THE DATA-AT-RISK INITIATIVE: ANALYZING THE CURRENT STATE OF ENDANGERED SCIENTIFIC DATA

Angela P. Murillo*, Cheryl A. Thompson, Nico Carver, W. Davenport Robertson, Jane Greenberg
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599
amurillo@email.unc.edu

William L. Anderson
School of Information
University of Texas at Austin
Austin, Texas 78712
band@ischool.utexas.edu

ABSTRACT
Examining fragile scientific data is crucial to ensuring that these data are not lost and can continue to be part of the scientific process. In order to investigate various aspects of endangered scientific data the Data-At-Risk Initiative (DARI) project has conducted three studies that provide insight into the current state of scientific data-at-risk. The three projects include (1) a metadata/system review of the of the to the Data-At-Risk Inventory, (2) focus groups with fourteen scientific scholars, and (3) surveys completed by 43 information custodians. This poster reports results from these studies, specifically how scientific scholars and information custodians view endangered data. The DARI project aims to understand the current state of endangered scientific data and to assist in the reduction of loss of valuable scientific data.

Keywords
Scientific data, endangered data, data at risk, data rescue.

INTRODUCTION
Valuable and unique scientific data are increasingly at risk of being lost forever due to deterioration, format obsolescence, and insufficient metadata for discovery and retrieval. The Data-At-Risk Initiative (DARI) is a project designed to understand the extent of this growing problem and to take action by helping the data rescue.

Over the last year the DARI research team has conducted three studies targeting metadata design, and scientists' and information custodian's perceptions about, and knowledge relating to, data at risk. This paper provides background for the DARI project, outlines the guiding research questions, reviews the research study methods, and presents some conclusions. The poster provides more detailed reporting of the studies.

ASIST 2012, October 28-31, 2012, Baltimore, MD, USA.

BACKGROUND
DARI is a collaboration among the Committee on Data for Science and Technology (CODATA) Data At Risk Task Group (DARTG); the Metadata Research Center (MRC), including the Center's supported DARI-SILS Student Learning Circle; and ibiblio. The DARTG defines Data-At-Risk as scientific data which are not in a format that permits full electronic access to the information which they contain. The data can be non-digital (paper, film, etc.), on near-obsolete digital media (magnetic tapes), or insufficiently described (lacking metadata). Data that are regarded as unusable are often considered useless and risk being destroyed, and thus their scientific content is lost. Most data at risk pre-date the digital era and can complement existing databases by extending the time-base. Some born-digital data can also be considered "at risk" if they cannot be ingested into managed databases due to the lack of adequate formatting or metadata. These data contain unique observations and information that are essential for studies of historical trends and have the potential to be combined with other data for new, interdisciplinary studies.

The data described above often retain significant scientific value and are crucial for future research. Examples of creating new scientific knowledge from old data collections are becoming more prevalent in the research literature and news (Griffin, 2005a; Griffin 2005b; Krotz, 2011; Rudin et al, 2011).

RESEARCH QUESTIONS
DARI aims to mitigate the risk of losing valuable scientific data through an assessment of the current state of at-risk data collections. In order to analyze the current state of this data, DARI has been investigating the below questions.

1. What is the most practical and effective metadata structure for a contributor-base data repository documenting data at risk?

2. What perceptions do scientists have on the topic of data at risk?
3. What perceptions do information custodians have toward data sharing and their knowledge and understanding of data at risk?

These three questions have guided DARI research and development and provide vital information for the understanding of endangered data, data reuse, and data sharing and inform the continuing effort to ensure that these valuable data are not lost.

METHODS

To address the guiding research questions several methods were used. These are outlined here:

- The Data-At-Risk Inventory development and metadata assessment including engaging scientists in metadata creation, and a team review and solicitation for participatory feedback from the formal CODATA-DARTG, which includes sixteen members (scientists and informaticians) http://ils.unc.edu/~janeg/dartg/).
- Focus groups with scholars in selected scientific disciplines to study their perception of endangered data, data reuse and sharing.
- Survey of information custodians at scientific centers and institutions to learn about the state of their institutions' data and curation practices.

CODATA charged the Data At Risk Task Group DARTG to inventory endangered, valuable data. During the summer 2011, UNC SILS graduate students affiliated with the SILS Metadata Research Center designed an Inventory prototype, and conducted a case study. Inventory development included soliciting feedback from DARTG and DARI members. A pilot test was conducted with scientists to identify additional refinements to the inventory. Updates were made based on feedback. Inventory development continues though eliciting feedback from the community and populating the inventory.

