6-21-2012

Data Curation as a Form of Collaborative Research

D. Scott Brandt
Purdue University Libraries, techman@purdue.edu

Data Curation as a Form of Collaborative Research

PURDUE UNIVERSITY LIBRARIES

D. Scott Brandt
assoc dean for research

June 24, 2012

ALA ANAHEIM

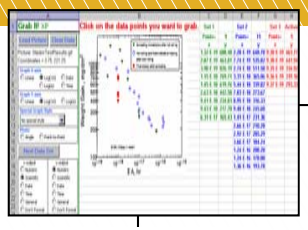
ACRL

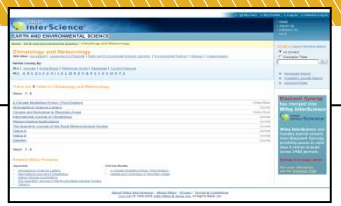
STS
Science & Technology Section

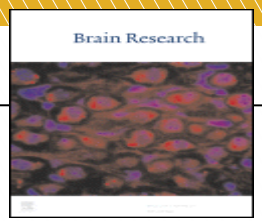
PURDUE
UNIVERSITY
LIBRARIES

2004	Exploration of campus research issues (“we need help organizing data!”)
2005	Strategic Plan → reorganization in Libraries creates position of Associate Deans for Research (ADR) to interface with Research Administration, support Libraries research
2006	e-Pubs IR (Digital Commons) launched and Distributed Data Curation Center (D2C2) created to leverage multi-disciplinary collaborations on funded research projects
2007	IMLS funded Data Curation Profile project (“who is willing to share what with whom, and when?”) and Data Research Scientist hired
2008	e-Data Task Force led to making data services a priority and demonstrated proof-of-concept data ingest into prototype data repository
2009	D2C2 continued collaborations with HUBzero® to investigate application of emerging library standards and technologies OAI-PMH, Linked Data
2010	NSF Data Mgmt Plan announcement, Libraries led meeting with IT, VPR, faculty
2011	Libraries/IT NSF DMP workshops, Purdue University Research Repository WG
2012	Four year budget to build out PURR, soft launch with data publishing functionality

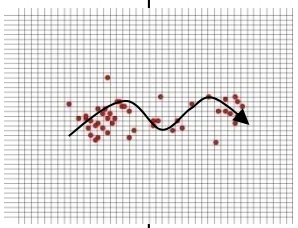
PURDUE UNIVERSITY LIBRARIES





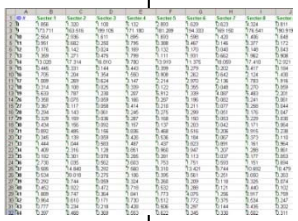


PURR | **e-Pubs** | **online research** Internet, OA, etc. | **published research** traditional | **secondary/ tertiary resources**

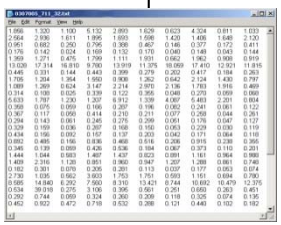


Librarians provide education on depositing and finding data.

Librarians help shepherd and curate datasets.



Librarians contribute to or collaborate on projects.



Librarians consult on data management plans.

PURDUE UNIVERSITY LIBRARIES

Modified from: Brandt, D.S. "Scholarly Communication" (in To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering.: Final Report of Workshop New Collaborative Relationships: Academic Libraries in the Digital Data Universe. ARL, Washington, DC, September 2006.)

Small science, single PI, small lab

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

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

PURDUE UNIVERSITY LIBRARIES

*http://units.sla.org/division/dst/Annual%20Conference/2009_Washington/Heidorn.pdf
photo from Science & Technology Review https://www.llnl.gov/str/December01/Orme.html

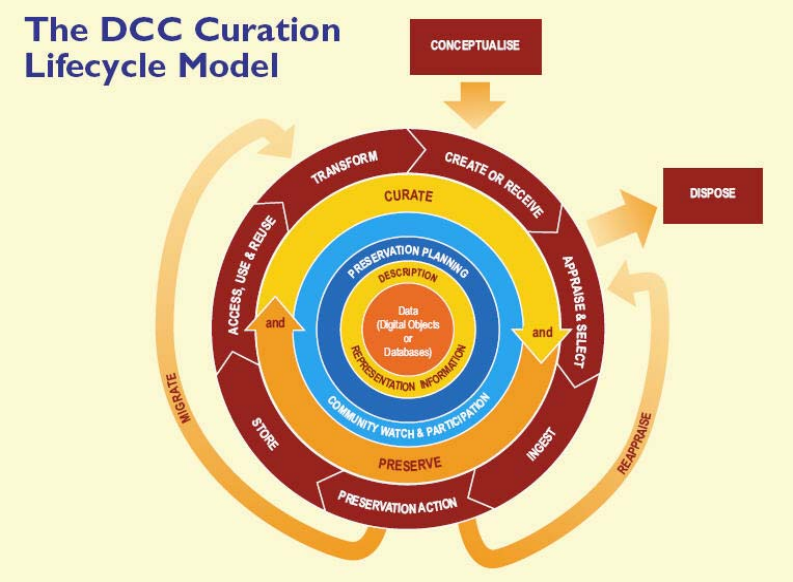
Small science can encounter unintentional consequences

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

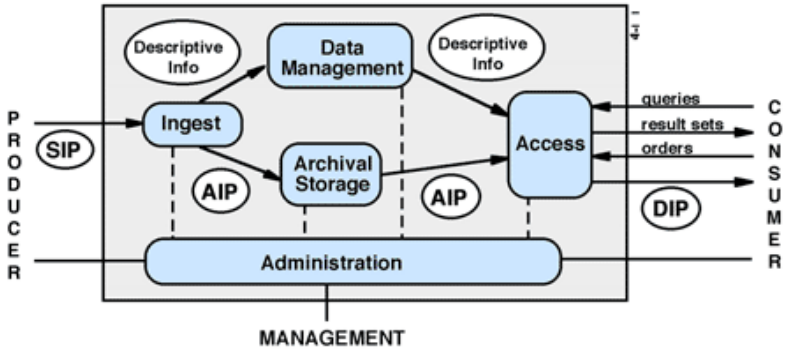
PURDUE UNIVERSITY LIBRARIES

*D. Scott Brandt. Provost Fellowship, Final Report 2009 (*unpublished*)

