# DATA AT RISK INITIATIVE: EXAMINING AND FACILITATING THE SCIENTIFIC PROCESS IN RELATION TO ENDANGERED DATA

***Angela P. Murillo***

*School of Information and Library Science, University of North Carolina at Chapel Hill, 216 Lenoir Drive • CB #3360 • 100 Manning Hall, Chapel Hill, NC 27599-3360*
*Email:* amurillo@email.unc.edu

## *ABSTRACT*

*Examining the scientific process in relation to endangered data, data reuse, and sharing is crucial in facilitating scientific workflow. Deterioration, format obsolescence, and insufficient metadata for discovery are significant problems leading to loss of scientific data. The research presented in this paper considers these potentially lost data. Four one-hour focus groups and a demographic survey were conducted with 14 scientists to learn about their attitudes toward endangered data, data sharing, data reuse, and their opinions of the DARI inventory. The results indicate that unavailability, lack of context, accessibility issues, and potential endangerment are key concerns to scientists.*

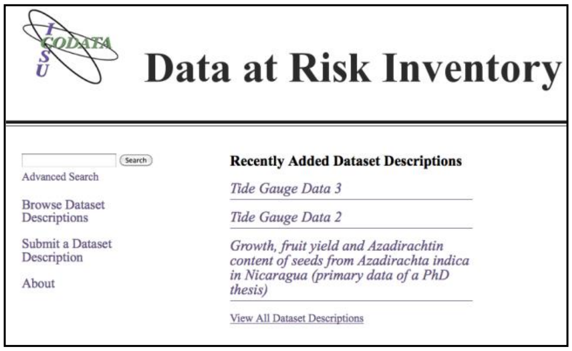
**Keywords:** Endangered data, Data at risk, Data reuse and sharing, Research process, Scientific workflow

## 1 INTRODUCTION

The Data at Risk Initiative (DARI) (DARI | Metadata Research Center, n.d.) was designed to understand the extent of the growing problem of endangered data and take action by assisting in data rescue missions. A key first step has been to investigate how to describe at risk data and prototype an inventory of endangered data as well as to gain feedback from scientists regarding data at risk. The 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 CODATA DARTG's objective is to seek out and prepare an inventory of at risk data. UNC's ibiblio hosts the inventory, and the MRC provides design assistance. UNC's DARI-SILS Student Learning Circle is a student driven group that conducts research on this topic and holds meetings throughout the year to discuss the progress of the DARI initiative.

The DARTG defines "data at risk" as scientific data that are not in a format that permits full electronic access to the information they contain (DARTG, n.d.). Data at risk are essentially endangered scientific data that are at risk of being lost. Data at risk can be inherently non-digital (paper, film, etc.), on near-obsolete digital media (magnetic tapes), or insufficiently described (lacking metadata). These data are regarded as unusable and are often considered useless and risk being destroyed. Much of the data in this at risk category pre-date the digital era; however, data at risk are not only non-digital data. Born-digital data can also be considered "at risk" if they cannot be ingested and managed by databases because they lack adequate formatting or metadata. These data at risk can be essential for studies of long-term trends as described in the literature review. Data at risk are a global concern that affects all scientific disciplines. However this concern is perhaps most pertinent to those fields that are attempting long-term trend analysis as will be shown in the literature review. Data at risk are sometimes called endangered data, and these terms will be used interchangeably throughout the paper.

The DARI research team has conducted several studies to gain a better understanding of these endangered scientific data to assist in facilitating the scientific process. In order to get a sense of how to describe at risk data, which may help with data rescue efforts, the DARI Initiative created an inventory (DARI, n.d.) in which scientists can deposit descriptions of data at risk (Figure 1). The purpose of the inventory is to create a catalog of valuable scientific data that are at risk of being lost. The inventory is open for scientists or information professionals to submit descriptions.


ICSU
CODATA
Search
Advanced Search
Browse Dataset Descriptions
Submit a Dataset Description
About
View All Dataset Descriptions

Data at Risk Inventory

**Recently Added Dataset Descriptions**

*Tide Gauge Data 3*

*Tide Gauge Data 2*

*Growth, fruit yield and Azadirachtin content of seeds from Azadirachta indica in Nicaragua (primary data of a PhD thesis)*

**Figure 1.** Data At Risk Inventory pre-redesign

Another way to gain insight into the current state of scientific data at risk is to get a sense of scientists' research processes in relation to endangered data. The research presented in this paper considered this need and reports on a focus group study conducted with scientists during the spring of 2012. Two DARI researchers conducted four one-hour focus groups. The focus groups gathered data on scientists' attitudes and perceptions about data reuse, data sharing, and endangered data, and their opinion of the initial Data At Risk Inventory (Figure 1). This paper presents the focus groups' findings. Overall, this paper furthers our understanding of scientists' perception of data reuse, sharing, and endangered data and how this understanding influences the scientific process.

