
RESEARCH PAPER

Risk Assessment for Scientific Data

Matthew S. Mayernik¹, Kelsey Breseman², Robert R. Downs³, Ruth Duerr⁴, Alexis Garretson⁵, Chung-Yi (Sophie) Hou⁴, Environmental Data Governance Initiative (EDGI) and Earth Science Information Partners (ESIP) Data Stewardship Committee

¹ National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR), US

² Environmental Data & Governance Initiative (EDGI), US

³ Columbia University, Center for International Earth Science Information Network (CIESIN), US

⁴ Ronin Institute for Independent Scholarship, US

⁵ George Mason University, Department of Biology, School of Systems Biology, US

Corresponding author: Matthew S. Mayernik (mayernik@ucar.edu)

Ongoing stewardship is required to keep data collections and archives in existence. Scientific data collections may face a range of risk factors that could hinder, constrain, or limit current or future data use. Identifying such risk factors to data use is a key step in preventing or minimizing data loss. This paper presents an analysis of data risk factors that scientific data collections may face, and a data risk assessment matrix to support data risk assessments to help ameliorate those risks. The goals of this work are to inform and enable effective data risk assessment by: a) individuals and organizations who manage data collections, and b) individuals and organizations who want to help to reduce the risks associated with data preservation and stewardship. The data risk assessment framework presented in this paper provides a platform from which risk assessments can begin, and a reference point for discussions of data stewardship resource allocations and priorities.

Keywords: risk assessment; data preservation; data stewardship; metadata

Introduction

At¹ the “The Rescue of Data At Risk” workshop held in Boulder, Colorado on September 8th and 9th, 2016, 2 participants were asked the following question: “How would you define ‘at risk’ data?” Discussions on this point ranged widely, and touched on several challenges, including lack of funding or personnel support for data management, natural and political disasters, and metadata loss. One participant’s organization’s definition of risk, however, stood out: “data were considered to be at risk unless they had a dedicated plan to not be at risk.” This simple statement vividly depicts how data’s default state is being in a state of risk. Thus, ongoing stewardship is required to keep data collections and archives in existence.

The risk factors that a given data collection or archive may face vary depending on the data’s characteristics, the data’s current environment, and the priorities and resources available at the time. Many risks can be reduced or eliminated by following best practices codified as certifications and guidelines, such as the CoreTrustSeal Data Repository Certification (2018) and the ISO 16363:2012. This ISO standard defines audit and certification procedures for trustworthy digital repositories (ISO 2012b). Both the CoreTrustSeal certification and ISO 16363:2012 are based on the ISO 14721:2012 standard that defines the Open Archival

¹ We list EDGI and the ESIP Data Stewardship Committee as authors due to the contributions of many individuals from both organizations to the work described in this paper. The named authors are the individuals involved in each organization who contributed directly to the paper’s text.

² The workshop was organized under the auspices of the Research Data Alliance (RDA) and the Committee on Data (CODATA) within the International Science Council, <http://www.codata.org/task-groups/data-at-risk/dar-workshops>.

Information System (OAIS) Reference Model (ISO 2012a). But these certifications can be large and complex. Additionally, many of the organizations that hold valuable scientific data collections may not be aware of these standards, even if the organizations are potentially resourced to tackle the challenge (Maemura, Moles & Becker 2017). Further, the attainment of such certifications does not necessarily reduce the risks to data that are outside of the scope of a particular certification instrument.

This paper presents an analysis of data risk factors that scientific data collections and archives may face, and a matrix to support data risk assessments to help ameliorate those risks. The three driving questions for this analysis are:

- How to assess what data are at risk?
- How to characterize what risk factors data collections and/or archives face?
- How to make risks more transparent, internally and/or externally?

The goals of this work are to inform and enable effective data risk assessment by: a) individuals and organizations who manage data collections, and b) individuals and organizations who want to help to reduce the risks associated with data preservation and stewardship. Stakeholders for these two activities include producers, stewards, sponsors, and users of data, as well as the management and staff of the institutions that employ them.

Background

This project has been coordinated through the Data Stewardship Committee within the Earth Science Information Partners (ESIP), a non-profit organization that exists to support collection, stewardship, and use of Earth science data, information, and knowledge.³ The immediate motivation for the project stemmed from the Data Stewardship Committee members engaging with groups who were undertaking grass-roots “data rescue” initiatives after the 2016 US presidential election. At that time, a number of loosely organized and coordinated efforts were initiated to duplicate data from US government organizations to prevent potential politically motivated data deletion or obfuscation (See for example Dennis 2016; Varinsky 2017). In many cases, these initiatives specifically focused on duplicating government-hosted Earth science data.

