Scientific Data Stewardship: Meeting the Challenge in Academic Libraries

Barb Losoff
Science Library
University of Colorado, Boulder, CO
March 2009
E-research—Data Intensive

• **Data Deluge**—both Big and Small data (Rhoten 2007)
• **Dark Data**—“doesn't yield a dramatic outcome — or, worse, the opposite of what researchers had hoped? It ends up stuffed in some lab drawer” (Goetz 2007)
• **Granting agencies** requiring access to archived data with a public emphasis access
Data

“Data are becoming an important end product of scholarship, complementing the traditional role of publications.” (Borgman et al 2007)

Human Genome Project data sets “have more value than any single publication that was derived from an analysis of them” (Witt in Carlson 2008).

“Opening up science data could speed discoveries, increase collaboration, and transform the field in unforeseen ways.” (Johnson 2008)
Data is the Driver

- Archivable
- Searchable with intelligent integration (metadata)
- Mined in conjunction with publications & other resources
- Public access
Data Controversy

- Young researchers sharing data—biological engineer at MIT posted raw data for thesis
- Open notebook—organic chemists, Drexel, PA & Univ. of South Hampton, UK
- Who’s data is it anyway? Being scooped (Helly et al. 2002)
- Promotion & tenure? Intellectual property rights?
- Lack of positive incentives to publish data—withholding data as if owned exclusively by researcher (Helly et al 2002)
Institutional Repositories (IR)

“university-based IR is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution.” (Lynch, 2003)
IRs

• **Curation**- “maintaining and adding value to a trusted body of digital information for current and future use; specifically…the active management and appraisal of data over the life-cycle of scholarly and scientific materials” (Gold 2007)

• **Access**

• **Quantitative measure** of productivity & invite dialog with tax payer “unlike many other enterprises, universities do a poor job of letting investors know what they get for their money” (Shulenburger 2008)
IR Controversy

- IRs born of ideology, OA & technology, and did not emerge from faculty needs (Weibel 2009)
- Low population—too passive in collection activities, “large digital card-board boxes without any specific mission, which faculty, unrealistically, were expected simply to fill.” (Salo 2008)
- Fuzzy on how to include data sets this comes from Boston University recent announcement for open access by faculty (Howard 2009)
IR Controversy

Peter Murray Rust (in Rusbridge 2007)

“I got the distinct impression that some IRs are run like Victorian museums—look but don’t touch. The very word repository suggests a funereal process—it’s no surprise that having put much of my stuff into Dspace I find it’s an enormous effort to get it out. Why don’t we build disseminatories instead?”
Libraries & IR

• 40% of the 250 doctoral granting institutions have some form of IR
• 88% of those that don’t have an IR are in planning process
• 80% of those with an IR indicated the library has the sole administrative responsibility (Lynch 2005)
Libraries Role & Data

Granting agencies requirements:

“all science & engineering data generated with NSF funding must be made broadly accessible and usable, while being suitably protected & preserved” (NSF 2007)

NSF Cyberinfrastructure Grants:

“The new types of organizations envisioned in this solicitation will integrate library and archival sciences, cyberinfrastructure, computer & information sciences, & domain science expertise…..” (NSF 2008)
Purdue University Libraries

Identified what researchers WANTED:

1. Consistent access to their data
2. Informed access to data of colleagues
3. Share data through distributed access or duplication of data-sets
4. Help in gaining this access

--James Mullins 2006
Purdue University Libraries 2004 Initiative

Libraries faculty collaborate across campus, discover problems & apply knowledge manage. expertise.

