The Data-at-Risk Initiative: A Metadata Scheme for Documenting Data Rescue Activities

Anona C. Earls*, Erin Clary*, Jane Greenberg, Aaron Kirschenfeld, Angela P. Murillo, W. Davenport Robertson, Shea Swauger
University of North Carolina, School of Information and Library Science, Metadata Research Center
300A Manning Hall, Chapel Hill, NC 27599-3360
*aearls@email.unc.edu; *eclary@live.unc.edu

William L. Anderson
School of Information
University of Texas at Austin
1616 Guadalupe
Austin, Texas, 78701
band@praxis101.com

ABSTRACT

The Data-At-Risk and Rescue Initiative (DARI), an extension of the international CODATA Data-at-Risk Task Group (DARTG), is investigating how to best document data rescue efforts. This poster reports on a metadata-driven content analysis, and presents a metadata scheme for documenting data rescue. Twenty data rescue projects were reviewed for background context, and seven metadata schemes in the areas of preservation and data description were analyzed via a content analysis. Version 1.0, Data Rescue metadata core, consisting of 13 core elements, is presented, and future directions are noted.

Categories and Subject Descriptors

D.3.3 **[Language Constructs and Features]**, *Data types and structures*; E.2 **[DATA STORAGE REPRESENTATIONS]** O*bject representation.*

General Terms

Documentation, Design, Standardization

Keywords

Data Rescue, Metadata Schemes, Documentation, Endangered Data, Scientific Data.

1. INTRODUCTION

The Data-at-Risk and Rescue Initiative (DARI) is a project under the Committee on Data for Science and Technology (CODATA) Data at Risk Task Group (DARTG) [3, 8]. Initial DARI activities focused on the development of a prototype inventory to document the existence of valuable scientific data that are at risk of being lost to posterity [1, 2, 10]. At-risk data are data that are fragile or deteriorating, data that are lacking sufficient metadata, or data that are not in formats that permit full electronic access.

As work on the data-at-risk inventory progressed, the DARI team recognized the need to understand scientists' perception of at-risk data [9] and provide an online resource where scientists, data custodians and other individuals could contribute descriptions of data rescue activities. The goal of the work reported on in this poster has been to address this latter need, and to contribute to DARI's effort to extend the data-at-risk inventory to include descriptions of data rescue activities. Documenting successful data rescue missions illustrates to scientists that data can be saved and made available, and provides an important record of work that can aid with planning future data rescue activities. The work presented draws from successful data rescue efforts and work conducted in preservation and metadata communities.

2. BACKGROUND WORK

Data rescue has been an important human endeavor throughout history. Perhaps the most profound 20th century event was the 1966 Flood of the Arno River in Florence, which drew attention to preservation challenges in libraries, archives, and museums. The international community gathered to conserve and restore historical treasures in many collections, including the Institute and Museum of the History of Science, which is known to house historical scientific instruments and significant scientific collections, including the works of Galileo.

Digital technology has enabled new methods of data rescue. Several notable efforts include: the Astronomical Plate Collection and Preservation in China project, which is an effort to rescue, catalog, and eventually digitize astronomical plates from several observatories in China; the Royal Observatory of Belgium project, which seeks to digitize astronomical plates from the 20th century; and the Dominion Astrophysical Observatory project, which is focused on the digitization of materials from Canada's largest optical astronomical observatory. Related projects focus on the planet's changing climate and ecosystem. For example, the Botanic Garden and Botanical Museum Berlin-Dahlem rescue effort uses reBIND workflows to transform biodiversity data stored in outdated database management systems into well-documented, standardized formats. These and other data rescue efforts are important if data significant to the pool of scientific knowledge are to be preserved. However, brief descriptions on a web page or project page may not be sufficient to highlight these data rescue efforts, for scientists and other researchers to find these data, or for sharing approach outcomes on a global scale.

The DARI team is extending the data-at-risk inventory to include descriptions of completed and ongoing data rescue efforts. Over the last several months, DARI researchers have engaged in discussions and an exploration of what elements are essential to simply, yet thoroughly, describe data rescue efforts. The lead author of this paper has also contributed to this undertaking via her master's paper research, and she has focused on the development of a core metadata scheme for describing data rescue efforts. The remainder of this paper presents the DARI team's overall work, and the work conducted by Ms. Earls to develop a prototype metadata scheme to document data rescue efforts.

3. OBJECTIVES

The goal of the research presented here was to design a core, functional metadata scheme for the description and documentation of endangered scientific data rescue activities, and to apply metadata in accordance with this scheme to known data rescue activities within a digital content management environment.

4. RESEARCH QUESTIONS

1) What are the main descriptive characteristics of known data rescue projects?
2) What existing metadata standards can be applied to describing a data rescue project as a whole?
3) What metadata elements are essential for describing data rescue projects in particular?