ESIP Data Stewardship Committee members wrote a white paper to provide the Earth science data centers’ perspective on these grass-roots “data rescue” activities (Mayernik et al. 2017). That document described essential considerations within day-to-day work of existing federal and federally-funded Earth science data archiving organizations, including data centers’ constant focus on documentation, traceability, and persistence of scientific data. The white paper also provided suggestions for how the grass-roots efforts might productively engage with the data centers themselves.

One point that was emphasized in the white paper was that the actual risks faced by the data collections may not be transparent from the outside. In other words, “data rescue” activities may have in fact been duplicating data that were at minimal risk of being lost (Lamdan 2018). This point, and the white paper in general, was well received by people inside and outside of these grass-roots initiatives (Cornelius & Pasquetto 2017; McGovern 2017). Questions then came back to the ESIP Data Stewardship Committee about how to understand what data held by government agencies were actually at risk.

The analysis presented in this paper was initiated in response to these questions. Since then, these grass-roots “data rescue” initiatives have had mixed success in sustaining and formalizing their efforts (Allen, Stewart & Wright 2017; Chodacki 2018; Janz 2018). The intention of our paper is to enable more effective data risk assessment broadly. Rescuing data after they have been corrupted, deleted, or lost can be time and effort intensive, and may be impossible (Pienta & Lyle 2018). Thus, we aim to provide guidelines to any individual or organization that manages and provides access to scientific data. In turn, these individuals and organizations can better assess the risks that their data face, and characterize those risks.

When discussing risk and, in particular, data risk, it is useful to ask the question: what is the objective that is being challenged by the possible risk factors? With regard to data, in general, discussions of risk might presume that “risks” threaten the current or future access to data by the potential data users. Currently, continuing public access to and use of scientific data is particularly relevant in light of recent open data and open science initiatives. In this regard, risks for scientific data include factors that could hinder, constrain, or limit current or future data use. Identifying such risk factors to data use offers further analysis opportunities to prevent, mitigate, or eliminate the risks.

³ http://wiki.esipfed.org/index.php/Preservation_and_Stewardship.

Data Risk Assessment

Risk assessment is a regular activity within many organizations. In a general sense, risk management plans are complementary to project management plans (Cervone 2006). Organizational assessment of digital data and information collections is likewise not new (Maemura, Moles & Becker 2017). The analysis presented in this paper builds on prior work in a number of areas: 1) research on data risks, 2) data rescue initiatives within government agencies & specific disciplines, 3) CODATA and RDA working groups & meetings, 4) trusted repository certifications, and 5) knowledge and experience of the ESIP Data Stewardship Committee members. **Table 1** summarizes data risk factors that emerge from these knowledge bases. The list of risk factors shown in **Table 1** is not meant to be exhaustive. Rather, it provides a useful illustration of the diverse ways in which data sets, collections, and archives might encounter risks to data usability and accessibility. The rest of this section details further key insights from the five areas of prior work noted above.

Research on data risks

A range of studies have explored the kinds of risks that scientific data may face, and potential ways to mitigate specific risk factors. Many of these studies touch on practices that are typical of scientific data archives. Metadata, for example, can be considered both a risk factor and a mitigation strategy. Insufficient metadata is itself a potential factor that can reduce the discoverability, usability, and preservability of data, particularly in situations where direct human knowledge of the data is absent (Michener et al. 1997). In fact, many data rescue projects find that the “rescue” efforts must be targeted much more toward metadata than data (see Knapp, Bates & Barkstrom 2007; Hsu et al. 2015). This might be the case for a couple of reasons. First, insufficient or missing metadata might prevent data from being usable regardless of the condition of the data themselves. Examples include missing column headers in tabular data that prevent a user from knowing what the data are representing, and insufficient provenance metadata that prevent users from trusting the data due to lack of context about data collection and quality control. Second, metadata are also central to documenting and mitigating risks as they manifest while preventing risks from becoming problematic in the future (Anderson et al. 2011). For example, documenting data ownership and usage rights is an essential step in mitigating the risk factor “Legal status for ownership and use” from **Table 1**.

