NSF Data Management Plan Requirements: Institutional Initiatives

Fall 2010 Membership Meeting

Coalition for Networked Information December 13-14, 2010 Arlington, VA

D. Scott Brandt
assoc dean for research
Purdue University Libraries

Data Mgmt Plan or Planning?

"I tell this story to illustrate the truth of the statement I heard long ago in the Army: Plans are worthless, but planning is everything."

Remarks at the National Defense Executive Reserve Conference, 11/14/57
http://www.eisenhower.utexas.edu/All_About_Ike/Quotes/Quotes.html
http://www.flickr.com/photos/nationalarchives/3158905675/lightbox/

Purdue context…

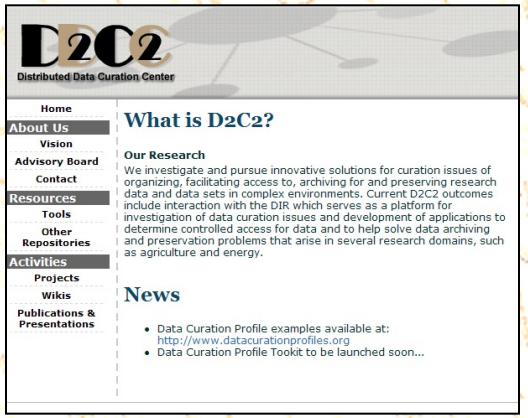
- Libraries leverages relationships with faculty, VPR and ITaP
- as well as pertinent projects of the Distributed Data Curation Center (D2C2) which investigates organizing, accessing, archiving research outputs
- to investigate and implement a response to the NSF data management plan "mandate" through Data Mgmt WG

2004: Purdue Libraries' Interdisciplinary Research Initiative revealed faculty data needs on campus

What they said…

- Not sure how or whether to share data
- Lack of time to organize data sets
- Need help describing data for discovery
- Want to find new ways to manage data
- Need help archiving data sets/collections