These findings provide a deeper understanding of endangered data that will facilitate and support the scientific process. Section 2 contains a literature review to analyze the current research in this area. Section 3 outlines the research questions of the study, and Section 4 describes the methods of the study. Section 5 describes the results of both the demographic survey and focus groups' results, and provides an analysis of the findings. Lastly, Section 6 provides a conclusion and possible future work.

## 2 LITERATURE REVIEW

There are many reasons to investigate data at risk. These data retain significant scientific value; they can be crucial for long-term observations; they can create new scientific knowledge, and they are inherent in the data reuse and sharing cycle within the scientific process. Recently, there have been efforts to use historic data in the scientific process as well as efforts to understand the importance of these historic data. These efforts are precisely what the DARI is facilitating through actively researching data at risk in order to gain an understanding of how scientists are utilizing at risk data. The DARI is also facilitating these efforts by providing an infrastructure, such as the DARI inventory for scientists, to which to contribute descriptions of at risk data.

The following literature describes efforts by scientists to use historic data in their research process and efforts to describe the importance of historic data. The literature below provides examples of how endangered data add crucial data to the scientific process, how scientists are using endangered data in their research process, and efforts to retain and rescue these endangered data.

Griffin (2005a) provides an overview of the importance of rescuing data at risk and the value of these data to the scientific process. She includes in her discussion specific disciplines that create irreplaceable historical records, such as astronomy, botany, zoology, and many others. This work describes how biodiversity and atmospheric historical data can be used for long-term trend analysis. Griffin suggests that efforts to rescue data at risk will allow modeling in both time directions. This is crucial for science because "Today's data can never recapture events of the past" (Griffin, 2005a). Additionally, Griffin provides examples of historic data rescue efforts, including GODAR (Global Oceanographic Data Archeology and Rescue), which located non-digital data to digitize and submit to world data centers, and the Working Group of the International Astronomical Union's plan for converting photographic observations into digital data.

Researchers reuse endangered data for long-term trend analysis in a variety of fields. Rudin et al. (2011) discuss the use of historic records for predicting events in electric grid systems and show that without access to these historic records, these predictions could not have been made. Historic records have also been used to conduct long term analysis of health related issues. For example, the longest analysis of cholesterol and heart disease was conducted by Krotz (2011) through the use of historic punch cards. In astronomy, Griffin (2005b) discusses the use of historic stellar spectra for determining telluric O3 column densities in order to investigate and compare ozone measurements. Axelsson (2001) describes the use of historic data in a retroactive gap analysis to analyze changes in forest conditions in Sweden.

Researchers have used multiple methods to locate these historic data. For example, Gizzi (2009) describes the use of eBay for obtaining research materials, such as manuscripts, photographs, and newspapers, for researching natural disasters, such as earthquakes, floods, and landslides. In each of the cases described above scientists would not have been able to conduct their analyses without the use of historic datasets, which shows the importance of endangered data to the scientific process.

In addition to scientific studies that have shown how historic data can be reused, there have been several articles discussing the importance of investigating endangered data. Similar to Griffin (2005a), Nordling (2010) discusses how important endangered scientific data are being lost and why these data are important to the scientific process and describes the CODATA effort to rescue these endangered data. As described by the author, these endangered data are being lost because they are fragile or obsolete or perhaps are destroyed by researchers who are unaware of their value. The author describes how endangered data are both physical data and digital data and suggests that digital data from between 1950 and 1980 are at a high risk. Furthermore, the author describes the CODATA - DARTG plans to create a catalogue for endangered data (Figure 1).

The studies above demonstrate the importance of historic data in the scientific process. These studies provide an overview of how endangered data impact the scientific process, why these data are important, how scientists are reusing these data in their research, and the importance of an inventory to catalogue these data. There are a very limited number of studies discussing this topic, and the majority of these studies show how these data are being used. However none of the studies asks the scientists directly how endangered data are part of their research process and what infrastructure is needed to facilitate the reuse of these data, therefore demonstrating the need for the research outlined and described below.

## 3 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 these data, the DARI focus group study investigated the questions below.

1. What perceptions do scientists have on the topic of data at risk?
2. What perceptions do scientists have of data reuse and sharing?
3. What opinions do scientists have in regard to the Data At Risk Inventory?

These questions guided the DARI research team.

## 4 METHODS

Four one-hour focus groups were conducted with scholars from selected scientific disciplines. The focus group method was pursued because it is a proven method for gathering nuanced perspectives and for its efficiency in generating new ideas from a group of people (Wildemuth, 2009). The focus group method is effective for gathering opinions and attitudes (Wildemuth, 2009), exploratory research particularly regarding shared and tacit beliefs (Macnaghten & Myers, 2004), and generating new ideas (Macnaghten & Myers, 2004). Although there are many strengths to focus groups, one weakness is that ideas shared in focus groups are sometimes not as fully developed as they are in interviews or observations (Macnaghten & Myers, 2004). However, the researchers did not observe any indications of this concern.

