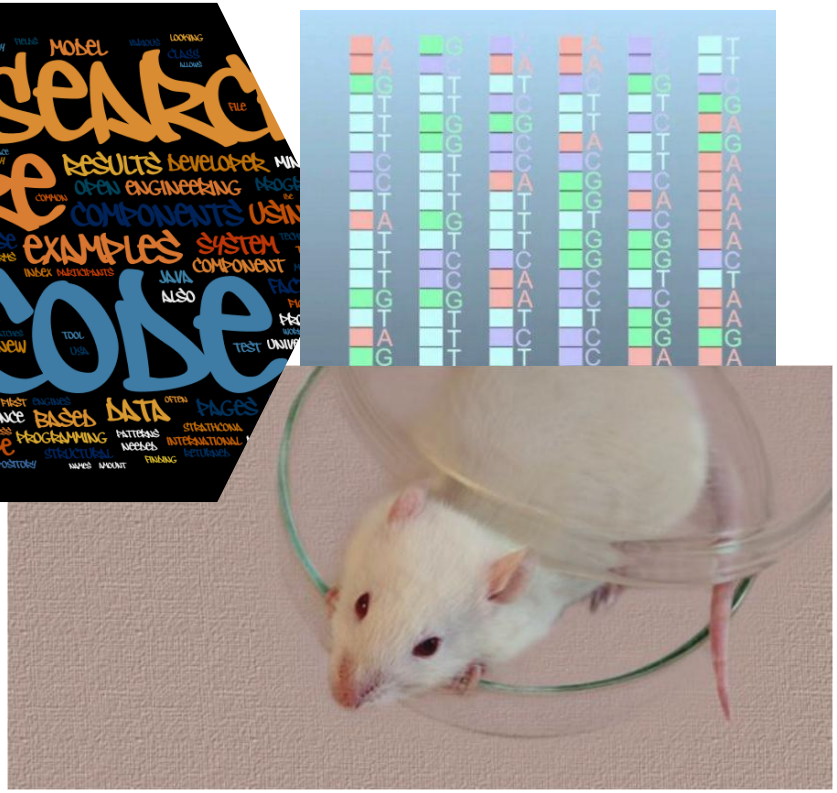
A DMP fulfills part of a central ethical principle and the responsible conduct of research.

PURDUE UNIVERSITY LIBRARIES

Basic Elements of an NSF DMP

- ✓ **Types of data**, samples, physical collections, software, curriculum materials, other materials to be produced



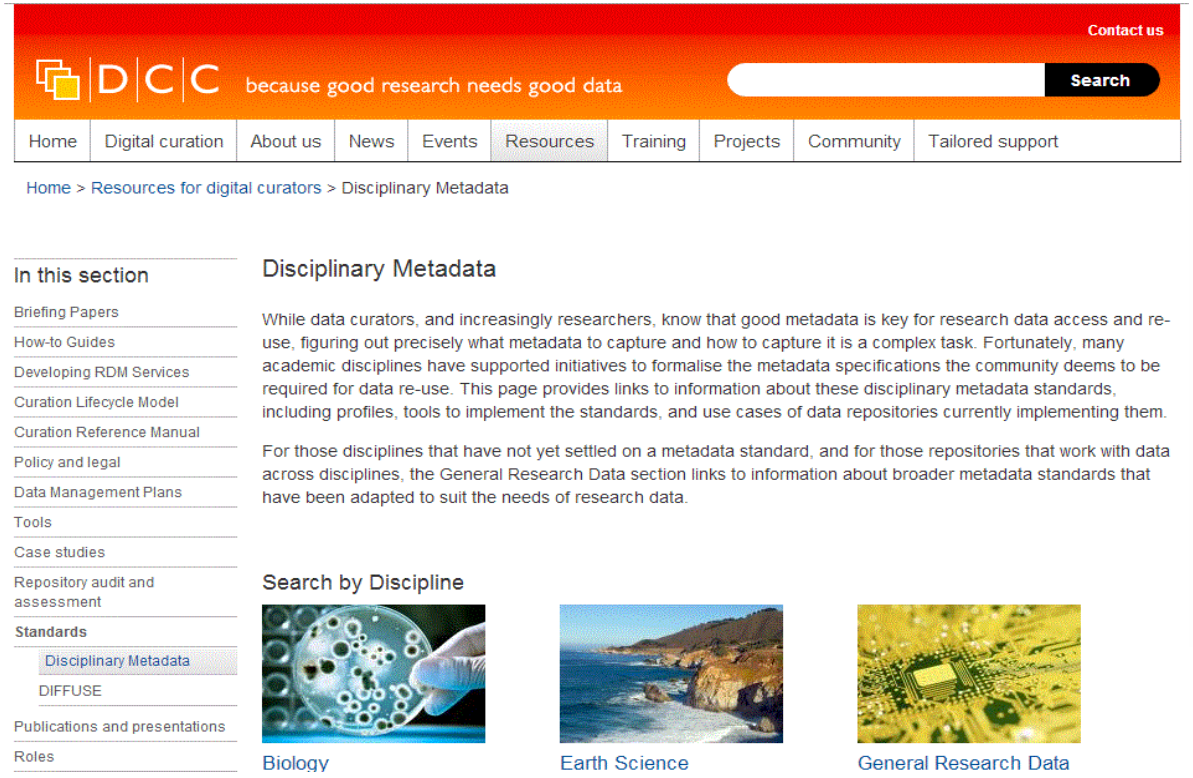


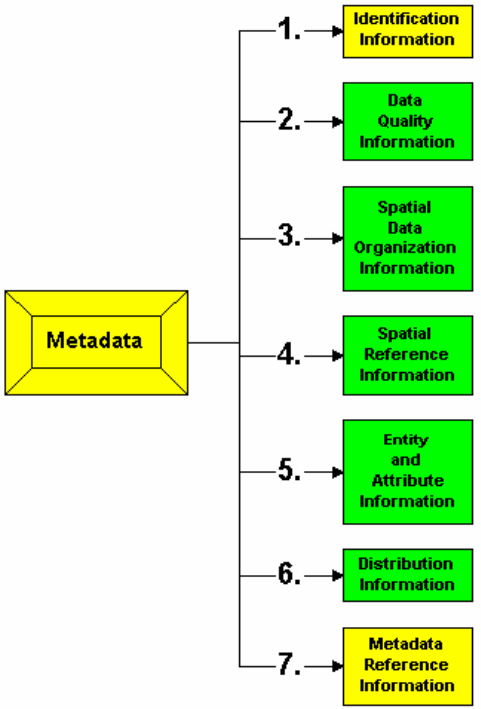
PURDUE UNIVERSITY LIBRARIES

http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp

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);



