Where Have All the Scientific Data Gone? LIS Perspective on the Data-At-Risk Predicament

Cheryl A. Thompson, W. Davenport Robertson, and Jane Greenberg

Scientists produce vast amounts of data that often are not preserved properly or do not have inventories, placing them at risk. As part of an effort to more fully understand the data-at-risk predicament, researchers who were engaged in the DARI project at UNC's Metadata Research Center surveyed information custodians working in a range of settings. The survey collected information on the data characteristics and preservation plans. Forty-three information custodians completed the survey. The results indicate that at-risk data include a variety of formats, subject areas, and ownership status, as well as compliance with a variety of standards. Although a majority of respondents agree that data preservation is important, they caution that time is the greatest barrier to sharing these data. The study has implications for data rescue and for training information custodians.

Experimental and observational data are central to scientific research. With a growing pool of scientific data, researchers have the potential to investigate new questions and use new analytical techniques. To enable this data revolution, the National Science Foundation has funded initiatives to develop cyberinfrastructure for managing, preserving, and sharing scientific data.[1] In addition to infrastructure development, there are increasing expectations of open access and new mandates for scientific data management and sharing. For instance, on February 22, 2013, the U.S. Office of Science and Technology Policy released a memo directing federal agencies to increase public access to federal research publications and digital data.[2] The memo called for federal agencies with over $100 million in research and development expenditures to develop policies and a plan for access and long-term preservation of government-sponsored research products. Open access presents many challenges for scientists. Scientists are so concerned about data sharing and preservation that the editors of *Science* recently devoted a special issue to the topic.[3]

Cheryl A. Thompson is a doctoral student in the Graduate School of Library & Information Science at the University of Illinois at Urbana-Champaign; e-mail: cathmps2@illinois.edu. W. Davenport Robertson is Professor of the Practice, and Jane Greenberg is Professor, both in the School of Information and Library Science at the University of North Carolina at Chapel Hill; e-mail: davrobertson1@gmail.com, janeg@email.unc.edu. *The survey was funded as part of the Data At Risk Student Learning Circle, UNC Center for Global Initiatives.*

doi:10.5860/crl.75.6.842

crl13-500

While the scientific process produces vast amounts of research data, often these data are not archived or inventoried properly, resulting in data at risk. At-risk data are often in fragile condition or stored on obsolete media, increasing the odds that these data will be ignored or lost forever. These data often contain valuable information that can be used for new investigations or historical analysis. To preserve data at risk, it is important to understand why data are at risk and the characteristics of these data.

Librarians and information science (LIS) professionals who are associated with collections containing scientific data can offer a valuable perspective on the data-at-risk predicament and preservation challenges. This paper reports on an exploratory and novel study of this population that explored how much and which types of data are at risk as well as data practices and factors that endanger these data. The paper presents a literature review of the research area, study methodology, findings, and a discussion of implications and future research.

Background

The Data-At-Risk Predicament

Recent technological advances have transformed scientific research into a largely digital process. From grant proposals to scholarly communication, new software and tools have allowed scientists to produce and store information in digital formats; in fact, manually collected data are usually transformed into a digital format for analysis.[4] However, vast amounts of scientific data, especially older studies, are in a format that does not permit full electronic access to the information that they contain. These data may be nondigital (such as photographic plate, plant specimen), stored on near-obsolete digital media (e.g., magnetic tapes), or insufficiently described, rendering them unusable. Data that are regarded as inaccessible tend to be at risk of being ignored and eventually destroyed.[5] According to Francine Berman and Vint Cerf, data are at risk "when economic models and infrastructure are not in place to ensure access and preservation."[6]

Perhaps most problematic is the expectation that all of the valued information is digitally accessible.[7] Scientists may be unaware of where valuable, older, and perhaps at-risk data reside. Exemplary work by Dr. R. Elizabeth Griffin shows that historical data about atmospheric ozone depended on her knowledge of the existence of these mostly unknown photographic plates.[8] Dr. Griffin, a Canadian astronomer, used historical plates of stellar spectra to measure telluric content and concentrations. These historical data enabled her to understand how the earth's ozone layer has evolved over time. Without her knowledge and perseverance in manually tracking down these plates, she would not have been able to conduct her research. Griffin's case illustrates how important it is to publicize historical, at-risk data.

Scientific data definitely have the potential to be used for research topics other than their initial purpose. Research literature is populated with data-reuse stories where new knowledge was gained.[9] For instance, John Gofman conducted a study of lipoproteins and heart disease, collecting data from almost 1,900 employees in 1956 and again in 1966.[10] In 1988 Paul Williams discovered the Gofman data, tracked down the participants to conduct a third follow-up, and created a data set with almost 1,900 participants spanning 29 years to investigate the role of lipoproteins in heart disease.[11] Williams's study exhibits how historical data can be repurposed for new analysis.

Data reuse enables scientific progress.[12] Carol Tenopir et al. conducted an international survey of scientists about their scientific data practices.[13] The survey included scientists from a variety of disciplines. Scientists reported being limited in answering research questions because of a lack of data sharing. In another example, PARSE Insight was a study of scientists, data managers, and publishers in Europe.[14] Similar to

Tenopir et al., the PARSE study found that both scientists and data managers cite the advancement of science as a reason for data preservation.

Most at-risk data predate the digital era, and preservation planning is vital and essential to prevent data loss in what may become obsolete formats. A few domains in science have developed data rescue efforts to preserve their historical data. In astronomy, data-driven initiatives have collected and rescued photoplates of the sky. Starting in the 1990s, both the Dominion Astrophysical Observatory in Canada and the Royal Observatory of Belgium started digitizing their collections to preserve the deteriorating plates.[15] In China, the recent Astronomical Plate Collection and Preservation project has created an environment that is conducive to plate preservation and has started to collect plates from observatories across China.[16] All of these projects have future plans for digitizing or creating metadata catalogs, but these plans have been hampered due to lack of funding.

