"Receptivity to Library Involvement in Scientific Data Curation: A Case Study at the University of Colorado Boulder," *portal: Libraries and the Academy*, Vol. 11, No. 4 (2011), pp. 915-937

---Katie Lage, Jack Maness, & Barb Losoff

Presented by Barb Losoff
University of Colorado Boulder
Science Libraries
April 2012

Presentation Overview

- **Motivation for this paper?**
- **Why Persona's as a methodology?**
- **What were the results?**
- **How to apply what was discovered?**


scientific
information
climate
science
research
data
new
many
access
researchers
analysis
visualization
sharing
model
knowledge
example
human
health
brain
challenges
results

E-science or E-research & Data

E-science “networked, data-driven science.”

---Tony Hey—Microsoft

Although mostly associated with ‘Big Science’ ex. Human Genome Project, the amount of data from ‘Small Science’ or laptop science is enormous.

Questions: Research Data at UCB & Libraries Role in Curation?

- Getting a snapshot of the research data produced at CU Boulder
- What do the researchers need?
- Libraries role? Harvard, Cornell, Georgia Tech?
- Common refrain in the Library Literature:

> "Librarians will have to embrace the role of data curator to remain relevant and vital to our scholars"—Joyce Ogburn

> "Data liaison services are a major component of libraries' future"—Tracy Gabridge

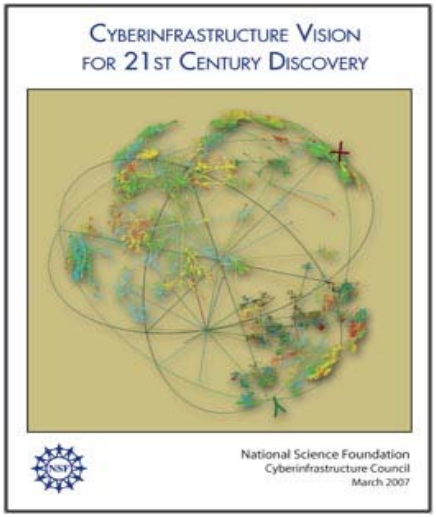
Letter of Invitation

- Short interview (15-30 min) regarding scientific data creation & use at the UC Boulder
- Conducting an organizational data inventory to gain an understanding of data production, use, storage & access
- Data inventory is motivated by both CU Boulder's proposed institutional repository & NSF mandates requiring grant recipients to archive and provide access to data

> *"Your contribution will inform the Libraries about data on this campus, offer insights for designing CU's institutional repository, and help define the role for the Libraries (if any) regarding data archiving, storage, and access."*

NSF & Data

- "All science & engineering data generated with NSF funding must be made broadly accessible and usable, while being suitably protected & preserved" (NSF 2007).
- "The new types of organization envisioned in this solicitation will integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise…" (NSF Cyberinfrastructure Grants 2008).
- "University-based research libraries and research librarians are positioned to make significant contributions in this area, where standard mechanisms for access and maintenance of scientific digital data may be derived from existing library standards developed for print material." (NSF Cyberinfrastructure Vision for the 21st Century, 2007).



Interview Questions

- Briefly describe your research.
- How long have you been conducting this type of research?
- Can you tell us a little about what sort of data your research produces?
- How is the data stored and accessed after it is produced?
- Who has access to this data?
- Does your department/lab have procedures in place for the preservation of researchers' data in the event they leave the university or pass away?
- Is storage space problematic?
- Would it be of interest to you for those responsibilities to be transferred to an entity within the university, such as the Libraries?
- Please rate on scale of 1-5 your receptivity to this question, 1 = least interested, 5= very interested.

