PURDUE UNIVERSITY Libraries *Access. Knowledge. Success.*

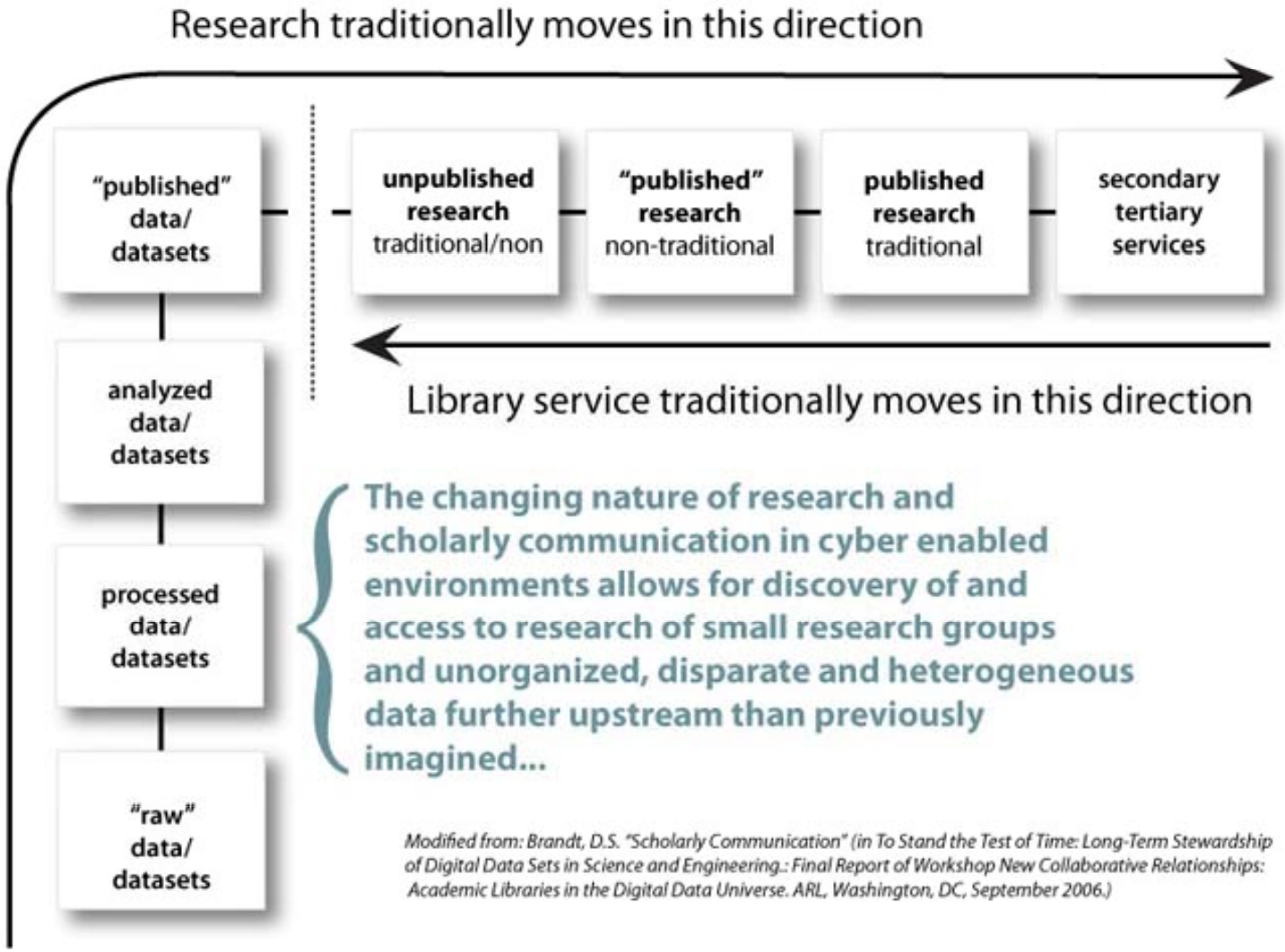
The Distributed Data Curation Center (D2C2)

Michael Witt
Interdisciplinary Research Librarian & Assistant Professor
Purdue University Libraries & D2C2
mwitt@purdue.edu