In the first decade of the twenty-first century, the environmental sciences started data preservation initiatives. The International Environmental Data Rescue Organization has focused on preserving weather data in Africa.[17] The project provides equipment and training on data rescue and imaging. Since 2000, African weather data have been submitted to the National Oceanic and Atmospheric Administration for preservation. Also, the U.S. Geological Survey has preserved legacy data and continues to seek out other geological data at risk.[18] These initiatives use a variety of digital preservation techniques including purchasing equipment to migrate data from obsolete media, scanning or digitizing, adding metadata, and sending collections to the National Archives and Records Administration if appropriate.

An example of a more recent data preservation program is found at the Botanic Garden and Botanical Museum Berlin-Dahlem. Their efforts include digitization of specimens and achieving network interoperability for their collection.[19] In 2011, the museum started the reBiND project to develop workflows for rescuing biodiversity data and transforming them into a well-documented, standardized, and preservable format. The target data for these initiatives are narrowly focused on biodiversity research.

Primarily, data libraries and archives have targeted digital data. The Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP) has funded several data rescue initiatives.[20] The goal of NDIIPP is to develop a national strategy for collecting, preserving, and sharing digital information for future use. This program has preserved over 1,400 collections of at-risk digital materials. NDIIPP was an original funder of the Data Preservation Alliance for the Social Sciences (Data-PASS). Data-PASS is a partnership of repositories that were created to archive, catalog, and preserve at-risk social science data.[21] The data included significant public opinion polls, voting records, large-scale surveys, and other significant social studies. Both NDIIPP and Data-PASS focused on resources that were digitized or were born digital.

A crucial shortcoming of these data preservation initiatives in general is the absence of attention to nondigital scientific data from a diversity of scientific domains. Francine Berman further advocates for a census of research data to plan for data preservation.[22] Efforts need to be directed toward understanding nondigital, fragile research data that are vital to the future of science.

Information Custodians Working with Data

Historically, librarians and information scientists have been responsible for the management and preservation of manuscripts, artifacts, and digital objects. Information custodians have also engaged in the curation of research data via traditional library and special collection practice; more recently, they see this practice as an important future area of engagement. Data curation is "a means to collect, organize, validate and preserve data."[23] The ability to reuse data for new inquiries depends on the quality of data curation:

> Managing data of this kind requires discipline if the results are to be scientifically useful …Managing your data properly simply means keeping the necessary context information and associated documentation to make sure you and others can make use of your data when the need comes.[24]

A well-prepared and trained information workforce is the key to data curation success as recommended in both the American Council of Learned Societies and Atkins reports.[25] Historically, librarians and archivists have organized and managed large collections of materials including both physical and digital formats. LIS professionals have the potential to play a vital role in data curation, and many academic and research libraries already offer data services.[26] Bowker and Star describe the information professional as an intermediary between the domain scientists and computer scientists, focusing on how to design systems and workflows to support data quality.[27] Information professionals working in science, research, or other special libraries offer a unique viewpoint on data at risk, given their position in the organization and their LIS training. Understanding the data-at-risk predicament from the information custodian perspective is vital for preservation planning and efforts. However, research to date has not examined this perspective. This paper aims to fill this research void.

The Data-At-Risk Initiative (hereafter referred to as DARI) is a research project designed to inventory scientific data that are at risk of being lost forever.[28] DARI is a partnership between the University of North Carolina (UNC) Metadata Research Center, ibiblio, and the Council for Scientific and Technical Data's Data At Risk Task Group (CODATA/DARTG). The project employs mixed methods to identify collections of endangered data and to investigate this predicament from various perspectives. This paper only includes the results from a survey of information custodians. The survey goals were to:

1. Identify where at-risk data exist and the characteristics of these data.
2. Understand the data practices of these information centers.

Methodology

This study used the survey method to understand the information custodians' perspective on the data-at-risk predicament. In developing the survey, the team reviewed questions from data management and preservation surveys conducted by UNC, CyAir, Yale University, and Cornell University.[29] The survey collected information on the at-risk data such as type, format, volume, risk level, reasons, and future plans. Data sharing practices and demographics were also captured. The survey was pretested with data archivists and information professionals. The final survey included 40 questions, and skip patterns were used to tailor the survey to the respondent's experience.

The survey was intended for librarians, archivists, and information custodians who are involved in any aspect of data curation. A survey invitation and reminder were sent to selected discussion lists of the Association for Information Science and Technology (formerly American Society for Information Science and Technology), the Society of American Archivists, the Special Libraries Association, and the Association of College and Research Libraries. The web survey was available from February to March 2012.

The survey was accessed by 109 custodians. A response was counted as a completed survey if the respondent had answered at least through question #6 and as an eligible response if the respondent had at-risk data at his or her institution. Forty-three participants completed the full survey. Since the study employed discussion lists for survey recruitment, there is no way to know exactly how many people received the invitation. We estimate that the distribution lists have a total of approximately 3,000 members, resulting in a response rate of 1.5 percent. This is fairly consistent with research showing that paper invite/paper surveys have significantly higher response rates than e-mail invite/web surveys do.[30]

Despite limited response rates, researchers have discovered that web surveys produce higher-quality responses than offline surveys do.[31] Since the survey was exploratory in nature, and since our primary goal was to find evidence of at-risk data collections, we believe the responses met this goal and are adequate to draw some conclusions. The survey responses present a rich source of information about the data-at-risk predicament from the information custodian perspective, which has been ignored by previous research. Moreover, these data provide valuable information for designing data rescue initiatives and a baseline for comparing future studies.

Results

The survey had two main goals: 1) to identify collections of at-risk data; 2) to understand the data practices of institutions that house data at risk. For the purpose of this paper, we use the data set of 43 respondents who completed the full survey. Percentages are based on the 43 respondents unless otherwise noted. The results are presented according to the research goals.

Demographics

Demographic data were collected on the survey respondents. Information custodians were predominantly female (65%). The median age category was 46 to 50 years. Respondents were highly educated and held a variety of degree including a master's in LIS (53%), master's in other field (33%), PhD in LIS (7%), and PhD in another field (13%). In terms of professional identity, 52 percent identified as information professionals, 38 percent identified as librarians, 31 percent identified as scientists, and 28 percent identified as archivists.