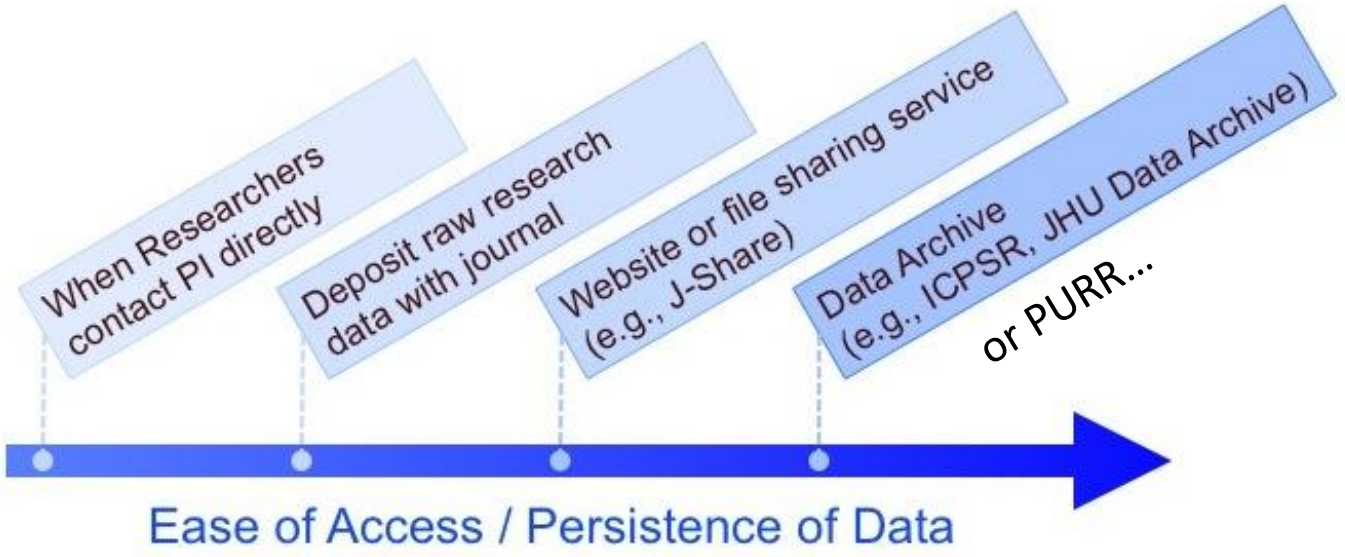


PURDUE UNIVERSITY LIBRARIES

http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp

Basic Elements of an NSF DMP

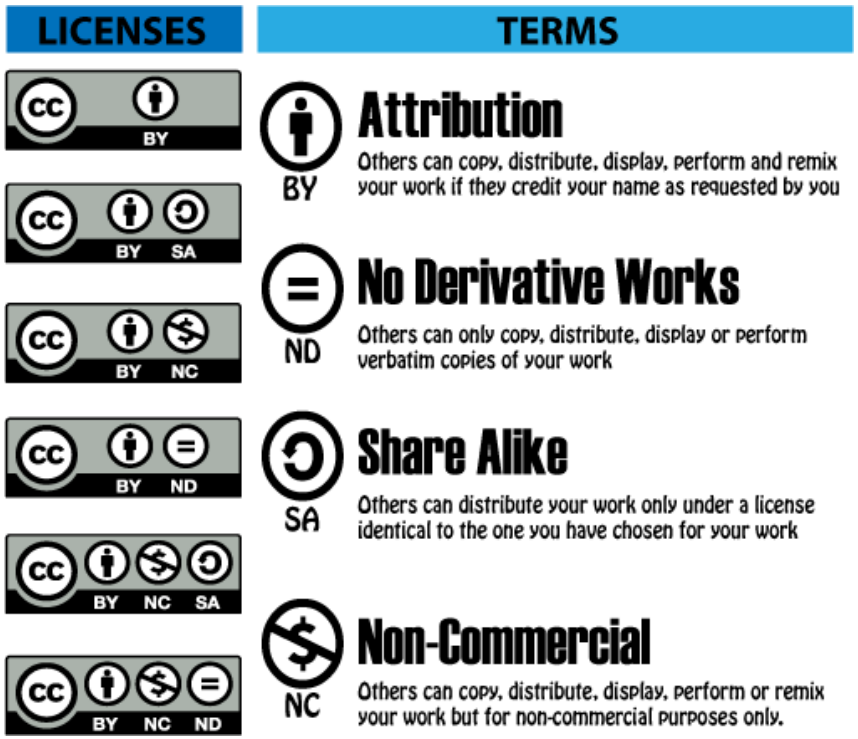
- ✓ **Policies for access and sharing** including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights;



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



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

PURDUE UNIVERSITY | Purdue University Research Repository

ICPSR A PARTNER IN SOCIAL SCIENCE RESEARCH | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

NSIDC National Snow and Ice Data Center

DRYAD

Requirements

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
  - http://nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_index.jsp
- Implementation as a basic requirement of all proposals: §II.C.2.j of the Grant Proposal Guide (GPG).
  - http://nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_index.jsp .
- Some Directorates and Divisions provide additional program specific instructions.
  - http://www.nsf.gov/bfa/dias/policy/dmp.jsp

PURDUE UNIVERSITY LIBRARIES

Other Agency Requirements

- NIH Data Sharing Policy and Implementation Guidance
  http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

- Data Management Plans for NEH Office of Digital Humanities
  http://www.neh.gov/files/grants/data_management_plans_2014.pdf

- EPA Funding Opportunities, G2010-STAR-N1 &N2 (Research & Data Plan: "describe plans to communicate and share data")
  http://www.epa.gov/ncer/rfa/2010/2010_star_nano.html

- 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

PURDUE UNIVERSITY LIBRARIES

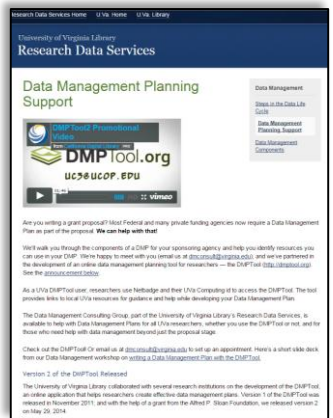
- Scott Brandt: Supporting DMPs

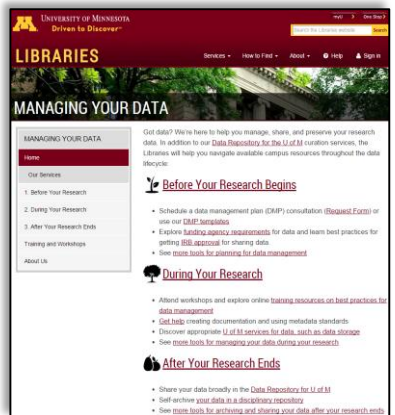
Data Sharing and Management Snafu in 3 Short Acts
By Karen Hanson, Alisa Surkis & Karen Yacobucci
NYU Health Sciences Libraries
August 3, 2012 (Last Update: December 12, 2012)
http://www.youtube.com/watch?v=N2zK3sAtr-4