Furthermore, this method has been used before to gather feedback from scientists as demonstrated by the work of Meyer et al. (2011) and Kuruppu and Gruber (2006). In addition to the focus groups, a demographic survey gathered participants' information, such as department, research area, position, years of research, and age.

## 4.1 Sample Population

Participants were recruited through departmental email listservs from physics and astronomy, biology, geography, geology, marine sciences, environmental sciences, and engineering departments and were faculty, post-doctoral researchers, and doctoral students. These disciplines were chosen 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.

## 4.2 Procedures

The study design was approved by the University of North Carolina Institute Review Board in April 2012, IRB # 11-2527. After organizing participants' availability, four focus groups were conducted during the spring of 2012. Each of the focus groups was conducted with two facilitators present, and each followed the same focus group guide (Appendix A). None of the participants overlapped from one focus group to another. Participants provided introductions to themselves and their research. Participants were then asked to discuss the data types they used in their research. They were asked to think about and discuss their dream data, data they wished they had, and what barriers they had to gathering this data. Participants were also asked about their data reuse and sharing practices. Within this context they were asked to think about what they considered at risk data and why as well as which data they prioritized above others. Lastly, participants were asked to analyze the initial DARI inventory (Figure 1) and were asked if they would use it and what changes they would make to it.

All focus groups were audio-recorded with the agreement of the participants. Audio recordings were fully transcribed by the researchers and were kept as Microsoft Word documents. The two researchers selected sample transcriptions and create a set of codes, which explained emerging patterns through inductive content analysis. Once all codes were developed, inter-coder reliability was tested. More than 85% inter-coder reliability was achieved. For systematic and efficient analysis, the software Nvivo 9 was used to assist with qualitative analysis and inter-coder reliability testing. In order to ensure that participants remained anonymous, all participants were assigned a nomenclature to link them to their data while de-identifying other personal information.

# 5 RESULTS AND ANALYSIS

## 5.1 Demographic Survey

Of the 14 participants, more than half had 5 to 10 years of active research in biology, physics, environmental science, geology, or marine science departments. The participants' specific research areas included ecological biogeography, climate change, physics education, cancer research, air and water quality, geologic mapping, and animal behavior. Slightly over half of the participants were doctoral students, 21.5 percent were post-doctorate researchers, and 21.5 percent were either faculty or research scientists. The average age of the participants was 41 years. Table 1 describes the participants.

**Table 1**. Description of Participants

| Department | Research Area | Position | Years of Research |
|---|---|---|---|
| Biology | Evolution/ Ecology | Doctoral Student | 5-10 |
| Biology | Biogeography/ Climate change ecology | Doctoral Student | 0-5 |
| Physics | | Doctoral Candidate | 5-10 |
| Physics | Magnetic materials | Post-Doc Researcher | 5-10 |

| | | | |
|---|---|---|---|
| Physics | Physics Education Research | Lecturer | 10-20 |
| Biology | Cancer Biology | Doctoral Candidate | 5-10 |
| Environmental Science & Engineering | Air quality modeling, atmospheric chemistry | Doctoral Student | 0-5 |
| Geology | Geologic mapping | Doctoral Candidate | 0-5 |
| Environmental Science & Engineering | Water + Sanitation; Water treatment processes | Post-Doc Researcher | 5-10 |
| Environmental Science & Engineering | Modeling/ Water Quality | Master's Student | 0-5 |
| Biology | Animal Behavior | Post-Doc Researcher | 20-30 |
| Marine Science | Ecology & public policy | Doctoral Student | 10-20 |
| Biology | Lemur Data Management/ life history | Research Scientist | 20-30 |
| Biology | | Post-Doc Researcher | 5-10 |

## 5.2 Focus Groups

Participants were asked to describe their research, the types of data they used in their research process, what they considered endangered data, and their opinions of the DARI inventory. For a full list of focus group questions see Appendix A. Topics and subtopics are listed in Table 2. Codes were developed during the coding process, as described in the procedures section. They were derived from inductive content analysis and were developed by the researchers as patterns emerged from the focus group transcripts. The main topics and subtopics discussed throughout the focus groups were: data types, data curation, endangered data, priority data, data reuse, data sharing, and the Data At Risk Inventory (Table 2). For example, as can be seen in column two, when participants discussed data curation, they typical considered data curation in reference to four categories: (1) data creation, (2) preservation action, (3) data storage, and (4) data transformation.

**Table 2.** Code words identified from focus groups

| **Data Types** | **Data Curation** | **Endangered Data** | **Priority Data** | **Data Reuse** | **Data Sharing** | **Inventory** |
|---|---|---|---|---|---|---|
| Digital | Create | Accessibility issues | Difficulty-Effort | Online | Incentives | Feature requests |
| Non-Digital | Preservation action | Lack of context | Valuable-Irreplaceable | Person | Disincentives | Reasons to use it |
| | Store | Potential endangerment | | Research group | | Reasons not to use it |
| | Transform | Unavailable | | | | |

The participants discussed these topics and subtopics throughout the focus groups. The results indicated that scientists are generally concerned with the possibility of data loss. Results also indicated that scientists view endangered data through multiple lenses, including lack of context and accessibility issues. Scientists communicated that they recognized the complexity of data at risk. Scientists also discussed when and how they reused and shared data as well as their opinions of the Data At Risk Inventory. The following section provides detailed results for the main topics participants discussed.