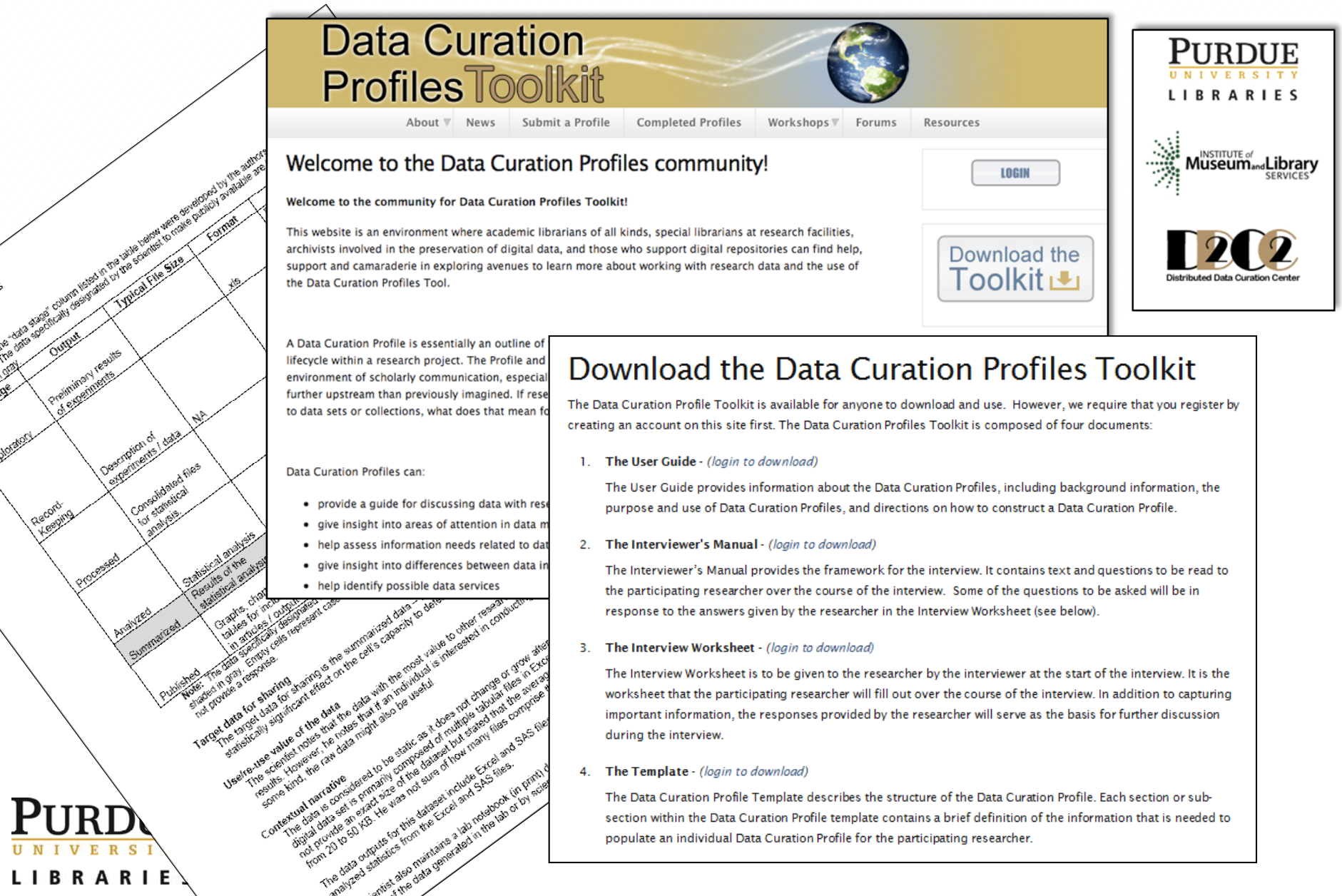
all pictures Purdue University



Because single PIs tend not to have mature data curation in place, they are not likely to see where to enter data curation continuum



PURDUE UNIVERSITY LIBRARIES


Data Curation Profiles Toolkit
About
News
Submit a Profile
Completed Profiles
Workshops
Forums
Resources
Welcome to the Data Curation Profiles community!
LOGIN
Welcome to the community for Data Curation Profiles Toolkit!
This website is an environment where academic librarians of all kinds, special librarians at research facilities, archivists involved in the preservation of digital data, and those who support digital repositories can find help, support and camaraderie in exploring avenues to learn more about working with research data and the use of the Data Curation Profiles Tool.
Download the Toolkit
A Data Curation Profile is essentially an outline of
lifecycle within a research project. The Profile and
environment of scholarly communication, especial
further upstream than previously imagined. If rese
to data sets or collections, what does that mean fo
Data Curation Profiles can:
provide a guide for discussing data with rese
give insight into areas of attention in data m
help assess information needs related to dat
give insight into differences between data in
help identify possible data services
Download the Data Curation Profiles Toolkit
The Data Curation Profile Toolkit is available for anyone to download and use. However, we require that you register by creating an account on this site first. The Data Curation Profiles Toolkit is composed of four documents:
1. The User Guide - (login to download)
The User Guide provides information about the Data Curation Profiles, including background information, the purpose and use of Data Curation Profiles, and directions on how to construct a Data Curation Profile.
2. The Interviewer's Manual - (login to download)
The Interviewer's Manual provides the framework for the interview. It contains text and questions to be read to the participating researcher over the course of the interview. Some of the questions to be asked will be in response to the answers given by the researcher in the Interview Worksheet (see below).
3. The Interview Worksheet - (login to download)
The Interview Worksheet is to be given to the researcher by the interviewer at the start of the interview. It is the worksheet that the participating researcher will fill out over the course of the interview. In addition to capturing important information, the responses provided by the researcher will serve as the basis for further discussion during the interview.
4. The Template - (login to download)
The Data Curation Profile Template describes the structure of the Data Curation Profile. Each section or sub-section within the Data Curation Profile template contains a brief definition of the information that is needed to populate an individual Data Curation Profile for the participating researcher.
PURDUE UNIVERSITY LIBRARIES
INSTITUTE of Museum and Library SERVICES
D2C2 Distributed Data Curation Center
Format
Typical File Size
Output
Preliminary results of experiments
Description of experiments / data
Consolidated files for statistical analysis
Statistical analysis
Results of the statistical analysis
Record-Keeping
Processed
Analyzed
Summarized
Published
NA
xls
Target data for sharing
Use/re-use value of the data
Contextual narrative
PURDUE UNIVERSITY LIBRARIES

What data sets do they have?
How are they managing these data sets?
What would they like to do with them?

Worksheet

2. How long would your data set be useful or have value for you or others if it were to be preserved?

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

Manual

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

- Why did you select this length of time?

Template

Section 13 - Preservation

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

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

13.1 - Duration of preservation
A statement about the length of time the data is to be preserved. The duration may be event based rather than time based, though estimation for the length of time related to the event should be noted where possible.

PURDUE UNIVERSITY LIBRARIES

U1N1S2 – (reviews the length of preservation question) – I think it is probably this.

I – Okay, 10 to 20 years.

U1N1S2 – Well I have data from my PhD that still would be useful if I could still access it. So that's from 1994, so yes, 10 to 20 years.

I- Okay. Is there a particular reason for the time frame as opposed to say 50 or 100 years?

U1N1S2 – I guess I'm hoping that we are able to do the modeling by then.

I – I see, so once the model has been worked out and you are very confident that the model does what you want it to do, then the underlying data used to get to that point would not be as valuable anymore?