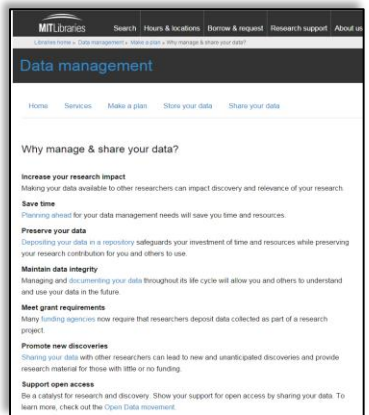
Research Data Mgmt & Curation

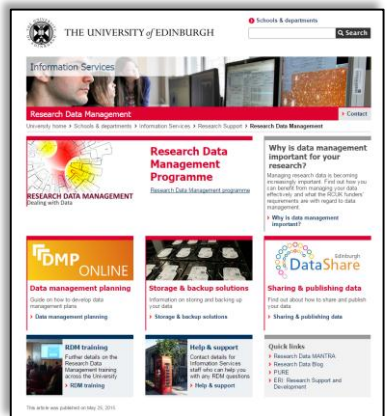
DMP Support
Other Peoples' tips

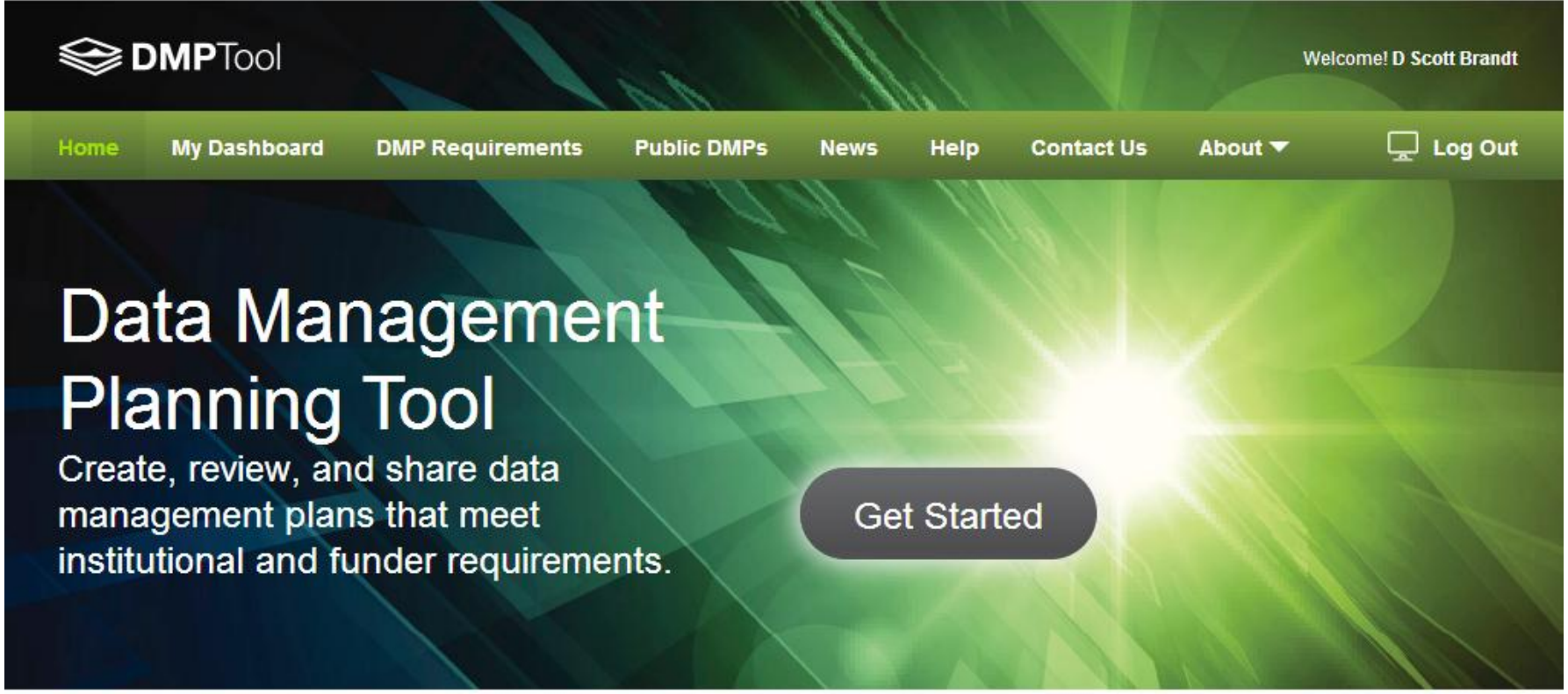
- *Why manage & share your data?*
  - MIT Libraries
- *Managing Your Data*
  - University of Minnesota Libraries
- *Data Management Planning Support*
  - University of Virginia Libraries
- *Research Data Management*
  - University of Edinburgh











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

View All

DMPTOOL NEWS

Latest information about data management and the DMPTool.

- US Dept of Energy data management requi...
- We need API use cases!
- DMPTool downtime this Saturday 7/12/14
- Misc Stats for DMPTool2
- DMPTool workshop at DataONE Users Group...

More News

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

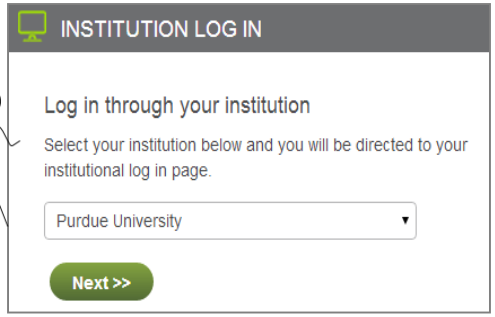
View All

DMPTool

Creating an account

As it implies, DMPTool is a tool you can use to create data mgmt plans

❶

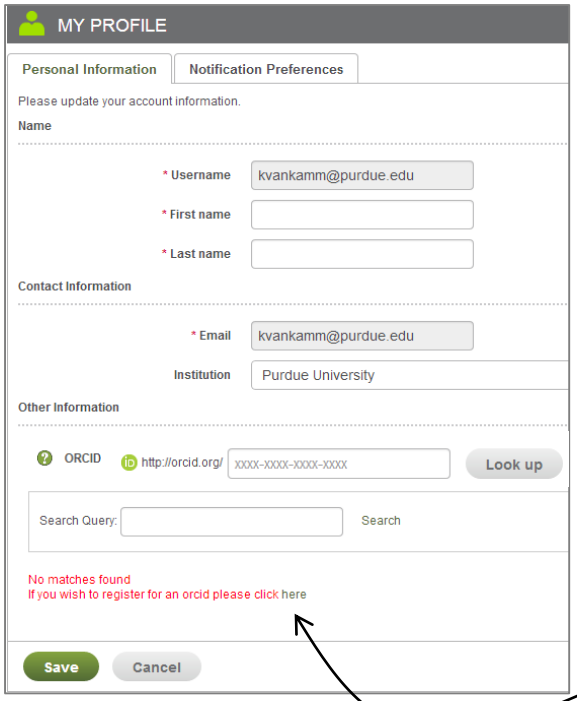


❷

You'll be directed to login via your career account…

❸

…then create your account.



❹

To register at ORCID site click **Look up** then click on the "here" link.

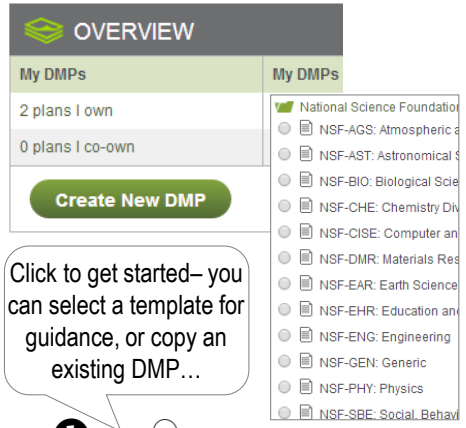
We suggest registering for a unique author ORCID, which helps publishers keep track of your work.

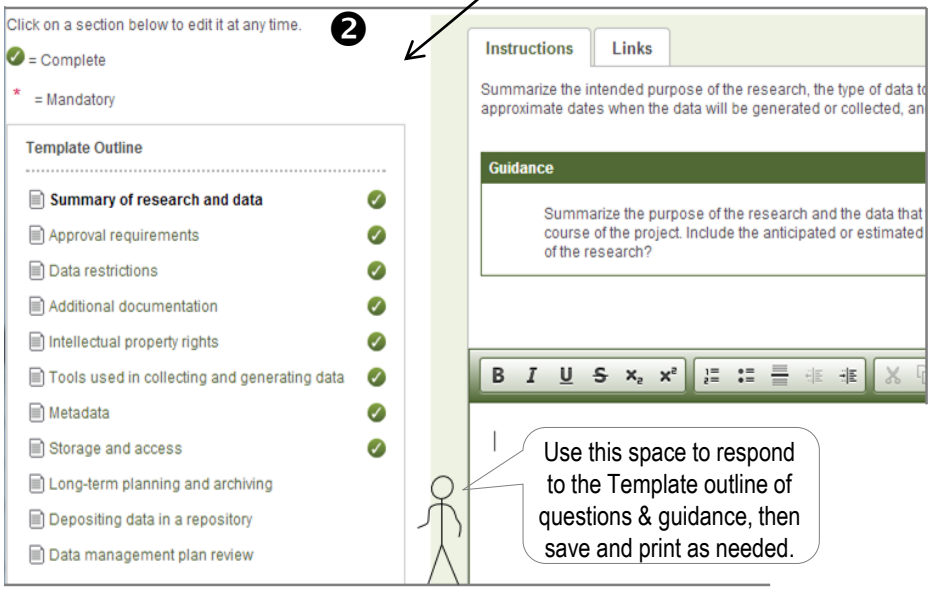
❺

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

DMPTool

Creating a DMP





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 PurdueDCC
- Data sets consisting of dance works generated primarily content deposited by creators INTO Purd

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

Output may need to be reformatted to fit two pages…

3

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.