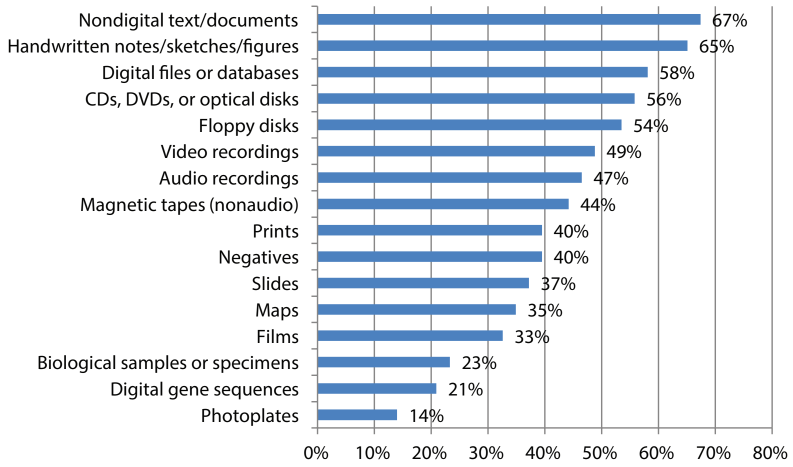
More than half of the respondents had worked in their current position for more than four years; the mean number of years working in their position was 8.9, with a range from less than a year to 32 years. In terms of current employer, 52 percent worked in academic institutions, 17 percent worked in corporations, 7 percent worked in government agencies, 3 percent worked in medical centers, and 20 percent worked in other types of institutions. Other institution types included research institutes and cultural heritage centers. The respondents performed a variety of duties such as reference and user services (70%) followed by arrangement and description (50%), outreach (50%), administration (47%), selection and appraisal (47%), preservation (47%), and systems and information technology (43%).

Data-at-Risk Characteristics

The survey gathered information about the formats of the endangered data in each respondent's institution. Overall, there was a wide range of formats, both digital and analog, reported to be at risk of being lost. The most common formats were nondigital text documents (67%), handwritten notes (65%), digital files or databases (58%), CD/DVD or optical disks (56%), and floppy disks (54%). Figure 1 provides a summary of the data-at-risk formats.

Furthermore, the survey asked custodians about the risk level of permanently losing these data formats. While more than half of the respondents rated all formats as very or somewhat likely to become lost, all respondents rated magnetic tapes, maps, negatives, photoplates, slides, and prints as very or somewhat likely to be lost. Floppy

FIGURE 1
Summary of the At-Risk Data Formats



disks (59%) and audio recordings (45%) topped the list of formats they rated as very likely to be lost forever (see figure 2).

Respondents were asked to indicate the various research areas of these at-risk data. The most common subjects were geology (42%), biology (39%), and climate (39%). The natural and social sciences were well represented in the data; however, few respondents reported topic areas from the humanities (see figure 3).

The survey asked respondents to indicate the current location of these at-risk data. A majority of data was stored in the institution's library or archive (75%) followed by

FIGURE 2
Likelihood of At-Risk Data Being Lost Forever by Format

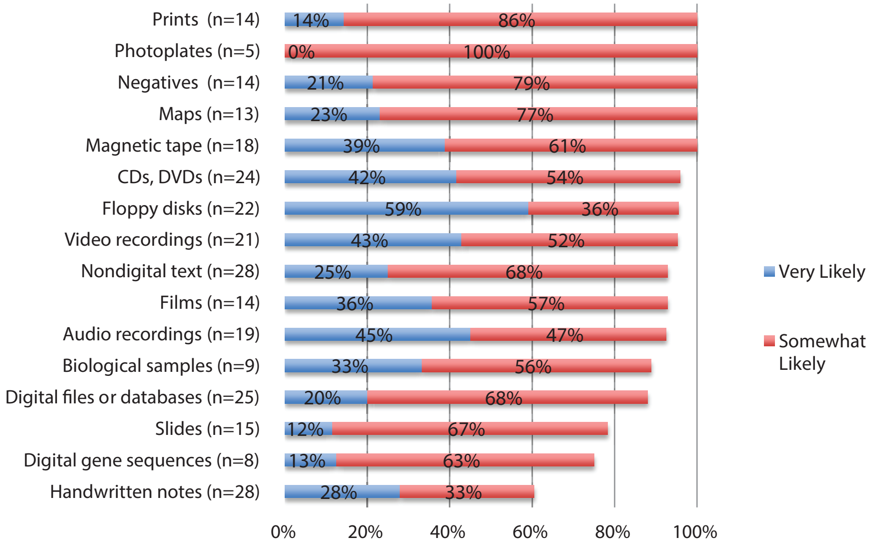
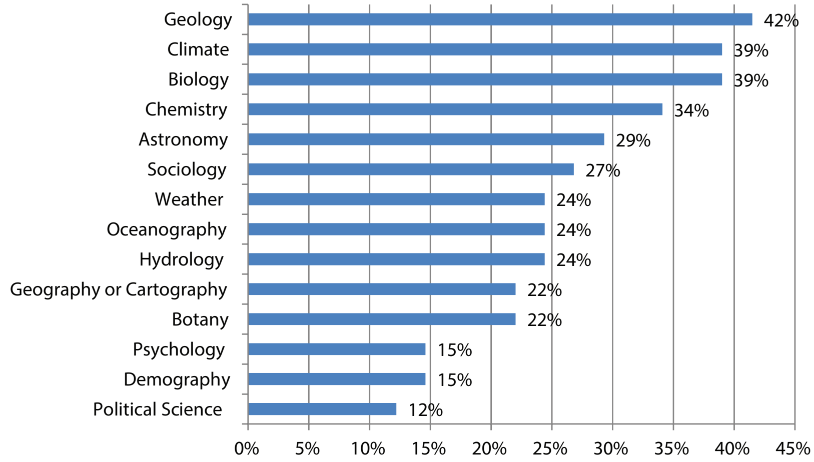


FIGURE 3
Research Area of At-Risk Data (n=41)



scientist's workspace (53%) and scientist's personal storage (40%). Only 25 percent of respondents reported storing data in external repositories.