U1N1S2 – I think so, yeah. And other things in that time period. actually used. So…

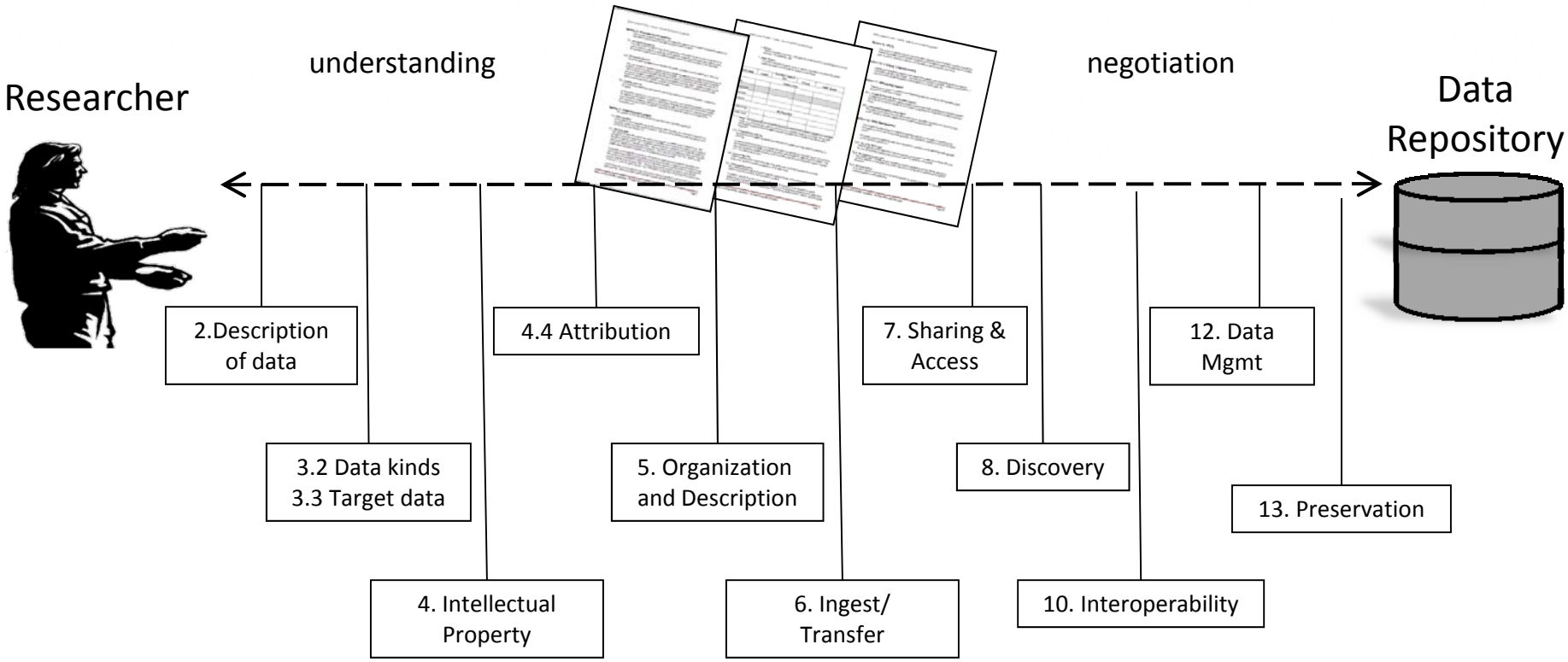
Transcript

13.1 - Duration of preservation

The scientist believes that her data should be preserved for 10 years or more but less than 20 years. The scientist believes that the mathematical model they are working on will be perfected within this timeframe. When the model is considered complete the data will lose much if not all of its value. In addition, the device used in generating the data may become obsolete within this time period and replaced, again diminishing the value of the data.

Profile

PURDUE UNIVERSITY LIBRARIES



PURDUE UNIVERSITY LIBRARIES

Data Curation Profiles Toolkit

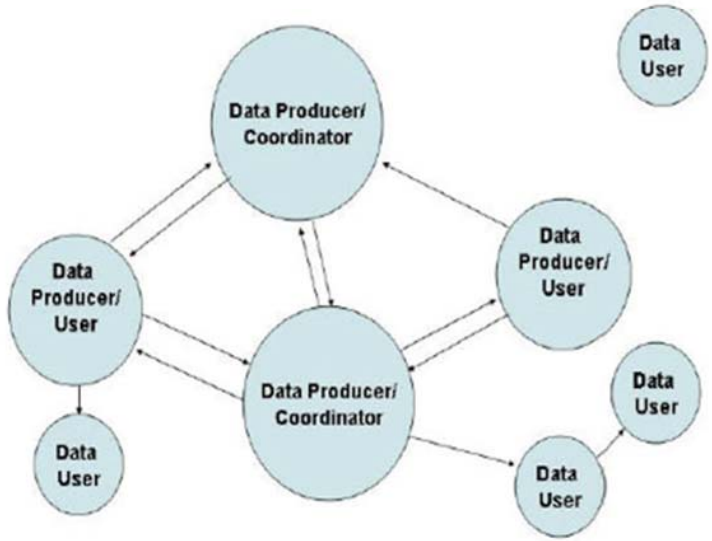
About | News | Submit a Profile | Completed Profiles | Workshops | Forums | Resources

Home

Module 10: Data Sharing

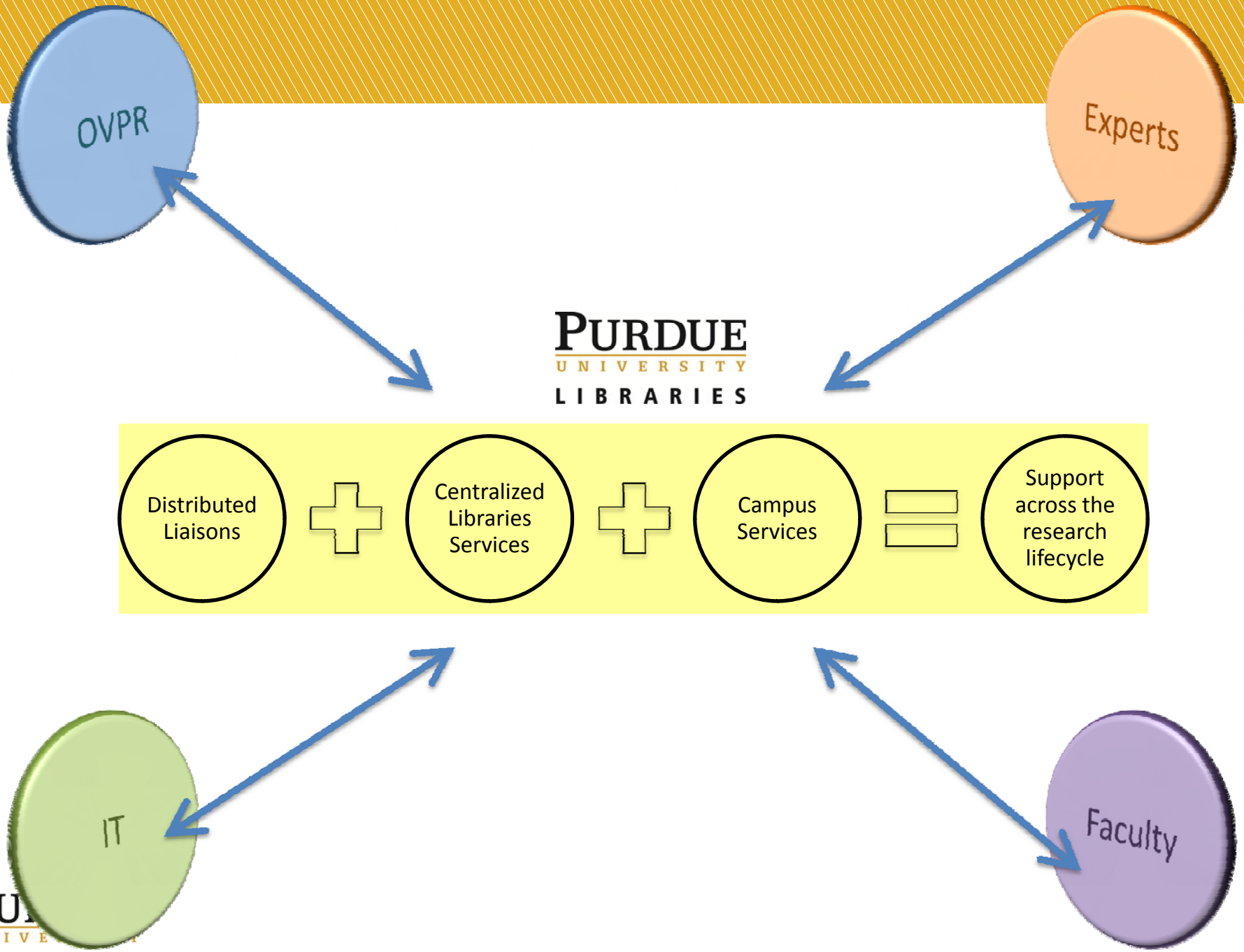
Intro | Importance | Data Sharing Table | Interviewing | Worksheet | Sharing | Review