### 5.2.1 Results: data types

Participants were asked to describe the types of data they used in their research. Non-digital data that participants discussed included physical samples, lab notebooks, and field notebooks. In regard to digital data, participants suggested a variety of ways they acquired these data. For example, one participant received all data online, “All the data…is actually international data sets that are already online” (P5). Some participants received their data from private industry, for example, one participant stated that she received “proprietary data from industry...the oil and gas companies...it’s either in spreadsheet form or csv” (P8). Participants discussed the format types of the data they received, for example, one participant stated, “Pictures and videos and some numerical data that we save in excel files… sometime we search all the type of the data in the notebook” (P10). Participants also discussed non-digital types of data that they used in their research including handwritten notebooks, paper maps, field samples, and animal specimens.

### 5.2.2 Results: data curation

Participants discussed various aspects of the data life-cycle including data creation, preservation actions, storing data, and transforming data. In order to organize these topics, they were categorized based on definitions of the digital curation lifecycle from the Digital Curation Center (Digital Curation Centre, n.d.).

For example, participants discussed preservation actions such as duplication. One participant discussed how she duplicate her films, “I always have the film and then I try to scan that and then keep that file and then save a separate file” (P 9). Participants also discussed data storage. For example one participant placed her field samples into “… individual Ziploc bags that were labeled with the location, year, sample number, and page number from field notebook” (P6). Other participants discussed duplication and storage of digital files, for example, “We back it up in mass storage … so we have to copies of it all the times” (P8). Lastly, participants discussed the transformation of their data, this included digitizing the data or changing them to another format for analysis. For example, as one participant described “… exporting the data as a ASCI file, and at some point convert it into comma separated values, which get imported into another computer that has a graphics program, which are now the third format, which is used for analysis” (P3). While these results did not specifically address endangered data, participants felt that these types of curation activities were important in ensuring that data did not become endangered in the future. From this DARI learned that scientists are in general thinking about how to ensure that data do not become endangered and discovered the types of activities scientists are engaging in to ensure this.

### 5.2.3 Results: endangered data

We asked participants what they considered to be endangered data. Four major areas of concern were identified as a result of this question.

- Unavailable: The data were restricted or did not exist.
- Lack of context: The data were lacking metadata or the record keeping was poor.
- Accessibility issues: The data were degrading or were in an old format.
- Potential endangerment: The data were not backed up or not kept properly, hypothetically.

Participants noted that there were data that had been restricted or were non-existent. For example, participants noted that there are data unavailable to the public with, “They haven’t released it to the public” (P14) and “…they don’t have any free data on it, they just say they own it” (P6). Participants considered these data to be at risk because they could not access the data for the moment and had no way of knowing if the data had at risk characteristics.

Participants also seemed very aware that data that lacked context, even if they had access to these data, would still be unusable. For example, one participant stated, “It was theoretically good data, but then the associated materials go missing, you can’t used it” (P12). Another participant discussed how researchers are part of this problem, “Not all researchers are going to have recorded all the associated data” (P13). In general all participants agreed that keeping good metadata records and record keeping was extremely important. However, most agreed that they knew their record keeping practices were not as good as they should be. One participant stated, “There is no metadata, even a logical naming scheme would be a lot to ask for” (P3).

Participants discussed accessibility issues with data by explaining that some data are endangered due to data degradation or old file formats. For example one participant stated, “Most of the data we have is stored in a format that can’t be used” (P3). Another participant discussed how sometimes data degrades due to the instruments they are using, “The analytical instrument broke down, my samples were waiting for the freezer, but they expired in the end” (P5). Another participant discussed how sometimes environmental factors cause degradation, for example, he received “a big box of moldy notebooks” (P14). Overall, participants had anxiety over receiving data that they could not use due to accessibility issues. One participant summarized this point very succinctly with, “What kind of format should data be stored so that it can be retrieved with confidence later…getting that terrible error ‘unable to retrieve data’” (P4). This is something that all participants seemed to anticipate and worry about.

Lastly, participants showed concerned for data they considered to be potentially endangered, such as lab notebooks and tissue samples. For examples one participant stated, “If my lab book were stolen or destroyed I’d have a lot of numbers that I could do nothing with” (P3), indicating how valuable this information is and that without it he would no longer be able to analyze his data. Participants also discussed how potential endangerment was somewhat dependent on the type of data. For example, one participant stated, “Tissue and biological samples are very vulnerable” (P11).