Sun PASIG Research Data Curation Working Group: May 29, 2008


D2C2 - Purdue University - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://d2c2.lib.purdue.edu/
Google
D2C2
Distributed Data Curation Center
PURDUE UNIVERSITY LIBRARIES
Access. Knowledge. Success.
Home
About Us
Vision
Advisory Board
Contact
Partners
Resources
Tools
Other Repositories
Activities
Projects
Wikis
Publications & Presentations
Events
What is D2C2?
Our Research
We investigate and pursue innovative solutions for curation issues of organizing, facilitating access to, archiving for and preserving research data and data sets in complex environments. Current D2C2 outcomes include interaction with the DIR which serves as a platform for investigation of data curation issues and development of applications to determine controlled access for data and to help solve data archiving and preservation problems that arise in several research domains, such as agriculture and energy.
News
The D2C2 has received the Sun StorageTek™ 5800 System courtesy of Sun Microsystems™. more...
Current Projects
Ingest, Preservation and Access for Water Quality Datasets in an Institutional Repository
Investigating Data Curation Profiles Across Multiple Research Disciplines
Metasearch Technologies for the NSDL Distributed Community
Investigate and Implement Persistence for HUB Resources
Featured Project
Investigating Data Curation Profiles Across Multiple Research Disciplines
Project goals are to enrich understanding of access to (or sharing of) data and related curation by conducting case studies of researchers' data practices, translating and comparing needs for archivin...
D2C2's philosophy is that data curation can be facilitated by a variety of services in a distributed environment...
"Data Curation is the activity of managing and promoting the use of data, starting from the point of creation, to ensure its fitness for contemporary purposes and availability for discovery and re-use."
(From: Lord, P. Macdonald, Lyon & Giaretta (2004) "From
Done
zotero

Library Science Opportunities in the Research Path



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.)

Distributed Institutional Repository: Purdue e-Scholar

1. **Purdue e-Pubs:** eprints, Digital Commons (bepress)
2. **Purdue e-Archives:** digitized archival collections, ContentDM (OCLC)
3. **Purdue e-Data:** research datasets, Fedora & ST5800, middleware

What librarians are asking our scientists to begin a conversation about data curation...

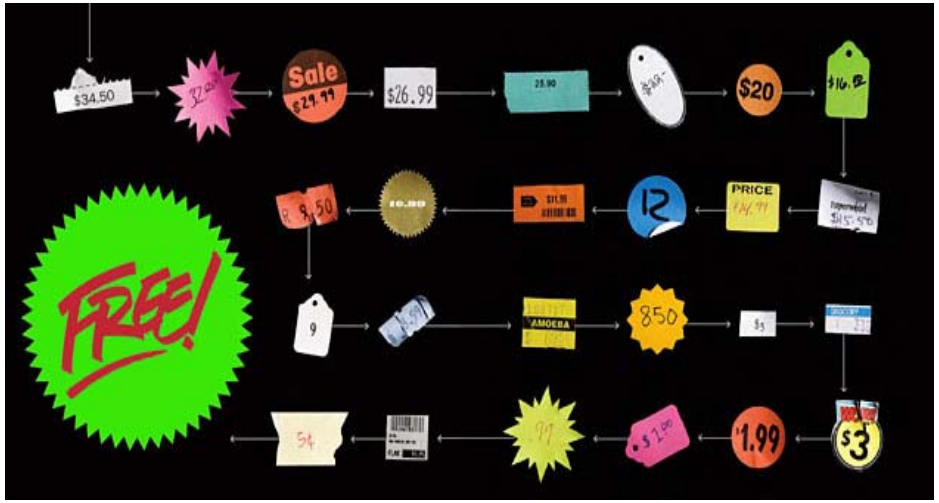
1. What is the story of your data?
2. What form and format are the data in?
3. What is the expected lifespan of your data?
4. How could your data be used, reused, and repurposed?
5. How large is your dataset, and what is its rate of growth?
6. Who are potential audiences for your data?
7. Who owns the data?
8. Does the dataset include any sensitive information?
9. What publications or discoveries have resulted from the data?
10. How should the data be made accessible?

Witt, M. & Carlson, J. (2007). Conducting a data interview. http://docs.lib.purdue.edu/lib_research/81/.

Some of the needs we're trying to understand…

Access
Persistence
Provenance
Ingest and scale
Intellectual property and permissions
Policies
Selection and appraisal
Metadata
Preservation

How can these things be free…?



- Web email
- Air travel
- Compact discs
- Digital video recorders
- Directory assistance

.

.

- Digital curation infrastructure?

Anderson, C. (2008). Free! Why $0.00 Is The Future of Business. *Wired, 16*(3). http://www.wired.com/techbiz/it/magazine/16-03/ff_free.