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



Source: http://www.urisa.org/publications/journal/articles/when_data_sharing

PURDUE UNIVERSITY LIBRARIES


OVPR
Experts
PURDUE
UNIVERSITY
LIBRARIES
Distributed Liaisons
Centralized Libraries Services
Campus Services
Support across the research lifecycle
IT
Faculty

OVPR

Experts

PURDUE UNIVERSITY Office of the Vice President for Research

Home | Integrity/Regulatory/QA | Research Development | Funding | Partnerships | Center

OVPR ▸ Research Development

Services & Resources

- Research Development Home
- Funding
- Limited Submissions
- Proposal Preparation

Research Development

The goal of the OVPR Research Development staff is to assist faculty in the
proposals. **Our staff** provide a broad range of services and resources related
some of the ways we can assist.

FUNDING
The funding page provides information on internal, ex

PURDUE UNIVERSITY Sponsored Program Services

General Info | Pre-Award | Post-Award | Contract Management | Data | Training | Links | SPS Ho

Business Services | Administrative Information Services (AIS)

Sponsored Program Services

News

- FY2011 A-133 Report (New)
- Archived News
- Subrecipient Monitoring Procedures

PURDUE UNIVERSITY Sponsored Program Services
Pre-Award Services

General Info | Pre-Award | Post-Award | Contract Management | Data | Traini

Pre-Award: | General Info | Compliance | Proposals | Sponsors

Proposal Preparation

- Interdisciplinary Proposals
- OVPR Proposal Preparation Resources
- OVPR Proposal Preparation Assistance

IT

Faculty

PU
UNIVE
LIBRARIES

What about my current proposal?

- **Option 1: use the Purdue Data Management Plan**
  - Libraries/OVPR will provide a citable ove
    basic, long term issues
  - You append details relevant to your proj
    guided by questions provided by Librari
  - Libraries are available to provide review
    Nelson, M. Stowell Bracke, and J. Carlso
- **Option 2: exploit data management**
  in place for your discipline
  - Example: data management resources c
    Consortium for Political and Social Rese
    http://www.icpsr.umich.edu/icpsrweb/ICPSR/
- **Option 3: personalized solution base**
  provide and control

What about my current proposal?

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

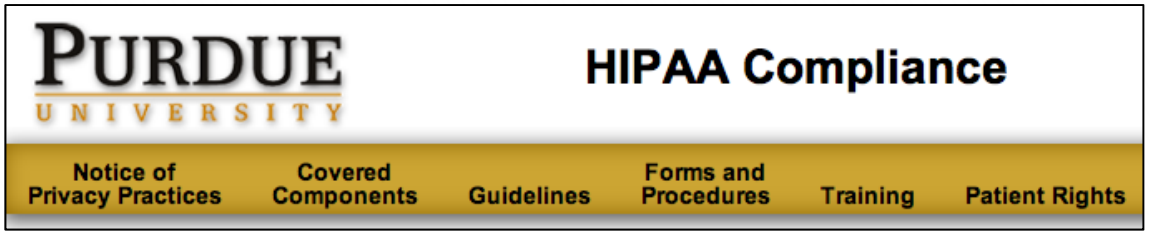
    It explains the current prospect of a centrally supported PURR resource available at no direct cost to the project.

	7a-d are related to Export Control.	
	a) Does this project require sponsor approval of publications?	☐ Yes ☐ No
	b) Does the announcement restrict participation based on citizenship?	☐ Yes ☐ No
	c) Does this proposal contain technical data within the statement of work that are restricted for purposes specific to military or space applications?	☐ Yes ☐ No
	d) Does this proposal contain technical data within the statement of work that are confidential or proprietary to a company or Purdue (i.e. requires a confidentiality agreement)? If yes, list the relevant pages.	☐ Yes ☐ No
8	Are you receiving private health information that is protected by HIPAA?	☐ Yes ☐ No
	Project Data	
1	Is there a conflict of interest?	☐ Yes ☐ No
2	Is this a Bowen Lab project?	☐ Yes ☐ No
3	Is this a Kepner Facility project?	☐ Yes ☐ No
4	List any approved University Centers affiliated with this proposal.	
5	Please identify all ITaP units which have or will contribute to this project: ☐ Envision Center for Data Perceptualization ☐ Rosen Center for Advanced Computing ☐ Other ITaP Units ☐ None	
6	Is this proposal a resubmission of a previously submitted application?	☐ Yes ☐ No
7	Are you an NSF Beginning Investigator?	☐ Yes ☐ No
8	Does the funding agency require a management plan for data or digital products produced using the grant? ☐ Yes, and we plan to use the Purdue University Research Repository (PURR) https://research.hub.purdue.edu/ ☐ Yes, and we plan to use another option ☐ No, funding agency does not require	
9	Will historical sites be affected?	☐ Yes ☐ No
	If Yes, please provide an explanation:	
10	If this project has an actual or potential impact on the environment, has an exemption been	☐ Yes ☐ No

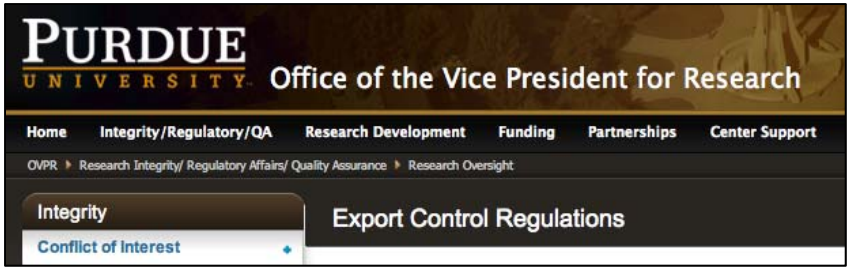
PURDUE UNIVERSITY LIBRARIES

OVPR

Experts



IT



Faculty

PURDUE
UNIVERSITY
LIBRARIES

Purdue University Research Repository | PURR

Will you be uploading any data that may be considered by the University to be sensitive or restricted? *REQUIRED*

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

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

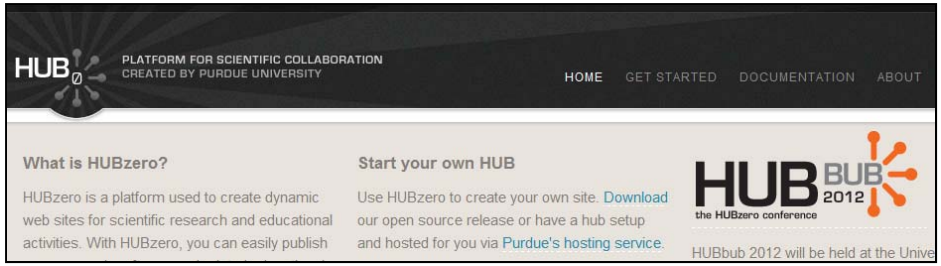
- ❑ This project may involve government-restricted, export-controlled, or proprietary company information (without permission).
- ❑ This project may involve data that is governed by an Institutional Review Board (IRB) Approval Protocol.
- ❑ This project may involve HIPAA data or Protected Health Information.
- ❑ This project may involve FERPA data or student records.