### 5.2.4 Results: priority data

We asked participants to think about what data they considered to be their top priority. Two themes emerged from this question:

(1) Difficulty and/or effort
(2) Highly valuable and/or irreplaceable

In regard to difficulty and effort, participants noted that time was a considerable factor. For example one participant stated while discussing his data collecting efforts, “It took forever to compile…I would be devastated if I lost it” (P14). Another participant discussed how it took over four summers to create a map, “It's all written down, but it's over 250 points. I don't want to re-enter them in my excel spreadsheet, and try to put them back in my map” (P6). Another participant discussed not only the difficulty in the gathering data but also the importance of that data for his research with, “I have one accumulated dataset. It's three years of measurements that were extremely difficult to take and are the basis for my entire thesis and any journal articles” (P3).

Another theme that emerged was highly valuable and irreplaceable data. Most participants discussed how it was not possible to duplicate some of their data or too expensive to duplicate. For example one participant stated that “It’s one of those cases where no one else in the world has all that information” (P14). Another participant described how the data was extremely valuable based on the time and money it took to create. The participant stated, “We brought it to the lab from places like Japan, Nova Scotia, and Seattle. So that is very valuable since we spent lots of money on sending me out to the field” (P1). Another participant discussed that data could expire with, “I honestly would have cried if I lost all the experimental points that I had. So many times when I was making my data, I had a set amount of time before the sample expired.” (P5).

While these results do not specifically address endangered data, participants felt that these types of data were important to discuss because these data were the data they would most likely want to ensure never becoming at risk, given their importance. This informs the DARI that scientists are in general thinking about how to prioritize their data and which data is of most value to them.

### 5.2.5 Results: data reuse and data sharing

In order to understand if scientists use or encounter endangered data in their research process, participants were asked to discuss their data reuse and sharing practices in relation to endangered data. Participants discussed receiving data for reuse from people, research groups, or online. For example one participant discussed being offered data for reuse by another colleague, “I was talking about my research, and he was like ‘Oh, I have these sequences lying around maybe you could take a look at them” (P14). In this case, these sequences would have been discarded had this conversation not occurred.

Participants also discussed disincentives and incentives in relation to sharing and reusing data. In regard to disincentives, participants suggested that scooping and competition, sharing outside of their research group, equipment and technical issues, and metadata issues were their main reasons for not sharing data. Participants reiterated that metadata and technical issues were main reasons for how data becomes endangered because of the loss of context and inaccessibility issues. In regard to incentives, scientists discussed that the possibility of collaboration, additional publication, and moving science forward were incentives to sharing data.

While much of this discussion did not address endangered data directly, it did provide the researchers with an understanding of how scientists try to locate data for reuse as well as some the incentives and disincentives for sharing and reuse of data and did provide a context for how this relates to endangered data.

### 5.2.6 Results: Data At Risk Inventory

The final part of the focus group involved researchers asking participants to provide feedback on the initial DARI inventory (Figure 1). This step was taken in order to gain a perspective of the system and design changes that should be implemented to make the inventory more effective for scientists. Participants discussed reasons to use the inventory and reasons not to use the inventory. Reasons to use the DARI inventory included for meta-analysis, comparison studies, overflowing labs with too much data, to make room for new data, and for studies with long-term analysis. Participants suggested that time constraints, inability to assess the true condition and size of the data, and how it is often easier to create new data were reasons not to use the DARI inventory. Also, several of the participants suggested they would not contribute to the inventory unless the data were already published. Participants also made many suggestions to improve the initial interface and had specific design and feature requests. Through this feedback, changes were made to the DARI inventory, which will be discussed in the following analysis section.

## 5.3 Analysis

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 these data, the DARI focus group study investigated the questions below.

1. What perceptions do scientists have on the topic of data at risk?
2. What perceptions do scientists have of data reuse and sharing?
3. What opinions do scientists have regarding the Data At Risk Inventory?

Each of these research questions addresses different aspects of examining and facilitating the scientific process in relation to data at risk. For research question number 1, the researchers were trying to gain an understanding of scientists' perceptions of data at risk. Prior to this study, it was unclear if scientists were concerned with at risk data and if so what they believed these data to be. Research question 2 attempts to address how and if at risk data are part of scientists' research process since at risk data could be used during the sharing and reuse cycle. Research question 3 attempts to investigate if the current Data At Risk Inventory is facilitating scientists in their ability to share and reuse at risk data as well as to gain feedback on their current design.

The following will address each of these research questions through an analysis of the results discussed above.

### 5.3.1 RQ1: What perceptions do scientists have on the topic of data at risk?

One of the main purposes of this study was to determine scientists' perception of data at risk and to understand the importance of these data in the scientific process. In order to investigate this, we asked scientists very specific questions regarding their perceptions of data at risk as well as questions about the data they use throughout their research process. All of the scientists we spoke to had a general concern for endangered data. Scientists were concerned about the overall risk of losing data that they could use for their current or future research.