2 **Expected data**

Expected Data, Formats and Dissemination- *Project Data*:

The following lists the types of data that might be generated during each stage of the

Generated by the DMPTool on August 27, 2014 13:04

Research Data Mgmt & Curation

"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?"

2. Expected data

The data generated by my research uses hospitality business (hotels and restaurants) data from the U.S. Census, hazard data from the Spatial Hazard Events and Losses Database (SHELDUS), socio-economic variables, and community vulnerability index and a hazard resilience index to form a computational database and website. It also functions as a natural hazard data collection repository and houses GIS-enabled community-specific vulnerability data. The website develops a framework for discerning natural disaster impacts on communities and their businesses so that adaptive decision making can occur.

3. Period of data retention

Data will be retained for at least three years beyond the award period, as required by NSF guidelines. In the event that discoveries are made in direct connection with this data, access will be granted upon request once appropriate disclosures are made. Key data relevant to the discovery will be preserved until timely publication and all issues of intellectual property are resolved.

4. Data formats and metadata

The data generated through the work described in this proposal will consist of the following data types: secondary data and sources such as County Business Pattern data, Spatial Hazard Events and Losses Database for the United States (SHELDUS), and various indexes (such as a social capital index); Primary data from focus groups, interviews, survey responses; Emergent data developed as a result of the project includes GIS-generated maps and a computational and collaborative website with GIS-enabled data.

Data Standards associated with the above data will conform to the following: secondary data will use NAICS, SPSS and / or Stata (statistical software) standards. Primary data will use interval scales; interview transcripts; qualitative data analyses (such as the qualitative Nvivo 10 software);Geospatial Data and Geospatial Technologies; Microsoft Word and Excel standard conventions; Network Authorization and Authentication Services, Portal, Schema, SSL, VPN, and Web Publishing, Web services, XML, and XQL standards for Website collaboration and usage.

The data to be acquired in the proposed project will include human subjects' data that require Institutional Review Board approval. As detailed in the human subjects section of the proposal, all rules and regulations related to privacy (i.e., HIPPA) will be observed with specific regard to collected data.

5. Data dissemination and policies for public access, sharing and publication delays

Data generated from the project may be accessible by email request as well as a public, open-access policy on the website. Data will be accessible immediately after timely publication.

6. Data storage and preservation of access

Records of primary research results involving human subjects will be temporarily labeled and stored as digitized recordings and images. The de-identified electronic data will be preserved on external hard drives. Copies of these data will also be preserved offsite at Purdue University's Research Repository (PURR). Completed questionnaires (as well as representative blank questionnaires) and human subject consent forms will be stored in a locked area of the PI's laboratory accessible

DMP 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

PURR

Researcher

PURDUE UNIVERSITY LIBRARIES

- Nastasha Johnson:

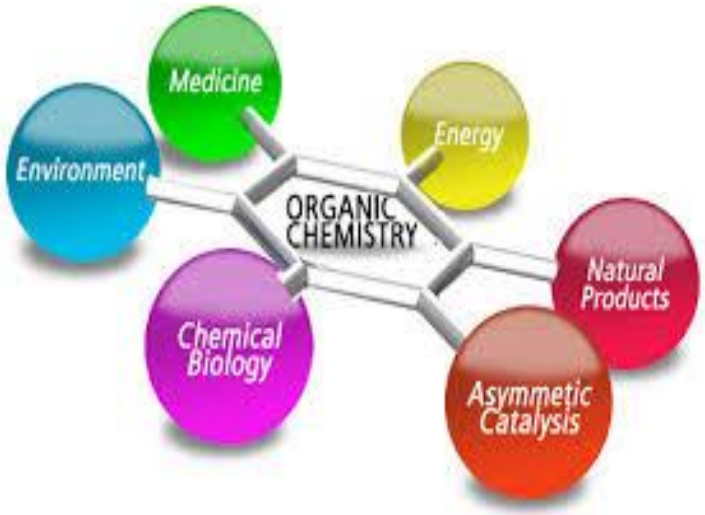
  Approaches in Science & Engineering

Research Data Mgmt & Curation

Sample Example: Chemistry

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

Research Data Mgmt & Curation

Sample Examples

Physics & Astronomy

- **national/international projects**
  - Dedicated data managers
  - Mediating between interdisciplinary and/or trans-disciplinary 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

Research Data Mgmt & Curation

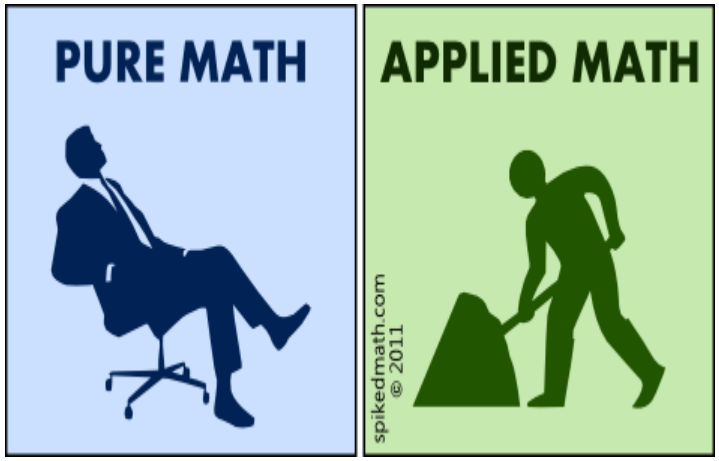
Sample Example: (Biomedical) Engineering

- 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

Image Source: http://whysoblu.com/wp-content/uploads/2011/06/Once-Upon-A-Time-In-The-West-standoff-www.whysoblu.com_.jpg