2006: Founded D2C2 to organize research and leverage collaborations

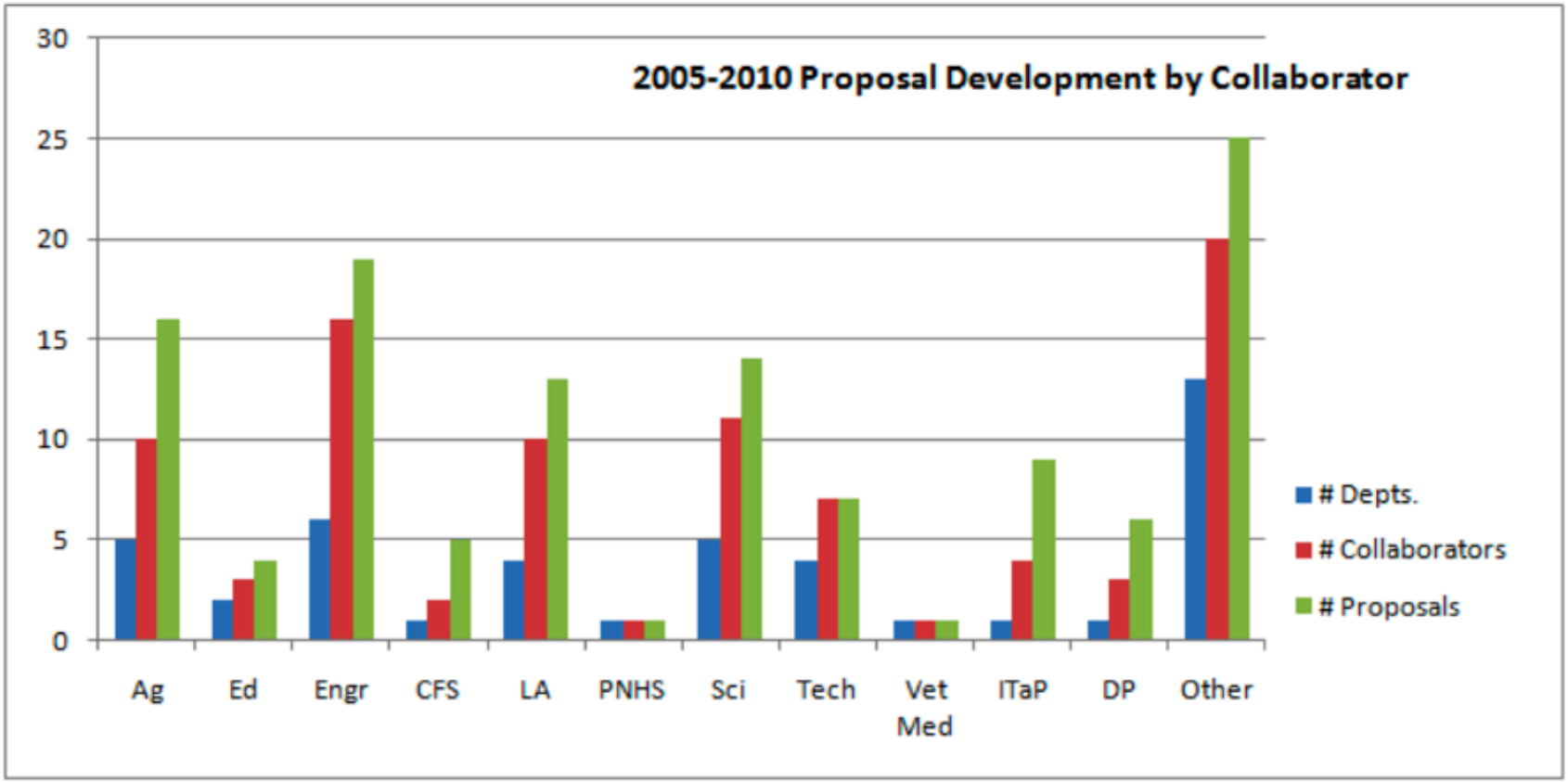


Distributed Data Curation Center:

- Created “research arm” for Purdue Libraries
- Established recognizable mode for collaboration with research faculty on campus
- Focused research on data curation problems, distributed environments, single PI/small lab

Faculty Collaborations around data

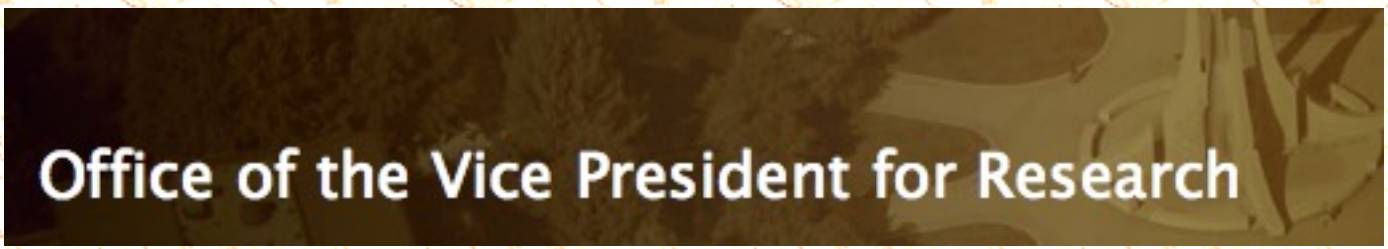
- *Developing a Data Management and Curation Workflow for Camp Calcium*—Consumer and Family Science faculty
- *Purdue University-Moi University Partnership to Investigate Water Data Curation*—Agronomy and Medical (IUPUI) faculty
- *Developing Content Organization Framework for Healthcare Delivery Hub*—Regenstrief Center for Healthcare Engineering



partnered with 68 Purdue faculty in 31 depts/units on 95 grant proposals

2009: Brandt's Fellowship in Purdue's Office of Vice President for Research

- Worked with compliance office
- Investigated peer institution research retention policies
- Analyzed guidelines and practices
- Surveyed departments re: research retention and raised awareness in OVPR of data mgmt needs



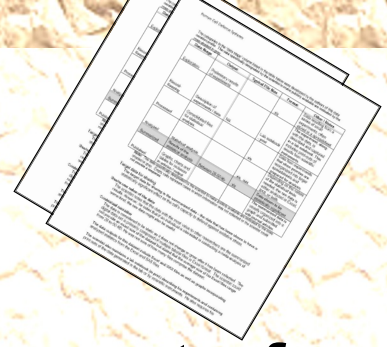
Data Mgmt Working Group

- Charged by OVPR to develop response to NSF data mgmt plan mandate
- Co-led by Dean of Libraries Jim Mullins and CIO Gerry McCartney
- Included faculty from Biology, Education, Engineering, Foods & Nutrition, Forestry, Pharmacy, Veterinary Science

Data Mgmt WG outcomes

- Libraries to modify Data Curation Profile to identify specific areas of need per the points listed in NSF's Grant Proposal Guide (GPG) Chapter II.C.2.j.
- OVPR identifies NSF proposals under development and contacts PIs to test the new instrument.
- As appropriate, Libraries help identify problem areas and ways to address them, for instance suggesting possible disciplinary standards.
- Libraries identify archiving and preservation services ("curation core") and work with ITaP to develop them in the HUB environment.