As discussed in the results section, four major areas of concern came from asking the scientists how they would describe endangered data. The areas of concern were (1) unavailability, meaning the data did not exist or were restricted, (2) lack of context, meaning the data were lacking metadata or the record keeping was poor, (3) accessibility issues, meaning the data were degrading or were in old format, and (4) potential endangerment, meaning the data were not backed up or not kept properly, hypothetically.

Each of these concerns points to considerations for understanding endangered scientific data and the scientific process. Scientists suggested that they worry that data will be unavailable. Several of the participants suggested that there were data they wished they could use; however, these data were not available to them due to proprietary restrictions. In reference to data not existing, in some cases, these data could be generated by experimentation or new observations; however, in some cases these data cannot be recreated and therefore once gone are lost forever. These types of data are key to considerations for data rescue missions, which is one of the purposes of the DARI initiative, to create an inventory of data at risk so that researchers do not lose valuable data that cannot be recreated.

Scientists' responses also indicated a concern for data having a lack of context. Most participants indicated that without appropriate metadata or record keeping, the data are essentially unusable. Metadata are a vital part of the scientific process because without metadata the data themselves are practically useless. Atkins, Hey, and Hedstrom (2011) describe, "Without such explicit schema and metadata, the interpretation is only implicit and depends strongly on the particular programs used to analyze it." .

Bruce and Hillman (2004) discussed similar concerns regarding accessibility as the scientists in this study did. They explained, "Metadata that cannot be read or understood by the user has no value….it may be unreadable for a variety of technical reasons, including the use of obsolete, unusual, or proprietary file formats." The scientists in this study indicated both a concern for lack of metadata as well as technical problems with file formats. The research conducted in this study provides a greater understanding of how endangered data impact the scientific research process. Finally, the study illustrates the importance for the DARI initiative to continue conducting research as well as provide support to scientists through the DARI inventory.

### 5.3.2 RQ2: What perceptions do scientists have of data reuse and sharing?

The second research question explored scientists' perceptions of data reuse and sharing. In order to investigate this, scientists were asked questions regarding their sharing and reuse practices.

When asked about their data sharing practices, scientists discussed incentives and disincentives. The disincentives included: scooping/competition, sharing outside of research group, equipment and technical issues, and metadata issues. Scooping and competition is a data sharing concern that has long been discussed in the scientific world (Borgman, Wallis, & Enyedy, 2007; Sayogo & Pardo, 2013; Sonnenwald, 2007), and the scientists in this study reinforced these ideas. The scientists' discussion of technical and metadata issues reinforced the above discussion; scientists are unable to reuse data that do not have appropriate metadata because without the metadata the data themselves are useless.

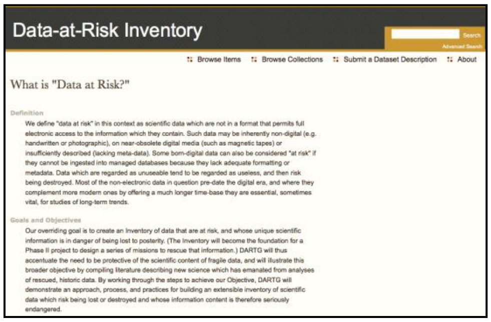
The incentives included: collaboration, additional publication, and moving science forward. These topics have been discussed thoroughly in data sharing literature. For example Lord and Macdonald (2003) suggested that data sharing reinforces scientific inquiry, encourages diversity in analysis, and promotes new ways to test hypothesis or methods of analyzing data. The scientists in this study echoed previous studies by suggesting that they believe data sharing should be the norm (Blumenthal et al., 2006; Blumenthal, Campbell, Anderson, Causino, & Louis, 1997; Ceci, 1988; McCain, 1995; Tenopir et al., 2011; Zimmerman, 2003). These findings indicate that many scientists believe that data reuse and sharing bring about important opportunities to the scientific process. The DARI research and subsequent inventory facilitate this process by exploring how endangered data are a part of scientific practice and providing scientists with the ability to deposit data descriptions into the DARI inventory.

### 5.3.3 RQ3: What opinions do scientists have in regard to the Data At Risk Inventory?

The final research question investigated the scientists' opinion of the Data At Risk interface design in order to ensure that the inventory met the needs of scientists who would use this resource. The initial Data At Risk Inventory (Figure 1) was created to provide an inventory of data at risk so that scientists would be able to access or deposit descriptions of these data.

The researchers asked the participants of the focus groups specifically if they would use this type of resource and also asked for design and function recommendations. Overwhelmingly, the scientists saw the importance of having this type of inventory available for their work and for data that could possibly be lost. One of the most common reasons discussed was the possibility of doing meta-analysis with the data that they could locate with this inventory.

Through feedback gathered from scientists during this study, significant changes have been made to the inventory. These updates include changes in search or browse features, changes in the submission process, and updates to the interface. More specific changes included adding a more detailed description of data at risk, the ability to browse by tags, browse by specific field, search by keyword, download associated files, such as PDFs, photographs, and handwritten notes as well as upload associated files, such as publications and presentations associated with the data, and the ability to submit descriptions of data rescue projects (Figures 2 and 3). Some of the features that were kept from the original design include citations, date, and size of collection. These changes were made in order to assist in the facilitation of the scientific process, and some were feature requests that the participants of the focus groups specifically asked for.