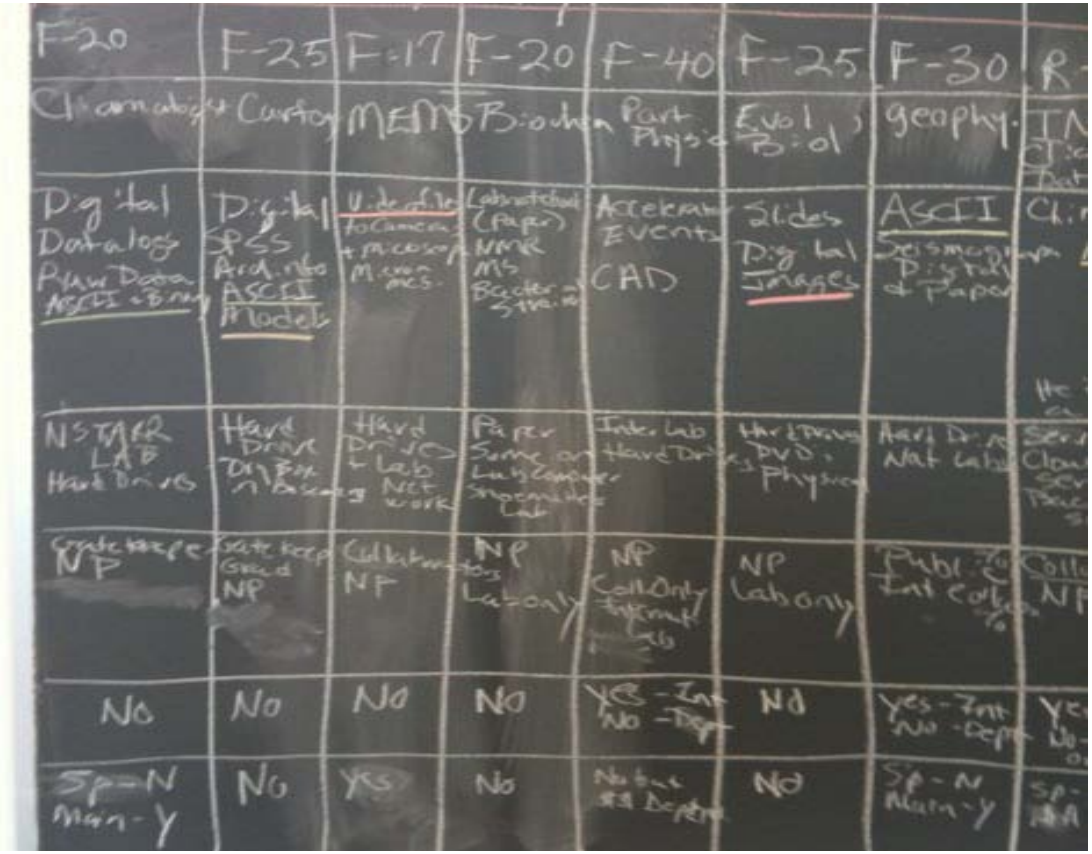
Researcher's Subject Areas

- Aerospace Engineering
- Civil, Environmental, & Architectural Engineering
- Mechanical Engineering
- Computer Science
- Ecology & Evolutionary Biology
- Geography
- Geological Sciences
- Chemistry & Biochemistry
- Molecular, Cellular, & Developmental Biology

Persona Definition

- Fictionalized aggregates of actual potential users of a product
- Allow designer to target users needs holistically in the design process by keeping the user front and center
- Driven by data collected from actual or potential users
- Interviews are a common methodology
- Conflate users with mutual needs into single persona with name, biographic info., intent of use of the product and an image for enhanced empathy
- Anonymous findings that are generalizable

35 Faculty invited, 24 Participated
74% Response Rate



Results of Survey

- Data ranged from paper notebooks, to digital ASCII files from field instruments, to large video and image files
- Data stored on researcher hard drives, lab servers, removable storage devices (CDs) and nat. & internal. discipline specific repositories
- Most said research data was non-public (20 out of 26) but share with others on as-need basis
- Digital storage space was not an issue, however server maintenance and management are on-going problems

Results of Survey

- Many researchers had curation plans for their data
- Many had orphan data without curation plans
- Few departments had procedure for data preservation, some participated in disciplinary based repositories supporting long-term storage
- Receptivity to a library role in data curation fell more in-line with the researchers disciplinary culture or philosophy regarding data sharing and collaborative projects.

Interest in the Library Role for Data Curation

Researchers expressed that:

- Data must be easy to access
- Researchers would maintain a level of control over who had access to the data
- Ease of use is a main requirement for a library-run repository
- All researchers expressed a strong reluctance to participate in a repository that was designed in a manner that did not fit their needs, requiring extra work on their part. It is imperative to engage researchers on their needs in designing a repository

Judy McDannell, "Very interested, has no support"

Very interested, has no support

Name: **Judy McDannell**
Age: **50**
Research: **Chemical reactions relating to the origins of life**

Judy McDannell is an associate professor in the Department of Molecular, Cellular, and Developmental Biology, at the University of Colorado at Boulder. Professor McDannell completed a PhD in biochemistry from Johns Hopkins University School of Medicine in 1988, followed by a post-doc at Stanford University. She has been a faculty member at the University of Colorado at Boulder for eighteen years. She runs a state-of-the-art laboratory conducting research on the chemical reactions that could be responsible for the origins of life. Professor McDannell consistently brings in NSF grant money in the hundreds of thousands of dollars which has helped her over the years to employ quality graduate students.

Professor McDannell's lab produces a variety of data: print laboratory notebooks, Nuclear Magnetic Resonance (NMR) or Mass Spectrometry (MS) data, chromatograms, and bacterial strains. Most of the data generated from her lab is in the form of paper laboratory notebooks. The spectral data produced in the laboratory is stored on computers in the lab and also stored on the hard drives in the NMR lab for the University. Bacterial strains are stored in the freezer with detailed information for each strain located on Judy's hard drive. Storage space is not a problem; however relying on paper lab notebooks is a major concern.

Professor McDannell is enthusiastic about a shared responsibility for her data. She is receptive to a system that would assist in organizing her data collections and wants desperately to move to a shared electronic management system. Professor McDannell simply does not have the time to do all this herself and does not have the departmental support either. Because she has "nothing" she would welcome an opportunity to share the responsibility of her data management with a trusted entity such as the library.

