CODATA Abstract for 2012 conference

CODATA 2012, Taipei, Taiwan

Authors: Angela P. Murillo[1], Nico Carver[1], Jane Greenberg[1], W. Davenport Robertson[1], Cheryl A. Thompson[1], and William Anderson[1]

_____

[1] School of Information and Library Science, University of North Carolina at Chapel Hill

[2] School of Information, University of Texas at Austin

Main Contact: Angela P. Murillo, amurillo@email.unc.edu

Title: Data At Risk Initiative: Scientists' perceptions of endangered data and data reuse.

Keywords:

Scientific data, Endangered data, Data at risk, Data reuse, Data sharing

Group (*):

Early Career Scientists (presentations to be given by researchers who recently received, or have yet to receive, PhD degrees)

Topic :

Information Science and Computational informatics, Earth and Environmental Data, Physical Science Data, Interoperability and Data Integration

Abstract:

Examining scientists' perceptions of endangered data, data reuse and data sharing is crucial in the understanding of the scientific process. Deterioration, format obsolescence, and insufficient metadata for discovery and retrieval are just some of the many elements that lead to possible loss of valuable scientific data. In order to investigate the point of view of the scientists, the Data-At-Risk Initiative (DARI) has conducted a research study to provide insight into the current state of scientific data-at-risk. In the spring of 2012, DARI members conducted four one-hour focus groups with scientists from selected departments. Focus groups participants were faculty, post-doctoral researchers, and Ph.D. students from

various scientific departments at the University of North Carolina at Chapel Hill and Duke University. Participants were recruited through email list serves. A total of four one-hour focus groups were conducted and a total of fourteen participants took part in the focus groups. The participants were asked a variety of 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. This talk presents the focus group findings and provides further understanding of how scientists perceive data reuse, sharing, and endangered data.

Introduction and Background

The Data-at-Risk Initiative (DARI) is designed to understand the extent of this growing problem and eventually take action by helping in the data rescue mission. DARI is a collaboration between the Committee on Data for Science and Technology (CODATA), Data At Risk Task Group (DARTG), the University of North Carolina at Chapel Hill (UNC) - Metadata Research Center (MRC), UNC's ibiblio, and UNC's -DARI-SILS Student Learning Circle.

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 inherently non-digital (paper, film, etc.), on near-obsolete digital media (magnetic tapes), or insufficiently described (lacking meta-data). Data which are regarded as unusable are often considered useless and risk being destroyed, and thus their scientific content is lost. Most data in this category 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 because they lack adequate formatting or metadata. These data can be essential for studies of long-term trend. Over the last year the DARI research team has conducted several studies to gain a better understanding of endangered scientific data. This talk presents the findings of the focus group study, which provides a deeper understanding of one aspect of endangered data, the scientists' perceptions.

Research Questions

The purpose of DARI is to understand the complexity of endangered data and to mitigate the risk of loss. In order to analyze the current state of this data, the DARI focus group study investigated the below questions.

1. What perceptions do scientists have on the topic of data at risk?
2. What perceptions do scientists have of data reuse?
3. What perceptions do scientists have of data sharing?
4. What opinions do scientists have in regards to the Data-At-Risk inventory?

The questions above have guided the focus group DARI research team to provide the vital information needed for the understanding of endangered data, data, reuse, and data sharing.

Methods

Four one-hour focus groups were conducted with scholars from selected scientific disciplines. Participants of these focus groups were faculty, post-doctoral researchers, and PhD students. Participants were recruited through departmental email list serves from physics and astronomy, biology, geography, geology, marine sciences, and environmental sciences and engineering departments from two major universities. These disciplines were chosen for this study in order to obtain a heterogeneous sample of the sciences. Due to geographic advantage, participants were recruited from two major universities, University of North Carolina at Chapel Hill and Duke University. A total of fourteen subjects participated in the study.

After organizing participants' availability, four focus groups were conducted during the spring 2012. All focus groups were audio-recorded with the agreement of the participants. Audio recordings were fully transcribed by the two principal investigators and were kept as Microsoft Word documents. For analyzing data, the two principal investigators read selected sample transcriptions and create a set of codes, which explained emerging patterns, as an inductive manner. Once all codes are developed, inter-coder reliability was tested. More than 85% of inter-coder reliability was achieved. For more systematic and efficient analysis, the software Nvivo 9 was used to assists with qualitative analysis. A demographic survey of all participants was also conducted to gather information such as department, research area, position, years of research and age.

Results

Participants discussed a variety of topics in relation to engendered data. Topics and subtopics are listed in Table 1, "Codes Identified from Focus Groups". These were the codes that were developed during the coding process. The main topics that were discussed throughout the focus groups were: data curation, data reuse, data sharing, data types, endangered data, data-at-risk inventory, and priority data. As can be seen in column one, when participants discussed data curation, they typical considered data curation to fall under four categories: (1) data creation (2) preservation action (3) data storage and (4) data transformation. Preliminary results indicate that scientists are generally concerned with the possibility of data loss. Preliminary results also indicate that scientists view endangered data through multiple lenses included lack of context and accessibility issues. Scientists demonstrated that they recognized the complexity of data-at-risk. Scientists also discussed when and how they reuse and share data, as well as their opinions of the Data-At-Risk Inventory. Data from these focus groups continues to be analyzed and the final results will be presented at this talk.

Conclusion

This oral presentation provides an overview of the Data-At-Risk Initiative (DARI) focus group project. This presentation will present the final results from the focus groups. This study informs DARI about the current state of endangered scientific data from the perspective of the scientists themselves and adds a valuable perspective to the discussion of how to ensure that endangered scientific data is not lost.

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.

Table 1 Codes Identified from Focus Groups

Data Curation	**Data Reuse**	**Data Sharing**	**Data Types**	**Endangered Data**	**Inventory**	**Priority**
Create	Online	Data Sharing-Incentives	Digital	Accessibility Issues	Feature Requests	Difficulty-Effort
Preservation Action	Person	Data Sharing-Disincentives	Non Digital	Lack of Context	Reasons to use it	Valuable-Irreplaceable
Store	Research Group			Potential Endangerment	Reasons not to use it	
Transform				Unavailable		