Secondly, focus groups were conducted with selected scientific scholars. The purpose of this activity was to gain an understanding of scientists' perceptions of data reuse, data sharing, and endangered data; and the capabilities of the Data-At-Risk Inventory. Participants for these focus groups were faculty, post-doctoral researchers, and PhD students from the University of North Carolina at Chapel Hill and Duke University. Participants were recruited through email listservs. A total of four one-hour focus groups were conducted. A total of fourteen participants took part in the focus groups. The participants were asked a variety of semi-structured questions including the types of data they use in their research, their perceptions towards data reuse and data sharing, their perceptions towards endangered data, and their opinion of the Data-At-Risk Inventory. All focus groups were audio recorded, fully transcribed, iteratively open-coded to examine emerging patterns, a codebook was created, and an 85% inter-coder reliability was reached.

Thirdly, a survey was conducted with information professionals to identify collections of research data known to be at risk. The survey was intended for librarians, archivists and information custodians who are involved in any aspect of data curation. The survey collected information such as type, format, volume, risk level, reason for risk, and future plans for these data at risk. The survey also examined the extent to which the collections have been cataloged or indexed using metadata, attitudes towards sharing data, and the professional qualifications of custodians. Participants were recruited from the American Society of Information Science and Technology, the Society of American Archivists, the Special Libraries Association, and the Association of College and Research Libraries. Forty-three (43) information custodians completed the survey. Data from this survey continues to be analyzed.

RESULTS

Data-At-Risk Inventory

Feedback from scientists, and DARTG and DARI members, have informed Data-At-Risk Inventory updates (Thompson, et al, 2011). The cycle of soliciting feedback and updating the inventory is an ongoing process. Focus group participants were also asked questions in regards to their opinion of the current inventory. These data will also be used to make updates to the inventory as the inventory continues to be developed and populated. The records are presented in a simple template in Omeka, based primarily on Dublin Core. Overall scientists have suggested that the inventory is simple and efficient.

Focus Groups

DARI members conducted focus groups with scientific scholars in order to gain an understanding of their perceptions of data reuse, data sharing, and endangered data. Participants discussed a variety of topics in relation to engendered data. These topics included data curation, data reuse, data sharing, data types, endangered data, the data-at-risk inventory, and priority data. Results indicated that that scientists view endangered data through personal and discipline perspectives including lack of context and inaccessibility issues. The results also indicated a range of opinions although all participants were concerned with the possibility of losing data and all recognized the complexity of the data-at-risk predicament.

Survey

DARI members conducted a web survey of information custodians to examine their perceptions of endangered data. 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. The survey results indicated that there is a diversity of data at risk in terms of domain science, format,

and ownership status. While about half of data at risk complies with a metadata or data curation standard, a variety of standards were reported. The survey highlights areas where information custodians can help resolve data at risk issues.

CONCLUSION

This poster paper provides of an overview of the Data-At-Risk Initiative (DARI) and the three current projects that the initiative in undertaking: 1) the Data-At-Risk Inventory 2) focus groups with scientific scholars 3) surveys of information professionals.

Research on matters relating to DARI are important for understanding how to inventory fragile and at risk research data. The results of these studies are informing the DARI about the current state of endangered scientific data. This work is providing the background needed for continual improvement to the inventory, as well as adds to the important discussion how to ensure endangered scientific data is not lost to prosperity.

ACKNOWLEDGMENTS

We would like to acknowledge the support of the SILS Carnegie Fund, the UNC Center for Global Initiatives, and CODATA.

REFERENCES

Data-at-Risk Inventory. (n.d.). Retrieved June 10, 2012, from http://www.ibiblio.org/data-at-risk/.

Griffin, R. E. (2005). Rescuing and recovering lost or endangered data. CODATA Data Science Journal, 4, 21-26. doi:10.2481/dsj.4.21.

Griffin, R. E. (2005). The Detection and Measurement of Telluric Ozone from Stellar Spectra. *Publications of the Astronomical Society of the Pacific, 117*(834), 885- 894.

International Council for Science: Committee on Data for Science and Technology. (2011, March 2) *CODATA Data At Risk Task Group (DARTG)*. Retrieved from http://ils.unc.edu/~janeg/dartg/.

Krotz, D. (2011). From Dusty Punch Cards, New Insights Into Link Between Cholesterol and Heart Disease. Retrieved from: http://newscenter.lbl.gov/featurestories/2011/01/04/cholesterol-heart-disease/.

Rudin, C., Passonneau, R.J., Radeva, A., Jerome, S., & Isaac, D.F. (2011) 21st-Century Data Miners Meet 19th-Century Electrical Cables. *Computer, 44*(6), 103-105. doi:10.1109/MC.2011.164.

Nordling, L. (2010). Researchers launch hunt for endangered data. Nature, 468: doi:10.1038/468017a.

Thompson C.A., Carver N., Collins K., Sinclair J., Veitch, J.M. (2011). Supporting scientists in data archiving: Emerging roles for information professionals. Poster presented at Digital Liaisons: Student Perspectives on Curating the Information Life Cycle session at the American Society of Information Science & Technology annual conference, New Orleans, LA.