Metadata are essential for the usability of these at-risk data. Respondents answered a series of questions about their metadata practices. As the most common response, 46 percent of respondents indicated that there was no metadata-enabled catalog or index for the endangered data. Of the custodians without a metadata catalog (n=22), 32 percent anticipated creating metadata, 18 percent did not anticipate producing metadata, and 50 percent were unsure. For those who were planning to create metadata (n = 7), most expected that a librarian, archivist, or information professional (57%) would be responsible for metadata generation, as well as expecting that the metadata catalog would be available for public use either at their institution (39%) or on their institution website (14%).

Of the respondents with existing data catalogs (n = 12), the metadata were usually created by information custodians (75%), scientists (42%), or other research staff (42%). For a majority of respondents, the catalogs were available at the institution (50%) or on the institution website (33%). A third of custodians did not make the metadata catalogs available to the public. A majority of custodians (64%) used metadata standards to produce their documentation. The standards included Dublin Core, Federal Geographic Data Committee (FGDC), Space Physics Archive Search and Extract (SPASE), Describing Archives: A Content Standard (DACS), Directory Interchange Format (DIF), and File Information Tool Set (FITS). In addition to these national and international standards, local and institution metadata standards were used.

Information custodians were asked for their opinion concerning the ownership of these endangered data. Most respondents thought the institution (73%) and the funding agency (41%) owned the at-risk data. Additional responses included the government (32%), public (32%), and the researcher (27%).

In designing data rescue efforts for these at-risk data, the ability to transfer preservation duties to another entity is an important consideration. The respondents were asked which entities they would allow to bear responsibility for preservation. The most frequently cited entities were a discipline-specific repository (47%), another research or academic center (44%), and an external library (44%). Over a third of the respondents would not allow another entity to bear responsibility for data at risk.

Data rescue efforts cannot save all of the scientific data that exist, so understanding the value of the data at risk is essential for preservation planning. The survey collected the custodians' opinions about how important it is to save these endangered data. A majority of respondents felt that at-risk data were very important (58%), followed by important (17%), somewhat important (19%), and not at all important (3%).

Data Practices

The final survey section inquired about the data practices in institutions that house data at risk. Respondents were asked if their organization complied with any standards or policies regarding data management, sharing, or archiving. The responses varied: 35 percent complied, 21 percent did not comply, and 30 percent were not sure. The guiding data standards or policies included the Open Archival Information System (OAIS) reference model, Preservation Metadata Maintenance Activity (PREMIS), and funder requirements.

Current data-sharing practices were collected. Professionals were asked to indicate how their institution shares data. A majority of custodians reported always or sometimes making scientific data available to the public (76%), upon request to the research staff (75%), or upon request to the library staff (77%). However, more than half (59%) indicated that sometimes data are not made publicly available.

Finally, custodians were asked to select factors that hampered the sharing of research data. The most common limits were the time involved in making data usable for others (72%) and accessing data files from storage media or repositories (59%). Other limits included maintaining human-subject confidentiality (35%) and protecting intellectual property rights (35%). A few custodians (3%) reported that no limits existed at their institutions.

Discussion

This section summarizes the findings, discusses their limitations, and suggests future research. This exploratory study highlights the diversity of scientific, at-risk data. While the endangered data represented a plethora of research areas, a majority of at-risk data were from the natural and social sciences. This is not surprising given the diversity of data that has been reported.[32] The at-risk data were stored in both analog and digital formats. The most at-risk data were stored on magnetic tapes and in analog formats such as maps, photoplates, or prints. Overall, custodians rated these data as important to preserve. Although the preservation of at-risk data is a valuable endeavor, the diversity of data characteristics will likely present challenges for preservation initiatives. This point is made clear by the Atkins report.[33]

The survey results also indicate that these data already comply with metadata and data management standards and policies. These standards encompass a variety of LIS and discipline-specific standards as well as international, national, and local mandates. The use of standards could assist information custodians in the inventorying and rescuing of at-risk data. In regard to data-sharing practices, the majority of at-risk data are made available to the public. However, data access may require contacting research or library staff or visiting the research center. The biggest threats to data sharing are accessing the data from storage and time to make the data usable. The scientists in the Tenopir et al. study also report insufficient time as a barrier to data sharing.[34] Data rescue efforts will need to secure funds to overcome these limitations.

It is important to note that this survey was exploratory in nature, and we recognize that the results only give the perspectives of those who participated in the survey. Although the survey data provide a rich source of information on the data-at-risk predicament, the data present certain limitations. The lack of generalizability of the data is a limitation. Little is known about librarians and information professionals who are working in data curation. It is not possible to compare our respondent demographics to the demographics of the population of data curation professionals. Geographic or discipline bias may exist in these data. Furthermore, the study design missed workers who were not members of professional LIS associations. The research provides a set of data for comparing future research that could extend to other populations, other geographic regions, and other methods. Additionally, this study is the first survey to ascertain the data-at-risk predicament from the perspective of information custodians.

Conclusion

In the quest for knowledge, scientists collect, process, and analyze data to answer research questions. Often, data are ignored after the original purpose has been fulfilled and are not properly preserved, which puts the data at risk. Old data often retain scientific value; stories of creating new scientific knowledge from older data collections are becoming more prevalent in the research literature and the news. Future research will be hampered if valuable historical data are lost.

This study provides an understanding of the data-at-risk predicament from the information custodian's perspective. Information professionals working in academic, science, or other special libraries reported that at-risk data include a variety of formats, subject areas, and ownership status as well as complying with a variety of standards. Although a majority of the respondents agreed that the data are important to preserve, time was the greatest barrier to sharing these data.

The National Data Stewardship Alliance advocates that research studies be broadened and replicated "to establish a robust evidence base from which generalizable guidance can be drawn."[35] There are several areas for future research, including understanding motivations for saving at-risk data and incentives for producing archive-ready data. DARI, the authors, are also interested in how to inventory fragile and at-risk data to understand the current state of data at risk and to inform data rescue efforts. Academic and research librarians will have expanded opportunities to design data rescue efforts and to shape data curation policies and practices, which will ensure that endangered scientific data are not lost to posterity.

These findings shed light on a topic of growing concern and are relevant for academic and research libraries. The understanding of the data-at-risk predicament can assist librarians, archivists, and scientists in designing and funding successful data rescue efforts. The study also has implications for LIS education. In North America, many universities have developed graduate programs to prepare information professionals for data curation.[36] An understanding of the data-at-risk predicament will enhance educators' ability to prepare and mentor students who want to pursue careers in data curation.