Data Curation Profile



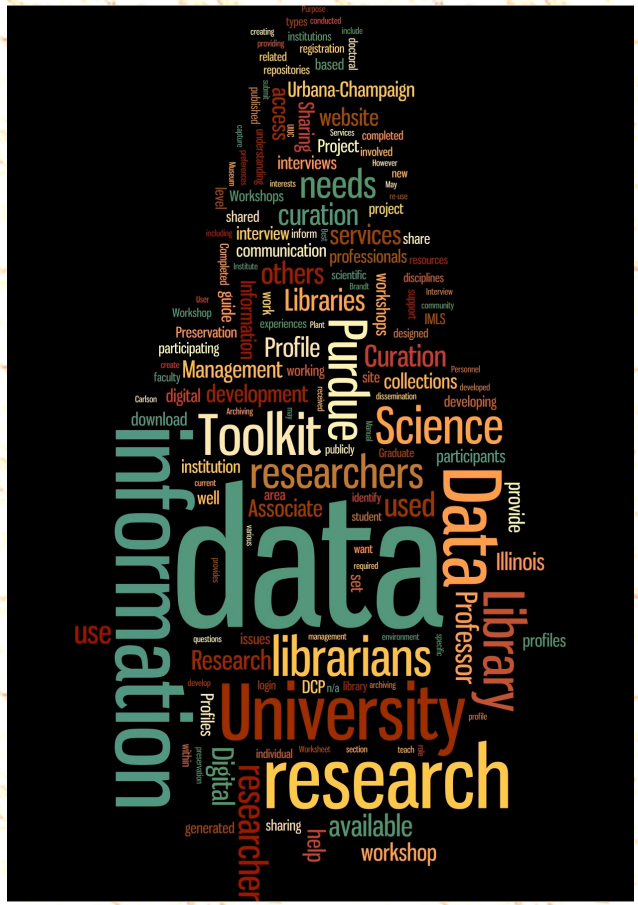
- A means to capture requirements for specific data generated by a single researcher or lab, based on their reported their needs and preferences for these data.
- A concise, structured document suitable for sharing and annotation.
- A resource for Librarians, Researchers, IT Professionals, Data Managers, and others.

Can the Data Curation Profiles address the NSF mandate?

*The Data Curation Profile is **not** a direct solution to a data management plan, nor a guide to curating data for ingest and archiving. However, it is a tool which **may** help facilitate these activities*

DCP helps identify

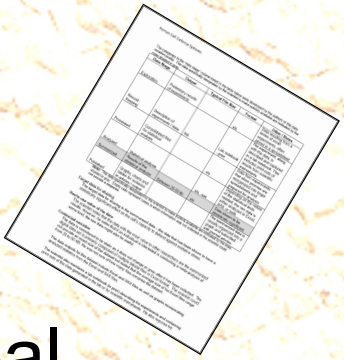
- Research Data Lifecycle
- Data Mgmt/Storage
- Disposition of the Data
- Data Dissemination and Sharing
- Data Preservation and Repositories



NSF wants:

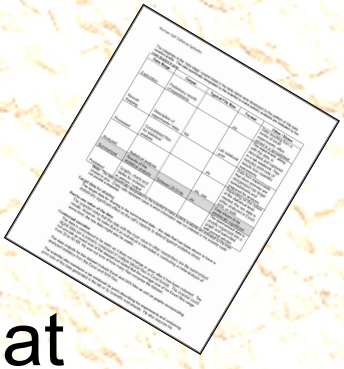
1. Types of data and samples to be produced in the project
2. Standards for data and metadata
3. Policies for access and sharing
4. Policies for provision and re-use, re-distribution and production of derivatives
5. Plans for archiving data, samples

describe data and samples to be produced



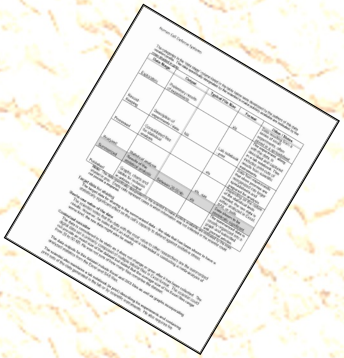
- “[O]bservational data collected during real weather occurrences, and ‘idealized’ data are derived computationally from observational data through the application of rule-based algorithms”
- “output of the model is three-dimensional floating point data using a post-processing tool called ‘Read/Interpolate/Plot’… produces the basic radar-like plots from the data” Formats: NetCDF, Raw binary .dat, Vis5D .v5d