Research Data Mgmt & Curation

Sample Example: Pure Mathematics



- **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

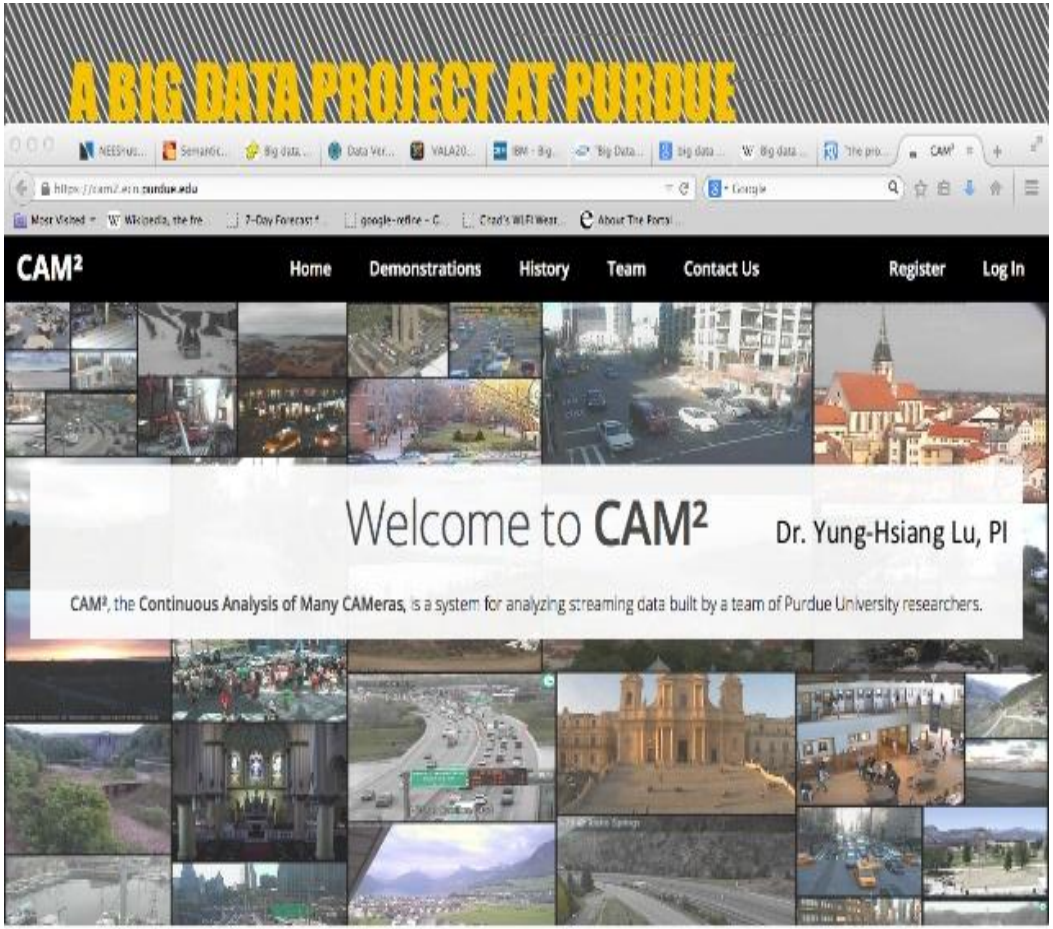
Research Data Mgmt & Curation

Sample Example: Geographic Information Systems

- 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

Research Data Mgmt & Curation

Summary: Interdisciplinary, Transdisciplinary

- 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: http://career-intelligence.com/wp-content/uploads/2014/07/Fast-Co-Negotiation.jpg
Image Source: https://ccph.memberclicks.net/assets/Images/ccph%20partnership%20graphic.png

Jud Harward: Approaches in the Humanities

Alex Caracuzzo and Rachel Wise: Approaches in Business

BREAK 2:30 - 2:45 p.m.

A Dozen Things We Do...

Things We Are Already Doing That Can Translate to Research Data Management

A Dozen Things We Do...

- Marianne Stowell Bracke: Intro & Things we do: reference and disciplinary expertise
- Line Pouchard: consultation & interventions
- Nastasha 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

PURDUE UNIVERSITY LIBRARIES

A Dozen Things We Do…

- Marianne Stowell Bracke: Intro & Things we do: reference and disciplinary expertise

Reference Interview Skills

PURDUE UNIVERSITY LIBRARIES

Data curation profile toolkit

Ask in-depth or follow-up questions:

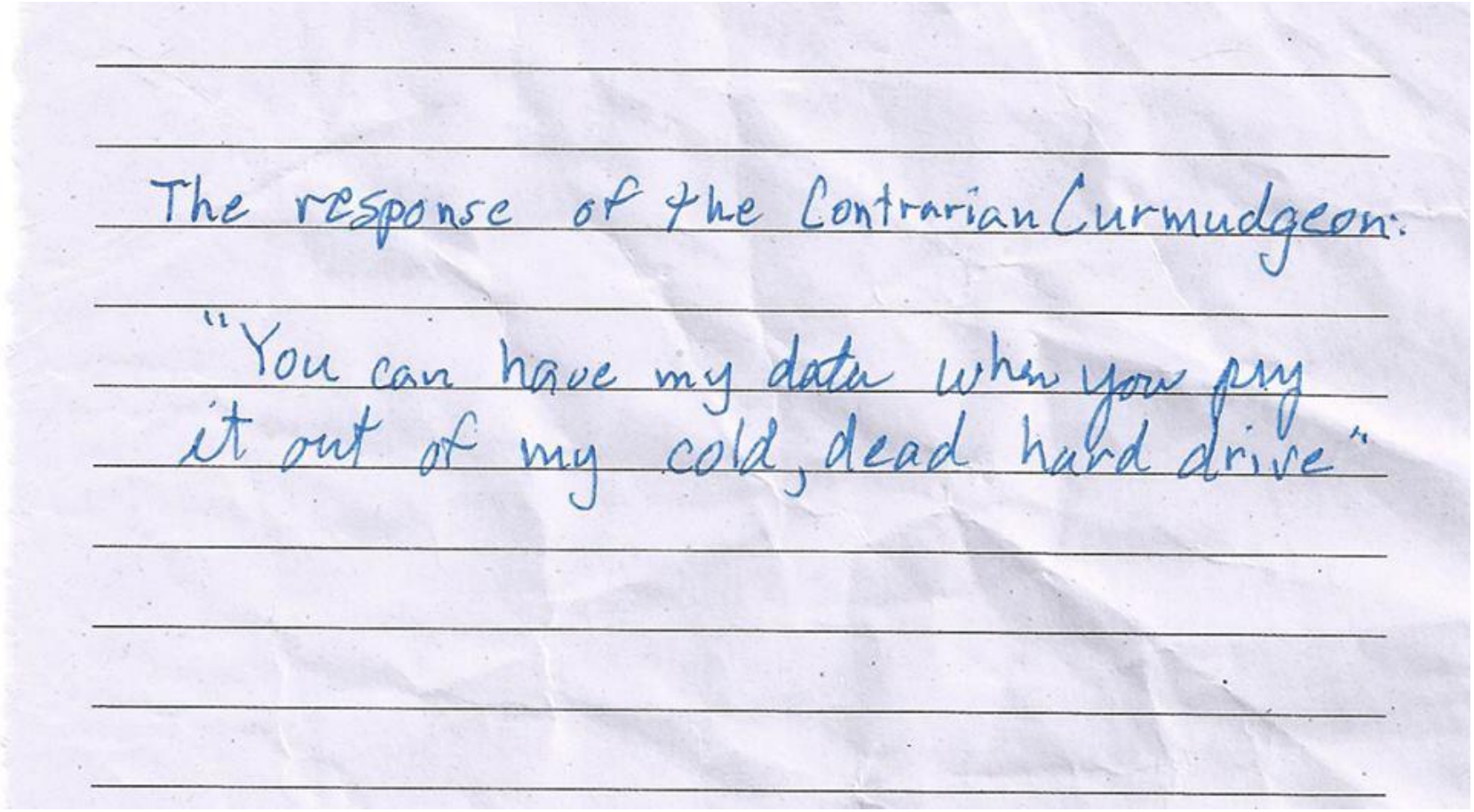
Current plans for storage?

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



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

A Dozen Things We Do…

- Nastasha Johnson: Things we do: instruction

A Dozen Things We Do...

Traditional Publication vs. Data Sets

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

PURDUE UNIVERSITY LIBRARIES

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

PURDUE UNIVERSITY LIBRARIES

A Dozen Things We Do...

Teaching Information Organizing

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

PURDUE UNIVERSITY LIBRARIES

A Dozen Things We Do...