Different kinds of metadata might be necessary to reduce specific data risks. For example, specifications of file format structures are a critical type of metadata for mitigating risks associated with digital file format obsolescence. Open specifications complement other critical mitigation practices and tools related to file format obsolescence. As one example, keeping rendering software available is an important way to retain access to particular file formats, but this typically also requires maintaining documentation of how the rendering software works (Ryan 2014).

Other risk factors (listed in **Table 1**) relate to the sustainability and transparency of the archiving organization. These factors are important in ensuring the accessibility of the data and the trustworthiness of the archive. As Yakel et al. (2013) note, “[t]rust in the repository is a separate and distinct factor from trust in the data” (pg. 154). For people outside of the repository, “institutional reputation appears to be the strongest structural assurance indicator of trust” (pg. 154). Effective communication about data risks and steps taken to eliminate problems is helpful in ensuring users that the archive is trustworthy (Yoon 2017).

Data that face extreme or unusual risks, however, may not be manageable via typical data curation workflows. Downs and Chen (2017) note that dealing with some data risk factors “may well require divergence from regular data curation procedures as tradeoffs may be necessary” (pg. 273). For example, Gallaher et al. (2015) undertook an extensive project to recover, reconstruct, and reprocess data from early satellite missions into modern formats that are usable by modern scientists. This project involved dealing with degrading and fragile magnetic tapes, extracting data from the tapes’ unusual format, and recreating documentation for the data. Natural disasters, fires, and floods also present unpredictable risk factors to data collections of all kinds. While these kinds of events can be planned for and steps can be taken to prevent the occurrence of some of them (e.g. fires), they can still cause major data loss and/or require significant recovery effort.

Mitigating risks, of whatever kind, takes effort and resources. The time required to create metadata, reformat files, create contingency plans, and communicate these efforts to user communities can be considerable. This time investment can be the greatest barrier to performing risk assessment and mitigation activities (Thompson, Robertson & Greenberg, 2014). Putting focus on assessment of data risk factors may mean that “certain priorities need to be re-ordered, new skills acquired and taught, resources redirected, and new networks constructed” (Griffin 2015, pg. 93). It can be possible to automate some components of risk assessment (Graf et al. 2017), but most of the steps require human effort. This intensive effort is vividly illustrated by the many data rescue initiatives that have taken place within government agencies and other kinds of organizations over the past few decades.

Data rescue initiatives within government agencies & specific disciplines

Legacy data are data collected in the past with different technologies and data formats than in use today. These data often face the largest numbers of risk factors that could lead to data loss. A wide range of government agencies and other organizations have conducted legacy data rescue initiatives to modernize data and make them more accessible and usable for today's science. Each data rescue project typically faces many different kinds of data risks. For example, a recent satellite data rescue effort had to address the "loss of datasets, reconciliation of actual media contents with metadata available, deviation of the actual data format from expectations or documentation, and retiring expertise" (Poli et al. 2017, pg. 1481). Data rescue projects typically involve work to prevent future risk factors from manifesting, in addition to modernizing data for accessibility and usability. For example, data rescue projects migrate data to less endangered data formats, and create new metadata and quality control documentation (Levitus 2012).

CODATA/RDA working groups & meetings

Relevant professional organizations, including the International Council for Science (ICSU) Committee on Data for Science and Technology (CODATA) and the Research Data Alliance (RDA), also have been actively identifying improvements for data stewardship practices that can reduce potential risks to data. For example, the former Data At Risk Task Group (DAR-TG), of CODATA, raised awareness about the value of heritage data and described the benefits obtained from several data rescue projects (Griffin 2015). This group also organized the 2016 "Rescue of Data At Risk" workshop mentioned in the introduction of this paper. That workshop led to a document titled, "Guidelines to the Rescue of Data At Risk" (2017). Subsequently, the Data Rescue Interest Group (2018) of the Research Data Alliance (RDA), spawned from the CODATA DAR-TG, also focuses on efforts to increase awareness of data rescue projects.

Repository certifications and maturity assessment

Many data repositories have conducted self-assessments and external assessments to document their compliance with the standards for trusted repositories and attain certification of their capabilities and practices for managing data. In addition to emphasizing organizational issues, repository certification instruments, such as ISO 16363 (2012b) and CoreTrustSeal (2018) certification, also focus on digital object management and infrastructure capabilities. Engaging in such assessments offers benefits to repositories and their stakeholders. A key benefit is the identification of areas where improvements have been completed or need to be completed to reduce risks to data (CoreTrustSeal 2018). In an examination of perceptions of repository certification, Donaldson et al. (2017) found that process improvement was often reported by repository staff as a benefit of repository certification.