What is "Data at Risk?"

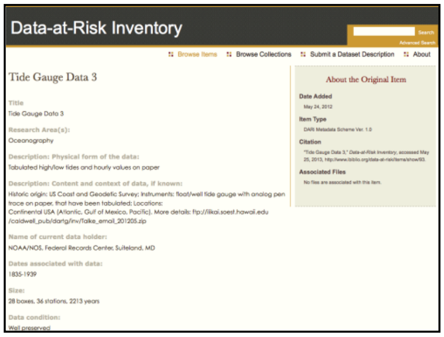
Definition

We define "data at risk" in this context as scientific data which are not in a format that permits full electronic access to the information which they contain. Such data may be inherently non-digital (e.g. handwritten or photographic), on near-obsolete digital media (such as magnetic tapes) or insufficiently described (lacking meta-data). 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. Data which are regarded as unuseable tend to be regarded as useless, and then risk being destroyed. Most of the non-electronic data in question pre-date the digital era, and where they complement more modern ones by offering a much longer time-base they are essential, sometimes vital, for studies of long-term trends.

Goals and Objectives

Our overriding goal is to create an Inventory of data that are at risk, and whose unique scientific information is in danger of being lost to posterity. (The Inventory will become the foundation for a Phase II project to design a series of missions to rescue that information.) DARTG will thus accentuate the need to be protective of the scientific content of fragile data, and will illustrate this broader objective by compiling literature describing new science which has emanated from analyses of rescued, historic data. By working through the steps to achieve our Objective, DARTG will demonstrate an approach, process, and practices for building an extensible inventory of scientific data which risk being lost or destroyed and whose information content is therefore seriously endangered.

**Figure 2**. Data at Risk Inventory redesigned homepage



Tide Gauge Data 3

Title
Tide Gauge Data 3

Research Area(s):
Oceanography

Description: Physical form of the data:
Tabulated high/low tides and hourly values on paper

Description: Content and context of data, if known:
Historic origin: US Coast and Geodetic Survey; Instruments: float/well tide gauge with analog pen trace on paper, that have been tabulated; Locations:
Continental USA (Atlantic, Gulf of Mexico, Pacific). More details: ftp://ilikai.soest.hawaii.edu /caldwell_pub/dartg/inv/Talke_email_201205.zip

Name of current data holder:
NOAA/NOS, Federal Records Center, Suitland, MD

Dates associated with data:
1835-1939

Size:
28 boxes, 36 stations, 2213 years

Data condition:
Well preserved

About the Original Item

Date Added
May 24, 2012

Item Type
DARI Metadata Scheme Ver. 1.0

Citation
"Tide Gauge Data 3," *Data-at-Risk Inventory*, accessed May 25, 2013, http://www.ibiblio.org/data-at-risk/items/show/93.

Associated Files
No files are associated with this item.

**Figure 3**. Data at Risk Inventory redesigned description page

## 6 CONCLUSION

This paper provides an overview of the Data At Risk Initiative (DARI) and reports on a focus group study conducted by the DARI team. These results informed the DARI about the current state of endangered scientific data from the perspective of scientists themselves. The purpose of this project was to gain an understanding of how endangered data are part of the scientific process, how this affects data reuse and sharing, and how the DARI inventory can help facilitate the scientific process. The findings of this research provide insight as to how scientists perceive data at risk and how such data affect their data reuse and sharing practices. The results also provide information as to how the DARI inventory can best fit scientists' needs. Another important contribution of this work is the design of a research study that can be used in future focus group studies to further knowledge in this area. A future focus group could ask scientists to discuss specific instances of data at risk that they have encountered. Through this research and the contribution of this work, the DARI, and other similar efforts, will be better able to assist in facilitating the needs of scientists and the scientific process in relation to endangered data.

## 7 ACKNOWLEDGEMENTS

I would like to acknowledge the support of the SILS Carnegie Fund, the UNC Center for Global Initiatives, and CODATA. Special acknowledgement to Jane Greenberg, Nico Carver, and W. Davenport Robertson. I would also like to thank the DARI team at http://ils.unc.edu/~janeg/dartg/index.html.

## 8 REFERENCES

Atkins, D., Hey, T., & Hedstrom, M. (2011) *Data and Visualization Final Report*. *Engineering*, 40. National Science Foundation - Advisory Committee for Cyberinfrastructure.

Axelsson, A.-L. & Östlund, L. (2001) Retrospective gap analysis in a Swedish boreal forest landscape using historical data. *Forest Ecology and Management*, *147*(2–3), 109–122.

Blumenthal, D., Campbell, E. G., Anderson, M. S., Causino, N., & Louis, K. S. (1997) Withholding research results in academic life science. Evidence from a national survey of faculty. *JAMA : the Journal of the American Medical Association*, *277*(15), 1224–8.