Thesis & Dissertation
Data Deposit

Depositing of Thesis/Dissertations

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

Depositing of Data

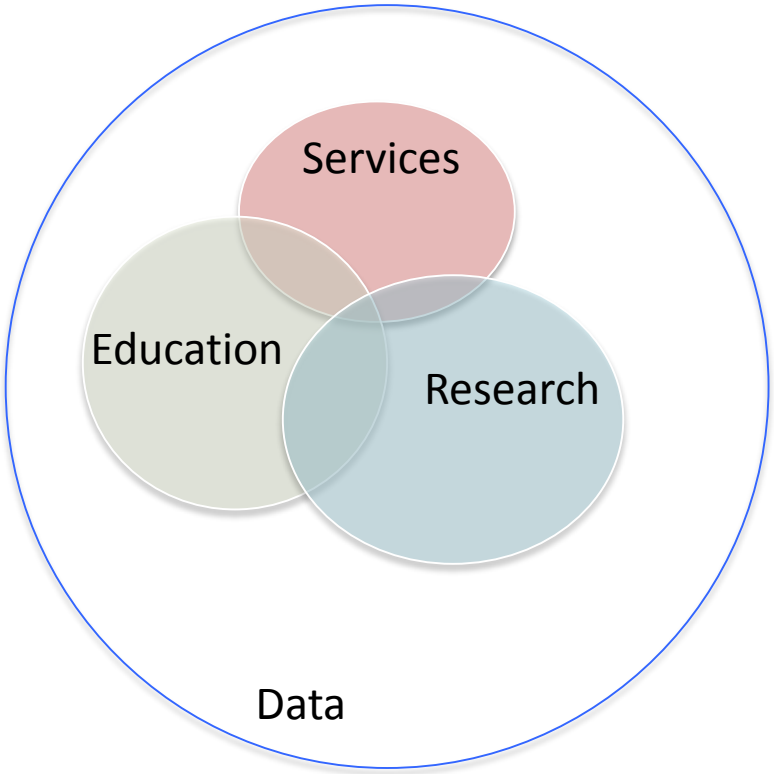
- Invited instructional sessions
- One-on-one consultations
- DMP follow-ups
- Lab consultations
- Knowledge Base

PURDUE UNIVERSITY LIBRARIES

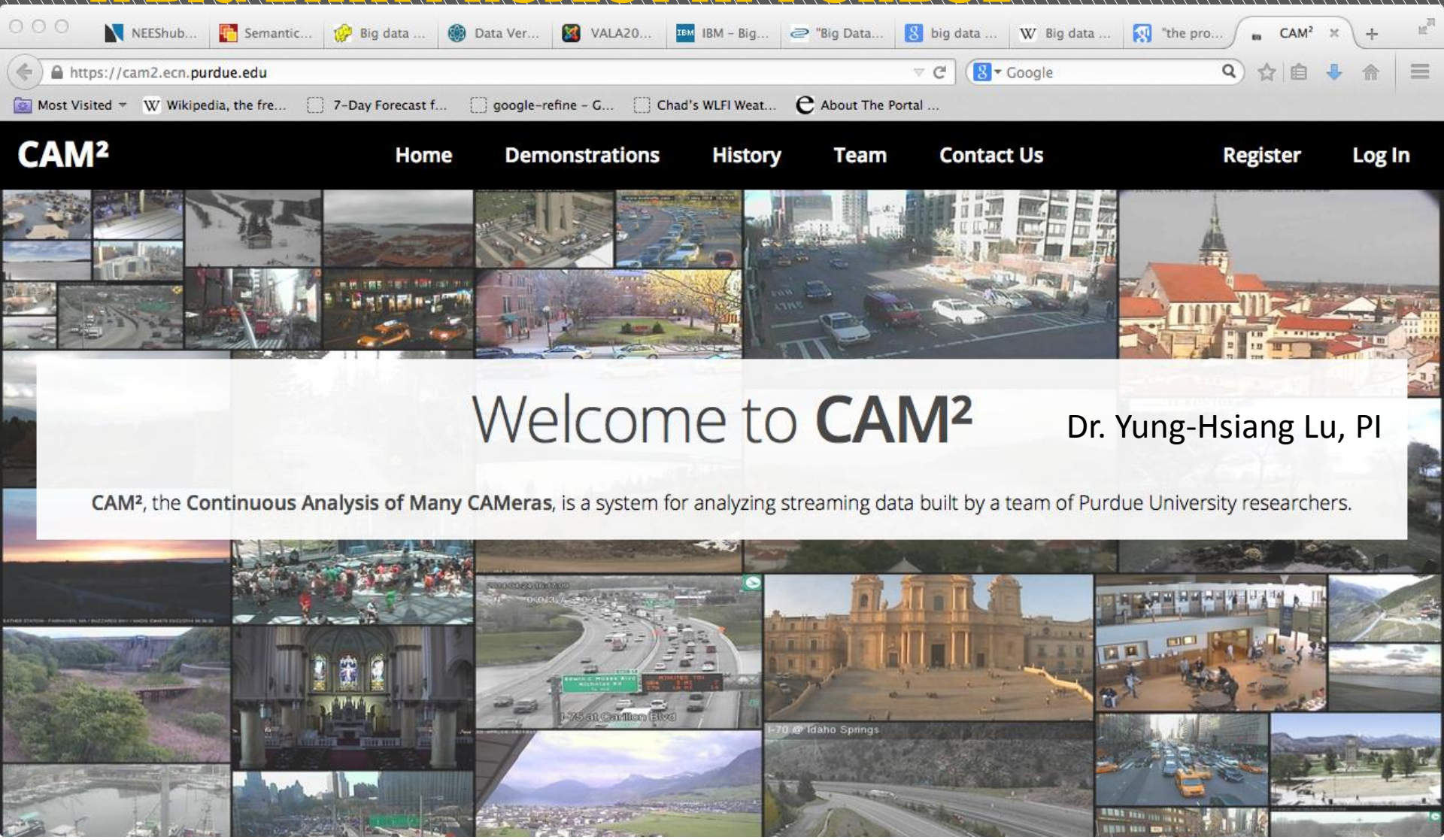
A Dozen Things We Do...

- Line Pouchard: Things We Do: consultation & interventions

Research Data management @ Purdue university Libraries



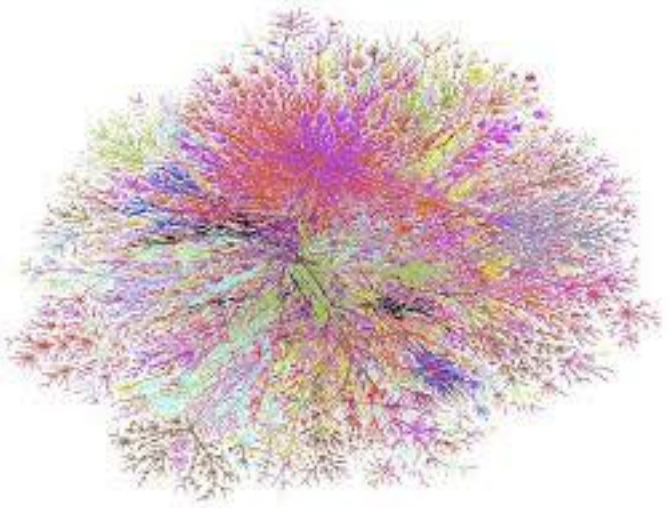
A BIG DATA PROJECT AT PURDUE



Dr. Yung-Hsiang Lu, PI

CURATION ISSUES IN CAM2 PROJECT

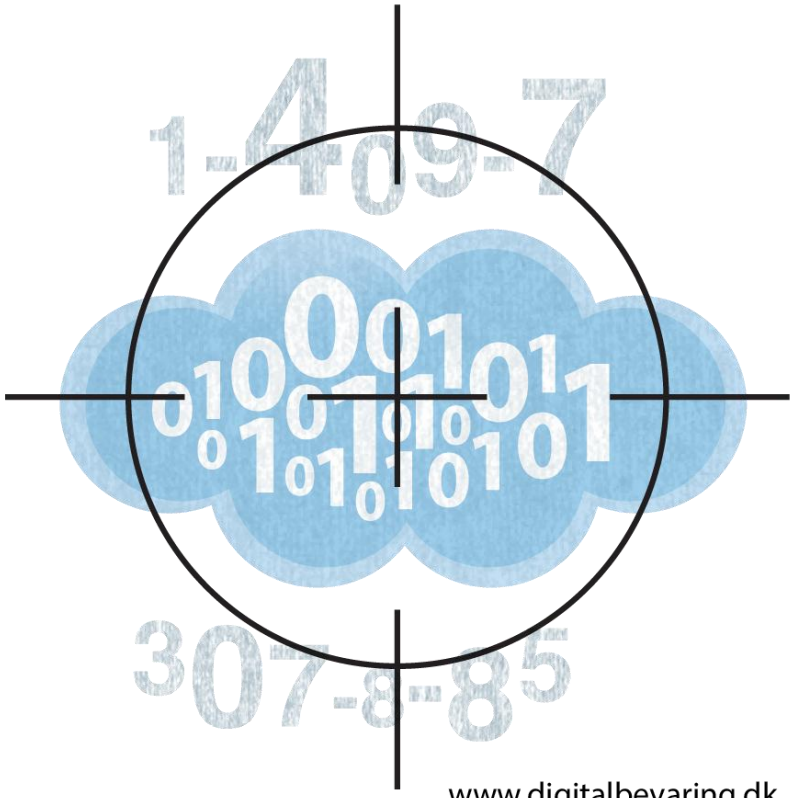
- Data access and re-use
  - policies of video streams and CCTV
  - Sparse or piecemeal legal framework US
  - Policies are mostly ad hoc
- Data ownership
- Data storage
- Data organization
  - naming scheme
  - metadata
- Protect metadata storage – where the intellectual property lies
- Data information literacy skills for Big Data



A Dozen Things We Do...