In addition to (or complementary to) formal certifications, data repositories may conduct data stewardship maturity assessment exercises to help in identifying data risks and informing data risk mitigation strategies (Faundeen 2017). "Maturity" is used in the sense presented by Peng et al. (2015), and refers to the level of performance attained to ensure preservability, accessibility, usability, transparency/traceability, and sustainability of data, along with the level of performance in data quality assurance, data quality control/monitoring, data quality assessment, and data integrity checks. Maturity at the institutional (or archive) level in areas such as policy, funding, and infrastructure does not necessarily translate to comprehensive maturity at the dataset level (Peng 2018). Data stewardship maturity assessment should therefore be performed both at the institutional level and at the dataset level. It is recognized that performing stewardship maturity assessments can be time consuming and resource intensive. However, the stewardship organizations are encouraged to perform self-assessment using "stage by stage" or "a la carte" approach (see example in Peng et al. 2019). Ultimately, both formal certifications and informal maturity assessments help organizations not only gain self-awareness, but also identify better solutions for their data that might be at risk of being lost or rendered unusable.

Developing a Data Risk Assessment Matrix

Risk assessment is a well-established field, with 30–40 years of history (National Research Council 1983; Aven 2016). However, the practice of applying risk assessment methodologies to scientific data collections is less formally established, though regular audits and reviews of data management systems are common in some organizations (Ramapriyan 2017).

The starting point for this project was to establish a process for categorizing the data risk factors shown in **Table 1**. The initial idea of our effort was that if data risk factors could be categorized into a logical structure, it would allow data managers to assess the risks to their data collections via a set of predefined and consistent categories. To develop a logical categorization, we held a session to conduct a "card sorting" exercise at the 2018 ESIP Summer Meeting, which took place in July 2018 in Tucson, Arizona. "Card sorting" is an established

method for developing categorizations of concepts, vocabulary terms, or web sites (Zimmerman & Akerelrea 2002; Usability.gov 2019). Following the card sorting methodology, participants in the 2018 ESIP meeting session were provided the list of data risks in **Table 1**, and asked to complete the following task: “Looking at the list of data risk factors, how would you group these factors, based on the categories you would define?”

Approximately 15 attendees engaged in the exercise. We used a combination of an online card sorting tool and hand-written recommendations to collect the completed card sorting categorizations. Following the completion of the exercise, the results were displayed in front of the session participants and a group discussion took place. The outcome of the card sorting exercise and subsequent discussion was a clear recognition that there could be many valid and useful ways of categorizing data risks. No single method for categorizing the risk factors would be sufficient to cover the diverse organizations and situations within which data collections exist. Depending on the situation(s), a data curation organization or individual is facing, they may

Table 1: Risk factors for scientific data collections.

	Risk Factor	Description
1.	Lack of use	Data are rarely accessed and dubbed ‘unwanted’, thus getting thrown away
2.	Loss of funding for archive	The whole archive loses its funding source
3.	Loss of funding for specific datasets	Lack of funding to monitor, maintain, and otherwise work with specific data
4.	Loss of knowledge around context or access	The loss of individuals who know how to access the data or know the metadata associated with these data that make the data useable to others, e.g. due to retirement or death
5.	Lack of documentation & metadata	Data cannot be interpreted due to lack of contextual knowledge
6.	Data mislabeling	Data are lost because they are poorly identified (either physically or digitally)
7.	Catastrophes	Fires, floods, wars/human conflicts, etc
8.	Poor data governance	Uncertain or unknown decision making processes impede effective data management
9.	Legal status for ownership and use	Uncertain, unknown, or restrictive legal status limits the possible uses of data
10.	Media deterioration	Physical media deterioration prevents data from being accessed (paper, tape, or digital media)
11.	Missing files	Data files are lost without any known reason
12.	Dependence on service provider	Risks due to potential single point of failure problems if a particular service provider goes out of business
13.	Accidental deletion	Data are accidentally deleted by a staff error
14.	Lack of planning	Lack of planning puts data collections at risk of being susceptible to unexpected events
15.	Cybersecurity breach	Data are intentionally deleted or corrupted via a security breach, e.g. malware
16.	Over-abundance	Difficulty dealing with too much data results in reduction in value or quality of whole collections
17.	Political interference	Data deleted or made inaccessible due to political decisions
18.	Lack of provenance information	Data cannot be trusted or understood because of a lack of information about data processing steps, or about data stewardship chains of trust
19.	File format obsolescence	Data cannot be accessed due to lack of knowledge, equipment, or software for reading a specific file format
20.	Storage hardware breakdown	Sudden & catastrophic malfunction of storage hardware
21.	Bit rot and data corruption	Gradual corruption of digital data due to an accumulation of non-critical failures (bits flipping) in a data storage device