Thinking holistically, what is the bottleneck in the "data curation" system right now?

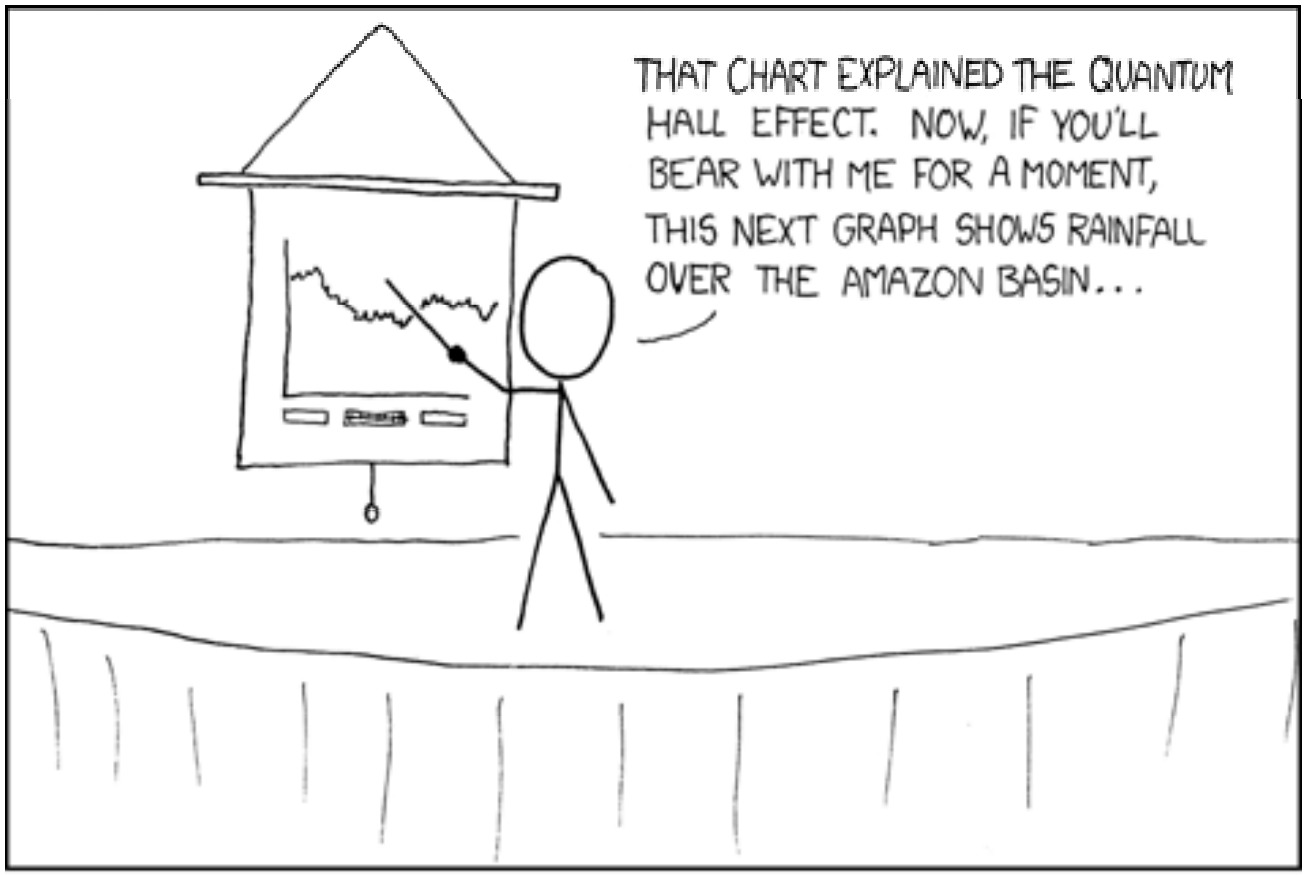
CPU? … not so much
RAM? … not so much
Bandwidth? … not so much
Storage? … not so much

Right now, we are the bottleneck: human beings

What are our roles in data curation? What resources do we need to execute and sustain them? What value do we add? How do we work together?

THANK YOU

PASIG Research Data Curation Working Group wiki:
http://tinyurl.com/4gyply



IF YOU KEEP SAYING "BEAR WITH ME FOR A MOMENT", PEOPLE TAKE A WHILE TO FIGURE OUT THAT YOU'RE JUST SHOWING THEM RANDOM SLIDES.

http://www.xkcd.com/365/