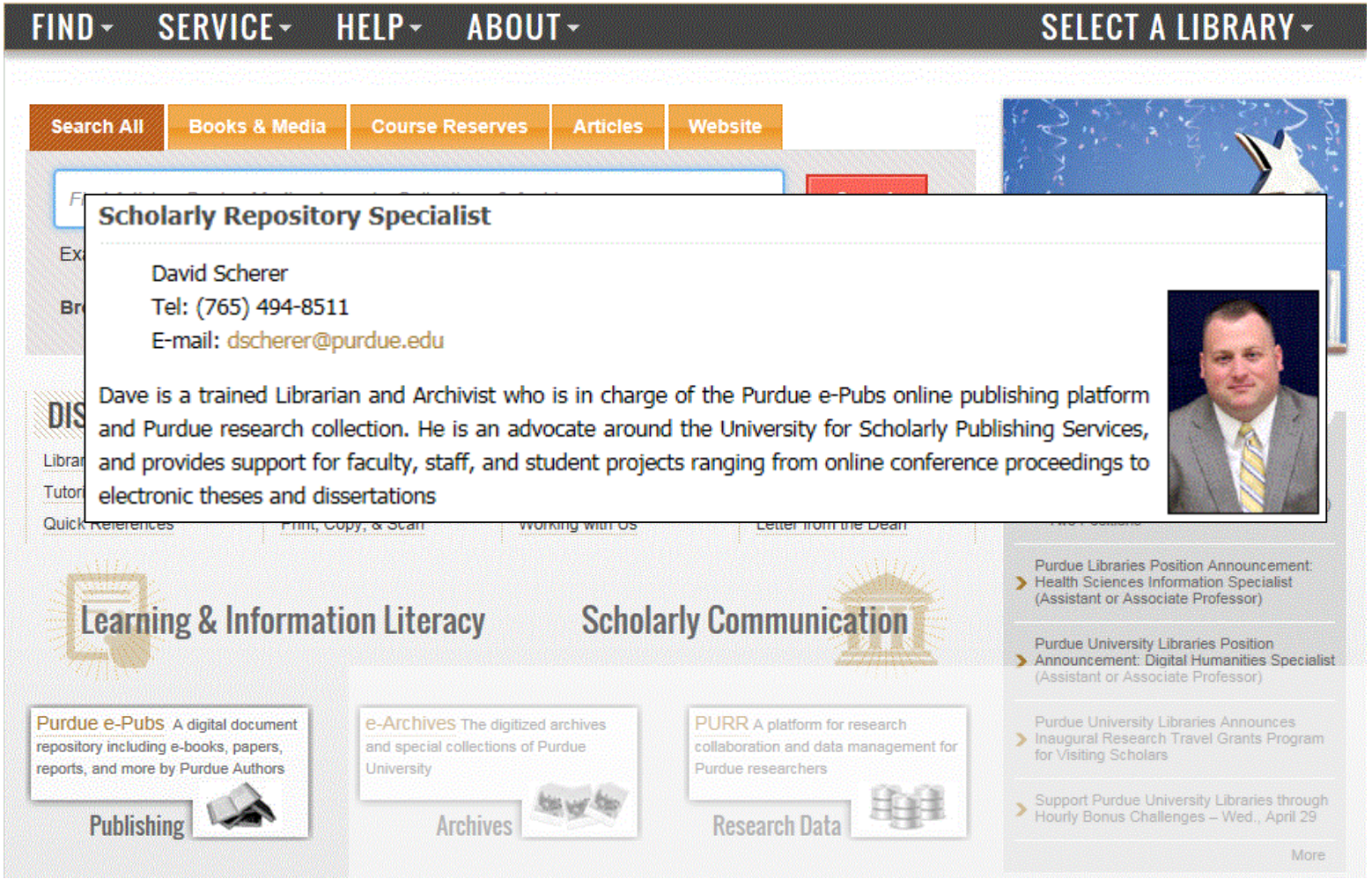
Research Data Metadata & Data Documentation

- Amy Barton: Things we do: repository intermediaries, metadata & data identification



A Dozen Things We Do...

E-Pubs Repository Intermediaries: Roles



A Dozen Things We Do...

E-Archives Repository Intermediaries: Roles

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

(765) 49-46766
cdearbor@purdue.edu

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

Publishing

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

Archives

PURR A platform for research collaboration and data management for Purdue researchers

Research Data

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

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

Archives

PURR A platform for research collaboration and data management for Purdue researchers

Research Data

A Dozen Things We Do...

USAID: Project Management & Data Sharing

PURDUE UNIVERSITY

Purdue University Research Repository PURR

Home | Publications | Projects | Get Started | Policies | Contact Us

You are here: Projects › Feed the Future Food Processing and Post-harvest ... › Files

FEED THE FUTURE FPL FOOD PROCESSING LAB

- Updates
- Info
- Team 28
- **Files** 21
- Publications
- To Do 2
- Notes 1

Feed the Future Food Processing and... (usaid)

Private project by Amy Barton

Files

Upload

Name
Branding-Marketing Materials
Forms
FPL-Docs
land_pop_paperregressions_cleaned.dta
Policy Documents
Templates

A Dozen Things We Do...

Traditional Cataloging Metadata

By subject

Browse | Simple Search | Advanced Search

File interchange handbook for images, audio, and metadata

Brad Gilmer; Society of Motion Picture and Television Engineers.; European Broadcasti

2004

Print Available: Engineering Engineering (006.696 F473 2004)

Find in Print | **Details** | Virtual Bookshelf

Title: File interchange handbook for images, audio, and metadata
Author: Brad Gilmer; Society of Motion Picture and Television Engineers
Subjects: Digital video -- Handbooks, manuals, etc; Digital television -- H
Handbooks, manuals, etc; Image processing -- Digital techniques -- Handb
and reproducing -- Digital techniques -- Handbooks, manuals, etc; Digital n
compression -- Handbooks, manuals, etc
Description: Convergence of information technology and traditional telev
organization of metadata and the SMPTE metadata dictionary -- The digit
360M : general exchange format -- The material exchange format -- Advar
systems format -- QuickTime file format.
Publisher: Burlington, MA : Focal Press
Creation Date: 2004
Format: xi, 307 p. : ill. ; 24 cm..
Identifier: ISBN0240806050
Language: English
Citation Source: PURDUE ALMA
Type: Book
MMS ID: PURDUE_ALMA21495255720001081

MARC 21

MARC 21 Format for
BIBLIOGRAPHIC DATA
Library of Congress
Network Development and MARC Standards Office