need to categorize data risks in different ways. This characteristic is common in risk assessments generally, as risk prioritization and categorizations are dependent on the phenomena being assessed, the characteristics of the situation, and the goals of the organizations or people performing the assessment (Slovic 1999).

Through subsequent discussion and analysis of the data risk assessment literature noted above, we identified at least ten different ways that data risk factors could be assessed. Many of these categorization methods are applicable to risk assessments of any kind (Cervone 2006). The list below is not meant to be exhaustive, and some methods are likely related. Data risk factors could be categorized or prioritized according to the methods listed in **Table 2**.

The lists shown in **Tables 1** and **2** offer characteristics on which data risk assessments can be built. Combining the categorization methods from **Table 2** with the selected risk factors from **Table 1** leads to a risk assessment matrix, as shown in **Table 3**. This figure shows an example of a selection of specific data risk factors and the categorization methods. Depending on the situation or data collection being assessed, different risk factors and/or categorization methods may be more applicable than the ones shown in **Table 3**. Those conducting a data risk assessment can then use the matrix as a way to organize, prioritize, or potentially quantify the selected risks according to the categorization methods that are most relevant for the specific case at hand. The next section provides more detailed illustrations of the use of the data risk assessment matrix. Appendix I shows the full data risk assessment template, with all risks and categorization methods from **Tables 1** and **2**.

Table 2: Methods for Categorizing Data Risks.

Categorization Method	Description
Severity of risk	How much impact could this risk factor have on the data itself, regardless of the current importance of data to the user?
Likelihood of occurrence	How likely a risk factor is to occur
Length of recovery time	How long it would take to recover data or re-establish data accessibility
Impact on user	How significantly data users are impacted by data loss or loss of data accessibility
Who is responsible for addressing the problem	Who has the expertise and responsibility to mitigate or respond to particular risk factors
Cause of problem	What caused a data risk factor to occur
Degree of control	How much control an organization or individual has over whether a risk factor is present or will occur
Proactive vs reactive response	Whether risk factors can be mitigated via preventative measures, or whether they must be responded to upon occurrence
Nature of mitigation	What steps must be taken or processes put in place to prevent a risk, or mitigate a risk after it has occurred
Resources required for mitigation	What time, money, or personnel resources will be necessary to mitigate risk factors

Table 3: Example of a blank data risk assessment matrix, after selection of specific risk factors and categorization methods of interest.

Risk Factors	Categorization Methods			
	Severity of risk	Likelihood of occurrence	Cause of problem	Resources req'd for mitigation
Lack of use				
Loss of knowledge				
Lack of docs & metadata				
Catastrophes				
Poor data governance				
Media deterioration				

Application of the Data Risk Assessment Matrix

Three case studies are described below in which the data risk assessment matrix was used to develop a better understanding of data risks for particular resources. These cases enable evaluation of the data risk assessment framework presented in this paper, clarifying its strengths and weaknesses, and pinpointing the situations in which it can be most useful (Becker, Maemura & Moles 2020).

Case 1 – NCAR Library Analog Data Collection

The National Center for Atmospheric Research (NCAR) Library maintains an analog data collection that consists of about 300 data sets in support of atmospheric and meteorological research conducted by NCAR scientists. These assets are largely compilations of measurements and statistics published by national and international meteorological services and other kinds of government entities. Many of these assets have been in the NCAR Library's collections for decades, and most were minimally cataloged when they were first brought into the collection. As such, the current usage of the collection is minimal. A prior assessment done by the NCAR Library and a student assistant sought to identify individual assets that were of higher potential value and interest for current science. This assessment effort resulted in a modernization prioritization based on a geographic and temporal framework, and improved metadata records for about 5% of the collection (Mayernik et al. 2018). This effort did not, however, include any kind of risk assessment related to the physical assets themselves.