HOW?
1. Attended seminars, learned the language
2. Built relationships
3. Identified problems & collaborated solutions
Purdue University Libraries
Successes

In 2006 success with 3 proposals & several in the works
Subject areas for collaboration included:

• Biology
• Chemistry
• Cancer Center
• Earth & Atmospheric Science
• Mechanical Engineering Tech. & Chemical Engineering
• English
Disciplinary Investigations

Small, Kushmerick & Benson (2008)

• Majority of scientists viewed their work as having some degree of social or political relevance based on question grouped by discipline of the researcher’s highly cited paper

• Surprising outcome: immunology & molecular bio. were less than 50%, chemistry 65%, geoscience, microbiology were 100%
Disciplinary Investigations
Borgman, Wallis and Enyedy (2007)

Habitat ecology data --- use by scientists, computer scientists and engineers

• Both sci and engin. teams used data in their research questions, but they did so for diff. purposes and diff. levels of granularity
• All felt that hand-collected data was ‘hard-won’ and less likely to share
• All would share scientific variables data
Disciplinary Investigations
Borgman, Wallis and Enyedy (2007)

“Found that researchers generally favor sharing data from their research, but do not agree on what those data are, on the conditions under which data should be released, or who has the authority to release them.”
Zuber (2008) compared IR content by discipline. Those disciplines that were early adopters of pre-prints, author self-archiving, etc. tend to populate IRs.

“Librarians, knowing national tendencies as well as philosophical positions with regard to open access, can more effectively solicit content by knowing beforehand which disciplines tend to embrace or reject IRs.”
Local Lens

Developing action plans will require a local lens, whereby librarians explore their role at the institutional level to discover researchers data management needs.

DePaul University “managers are taking small bites: starting repositories with a specific project.” (Albanese 2009)

Position paper John Hopkins: “it’s essential to develop prototype systems that demonstrate both technical and organizational infrastructure to support data curation” (Choudhury et al 2007)
National Snow & Ice Data Center

NSF grant developed with the University Libraries, University of Colorado at Boulder, as trusted partner

Collaborative venture both across campus & with other universities

Libraries are critical partners for preservation and access

“Prominence of the institution will drop without the libraries involvement.”
Best Practices

• Make contact with key researchers (Rusbridge 2007)
• Needs assessment
• Encourage curation into science workflow and keep data. (Rusbridge 2007)
• Create policies about access to what data, what form, and under what conditions (Borgman et al 2007)
• Outreach to populate IR
• Advertise most popular paper/data set each month
• Create simple metadata (Allard et al 2005)
• Train authors and the public to use IR (Allard et al 2005)
Best Practices

• What type of data is produced?
• How is data used?
• How is the data accessed?
• Is anyone in the department responsible for curation?
• What software is used?
• How is the data managed?
• Should the data be shared?
• What is the criteria for sharing & who can authorize sharing?
• Is there metadata and/or associated ontologies?
• Are there other resources (publications, teaching materials, etc.) that would be useful to integrate?
Why Librarians?

- Generalists—even as subject specialists, librarians offer a broad overview of research campus-wide
- “Natural role in building & supporting information infrastructures of local data centers—a role that draws on and adapts the mediating practices of the library profession to the world of data.” (Gold 2007)
- Librarians can be the connection between the IR and users (Jenkins 2005)
Why Librarians?

• New role in Knowledge Management (emphasizes the human side of knowledge) "offers academic libraries the opportunity to create knowledge to improve organizational effectiveness, for both themselves and their institution." (Townley 2001)

• Role of collection developer—readjusted from selector to expert helping university communities define what to collect in the IR (Allard et al 2005)
Why Librarians?

• Librarians provide professional expertise ensuring the principles guiding academic education and scholarship prevail
• Librarians are knowledge disseminators, supporters of open access and open data
Conclusion

Whether attending data management seminars or conversing with researchers over coffee, librarians need to learn the language of data. As bibliographers, librarians liaise with departments, providing that opportunity to dialog with researchers about their data management needs. “Designing a new model for academic communication requires collaboration from the authorities, from universities, from librarians and also support from researchers themselves.” (Bravo & Diez 2007) As professional information mediators, librarians must explore this new role as they forge collaborative relationships in the pursuit of data stewardship at the local level.


References continued


References continued


References continued


References continued


References continued