Chen Ming, "Very interested, space issues, open to data sharing"

Persona 2

Very interested, space issues, open to data sharing

Name: **Chen Ming**
Age: **28**
Research: **Systems ecology; Non-point source pollution**

Chen Ming is a PhD candidate in physical geography. He has a passion for teaching that began at MIT, where he obtained his BS and MS in the first class to graduate in the new 5-year Bachelors to Masters environmental engineering program. During his fourth and fifth years he was the primary teaching assistant for the core ecological engineering courses. Ming's specialty is hydrology and systems ecology. He was interested in continuing his studies in the Department of Geography in order to include the study of human-induced changes to ecological systems. His area of research is non-point source pollution generation, focusing on mercury.

Ming's research produces ASCII files, imagery (.jpg files), and simulations (MATLAB files and movie files). For his analysis, he manipulates some of his data using spreadsheets. His data are all electronic. Ming stores his data on the small network of computers in his advisor's lab. He also uses Google Docs to store and share spreadsheets with his students and collaborators. The lab network is maintained by another graduate student. The computers are backed-up, but occasionally there are technical problems, and support for resolving technical issues can be problematic. Space is never a problem for the smaller ASCII files and spreadsheets, but he is always running into storage space problems for his larger imagery and simulation files—the core of his research data.

Ming is very interested in a library-run data repository. He sees it as a solution to his space problems and as a means to more easily share data with colleagues. He would want the interface to be simple and the system to be robust and to support a variety of file types and access modes. His data is not all open access—some of it may be patentable or not able to be shared at all levels, so he would need the ability to control who has access to the data.

Professor Mel Hampton, "Interested, has robust support (graduate students), however maintenance is a problem"

Persona 4

Interested, has robust support (graduate students), however maintenance is a problem

Name: **Professor Mel Hampton**
Age: **49**
Research: **Climate changes at the land-atmospheric boundary**

Mel Hampton is a professor in the Department of Geography and has been a faculty member at the University of Colorado at Boulder for ten years. Professor Hampton completed his PhD at Stanford in 1993, followed by a post-doc at Harvard. For his sabbatical, Professor Hampton spent a semester in Spain at the University of Vigo in Ourense writing a book on the trends in climate change. He collects data at the land-atmosphere boundary at locations in the U.S. and abroad.

Professor Hampton is comfortable in the digital world using products that translate ASCII to binary formats. Most of his data is raw, collected from data recorders and loggers at the various collection sites. He also has digital data that is produced by his graduate students using a variety of instruments stored on his hard drive. Professor Hampton is not concerned with data storage since hard drives are cheap; however changing technology is an issue. Another concern is his reliance on NSF grants or soft money to support his research. If the NFS money were to disappear he would not be able to employ the research scientists or the number of graduate students that currently assist him in managing the data.

Professor Hampton is interested in exploring a shared responsibility for his data with the library. At the moment, he has enough graduate students and funding to support in-house data management, but the sheer growth of data will soon outpace his ability to manage it. Professor Hampton finds that he is spending more time in the day-to-day management of the data, which impacts the time available for his research. He is interested in the possibility of a trusted entity, like the library, to provide backup, archiving, and searchability across diverse data sets.

Positive Correlation—Receptivity to Library

- Close prox. to data curation activities, those sharing responsibilities for the task
- Lack of existing curation support either disciplinary repository or departmental
- Personal ideology disposed toward sharing, sharing data as part of mission & social obligation as scientists
- Earth Sciences—geologists, environmental engineers, share disciplinary culture more likely to lend itself to partnering with librarians, other labs and researchers

Negative Correlation—Receptivity to Library

- Research involves proprietary data from funding agency that require non-disclosure agreements
- Inability to share data bec. of ethnographic research or using human subjects
- Extremely competitive field or disciplinary culture that discourages outside involvement
- Existing repository, discipline-based

By offering data curation services to researchers who share traits with the personas, libraries can expedite the partnerships necessary to begin fostering a new form of library service.

Applying the Personas

- Data Management Task Force which from Research Computing and is chaired by the Associate Vice Chancellor for Research—membership
- Laid the groundwork for much of the strategic planning, resulting from participating in the ARL (Association of Research Libraries) E-Science Institute for developing Data Management Plans
- New IR and how to populate—seek partnerships with researchers sharing the traits of the personas with positive correlation as the early adopters.

Relationships

Images

1. http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2011_02_11/caredit.a1100013 (Cloud, Data, cover *Science*)
2. http://kdpaine.blogs.com/themeasurementstandard/2012/03/todays-data-deluge-makes-measurement-anything-but-easy.html?cid=6a00d83451658a69e20168e998f5de970c (Laptop with Wave)
3. http://www.nsf.gov/pubs/2007/nsf0728/ (Cover NSF Cyberinfrastructure Report
4. http://sirls.arizona.edu/ (Cloud, Digital, Data Curation, Libraries)

Questions?