5. RESEARCH METHODS

Scheme development was pursued using a mixed methods approach. First, 20 data rescue projects were reviewed for contextual background. Second, a content analysis was conducted to further examine seven metadata schemes in the areas of data description and preservation [6, 7]. The background review of existing data rescue projects included the identification of existing metadata used to describe or report on the effort, and a review of literature that reported on the effort [4, 7]. The content analysis compared schemes via a crosswalk to identify similarities and differences. Basic, core metadata elements became evident and form the basis for the Data Rescue metadata core, version 1.0.

Descriptive Elements	Archaeology Data Service Guidelines	Data-PASS	DOAP	Dublin Core, v. 1.1	DCMI-TERMS	Goddard Core	IMDI (ISLE Metadata Initiative)	RSLP Collection Description Schema
Title/Name	x	x	x	x	x	x	x	x
Description	x	x	x	x	x	x	x	x
Methods								
Notes		x						x
Creator/Author	x	x		x	x	x	x	
Sponsor								
Contributor	x			x	x			
Dates	x	x	x	x	x	x	x	
Geographic Location		x			x	x		x
Associated Resources	x	x			x	x		
URL/citation			x		x			
Subject Keywords	x	x		x	x	x		x
Unique ID		x		x	x	x	x	x

Table 1. Comparison of metadata elements across schemes.

6. RESULTS

The specific outcomes of this work to date include: 1) a prototype inventory for documenting data rescue activities that will serve a reference function similar to that provided by the descriptions of at-risk datasets, 2) version 1.0 Data Rescue metadata core for describing data rescue activities, and 3) a selected set of data rescue activities that are described in the inventory. The current scheme is heavily Dublin Core based, and future work will explore the value of advancing this work toward an endorsed Dublin Core Application Profile [5].

6.1 Metadata for data rescue, version 1.0

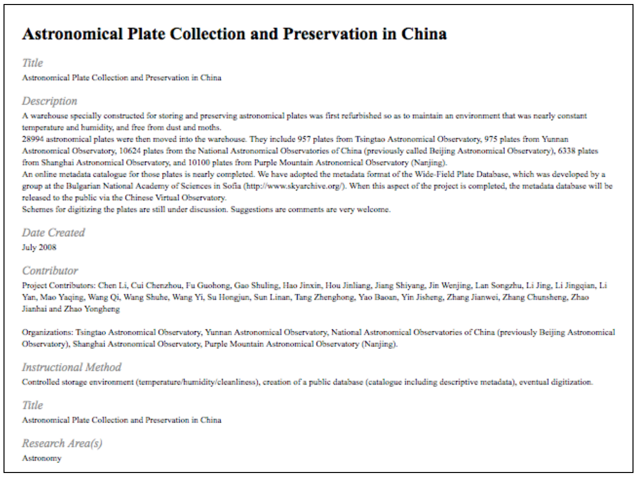
A proposed metadata scheme (**Table 2**) of thirteen elements has been developed. This scheme includes core elements for describing data rescue. The goal of the scheme is to facilitate consistent description of data rescue activities. The scheme forms the basis of an input template and has been through base-level testing. The scheme has been integrated into the DARI inventory, a publicly accessible metadata repository, developed via Omeka and located at http://ibiblio.org/data-at-risk/.

Element Name	**Element Description**
Title*	The title (and any alternatives) for the project.
Description	A brief summary of the main focus, goals, aims, and/or objectives of the project.
Methods	A brief summary of the approach, methods, techniques, and/or processes (including tools, software, etc.) being used for the data rescue.
Notes	Other details pertinent to the project, such as background information or project history.
Creator*	Individual(s) or organization(s) who initiated and have overseen the data rescue effort. May include contact information.
Sponsor	Individual(s) or organization(s) who have contributed financially or otherwise endorsed the project.
Contributor	Other individual(s) or organization(s) who have contributed to the project; for example, project partners/collaborators (physical or intellectual efforts), contributors of data/materials, etc.
Dates*	Dates indicating when the project was initiated and when the project was completed. May also include important milestones or other significant dates associated with the project.
Location	Location where the project was/is being carried out (if applicable).
Associated	Any other important projects or work (in particular, other data rescue initiatives) associated with this project, or upon which this project has been built.
URL	A link to the project website and/or online documentation of the project.
Keywords	Keywords indicating subject content of the project.
Project ID	A unique ID# assigned to the project by the repository (optional).

Table 2. DARI proposed metadata scheme for the description of data rescue activities. * Indicates a required element.

6.2 DARI data rescue description

The DARI team is at the early stage of testing the scheme's ability to represent a range of data rescue activities. A screen capture for one of the rescue projects is presented below in Figure 1. To date, we have tested two rescue projects thoroughly, and work will continue over the coming months.