Appendix. Questionnaire Specs

Survey Title: Data At Risk Survey

Support contact information:
dari-sils@listserv.unc.edu, http://www.ibiblio.org/data-at-risk/

Question Mandatoriness: All questions are optional with a soft prompt included if the respondent does not provide an answer.
Soft Prompt Text:
We noticed that you did not answer a question on the previous page. To return to the last question, please click Previous and select an answer. Otherwise, click Next and you will advance to the next page.

Survey Sections:
About Your Data
About Your Data Practices
About You

Welcome page text:
Welcome to Data At Risk Survey!

The purpose of this survey is to gather information on endangered data and preservation practices in order to identify and located at-risk data. Your responses will help to inform the Data At Risk Inventory (DARI), a project of the *Committee on Data for Science and Technology (CODATA) Data at Risk Task Group*.

Thank you for taking the time to complete this survey. Please click on the START SURVEY to begin.

End Page Text:
Thank you for contributing to this important survey! By sharing your story you are making a positive contribution to the future of research data preservation.

To contribute to the Data At Risk Inventory, please click here to submit a description. Your submission will help to identify endangered data and inform data rescue efforts. The Data At Risk Inventory (DARI) is a project of the Committee on Data for Science and Technology (CODATA) Data at Risk Task Group.

Programming Notes:

Section I. About Your Data

Data come in many formats and each discipline has its own definition of data.

We define "Data at risk" as endangered data due to:

1. **A format that is deteriorating (e.g., photoplates) or near-obsolete (e.g., magnetic tapes).**
2. **Insufficient documentation or description for use.**
3. **Planned for destruction if not relocated.**
4. **Unknown to the scientific community.**

Q 1.1 If you have other view(s) of "data at risk," please share:
[TEXT BOX]

Q1.2 Please select any of the following format(s) that qualify as at-risk data at your institution.
Select all that apply.

1. Non-digital text/documents
2. Hand-written notes/sketches/figures
3. Prints
4. Negatives
5. Photoplates
6. Slides
7. Other non-digital format(Specify): [OPEN TEXT]
8. Audio recordings
9. Video recordings
10. Films
11. Maps
12. Digital files or databases
13. CDs, DVDs, or optical disks
14. Floppy disks
15. Magnetic tapes (non-audio)
16. Biological/organic/inorganic samples or specimens
17. Digital gene sequences or similar digital renditions of biological/organic/inorganic samples or specimens
18. Other (Specify): [OPEN TEXT]

[PRG: ONLY SHOW ITEMS SELECTED IN Q1.2]
Q1.3 Please indicated the risk level of permanently losing these endangered data at your institution.

1. Not at all likely
2. Somewhat likely
3. Likely
4. Very Likely

a. Non-digital text/documents
b. Hand-written notes/sketches/figures
c. Prints
d. Negatives
e. Photoplates
f. Slides
g. Other non-digital format: [PRG: SHOW TEXT RESPONSE FROM Q1.2_OTHER_NONDIGITAL]
h. Audio recordings
i. Video recordings
j. Films
k. Maps
l. Digital files or databases
m. CDs, DVDs, or optical disks
n. Floppy disks
o. Magnetic tapes (non-audio)
p. Biological/organic/inorganic samples or specimens

q. Digital gene sequences or similar digital renditions of biological/organic/inorganic samples or specimens
r. Other format [PRG: SHOW TEXT RESPONSE FROM Q1.2_OTHER]

[PRG: Q1.4-6 appear in table on one screen]

Q1.4 Please describe the reason(s) for the risk rating of these data.
[TEXT BOX]
a. Non-digital text/documents
b. Hand-written notes/sketches/figures
c. Prints
d. Negatives
e. Photoplates
f. Slides
g. Other non-digital format: [PRG: SHOW TEXT RESPONSE FROM Q1.2_OTHER_NONDIGITAL]
h. Audio recordings
i. Video recordings
j. Films
k. Maps
l. Digital files or databases
m. CDs, DVDs, or optical disks
n. Floppy disks
o. Magnetic tapes (non-audio)
p. Biological/organic/inorganic samples or specimens
q. Digital gene sequences or similar digital renditions of biological/organic/inorganic samples or specimens
r. Other format [PRG: SHOW TEXT RESPONSE FROM Q1.2_OTHER]

Q1.5 What is the expected life span of these endangered data?
[TEXT BOX]
a. Non-digital text/documents
b. Hand-written notes/sketches/figures
c. Prints
d. Negatives
e. Photoplates
f. Slides
g. Other non-digital format: [PRG: SHOW TEXT RESPONSE FROM Q1.2_OTHER_NONDIGITAL]
h. Audio recordings
i. Video recordings
j. Films
k. Maps
l. Digital files or databases
m. CDs, DVDs, or optical disks
n. Floppy disks
o. Magnetic tapes (non-audio)
p. Biological/organic/inorganic samples or specimens
q. Digital gene sequences or similar digital renditions of biological/organic/inorganic samples or specimens
r. Other format [PRG: SHOW TEXT RESPONSE FROM Q1.2_OTHER]

Q1.6 If possible, please estimate how much endangered data you have (e.g., 25 specimens, 35mm film, 2 GB, etc.).
[TEXT BOX]

a. Non-digital text/documents
b. Hand-written notes/sketches/figures
c. Prints
d. Negatives
e. Photoplates
f. Slides
g. Other non-digital format: [PRG: SHOW TEXT RESPONSE FROM Q1.2_OTHER_NONDIGITAL]
h. Audio recordings
i. Video recordings
j. Films
k. Maps
l. Digital files or databases
m. CDs, DVDs, or optical disks
n. Floppy disks
o. Magnetic tapes (non-audio)
p. Biological/organic/inorganic samples or specimens
q. Digital gene sequences or similar digital renditions of biological/organic/inorganic samples or specimens
r. Other format [PRG: SHOW TEXT RESPONSE FROM Q1.2_OTHER]

Q1.7 Please select where your endangered data are located.