PU
UNIVE
LIBRARIES

OVPR

Experts

IT



Faculty

PU
UNIVE
LIBRARIES

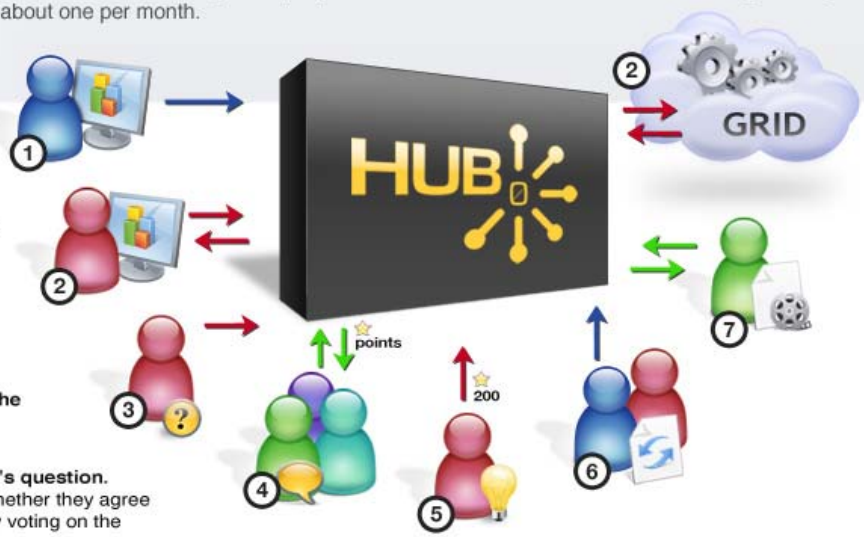
HOW HUBS WORK

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

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

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

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



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

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

nd Sue decide to collaborate on their next project, which is successful, so they write a research paper and give a presentation at a . They post the presentation slides on the hub and then track how many users view the materials.

s the presentation and uses some of the slides in his class to explain this new area of research to his students. He creates a homework which his students use Hugh's simulation tool to investigate the phenomena, and then Drew posts the assignment on the hub for other educators

Storage space for PURR projects currently covered by central resources

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

Estimated cost of additional space *

Option:	*Space Available:*	*For how long:*	*For who:*	*Cost:*
Extra project space	As needed	Life of project	Anyone with a Purdue Career Account	$2,102.57 per TB per year **
Extra publication space	As needed	10 years	Anyone with a Purdue Career Account	$14,297.48 per TB

PU
UNIVE
LIBRARIES

Applying library science to information problems

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

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

Identify

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

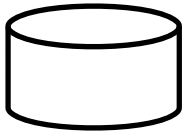
Assess

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

Analyze

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

Model



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

Demonstrate

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

***Problem*: Water quality research is conducted by several groups which results in various amounts and sets of data which need to be organized**

Curiosity + Collaboration = Choreographic Expression

A History of Concert Dance at Purdue University 1931-2011

- The Queen is Dead -

Excerpts from the 2009 Purdue Contemporary Dance Company performance of "The Queen is Dead".

Choreographer's notes: *"The Queen is Dead was created in about 3 weeks-- a very fast process for me. I started from a very different premise than where I ended. I used some concepts for the original movement vocabulary creation from the early ideas, and then used a different concept when actually constructing the piece. I used images and text from a collage on a table in a coffee shop where I was sitting to prepare for rehearsals, plus ideas about falling and physical limitations, and then it all came together when a mentor suggested using the actual music used for coronations in England. It was created for a different cast, in Milwaukee, Wisconsin, and then set on the PCDC cast. The original work had 2 men and 1 woman, but I think it totally worked with the 2 women and 1 man."*

Date:	2009	Performers:	Mandy Hampton Carrie Meyer Paul Rutz
Venue:	Hansen Theatre, Yue-Kong Pao Hall, Purdue University	Musicians:	N/A
Choreographer:	Holly Jaycox (Division of Dance faculty)	Lighting Designer(s):	Krystle Smith
Composer:	N/A	Costume Designer(s):	Holly Jaycox
Sound Designer(s):	Seth Warren-Crow	Set Designer(s):	N/A
Source Music:	William Boyce Propellerheads	Videographer:	Kathy Evans (Visual Arts Librarian)
Other Collaborator(s):			

Subject: Faculty choreography, Dance performance
Type of content: Dance performance
Content format: H.264/MPEG-4
Content editor: Eugenia Kim
Content source: Division of Dance video archives
Publisher: Virginia Kelly Karnes Archives and Special Collections Research Center, Purdue University Libraries
Rights Policy:
This video is for recreational viewing only and is not meant to be reproduced nor used for research purposes. Please contact the Virginia Kelly Karnes Archives and Special Collections Research Center for inquiries about the original media.

ChoreoSave: A Digital Dance Preservation System Prototype

Eugenia S Kim
Purdue University Libraries, Purdue University
West Lafayette, IN, USA
eugeniakim@purdue.edu

ABSTRACT

This paper presents the design and implementation of a short-term digital dance preservation solution called *ChoreoSave*. The author created a prototype system that identifies what components comprise a dance work and how such components can be represented using EPrints software hosted on an Amazon EC2 instance. Suggestions for future development are discussed based on evaluation and challenges encountered in the course of research.

such as music through documentary footage and contextual information. This approach challenges traditional notions of capturing movement through the use of information organization principles. By forming a standard submission template with customization options for users, the appraisal, selection, and ingestion processes will become the users' responsibility, thereby streamlining the flow of content from creator to repository. The short-term characteristic coupled with a peer-based model facilitates a more

Problem: How are digital objects of dance identified, accessioned, and prepared for distribution? And what are the best practices and policies required to produce a multimedia-based dance history website?