The data risk assessment matrix was therefore helpful in doing a second-level priority analysis for these NCAR Library analog data assets. We used the matrix as a way to identify which risk factors were most important for these materials, and to characterize the mitigation efforts that were needed for each risk factor. In particular, we focused the risk assessment on the data assets that were previously identified as having high geospatial and temporal interest. The NCAR Library use of the matrix involved a series of steps:

- Step 1 – A number of risk factors listed in the matrix were identified as being of most importance, with the focus being on factors that prevented or impeded the use of these data within current scientific studies. The most immediate risk factors were identified to be the “lack of use” and the “lack of documentation/metadata” for these assets. Other risks that were secondary in immediacy, but still potentially important, were: Data mislabeling, the questionable legal status for ownership and use, media deterioration, lack of planning, and poor data governance.
- Step 2 – The second step was to identify which categorization methods shown in the matrix were most applicable/appropriate for the NCAR Library's management and maintenance of this collection. The methods selected were: a) Length of recovery time, b) Who is responsible for addressing the problem, c) Nature of mitigation, and d) Resources required for mitigation.
- Step 3 – The third step was to fill in the boxes in the matrix for the risk factors and categorization methods. For example, for the “Length of recovery time” question, we used a simple 1–3 scale to indicate relative differences in how long it would take to mitigate the two most important risk factors: “lack of use” and the “lack of documentation/metadata”. As one example, some data assets were published by international agencies and therefore have title pages and documentation that are not in English. In turn, due to the lack of relevant foreign language expertise in the NCAR Library staff, developing new metadata for these resources will take more effort than for those assets that were published by English-speaking countries. For the “Resources required for mitigation” categorization method, a numerical scale was not as appropriate. Instead, we filled in the matrix with text descriptions of the resources required to mitigate the risk factors. An example entry under the “lack of documentation & metadata” risk factor was: “We would need to create new metadata for the library catalog, then transform to ISO for inclusion in NCAR DASH Search, with added challenge of needing to look at microfilm files (no current working reader in Library).”

In summary, the matrix was very useful as “something to think with.” In other words, it jump-started the process for doing the risk assessment because the NCAR Library staff did not need to spend time developing a comprehensive list of risk factors that may apply for these data, or brainstorm about how to categorize those risks. The risk factor matrix provided a ready-made starting point for the assessment. Because the matrix does not dictate how the cells should be filled in, the NCAR Library staff made decisions about how to apply the matrix for each categorization method that was chosen. The matrix structure could potentially be applied or customized to create a prioritization rubric, by supporting the creation of a numeric scoring process for categories where that is appropriate.

Case 2 – Mohonk Preserve Daniel Smiley Research Library

Mohonk Preserve is a land trust and nature preserve in New Paltz, New York covering more than 8,000 acres of a northern section of the Appalachian Mountains known as the Shawangunk Mountains. Mohonk Preserve's conservation science division, the Daniel Smiley Research Center (DSCR), is affiliated with the Organization of Biological Field Stations (OBFS) and acts as a NOAA Climate Observation Center. DSCR staff and citizen scientists carry out a variety of long-term monitoring projects and manage an extensive archive of historical observations. The archive houses 60,000 physical items, 9,000 photographs, 86 years of natural history observations, 123 years of daily weather data, and a research library of legacy titles. The physical items include more than 3,000 herbarium specimens, 107 bird specimens, 140 butterfly specimens, 139 mammal specimens, 400 arthropod specimens, and over 14,000 index cards with handwritten and typed observations. The digitization process of the archive holdings is ongoing, but the packaging and publishing of datasets in the Environmental Data Initiative is a priority (Mohonk Preserve et al. 2018a, 2018b, 2019). These data and natural history collections underpin the Mohonk Preserve's land management and stewardship and have been crucial to an increasing number of scientific publications (e.g., Cook et al. 2009; Cook et al. 2008; Charifson et al. 2015; Richardson et al. 2016), but the collections remain underutilized.

The data rescue effort for the archives has largely consisted of digitization and cataloging. Hence, the data risk assessment matrix was used to guide the prioritization of datasets for publication and assess other data rescue needs and considerations for the archives. The most critical risk factors identified through the process were 'lack of documentation & metadata', 'loss of knowledge', and 'lack of use.' In order to address the lack of use, we collaboratively developed a prioritization of the data holdings for publication in a repository, particularly based around the value of data collected for scientific investigations, the temporal coverage of the dataset, and an assessment of the resources required for the digitization, packaging, and publishing of the relevant dataset.