Blumenthal, D., Campbell, E. G., Gokhale, M., Yucel, R., Clarridge, B., & Hilgartner, S. (2006) Data Withholding in Genetics and the Other Life Sciences : Prevalences and Predictors. *Academic Medicine*, *81*(2), 137–145.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007) Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, *7*(1-2), 17–30. doi:10.1007/s00799-007-0022-9

Bruce, T. R. & Hillmann, D. I. (2004) The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In D. I. Hillmann & E. L. Westbrooks (Eds.), *Metadata in Practice*, 238–256. Chicago: American Library Association. doi:0838908829

Ceci, S. J. (1988) Scientists' Attitudes toward Data Sharing. *Science Technology And Human Values*, *13*(1/2), 45–52. doi:10.2307/690052

DARI. (n.d.) Data-at-Risk Inventory. Retrieved January 1, 2014 from the World Wide Web: http://www.ibiblio.org/data-at-risk/

DARI | Metadata Research Center (n.d.) Retrieved October 28, 2013 from the World Wide Web: http://ils.unc.edu/mrc/dari-2/

DARTG. (n.d.) CODATA "Data At Risk" Task Group (DARTG). Retrieved June 10, 2012 from the World Wide Web: http://ils.unc.edu/~janeg/dartg/

Digital Curation Centre. (n.d.) DCC Curation Lifecycle Model.

Gizzi, F. T. (2009) The electronic trading site eBay as a useful tool for obtaining historical data on natural events. *Computers & Geosciences 35*(9), 1950–1957.

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

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

Krotz, D. (2011) From Dusty Punch Cards, New Insights Into Link Between Cholesterol and Heart Disease. *Berkeley Lab News Center*. Retrieved July 11, 2012, from the World Wide Web: http://newscenter.lbl.gov/featurestories/2011/01/04/cholesterol-heart-disease/

Kuruppu, P. U., & Gruber, A. M. (2006) Understanding the Information Needs of Academic Scholars in Agricultural and Biological Sciences. *The Journal of Academic Librarianship 32*(6), 609–623. doi:10.1016/j.acalib.2006.08.001

Lord, P. & Macdonald, A. (2003) *e-Science Curation Report Data curation for e-Science in the UK : an audit to establish requirements for future curation and provision*. *Intellectual Property,* 1–84. The JISC Committee for the Support of Research (JCSR). Retrieved January 1, 2014 from the World Wide Web: http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Macnaghten, P., & Myers, G. (2004) Focus Groups. In C. Seale, G. Gobo, J. F. Gubrium, & D. Silverman (Eds.), *Qualitative Research Practice*. London ; Thousand Oaks, Calif: SAGE.

McCain, K. W. (1995) Mandating Sharing: Journal Policies in the Natural Sciences. *Science Communication 16*(4), 403–431. doi:10.1177/1075547095016004003

Meyer, E. T., Bulger, M. E., Zacharoudiou-Kyriakidou, A., Power, L., Williams, P., Venters, W., …, Wyatt, S. (2011) *Collaborative yet independent: Information practices in the physical sciences*. *Research Information Network*. Oxford.

Nordling, L. (2010) Researchers launch hunt for endangered data. *Nature 468*(7320), 17.

Rudin, C., Passonneau, R. J., Radeva, A., Ierome, 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

Sayogo, D. S. & Pardo, T. A. (2013) Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly 30*, S19–S31. doi:10.1016/j.giq.2012.06.011

Sonnenwald, D. H. (2007) Scientific collaboration. *Annual Review of Information Science and Technology 41*(1), 643–681. doi:10.1002/aris.2007.1440410121

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., …, Frame, M. (2011) Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE 6*(6), 21.

Wildemuth, B. M. and Jordan, M. W. (2009). Focus Groups. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (pp. 242–255). Westport, CT: Libraries Unlimited.

Zimmerman, A. S. (2003) *Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists*.

# 9 APPENDIX A

Focus Group Guide

1) To introduce yourself to the group, could you describe your research?

2) What kind of data do you use in your research?
- What formats? Describing the physical and digital aspects?

3) Are there any data you wish you had? But don't know how or where they are?
- Describe your "dream data" and what are the barriers to obtaining/finding them.

4) Are you reusing each other's data? How are you gaining access to the data? What types of data are you reusing? If you haven't reused data yourself, have you seen or heard of colleagues reusing data, if so describe this.

5) How would you describe data that are endangered or at risk of being lost? Which of your data do you consider your top priority for not wanting to lose? What measures are you using to protect and preserve your data? Could you describe any data (yours or your colleagues) that you consider highly at risk?

6) [Moderators show participants DARI inventory, http://www.ibiblio.org/data-at-risk]
Would you use this inventory? What don't we have here that you would need? Would this help your research? How would this help your research?

(Article history: Received 26 December 2013, Accepted 31 December 2013, Available online 7 January 2014)