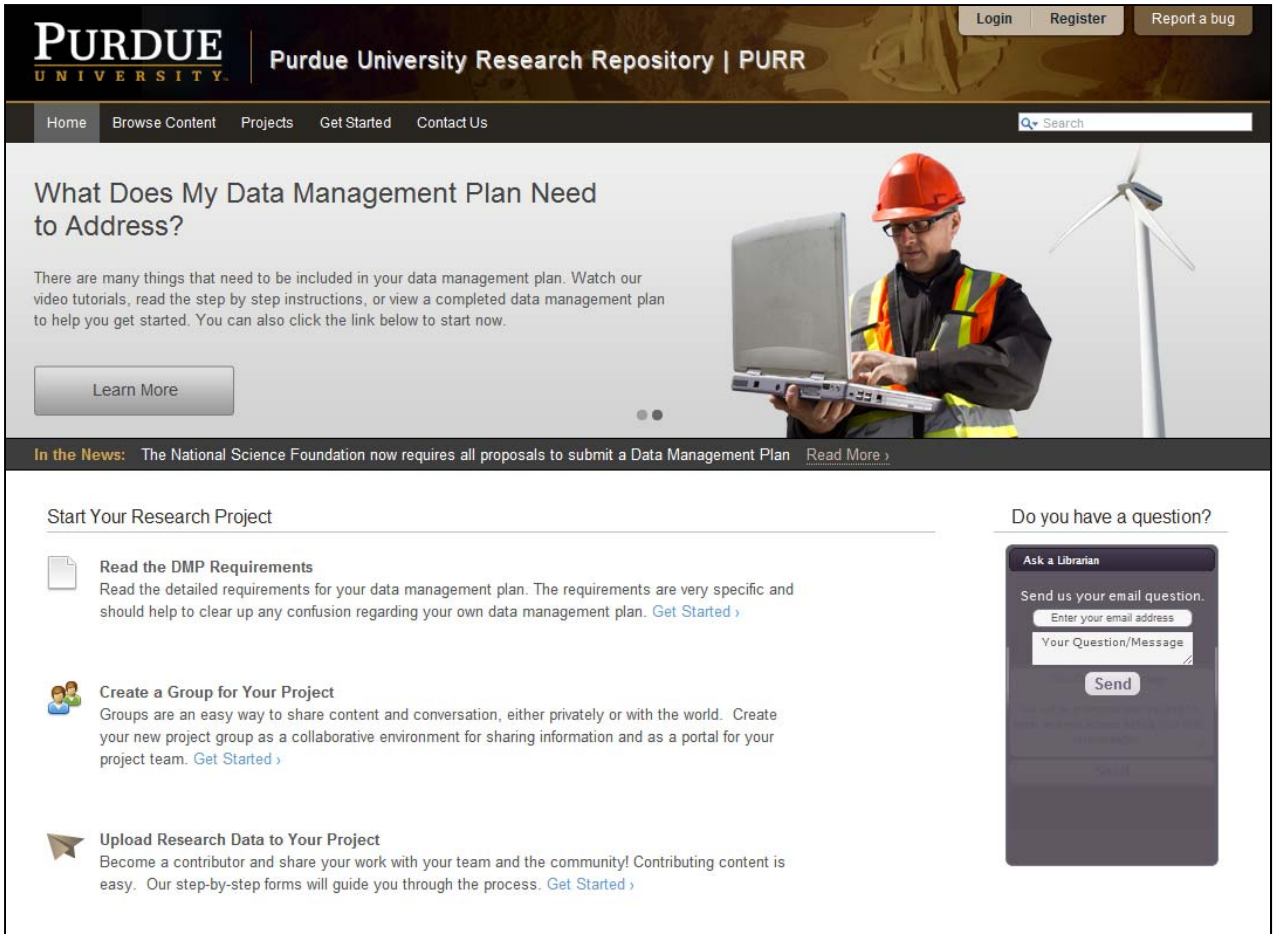
Faculty

PU
UNIVE
LIBRARI

Data curation is a collaborative process…

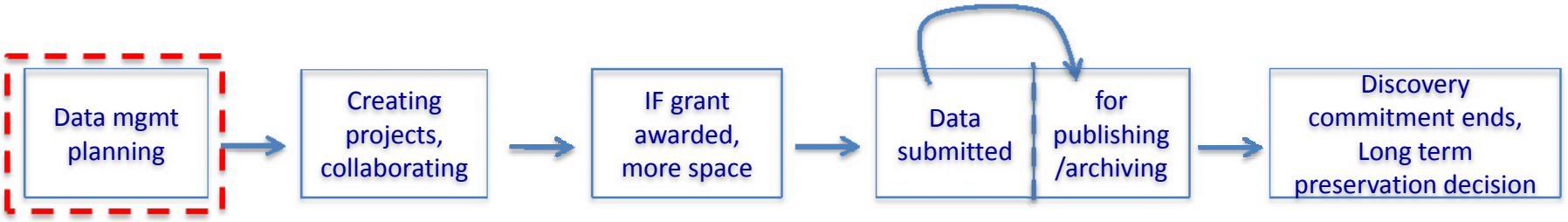
How this is manifested in PURR

http://research.hub.purdue.edu



PURDUE UNIVERSITY LIBRARIES

PURR functionality & service

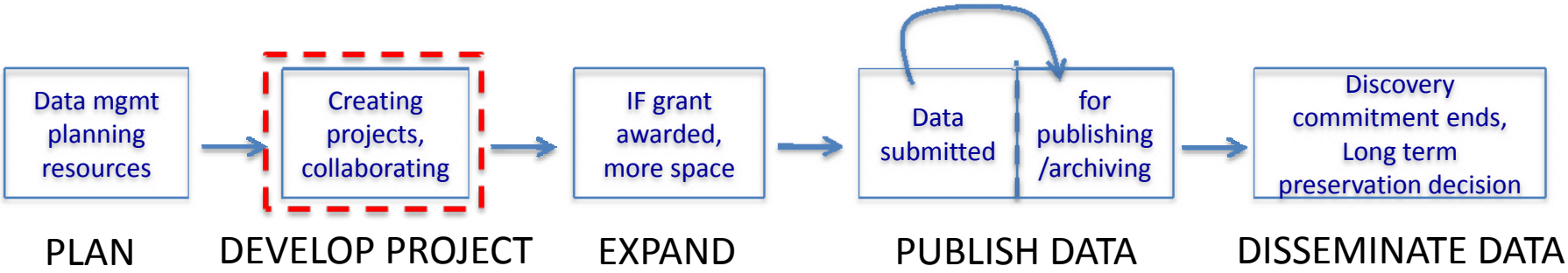


Researchers are guided to PURR for help with data mgmt plans by Pre-Awards, workshops and promotion, and by word-of-mouth

Librarians consult on data management plans.

PURDUE UNIVERSITY LIBRARIES

PURR functionality & service

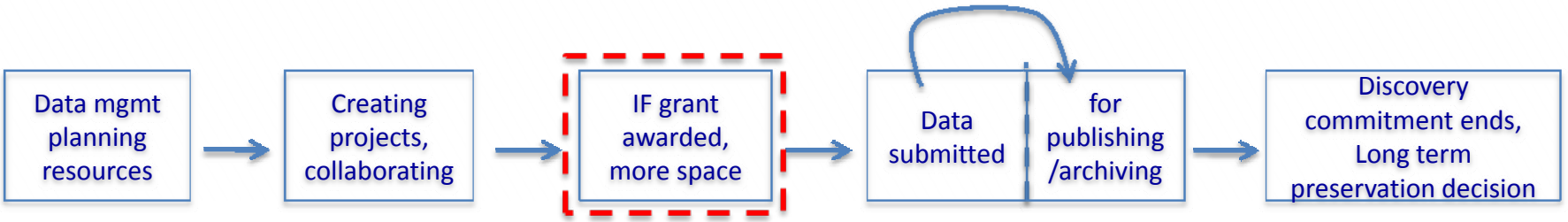


Researchers can create projects at any time, invite others to join… the goal is to help facilitate research development

Librarians can contribute to or collaborate on projects.

