Paper presented at the 2011 Conference on Ensuring Long-Term Preservation in Adding Value to Scientific and Technical Data (PV2011), Toulouse, France, 15-17 November 2011 (http://www.pv2011.com).

Metadata for Data Rescue and Data at Risk

William L. Anderson (1), **John L. Faundeen** (2), **Jane Greenberg** (3), **Fraser Taylor** (4)

(1) ***University of Texas at Austin School of Information***

1616 Guadalupe Suite #5.502, Austin, TX 78701-1213, USA

EMail: band@ischool.utexas.edu

(2) ***U.S. Geological Survey, Earth Resources Observation and Science Center***

47914 252nd Street, Sioux Falls, SD, 57198-9801, USA

EMail: faundeen@usgs.gov

(3) ***School of Information and Library Science, University of North Carolina***

216 Lenoir Drive • CB #3360 • 100 Manning Hall, Chapel Hill, NC 27599-3360, USA

EMail: janeg@email.unc.edu

(4) ***Geomatics & Cartographic Research Centre***

B349 Loeb Building 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada

EMail: Fraser_Talor@carleton.ca

ABSTRACT

Scientific data age, become stale, fall into disuse and run tremendous risks of being forgotten and lost. These problems can be addressed by archiving and managing scientific data over time, and establishing practices that facilitate data discovery and reuse. Metadata documentation is integral to this work and essential for measuring and assessing high priority data preservation cases. The International Council for Science: Committee on Data for Science and Technology (CODATA) has a newly appointed Data-at-Risk Task Group (DARTG), participating in the general arena of rescuing data. The DARTG primary objective is building an inventory of scientific data that are at risk of being lost forever. As part of this effort, the DARTG is testing an approach for documenting endangered datasets. The DARTG is developing a minimal and easy to use set of metadata properties for sufficiently describing endangered data, which will aid global data rescue missions. The DARTG metadata framework supports rapid capture, and easy documentation, across an array of scientific domains. This paper reports on the goals and principles supporting the DARTG metadata schema, and provides a description of the preliminary implementation.

Keywords: science, data, rescue, metadata, inventory, preservation, ICSU, CODATA

INTRODUCTION

At the 22nd International Council of Science CODATA Conference in October 2010, a special session was held to examine the growing volume of "endangered data", described as unique scientific data that are at risk of permanent loss. As a result of this session a formal CODATA Data-at-Risk Task Group (DARTG) was established with a mission of planning and developing a data-at-risk inventory [1]. This paper describes the preliminary work of the task group in building this inventory.

The primary DARTG goal is to create an inventory of data whose unique scientific information and value is in danger of being lost to posterity. A secondary goal is that for the inventory to provide descriptive information that is useful to projects designed to rescue that at-risk information. The work of the DARTG will demonstrate an approach, a process, and a set of practices for building an extensible inventory of valuable scientific data, which are at risk being lost or destroyed, and whose information content is therefore seriously endangered.

The work plan for the task group has three elements:

- Define a set of core metadata properties essential for a data-at-risk inventory.
- Prototype a system to support inventory data collection and maintenance.
- Populate the inventory with data at risk in selected target disciplines.

The following sections define the scope and context of the Data-at-Risk Task Group initiative as part of data rescue needs, describe the preliminary inventory metadata development work, and describe the relation of this work to other scientific data management initiatives.

THE DATA RESCUE AND DATA-AT-RISK CONTEXT

In a 2005 essay, Griffin makes a strong case for the scientific value of rescuing and recovering endangered scientific data [2]. The insight that a collection of data is a scientific instrument is consonant with the recognition that much scientific research in the 21st Century will be data-driven [3]. But modern research will be hampered if valuable historical scientific data are ignored or lost. The CODATA DARTG was formed to mitigate the risks of loss.

The DARTG defines "data at risk" as scientific data, which are not in formats that permit full electronic access to the information that they contain. Such at-risk data may be primarily non-digital (e.g., handwritten, photographic, or physical specimens), on near-obsolete digital media (such as magnetic tapes), or lacking adequate description (metadata). These data are not inventoried and cannot be easily discovered, accessed, shared, and used by research communities. Data that are regarded as unusable tend to be regarded as useless and risk being lost or destroyed.

Data that are old often retain scientific value. Stories of creating new scientific knowledge from old data collections are becoming more prevalent in the research literature and news. Some recent notable examples include:

- A research study of historic stellar data providing historic data on atmospheric ozone that augments current models of ozone changes [4].
- A Berkeley Lab News article that describes new science on the link between cholesterol and heart disease that used old data on old media, in this case, computer punch cards [5].

- An IEEE Computer article on a project between Columbia University scientists and Consolidated Edison (ConEd) engineers that determined that old, raw ConEd data on low-voltage failures (manhole fires, explosions, etc.) can be used in predicting and preventing future events [6].