- In the scientist's laboratory or workspace
- In the scientist's personal storage
- In my institution's library or archive
- In an external repository
- Other (Specify): [TEXT BOX]

Q1.8 Please select any of the following research areas that describe your endangered data.

1. Astronomy
2. Biology
3. Botany
4. Chemistry
5. Climate
6. Demography
7. Hydrology
8. Geography or Cartography
9. Geology
10. Oceanography
11. Political Science
12. Psychology
13. Sociology
14. Weather
15. Other (Specify): [TEXT BOX]

Q1.8a What is your opinion about ownership of the endangered data (i.e., who owns the data)?

1. Researcher

2. Institution
3. Funding agency
4. Government
5. Public (public domain)
6. Other (Specify): [OPEN TEXT]
7. No opinion [EXCLUSIVE]

Q1.9 For your endangered data, please indicate whether you would consider allowing any of the following entities to bear responsibility for data preservation activities? *Select all that apply*.

a. Discipline/domain specific repository
b. Publisher repository
c. Another research center or academic university
d. An external library or archive
e. I would not allow an entity outside of my organization to bear responsibility for these endangered data. [EXCLUSIVE]
f. Additional comments. [text box]

Q1.10 Please describe any efforts that your organization is taking or plans to take to save these endangered data.
[TEXT BOX]

Q1.10a In your opinion, please indicate how important it is to save these endangered data.

1. Not at all important
2. Somewhat important
3. Important
4. Very important

Please explain your rating. [text box]

"Metadata" refers to descriptive information or documentation about data.
Q1.11 At your institution, is there a catalog or index to the endangered data that uses metadata?

1. Yes
2. No
3. I'm not sure

Additional comments. [text box]

(Programming: If Q1.11 = 1, SHOW Q1.12A-B. OTHERWISE, SKIP to Q1.14)
Q1.12a At your institution, who creates the metadata for this catalog?

1. Scientist
2. Graduate/post-doctoral students
3. Other research staff
4. Information custodians/librarians
5. No one
6. I'm not sure

Additional comments. [text box]

Q1.12b Is this catalog or index publicly available:

a. At your institution

b. On your institution's website
c. Not publicly available [EXCLUSIVE]
d. Other (specify): [text box]

(PRG: ONLY SHOW Q1.13a-Q1.13c, IF Q1.11 = 2,3. OTHERWISE, SKIP to Q1.14)
Q1.13a Do you anticipate producing metadata for these endangered data?
1. Yes
2. No
3. I'm not sure

Additional comments. [text box]

(PRG: ONLY SHOW 1.13b-c, IF Q1.13a=1. OTHERWISE, SKIP to Q1.14)
Q1.13b Who will be primarily responsible for creating metadata for the endangered data?
1. Scientist
2. Graduate/post-doctoral students
3. Other research staff
4. Information custodians/librarians
5. No one
6. I'm not sure

Additional comments. [text box]

Q1.13c Will the metadata be publicly available:
- At your institution
- On your institution's website
- Not publicly available [EXCLUSIVE]
- Other (specify): [text box]

Q1.14 Does the metadata you have produced or intend to produce conform to known standards in your discipline?
1. Yes
2. No
3. I'm not sure

Please specify the standard(s) that you are using. [text box]

Section II. About Your Data Sharing Practices

This section is about your institution's data sharing practices.

Q 2.1 Do the data that your organization produces conform to any standards or policies regarding data management, sharing or archiving?
1. Yes
2. No
3. I'm not sure

Please specify the standard(s) that you are using and briefly describe them. [text box]

Q2.2. For each of the following statements, please indicate how often each statement is applicable to your organization's data.
1. Always
2. Sometimes
3. Never
4. Not Sure

a. A version of the data is made publicly available.
b. The data are not made publicly available, but **research staff** respond to individual requests.
c. The data are not made publicly available, but **librarians or information custodians** responds to individual requests.
d. The data are not made publicly available beyond members of the research team.

Q2.3. Which of the following limits data sharing after completion of the research project? *Select all that apply*.
1. Maintaining confidentiality of research participants
2. Gaining appropriate intellectual property rights protection
3. Accessing data files from storage media or data repositories/archives
4. Spending time involved in making data usable for others
5. Other (Specify): [text box]
6. No limits exist to data sharing. [RESPONSE IS EXCLUSIVE]

Section III. About You

In order to help us describe our sample and understand our findings, please tell us a bit about yourself.

D1. What is the title of your current job?
[TEXT BOX]

D2. How many years have you been working in your current position?
[Numeric, XX.XX] Years

D3. Which of the following best describes the type of institution you work in:
1. Academic
2. Health/medical
3. Corporation
4. Federal, state or local government
5. Other (specify): [TEXT BOX]

D4a. Please indicate the areas you have responsibilities within your current job. *Select all that apply*.
a. Administration
b. Selection or appraisal
c. Arrangement and description
d. Reference and user services
e. Preservation
f. Systems, information technology or web access
g. Outreach, advocacy or promotion
h. Other (Specify): [text box]

D4b. If you were going to provide someone with a brief overview of your current job, what would you tell them? Please include your daily duties or responsibilities.
[TEXT BOX]

D4c. For each of the following decisions regarding research data, please indicate whether you have complete or share decision-making authority at your institution.

1. Complete authority
2. Share authority
3. No authority

a. Deciding which data are important to preserve
b. Deciding whether data can be safely shared
c. Determining standards for de-identifying sensitive data
d. Determining what constitutes compliance with commercial licenses, government regulations, funding agency mandates, etc.
e. Determining the appropriate metadata to describe data sets (i.e., descriptive information to enable others to reuse data)
f. Determining provisions for short-term data preservation (5 years or less)
g. Determining provisions for long-term data preservation (more than 5 years)
h. Deciding the circumstances under which data should be submitted to a long-term preservation provider

D4d. Please describe any other duties or responsibilities that you have regarding research data.
[TEXT BOX]

D4e. Please describe any special talents, skills, prior education, or experiences that helped you get your current job. [OPEN END RESPONSE]

D5a. Please indicate the educational degree(s) that you hold.