PURDUE UNIVERSITY LIBRARIES

PURR functionality & service

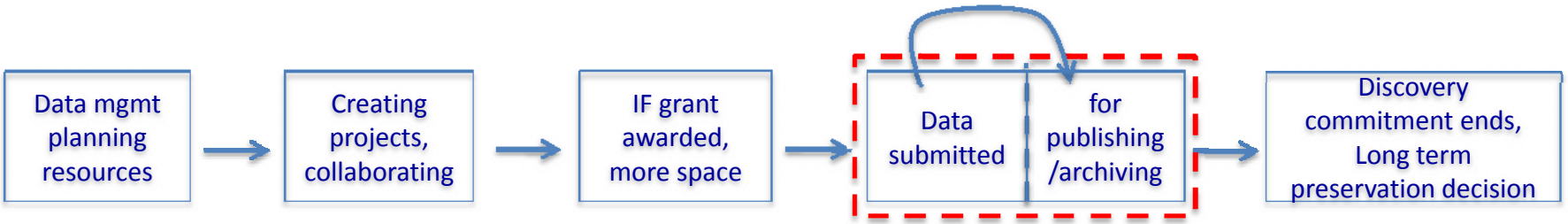


Once a grant is awarded, researchers get an increase in space allocation and length of time for project and data

Project space increases from 50 MB to 100 GB

PURDUE UNIVERSITY LIBRARIES

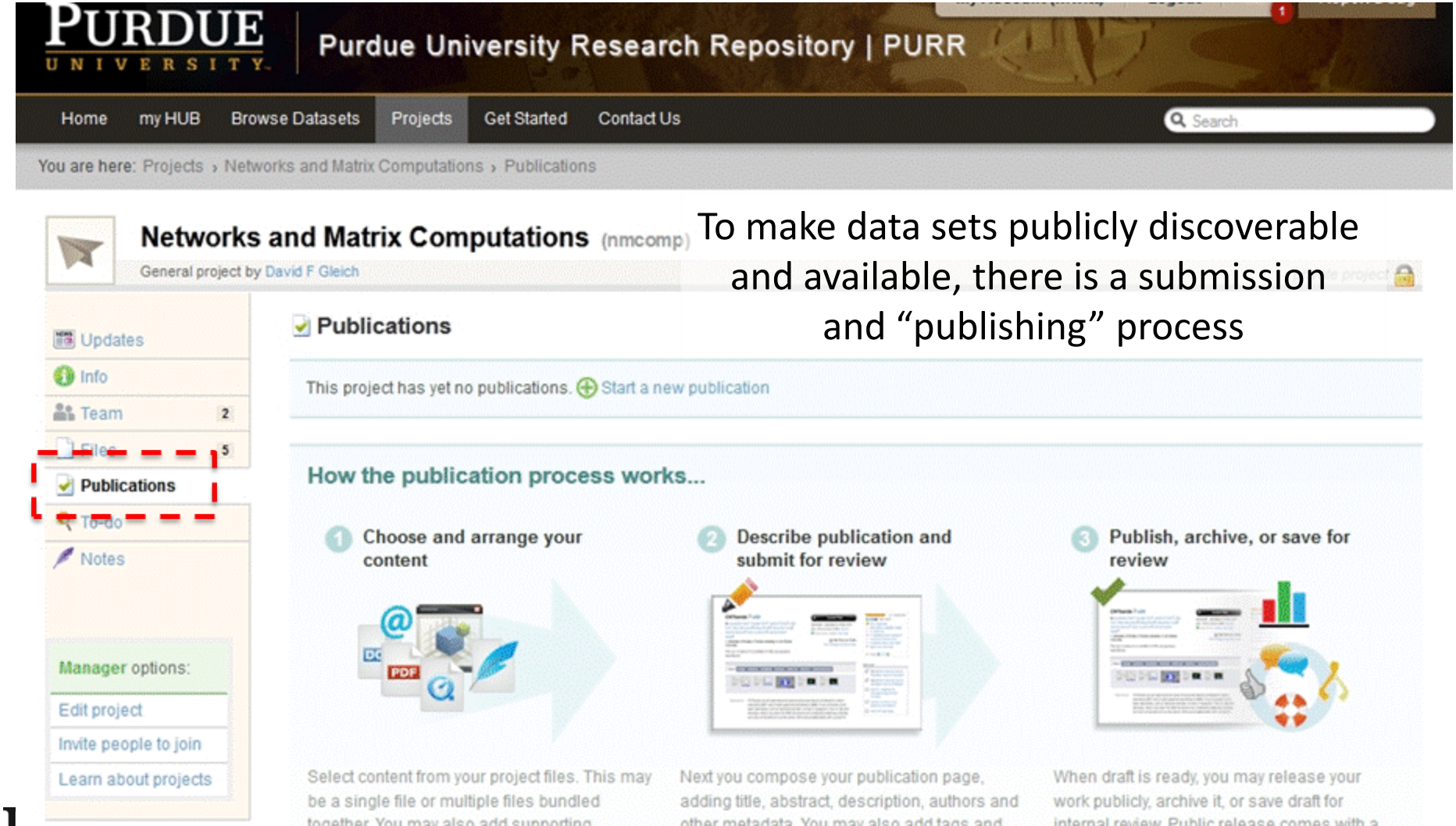
PURR functionality & service



To make data sets publicly discoverable and available, there is a submission and "publishing" process

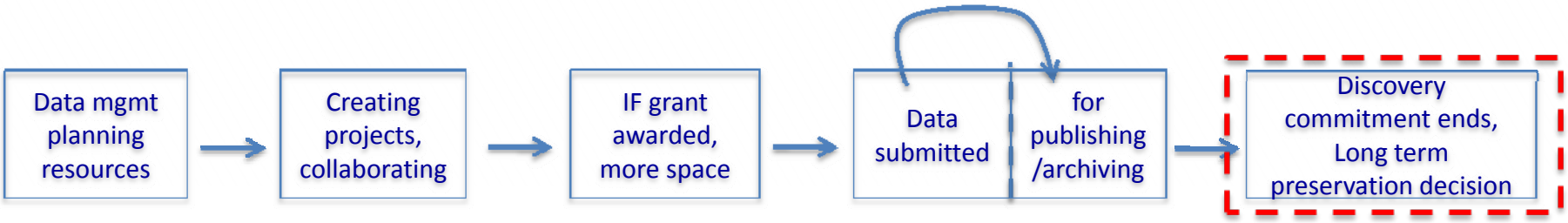
Librarians help shepherd and curate datasets.

PURDUE UNIVERSITY LIBRARIES



To make data sets publicly discoverable and available, there is a submission and "publishing" process

PURR functionality & service



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

Librarians provide education on depositing and finding data.

PURDUE UNIVERSITY LIBRARIES

General Library Services

Reference: assessing complex information needs

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

Instruction: enhance finding, evaluating, using research

- Explore activities related to creating and using information
- Partner with faculty to teach various illiteracies (info, data, etc.)

Collection mgmt: appraising local collections, including data

- Identify content for different dissemination modes
- Determine selection practices for new collections

Liaison: engaging researchers in new ways, in their environments

- Discuss research initiatives, projects, outputs…
- Collaborate on funded research projects— apply library science