Today many scientific endeavors are restricted to digital data as our culture becomes immersed in computer-networked technology. Practical research constraints, such as time, funding, and the pressures related to scholarly output and academic promotion, increase the reliance on digital resources. It is easier to work with easily accessible and machine-readable data. Data not in electronic forms cannot by copied – physical specimens are a notable example. In addition, the skills and techniques needed to look at pre-digital, historical forms of data are time consuming and not convenient to the modern day research pace. Furthermore, these skills and techniques themselves are at risk of being lost to younger generations of researchers. As more research is carried out online, less knowledge is shared about non-digital data. Data that are not visible are not used, and resources that are not used tend to be lost or destroyed.

While paper and older analog media-based data are hard to find, and once found, hard to use, digital data are not immune to loss or usability problems. Some born-digital data can also be considered "at risk" if they cannot be ingested into managed databases because they lack adequate formatting or descriptive metadata. The threats to digital data preservation and use are not restricted to technical problems, but include economic factors, as well as a lack of effective archival policies and practices [7]. Furthermore, the volume of research data being collected and generated is growing, and this growth is challenging current practices and resources to support long-term access and use. The articles in the 11 February 2011 issue of Science address many of the challenges raised, and opportunities afforded, by modern, data-driven science [8].

Establishing a measure or seeking a picture of the full extent of the scientific data-at-risk situation is daunting. Given the growing awareness of fragile and at risk data, acknowledging the size of the challenge is a first step in finding solutions. Digital cataloging technologies and networking capabilities provide the tools and applications that allow for collective reporting of this problem; and the DARTG is taking on this challenge by pursuing an infrastructure supporting a global data-at-risk inventory. The primary objective of the DARTG is to develop a simple and robust metadata scheme that will enable collecting documentation to get some scope of the endangered data problem in selected fields. Identifying key datasets at risk is a necessary step in developing strategies to mitigate losses and to inform recovery and rescue activities.

METADATA: DATA DOCUMENTATION AND FUNCTIONS

The primary purpose for recording metadata is to support a function or set of functions on the data that is described. A fairly universal function is *resource* discovery: metadata searched by a user and manipulated by a machine to aid in finding or discovering a dataset. Cataloging schemes by which libraries organize their holdings, with standard descriptions for author, title, and subject, are familiar examples of metadata designed to support discovery. Other common metadata functions include supporting resource access, resource management, and usage rights.

Metadata Essential for Data Rescue

A good metadata practice is to define the metadata functions that will support a project's overall goals and mission. The DARTG's mission is to produce an inventory that effectively describes the risk associated with scientific data, and can serve as a starting point for data rescue.

The DARTG's prototype inventory will include data in a range of disciplines and research areas (Astrophysics, Botanical science, Climatology, Oceanographic studies, and other areas). Developing a metadata scheme that represents multiple disciplines and informs the data rescue mission presents a two-fold challenge. Luckily, cross-disciplinary metadata description is an area of active research that has progressed in response to the evolution of the web. The DARTG is able to benefit from developments in this area, most notably, the Dublin Core Metadata Initiative (DCMI)'s development and promotion of universal, cross-domain metadata properties for resource description [9].

The DARTG members articulated two key metadata requirements. The DARTG inventory metadata needs to:

- Be applicable across a range of discipline and scientific research areas.
- Sufficiently support the data rescue mission.

Metadata Frameworks Useful for Data-at-Risk

The DARTG chair, Elizabeth Griffin, proposed an initial set of metadata properties. These elements were informed by her research with abandoned photographic observations of stellar spectra that support research on the Earth's stratosphere, a general knowledge on a range of scientific disciplines, and the goal to extend the DARTG's effort across disciplines. The proposed scheme included the seven general properties presented in Table 1.

Metadata Property
1. Science area
2. Nature of data
3. Date or date-span
4. Location of original
5. Present location
6. Expected future
7. Risk level

Table 1: Initial proposed DARTG Inventory Metadata Properties

The DARTG members have used a series of online and phone meetings to work on metadata definition challenges. In addition to the disciplinary expertise of the DARTG, the U.S. Geological Survey (USGS) has conducted its own work on defining information about data for rescue [10]. The USGS inventory template gathers data in a succinct survey to aid the USGS in setting data rescue priorities. The key factors used to prioritize data rescue projects are continuing scientific value and the condition of the data collection. Portions of the template include factors to help assess the data impact, condition and value. Examples of the criteria include: climate variability and change; energy and minerals for America's future; and a national hazards, risk and resilience assessment program. The USGS form requests the person proposing a data rescue to "Describe the rescue need and how it relates to the agency's Science Strategies, and also asks, "Will the rescued data be digital?" One of the last elements of the USGS template is "Estimated Cost (14 char max)." The USGS template provided insight into the DARTG metadata development efforts.

THE DARTG METADATA SCHEME

The DARTG Data-at-Risk Inventory work is aimed at developing a minimal number of metadata elements that can serve to describe the ways in which a given dataset is at risk, and to provide enough information for any future rescue efforts. The DARTG inventory metadata needs to be applicable across domains. The intent is to use properties from existing schemes, rather than invent new ones. The DARTG seeks a scheme that is interoperable with existing standards.