1. Associate Degree
2. Bachelor in LIS
3. Bachelor of Arts
4. Bachelor of Science
5. Masters in library science and/or information science
6. Master of Arts
7. Master of Science
8. PhD in LIS
9. PhD in other field
10. Professional degree, please specify [TEXT BOX]
11. Other degree, please specify [TEXT BOX]

(Programming: If D5a = 1, SHOW D5b_1)
D5b_1. What was your major area of study for this Associate degree? [TEXT BOX]

(Programming: If D5a = 3, SHOW D5b_3)
D5b_3. What was your major area of study for this Bachelor of Arts degree? [TEXT BOX]

(Programming: If D5a = 4, SHOW D5b_4)
D5b_4. What was your major area of study for this Bachelor of Science degree? [TEXT BOX]

(Programming: If D5a = 6, SHOW D5b_6)
D5b_6. What was your major area of study for this Masters of Arts degree? [TEXT BOX]

(Programming: If D5a = 7, SHOW D5b_7)

D5b_7. What was your major area of study for this Masters of Science degree? [TEXT BOX]
(Programming: If D5a=9, SHOW D5b_9)
D5b_9. What was your major area of study for this PhD degree? [TEXT BOX]

D6. Do you currently consider yourself to be:
1. A librarian
2. An information professional
3. An archivist
4. A scientist
5. Other (specify): [text box]

D7. What is your age category?
1. 25 years or younger
2. 26-30 years
3. 31-35 years
4. 36-40 years
5. 41-45 years
6. 46-50 years
7. 51-55 years
8. 56-60 years
9. 61-65 years
10. 66 years or older
11. Prefer not to answer

D8. What is your sex?
1. Male
2. Female
3. Prefer not to answer

D9. Please provide any additional comments regarding the management or endangerment of research data.
[TEXT BOX]

D10. if we have additional questions about your survey responses, may we contact you?
1. Yes
2. No

(PRG: SHOW D11 IF D10=1; OTHERWISE GO TO END SCREEN)
D11. Please give us your name, current email address or phone number where our research staff can contact you. Your contact information will be stored separately from the survey data.

D11a. Name: [TEXT BOX]
D11b. Email or phone number: [TEXT BOX]

END OF SURVEY

Notes

1. National Science Foundation, "Press Release 12-060 NSF Leads Federal Efforts In Big Data," available online at www.nsf.gov/news/news_summ.jsp?cntn_id=123607 [accessed 1 August 2013].

2. John P. Holdren, "Public Access Memorandum from the Office of Science and Technology Policy," (2013), available online at www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf [accessed 1 August 2013].

3. Science Staff, "Dealing with Data (Special Issue)," *Science* 331, no. 6018 (2011): 692–93.

4. Jane Greenberg, Hollie C. White, Sarah Carrier, and Ryan Scherle, "A Metadata Best Practice for a Scientific Data Repository," *Journal of Library Metadata* 9, no. 3/4 (2009): 194–212.

5. P. Bryan Heidorn, "Shedding Light on the Dark Data in the Long Tail of Science," *Library Trends* 57, no. 2 (2008): 280–99.

6. Francine Berman and Vint Cerf, "Who Will Pay for Public Access to Research Data?" *Science* 341, no. 6146 (2013): 616–17.

7. Paige Shaughnessy, "Reaching Millennials through Innovations in Teaching," *Perspectives on Issues in Higher Education* 12, no. 1 (2009): 4–15; Susan Gardner and Susanna Eng, "What Students Want: Generation Y and the Changing Function of the Academic Library," *portal: Libraries and the Academy* 5, no. 3 (2005): 405–20.

8. Elizabeth Griffin, "Rescuing and Recovering Lost or Endangered Data," *Data Science Journal* 4 (2005): 21–26; Elizabeth Griffin, "Detection and Measurement of Total Ozone from Stellar Spectra: Paper 2. Historic Data from 1935–1942," *Atmospheric Chemistry and Physics* 6, no. 8 (2006): 2231–40.

9. Griffin, "Detention and Measurement"; Dan Krotz, "From Dusty Punch Cards, New Insights into Link Between Cholesterol and Heart Disease," (2011) available online at http://newscenter.lbl.gov/feature-stories/2011/01/04/cholesterol-heart-disease/ [accessed 1 August 2012]; Cynthia Rudin, Rebecca J. Passonneau, Axinia Radeva, Steve Ierome, and Delfina F. Isaac, "21st-Century Data Miners Meet 19th-Century Electrical Cables," *Computer* 44, no. 6 (2011): 103–05.

10. John W. Gofman, Wei Young, and Robert Tandy, "Ischemic Heart Disease, Atherosclerosis, and Longevity," *Circulation* 34, no. 4 (1966): 679–97; Krotz, "Dusty Punch Cards."

11. Paul T. Williams and Daniel E. Feldman, "Prospective Study of Coronary Heart Disease vs. HDL2, HDL3, and Other Lipoproteins in Gofman's Livermore Cohort," *Atherosclerosis* 214, no. 1 (2011): 196–202; Krotz, "Dusty Punch Cards."

12. Tom Kuipers and Jeffrey van der Hoeven, "Insights into Digital Preservation of Research Output in Europe: PARSE-Insight Survey Report," (2009), available online at www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf [accessed 12 October 2010]; Carol Tenopir, Suzie Allard, Kimberly L. Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame, "Data Sharing by Scientists: Practices and Perceptions," *PLoS One* 6, no. 6 (2011): e21101.

13. Tenopir et al., "Data Sharing by Scientists," e21101.

14. Tom Kuipers and Jeffrey van der Hoeven, "Insights into Digital Preservation of Research Output in Europe," available online at www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf [accessed 1 August 2012].

15. National Research Council Canada, "Dominion Astrophysical Observatory," available online at www.nrc-cnrc.gc.ca/eng/facilities/hia/astrophysical-observatory.html [accessed 1 August 2012]; Royal Observatory of Belgium, "Astronomical Events and News," available online at www.astro.oma.be/EN/hotnews/index.php [accessed 1 August 2012].

16. Wen-Jing Jin, Zheng-Hong Tang, Shu-He Wang, Bao-An Yao, Kai-Ping Tian, Li Chen, Yong-Heng Zhao, and Shi-Yang Jiang, "Review on IAU Work for Preservation and Digitization of Astronomical Photographic Plates and Suggestions of Plates Digitization in China," *Progress in Astronomy* 25 (2007): 1–12.