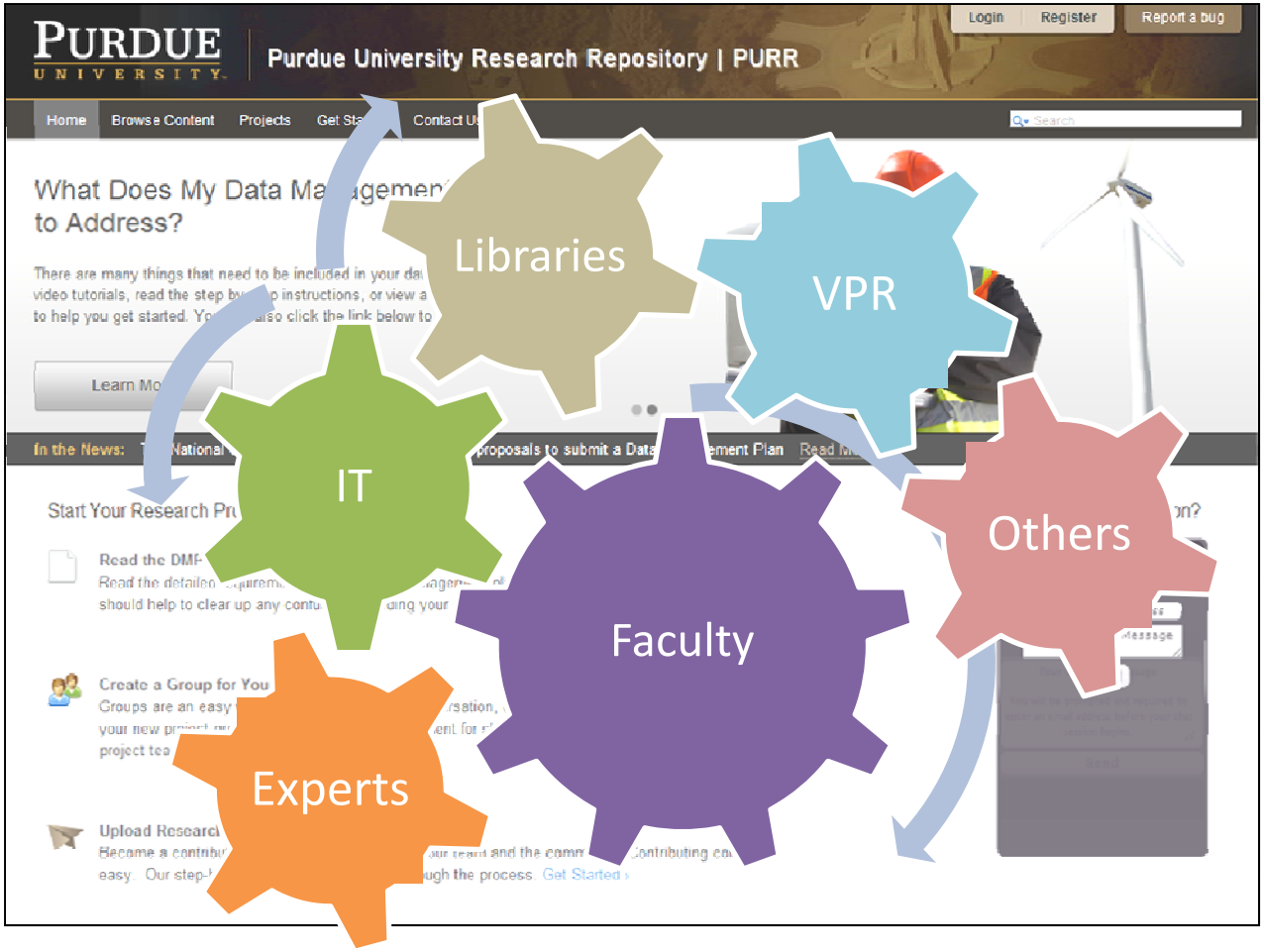
Specific Data Services

- Data reference
- Data mgmt planning
- Data consultation (may lead to collaborations/grants)
- Using PURR
- Promoting data DOIs
- Data mgmt education and information literacy
- Finding and using data
- Developing tools (DCP 2.0, DataBib, DMP-SAQ)
- Data visualization/GIS
- Developing data resources (LibGuides, tutorials)
- Linking data to articles and dissertations
- Promoting open access (Authors rights, IR deposit)*
- Leveraging publishing opportunities*
- Developing local collections*
- Collection mgmt of “e” (journals, data, archives)*
- Integrating systems * (i.e., finding data in Primo)

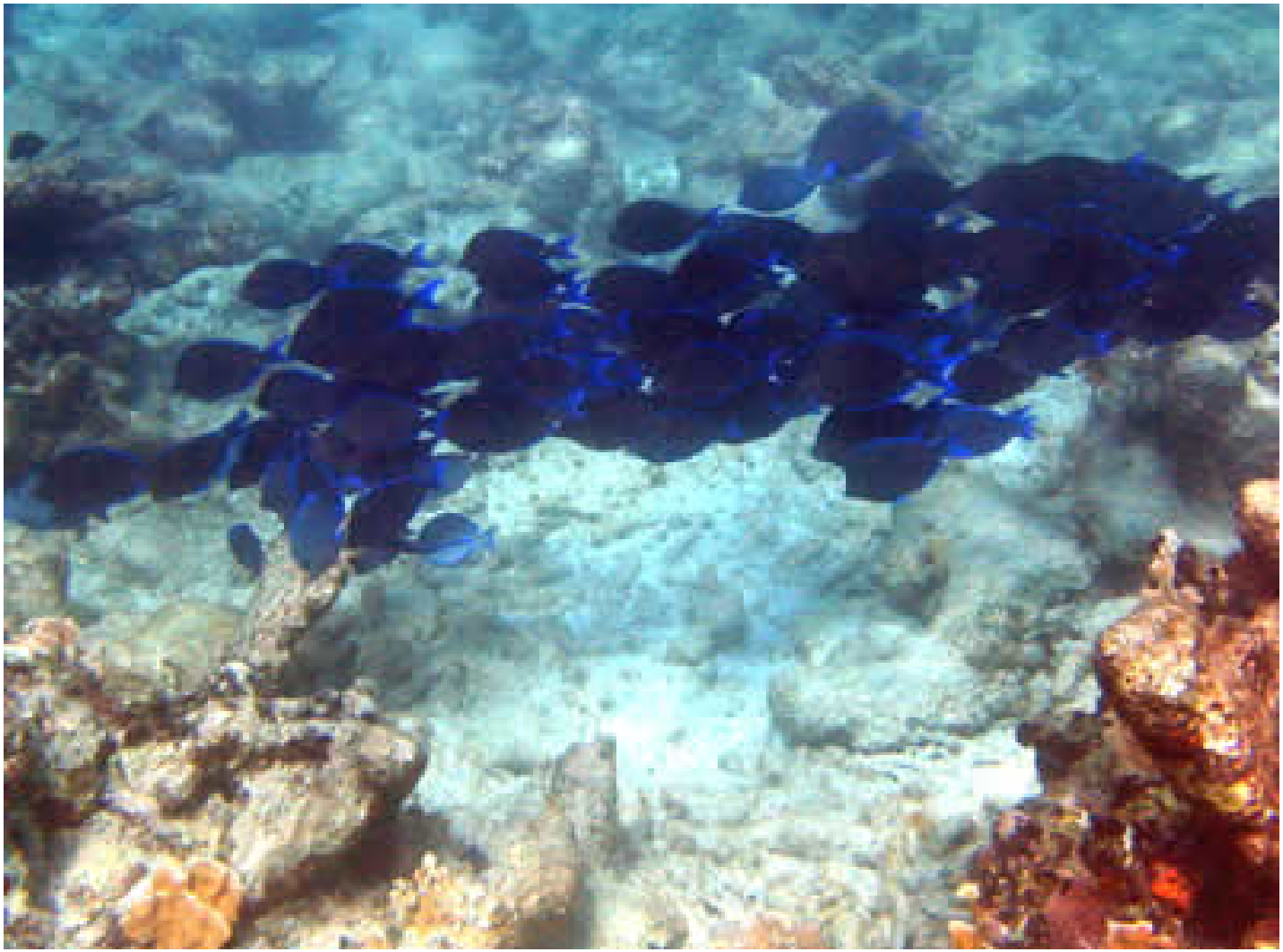
* As relates to data

PURDUE UNIVERSITY LIBRARIES

Data curation is a process based on partnerships & collaborations



PURDUE UNIVERSITY LIBRARIES



PURDUE
UNIVERSITY
LIBRARIES