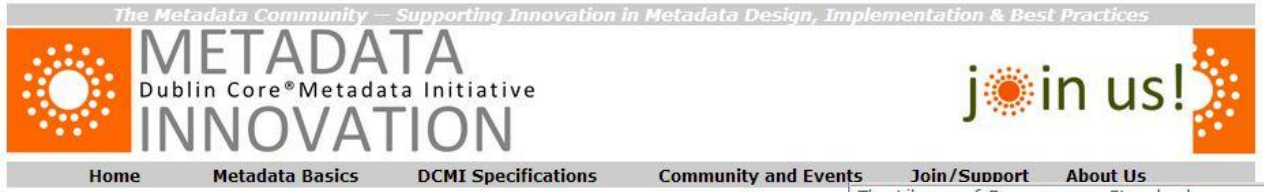
21495255720001081 - Google Chrome

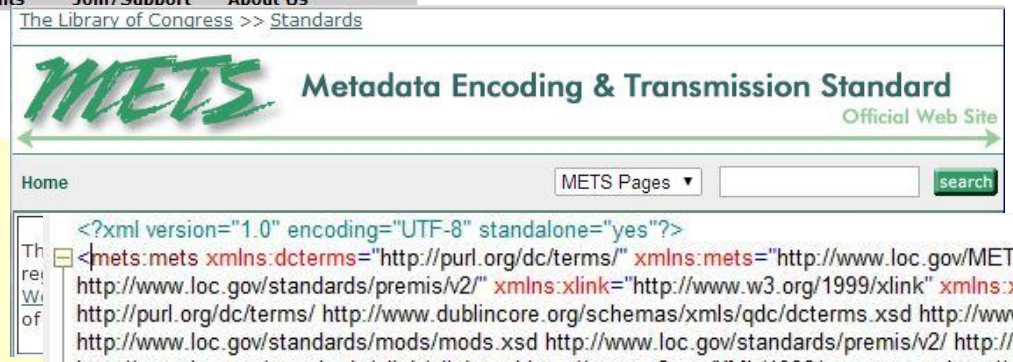
purdue-primo-prod.hosted.exlibrisgroup.com/primo_library/libweb/customized/puwl/jsp/marc.jsp?id=21495255720001081

```
000 01820cam a2200373Ia 4500
005 20101014115855.0
008 040514s2004    maua     bf   001 0 eng d
001 991348627200001081
010   |a  2004300184
020   |a0240806050
035   |a(OCoLC)ocm55471058
035   |a1380469
035   |a(InLP)1380469-wlafdb-Voyager
040   |aVYF|cVYF|dDLC|dOCLCQ|dIBS
049   |aIPL1
082 00|a006.6/96|222
090   |aTK6680.5|b.F55 2004
245 00|aFile interchange handbook for images, audio, and metadata /|ceditor in chief, Brad
    Gilmer.
260   |aBurlington, MA :|bFocal Press,|cc2004.
300   |axi, 307 p. :|bill. ;|c24 cm.
504   |aIncludes bibliographical references and index.
505 0 |aConvergence of information technology and traditional television production --
    Structure and organization of metadata and the SMPTE metadata dictionary -- The digital
    picture exchange file format -- SMPTE 360M : general exchange format -- The material
    exchange format -- Advanced authoring format -- Advanced systems format -- QuickTime file
    format.
650  0|aDigital video|vHandbooks, manuals, etc.
650  0|aDigital television|vHandbooks, manuals, etc.
650  0|aMetadata|vHandbooks, manuals, etc.
650  0|aImage processing|xDigital techniques|vHandbooks, manuals, etc.
650  0|aSound|xRecording and reproducing|xDigital techniques|vHandbooks, manuals, etc.
650  0|aDigital media|vHandbooks, manuals, etc.
650  0|aVideo compression|vHandbooks, manuals, etc.
```

A Dozen Things We Do...

Research Data Metadata Considerations



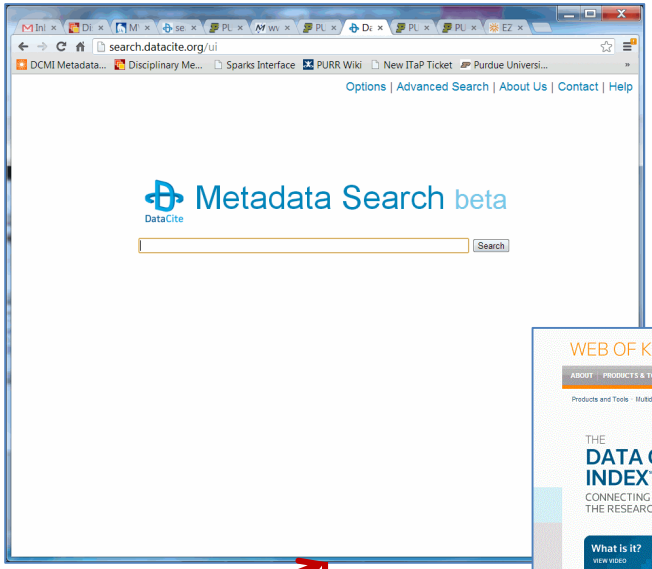


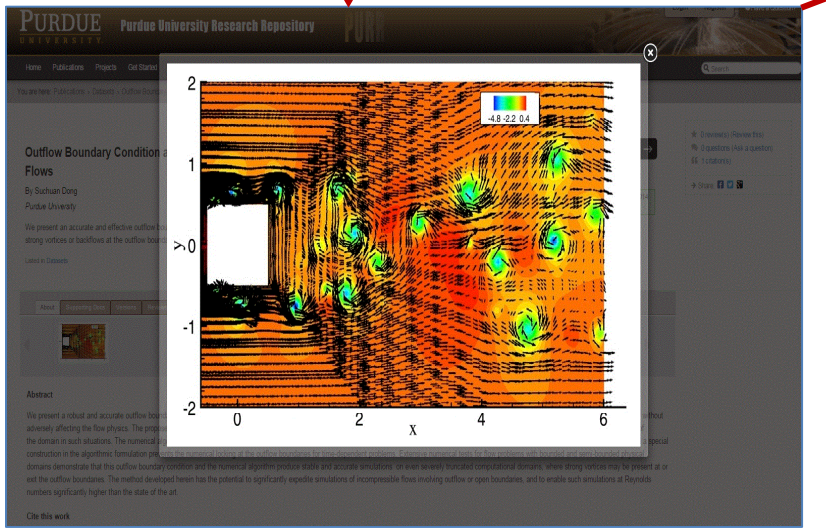
A Dozen Things We Do...

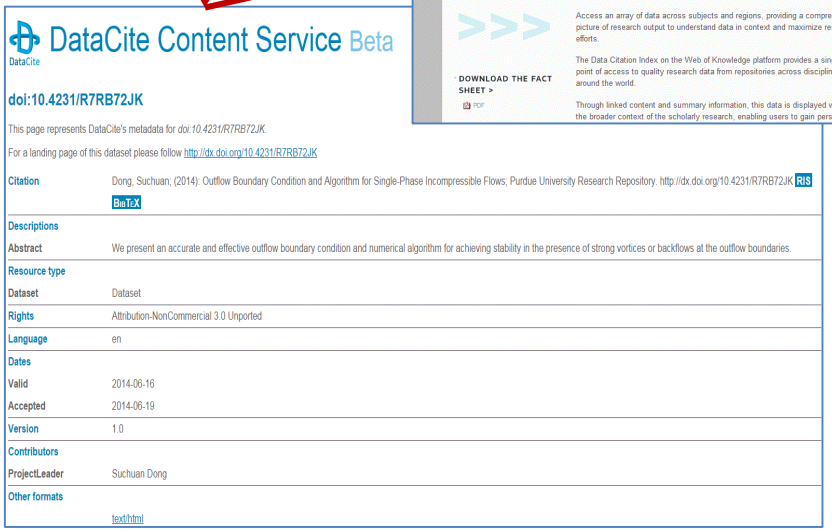
Digital Object Identifier (DOI)

A Dozen Things We Do...

Dataset Persistent Identifier







DATA → PUBLICATION → DISCOVERABILITY → ATTRIBUTION → IMPACT FACTOR!

Wrap up

- Connie Rinaldo: Wrap-up