17. IEDRO, "Data Rescue," available online at www.iedro.org/en/datarescue/index.html [accessed 1 August 2012].

18. U.S. Geological Survey, "US Geological Survey," available online at www.usgs.gov/ [accessed 1 August 2012].

19. Free University of Berlin, "Botanic Garden and Botanical Museum Berlin-Dahlem Projects," available online at www.bgbm.org/BioDivInf/projects-e.htm [accessed 1 August 2012].

20. Library of Congress, "Digital Preservation: Program Background," available online at www.digitalpreservation.gov/about/background.html [accessed 1 August 2012].

21. Micah Altman, Margaret O. Adams, Jonathan Crabtree, Darrell Donakowski, Marc Maynard, Amy Pienta, and Copeland H. Young, "Digital Preservation through Archival Collaboration: The Data Preservation Alliance for the Social Sciences," *American Archivist* 72, no. 1 (2009): 170–84; Myron P. Gutmann et al., "From Preserving the Past to Preserving the Future: The Data-PASS

Project and the Challenges of Preserving Digital Social Science Data," *Library Trends* 57, no. 3 (2009): 315–37.

[22] Francine Berman, "We Need a Research Data Census," *Communications of the ACM* 53, no. 12 (2010), 39–41.

[23] Sayeed Choudury, "Data Curation: An Ecological Perspective," College *& Research Libraries News* 71, no. 4 (2010): 194–96.

[24] Chris Rusbridge, "Create, Curate, Reuse: The Expanding Life Course of Digital Research Data," paper presented at EDUCAUSE Australasia (2007): 2–3, available online at http://hdl.handle.net/1842/1731 [accessed 1 August 2012].

[25] American Council of Learned Societies, "Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences" (2006), available online at *www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf* [accessed 1 August 2012]; Daniel Atkins, "Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure" (2003), available online at www.nsf.gov/cise/sci/reports/atkins.pdf [accessed 1 August 2012].

[26] Association of Research Libraries Digital Repository Issues Task Force, "The Research Library's Role in Digital Repository Services: Final Report of the ARL Digital Repository Issues Task Force," Association of Research Libraries (2009), available online at www.arl.org/storage/documents/publications/repository-services-report-jan09.pdf [accessed 1 August 2012]; Sarah Higgins, "Digital Curation: The Emergence of a New Discipline," International Journal of Digital Curation 6, no. 2 (2011): 78–88; Christopher Prom, "Making Digital Curation a Systematic Institutional Function," *International Journal of Digital Curation* 6, no. 1 (2011): 139–52; Donald J. Waters, "Doing Much More Than We Have So Far Attempted," *EDUCAUSE Review* 42, no. 5 (2007): 8; Anna K. Gold, "Cyberinfrastructure, Data, and Libraries," *D-Lib Magazine* 13, no. 9/10 (2007); Patricia Hswe and Ann Holt, "Guide for Research Libraries: The NSF Data Sharing Policy ARL Transforming Research Libraries," Association of Research Libraries (2011), available online at http://old.arl.org/rtl/eresearch/escien/nsf/index.shtml [accessed 1 August 2012].

[27] Geoffrey C. Bowker and Susan Leigh Star, "Cyberscholarship; or , 'A Rose Is a Rose Is a…'," *EDUCAUSE Review* 44, no. 3 (2009): 6–7.

[28] UNC Metadata Research Center, "Data-At-Risk Initiative," available online at http://ils.unc.edu/mrc/dari/ [accessed 1 August 2012].

[29] Gail Steinhart, Eric Chen, Florio Arguillas, Dianne Dietrich, and Stefan Kramer, "Research Data Management Service Group Survey of NSF Principal Investigators at Cornell University" (2011), available online at http://hdl.handle.net/1813/25624 [accessed 1 December 2011]; CyAir, "Air Quality Data Questionnaire" (2010), available online at http://cyair.net/node/46 [accessed 1 December 2011]; Yale University, "Research Data Task Force Report: Yale University Research Data Interviews" (2010), available online at http://ydc2.yale.edu/projects/research-data-task-force [accessed 1 December 2011]; University of North Carolina, "Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership" (2012), available online at http://sils.unc.edu/sites/default/files/general/research/UNC_Research_Data_Stewardship_Report.pdf [accessed 15 February 2012].

[30] Michele M. Hayslett and Barbara M. Wildemuth, "Pixels or Pencils? The Relative Effectiveness of Web-based versus Paper Surveys," *Library & Information Science Research* 26, no. 1 (2005): 73–93.

[31] Barrie Gunter, David Nicholas, Paul Huntington, and Peter Williams, "Online versus Offline Research: Implications for Evaluating Digital Media," in *Aslib Proceedings,* vol. 54, no. 4 (2002): 229–39.

[32] Atkins, "Revolutionizing Science and Engineering through Cyberinfrastructure"; Douglas Laney, "3D Data Management: Controlling Data Volume, Velocity and Variety," Application Delivery Strategies, File 949 (2001).

[33] Atkins, "Revolutionizing Science and Engineering through Cyberinfrastructure."

[34] Tenopir et al., "Data Sharing by Scientists," e21101.

[35] National Data Stewardship Alliance, "2014 National Agenda for Digital Stewardship" (2013), 23, available online at www.digitalpreservation.gov/ndsa/documents/2014NationalAgenda.pdf [accessed 15 August 2013].

[36] Carolyn Hank, Helen R. Tibbo, and Christopher A. Lee, "DigCCurr I Final Report, 2006–09: Results and Recommendations from the Digital Curation Curriculum Development Project and the Carolina Digital Curation Fellowship Program" (2010), available online at www.ils.unc.edu/digccurr/digccurr_I_final_report_031810.pdf [accessed 1 August 2012]; Carole Palmer, Bryan P. Heidorn, Dan Wright, and Melissa H. Cragin, "Graduate Curriculum for Biological Information Specialists: A Key to Integration of Scale in Biology," *International Journal of Digital Curation* 2, no. 2 (2008): 31–40.