The specific principles guiding the DARTG metadata development are that the scheme is:

- Simple: Enable non-scientists to describe the range, scope, and extent of data sources identified as "at risk."
- Broadly applicable: Cover essential properties for diverse data sources identified as "at risk".
- Extensible: Support metadata extensions over time.

To move forward, the CODATA DARTG engaged in a partnership with the University of North Carolina at Chapel Hill, School of Information and Library Science, Metadata Research Center (UNC-CH/SILS/MRC[1]) and ibiblio[2]. UNC-CH SILS students and DARTG members developed a prototype inventory using Omeka[3], an open source cataloging and exhibition software package. The UNC working group is referred to as DARI (Data-At-Risk Inventory). The DARI team conducted a case study of the preliminary metadata framework proposed by DARTG members [11]. These early results provided informative feedback regarding inventory metadata element names, as well as survey usability. The work under way has been guided by DARTG's principles and goals, and an effort to define the functions essential for data rescue. The Dublin Core Metadata Initiative, the above noted endeavors at the USGS, and the work of Elizabeth Griffin inform the effort. Table 2 provides the most current version of the DARTG metadata scheme that has emerged from the DARTG/UNC collaboration.

Metadata Element Name	**Element Description**
Research Area(s)	The domains represented by DARTG experts and the more general category of "Other".
Title	The name associated with the collection.
Description: Physical form of the data	Paper, photograph, specimen, record book, magnetic tape, etc.
Description: Content and context of the data	History, topic, etc. -- if known
Name of current holder	Institution, organization or individual.
Dates associated with data	Time period when data were collected.
Size	Extent, volume, size.
Data condition	Stable, deteriorating, etc.
Risk level	Poor storage conditions, limited storage time, etc.
Known access and restrictions	Public domain, private collection, etc.
Notes	Any additional information.
Contact information	Address or other contact information for the institution, organization or individual.

Table 2: DARTG DARI Metadata, Version 1.0

[1] UNC-SILS Metadata Research Center, http://ils.unc.edu/mrc/

[2] ibiblio: The Public's Library and Digital Archive, http://www.ibiblio.org/

[3] Omeka: Serious Web Publishing, http://omeka.org/about/

CONCLUSION

In order to gather the knowledge needed to institute a Data-at-Risk Inventory for a larger set of scientific disciplines, the DARTG, in collaboration with the UNC-DARI group, will continue soliciting contributions from scientists and science librarians. Contributors of at-risk data bring their own expectations and ideas about how to mitigate those risks. By working with a small but diverse set of disciplines and types of data, the inventory metadata will come to represent pan-disciplinary data description requirements. The experience gained in documenting at-risk data in ways that satisfy diverse contributors will help define what is practicable for an inventory endeavor to grow.

ACKNOWLEDGMENTS

The DARTG would like to acknowledge University of North Carolina Center for Global Initiatives support of the DARI SILS Student Learning Circle.

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

[1] – L. Nordling, “Researchers launch hunt for endangered data”, Nature 468, 17 (2010), doi:10.1038/468017a

[2] – R.E. Griffin, “Rescuing and recovering lost or endangered data”, CODATA Data Science Journal, Vol. 4 (17 July 2005), pp. 21-26. doi:10.2481/dsj.4.21

[3] – T. Hey, S. Tansley, K. Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, (2009), URL: http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx, Retrieved on 2011-09-13.

[4] – R.E. Griffin, “The Detection and Measurement of Telluric Ozone from Stellar Spectra”, *Publications of the Astronomical Society of the Pacific*, Vol. 117, No. 834 (August 2005), pp. 885-894. Article Stable URL: http://www.jstor.org/stable/10.1086/431935, Retrieved 2011-09-07.

[5] – D. Krotz, “From Dusty Punch Cards, New Insights Into Link Between Cholesterol and Heart Disease”, URL: http://newscenter.lbl.gov/feature-stories/2011/01/04/cholesterol-heart-disease/, Retrieved 2011-09-07.

[6] – C. Rudin, R.J. Passonneau, A. Radeva, S. Jerome, D.F. Isaac, "21st-Century Data Miners Meet 19th-Century Electrical Cables," Computer, vol. 44, no. 6, pp. 103-105, June 2011, doi:10.1109/MC.2011.164

[7] –“It’s About Time”, Final Report of the Workshop on Research Challenges in Digital Archiving and Long-term Preservation, April 12-13, 2002. Published August 2003. P. vii., URL: http://www.digitalpreservation.gov/library/resources/pubs/docs/about_time2003.pdf , Retrieved 2011-09-11.

[8] – Science: Special Online Collection: Dealing with Data. URL: http://www.sciencemag.org/site/special/data/, Retrieved on 2011-09-07.

[9] – Dublin Core Metadata Initiative Metadata Terms, URL: http://dublincore.org/documents/dcmi-terms, Retrieved on 2011-09-13.

[10] – U.S.Geological Service: “Create a Rescue Request”, URL: http://eros.usgs.gov/government/archive_rescue/archive_request.php, Retrieved on 2011-09-13.

[11] – C.A. Thompson, N. Carver, K. Collins, J. Sinclair, J., M. Veitch, “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 (2011)