WHERE HAVE ALL THE SCIENTIFIC DATA GONE?
LIS PERSPECTIVE ON THE DATA AT RISK INITIATIVE

Cheryl A. Thompson[1], W. Davenport Robertson[2] and Jane Greenberg[2]

[1] Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign | [2]School of Information & Library Science, University of North Carolina at Chapel Hill

ABSTRACT

In the quest for knowledge, scientists produce vast amounts of data and often these data are not properly preserved or archived. **Data at risk** are often fragile, deteriorating, insufficiently described or simply unknown to other scientists. Data that are old often retain scientific value. Future research will be hampered if valuable, historical scientific data are ignored or lost.

In conjunction with CODATA, the UNC Metadata Research Center's Data At Risk Initiative (DARI) aims to mitigate the risk of loss. DARI has conducted **a web survey of information custodians** to understand the data at risk predicament from their perspective. Librarians and information professionals working in science, research or other special libraries offer a unique viewpoint given their position in the organization and their LIS training. This poster presents the survey findings.

METHOD

We conducted a survey of LIS professionals who are involved in any aspect of data curation. The survey collected information on the data at risk such as type, risk level and future plans for these data. Data sharing practices and the demographics also were captured. A survey invitation was sent to the discussion lists of the ASIS&T, SAA, SLA and ACRL. Forty-three information custodians completed the survey. Quantitative analyses have been conducted.

RESULTS

Demographics of respondents

Gender: 65% female

Age category: median 46-50 year

Years in current position: mean 7.6, sd. 8.9

Education: 53% Masters in LIS, 33% Masters in other

- **Professional identity:** 52% identified as an information professional, 38% identified as a librarian, 31% identified as a scientist and 28% identified as an archivist*
- **Work setting:**
  - 52% academic institutions
  - 17% corporations
  - 7% government agencies
  - 3% health/medical centers
  - Other institutions: research and cultural heritage centers

Credit: http://johnkingworld.com/

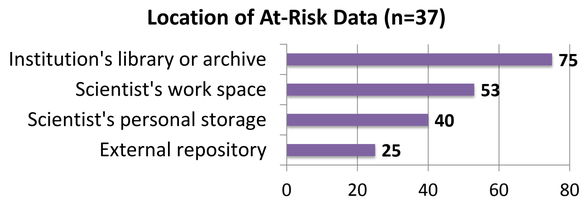
Data at risk characteristics

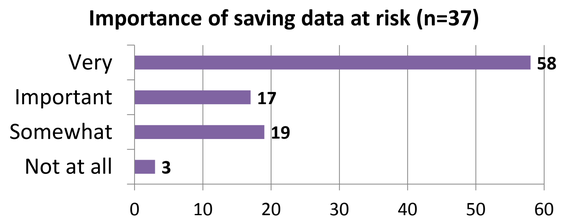
- **Research area:** A variety of research areas were reported. The most common were geology (42%), biology (39%) and climate (39%) *.
- **Formats**: There were a wide range of the formats. The most common were non-digital text documents (67%), handwritten notes (65%), digital files (58%), CDs (56%) & floppy disks (54%) *.

**Respondents were able to check all that apply. The percents will not equal 100.*

Data practices

- **Metadata**: 54% have a data catalog using metadata. Metadata standards used were Dublin Core, FGDC, SPASE, DACS, DIF, FITS and local/institutional standards.
- **Data management:** 44% comply with a data curation standard. Started reported were OAIS, PREMIS and funder requirements.
- **Ownership:** 73% institution, 41% funding agency, and 32% government*.
- **Data sharing:** 76% make a version of the data publicly available. The most common limits to data sharing were time involved in making data usable (73%) and accessing files from storage (59%).





Data At Risk Inventory

DARI is creating an inventory of valuable scientific data that are at risk of being lost to posterity. It is not a repository for data. It is a descriptive inventory of endangered data that are held by others: individuals and research institutions. A goal of DARI is to design rescue efforts to save these data. To contribute to the Inventory, please submit a description at:
http://ibiblio.org/data-at-risk/contribution

Custodians' data responsibilities

Data duties (n=35)	Complete or shared responsibility
Determining the appropriate metadata to describe data sets (i.e., descriptive information to enable others to reuse data)	68%
Determining provisions for short-term data preservation (5 years or less)	57%
Determining provisions for long-term data preservation (more than 5 years)	56%
Deciding which data are important to preserve	47%
Determining what constitutes compliance with commercial licenses, government regulations, funding agency mandates, etc.	43%
Deciding whether data can be safely shared	33%
Determining standards for de-identifying sensitive data	31%

CONCLUSION

A diversity of data at risk was found. These findings shed light on a topic of growing concern and they are relevant for research institutions and preservation planners. The understanding of the data at risk predicament can assist librarians, archivists and scientists in designing and funding data rescue efforts that are successful. An understanding of the data at risk predicament will enhance educators' ability to prepare and mentor students who want to pursue careers in data curation.

Acknowledgements | We would like to acknowledge the support of Paul Jones & John Reuning at ibiblio, UNC SILS Carnegie Fund and the UNC Center for Global Initiatives for the work presented here.