Astronomical Plate Collection and Preservation in China

Title

Astronomical Plate Collection and Preservation in China

Description

A warehouse specially constructed for storing and preserving astronomical plates was first refurbished so as to maintain an environment that was nearly constant temperature and humidity, and free from dust and moths.
28994 astronomical plates were then moved into the warehouse. They include 957 plates from Tsingtao Astronomical Observatory, 975 plates from Yunnan Astronomical Observatory, 10624 plates from the National Astronomical Observatories of China (previously called Beijing Astronomical Observatory), 6338 plates from Shanghai Astronomical Observatory, and 10100 plates from Purple Mountain Astronomical Observatory (Nanjing).
An online metadata catalogue for those plates is nearly completed. We have adopted the metadata format of the Wide-Field Plate Database, which was developed by a group at the Bulgarian National Academy of Sciences in Sofia (http://www.skyarchive.org/). When this aspect of the project is completed, the metadata database will be released to the public via the Chinese Virtual Observatory.
Schemes for digitizing the plates are still under discussion. Suggestions are comments are very welcome.

Date Created

July 2008

Contributor

Project Contributors: Chen Li, Cui Chenzhou, Fu Guohong, Gao Shuling, Hao Jinxin, Hou Jinliang, Jiang Shiyang, Jin Wenjing, Lan Songzhu, Li Jing, Li Jingqian, Li Yan, Mao Yaqing, Wang Qi, Wang Shuhe, Wang Yi, Su Hongjun, Sun Linan, Tang Zhenghong, Yao Baoan, Yin Jisheng, Zhang Jianwei, Zhang Chunsheng, Zhao Jianhai and Zhao Yongheng

Organizations: Tsingtao Astronomical Observatory, Yunnan Astronomical Observatory, National Astronomical Observatories of China (previously Beijing Astronomical Observatory), Shanghai Astronomical Observatory, Purple Mountain Astronomical Observatory (Nanjing).

Instructional Method

Controlled storage environment (temperature/humidity/cleanliness), creation of a public database (catalogue including descriptive metadata), eventual digitization.

Title

Astronomical Plate Collection and Preservation in China

Research Area(s)

Astronomy

Figure 1. Screenshot, "Astronomical Plate Collection and Preservation in China," *Data-at-Risk Inventory*, accessed April 26, 2013, http://www.ibiblio.org/data-at-risk/items/show/94.

7. CONCLUSION

This paper reports on initial work to extend the DARI inventory and capture descriptions of data rescue activities. A core metadata scheme for documenting major identifying features of data rescue efforts, version 1.0 of the Data Rescue metadata scheme is presented, along with an example of a data rescue effort described with this scheme. A finalized scheme will serve to support knowledge and discovery of how endangered scientific data have been rescued in various cases. Future work will include further development of the inventory to support documentation of data rescue activities, and will seek to engage scientists, collection custodians, and other individuals in the documentation effort.

8. ACKNOWLEDGMENTS

We would like to acknowledge the support of CODATA and thank ibiblio for hosting the DARI Inventory.

9. REFERENCES

[1] Anderson, W., Faundeen, J., Greenberg, J., & Taylor, F. (2011). Metadata for data rescue and data at risk: ensuring long-term preservation and adding value to scientific and technical data. PV2011, 17 November 2011, Toulouse, France. Proceedings paper retrieved 2013-06-24 from http://hdl.handle.net/2152/20056. Conference presentation retrieved 2013-04-26 from http://www.slideshare.net/2ghouls/metadata-for-data-rescue-and-data-at-risk.

[2] Carver, N., Collins, K., Greenberg, J., Sinclair, J., Thompson, C., Veitch, M., & Anderson, W. (2011). Identifying endangered data: a case study supporting inventory design and implementation. ASIST 2011, October 9-13, 2011, New Orleans, LA, USA.

[3] Data-at-Risk Inventory. (n.d.). Retrieved 2013-04-26 from http://www.ibiblio.org/data-at-risk/

[4] Downs, R. R. (2009). Managing risks to scientific data. Prepared for presentation to the NYU/IBM Workshop on managing data risk: acquisition, processing, retention and governance, New York University, New York, NY April 24, 2009. Retrieved 2013-04-26 from http://w4.stern.nyu.edu/emplibrary/Downs-ManagingRisksSciData.pdf.

[5] Dublin Core Application Profile, URL: http://dublincore.org/documents/profile-guidelines/.

[6] Dublin Core Metadata Initiative Metadata Terms, URL: http://dublincore.org/documents/dcmi-terms. Retrieved on 2013-04-26.

[7] Hodge, G., Templeton, C., & Allen, R. (2005). A metadata element set for project documentation. Science & Technology Libraries, 25:4, 5-23. doi: http://dx.doi.org/10.1300/J122v25n04_02

[8] International Council for Science: Committee on Data for Science and Technology. CODATA Data At Risk Task Group (DARTG). Retrieved 2013-04-26 from http://ils.unc.edu/~janeg/dartg/.

[9] Murillo, A. P., Carver, N., Greenberg, J., Robertson W. D., Thompson C. A., & Anderson, W. (2012). Data At Risk Initiative: Scientists' perceptions of endangered data and data reuse. 23rd International CODATA Conference, 29-30 October 2012, Taipei, Taiwan.

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