address standards for data & metadata



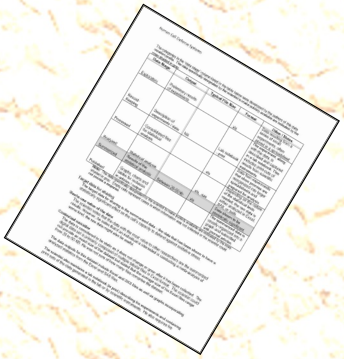
- “a locally developed metadata schema that includes information about the plant sample, the conditions under which it was grown, and the resulting data is captured as a part of the workflow within the information management system;” this system has been published and “information as to how the data and metadata are generated, collected and disseminated.”

ID policies for access/sharing



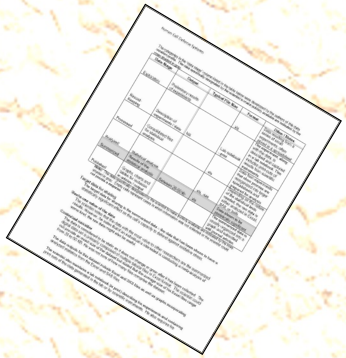
- “For reliability and to facilitate accuracy measures, the spreadsheet includes 3-10 observations for each variable. The replicate data are then reduced by combining (usually with statistical averaging) the multiple observations." The resulting “Reduced spreadsheet” contains the reduced data, which, along with related images, has been identified as having the most value for sharing.”

ID provision for re-use, re-distribution



- “system allows for the searching of data through a basic and an advanced search” of “gene type/ATG number, parent line, line name, tray number(s)” and “Boolean operators to search for data and allows searching for a range of values.”

ID plans to archive data/samples



- The “scientist rents server space to host the information management system and pays a fee for the maintenance. Maintenance includes migration to the most current version of Postgress.” And “a cost benefit analysis a means to determine the duration of preservation” will be developed.

testing new instrument

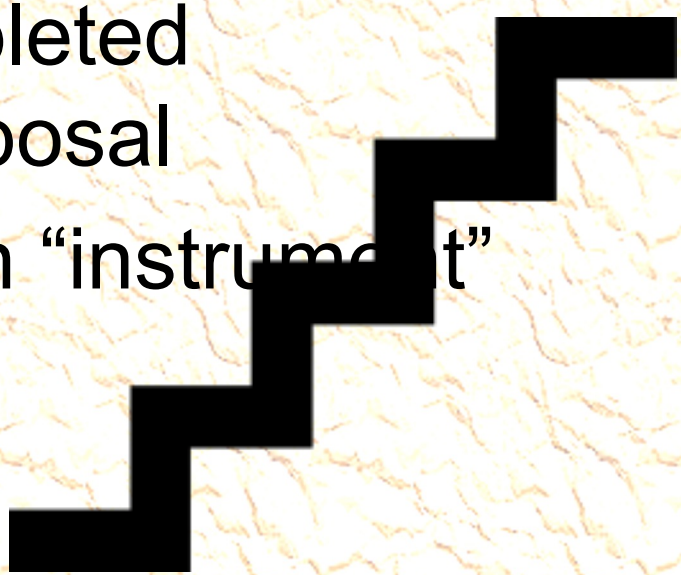
- Sections with the Profile were extracted and interviews asked detailed questions
- Focused on data mgmt questions we have dealt with on other projects
- Drilled much deeper into practice
- Also specified who and how, not just what was going to be done

initial DMP interviewing

- Identified four researchers submitting proposals in January
- Met them to initiate first interview of about an hour long
- Meeting included researcher, data research scientist (Jake Carlson), a subject librarian and grant coordinator
- Analysis and draft of plan

next steps

- Developing an “instrument” that can be used by researchers directly
- Providing consulting “workshops”
  - for those who have completed technical sections of proposal
  - and have worked through “instrument”



implementation of research hub

- Using HUBzero platform
- Designing repository services that support a “curation core”
- Developing related resources
- Integrating library support for consulting, assisting

Dilbert. Scott Adams. May 28, 2010
http://dilbert.com/strips/comic/2010-05-28/