We also realized through the data risk matrix process that many of our risk factors are interdependent – for example, the lack of documentation may not be because the documentation does not exist in the library, but rather that it may not be discoverable in the archives due to incomplete digitization or cataloging of the relevant records or field notes. For example, during the assessment process for our vernal pool monitoring dataset (Mohonk Preserve et al. 2019), we discovered previously unknown environmental quality notes in narrative sections of an undigitized collection of field notes. This supported the current emphasis on the digitization and cataloging of the holdings and suggested areas of high importance, particularly the narrative sections of field notebooks. Additionally, the lack of documentation and metadata is directly related to the loss of knowledge through leadership transitions. Like many long-term ongoing collections projects, metadata and documentation – particularly related to data collection protocols – are held as tacit knowledge by key stakeholders who have been involved with the project for an extended period of time. The loss of those stakeholders or their knowledge, through retirement or employment changes, poses a significant risk to the long-term value of the associated data (Michener et al. 1997).

Because the holdings largely consist of physical items, a subset of the risk factors in the matrix were not directly applicable to the collections but had corollaries in physical collections management. For example, bit rot and data corruption are not a concern for the physical items, but pests present a similar concern that needs mitigation in a physical archive setting. Additionally, storage hardware breakdown is not directly applicable to herbarium collections but ensuring that the mounting sheets are acid-free is key to ensuring the protection of the specimens and preventing deterioration over time. Considering physical risks to the collection media remains a crucial aspect of managing and planning for the future of physical specimen holdings. Through the data assessment process, one of the key risk areas identified through the assessment was the loss of knowledge and documentation due to retirement, so planning for mitigating this risk is ongoing. Overall, the matrix provided a helpful starting point for guiding conversations relating to the stewardship of the archives and proactively planning and allocating resources to make the data more accessible to scientists and researchers.

Case 3 – EDGI Response to the Deer Park Chemical Fire

On March 17 2019, tanks of chemicals at the Intercontinental Terminals Company (ITC) in Deer Park, Texas, caught fire and began a blaze that would last several days, emitting a chemical-laden plume of smoke over surrounding communities. The Environmental Data & Governance Initiative (EDGI) was approached for assistance in rapid-response archival backup of digital environmental data relevant to the fire in case of future tampering or loss of availability.

There were two major causes for concern: (1) evident tampering– the closest air quality monitor was taken down during the fire, and (2) potential conflict of interest– the entity furnishing the data might have some culpability in a future legal case using the data. The approaching organization hopes to use saved data as evidence in legal cases that may take several years (potentially due to the long timespan for benzene-related illnesses to surface in then-students at the local school and workers in nearby factories).

On a limited timeline and with little capacity, EDGI needed to downselect from hundreds to thousands of possibly relevant data sources (including air and water quality monitors affected by the fire's plume, and plans and response documents surrounding handling of the fire). The primary mission: ensure the data of primary concern is backed up in a legally legitimate (traceable) format that will be usable in a decade or more.

The data risk matrix from this paper was not used at the time **Prioritization**

The approaching organization suggested a few directories of static data to archive. With additional investigation, EDGI also found some API-accessible structured data from the air monitors that was updated daily.

The information proposed for potential rescue included:

- Data from the Deer Park air quality monitor data that was taken down
- Data from other nearby air quality monitors
- Air quality monitors downstream of the plume (potentially very many of them, as the plume traveled more than 20 miles)
- Three years of back-data from any air quality monitors, to establish baseline
- Water quality monitors– local, downstream, and down-plume, in case relevant (no evidence of contamination yet, but the situation still developing), and three years of back data to establish baseline
- Future data from any monitors, to track the still-developing situation and archive it in case of any present risk
- Contextual information: air sampling plans, disaster response plans, air and water quality sampling maps, PDFs of additional air and water quality sampling from different entities than provide the API-callable data

There was no formal review process for deciding what to save. There was some brief discussion internally around technical feasibility and potential environmental justice-focused mapping efforts, but the major use anticipated for the saved data was the legal case. The whole process from request to data backup took just a few days. Ultimately, EDGI's choices of data to save depended primarily on the abilities and assessment of the two volunteers available. The volunteers used the skills they had and their best intuitions– lacking a clear prioritization between different data that could be saved.

Applying the data risk matrix to this situation, the two major risk factors can be immediately identified as “catastrophe” and “political interference”. Both risk factors are relevant, likely, and potentially catastrophic in effect. This highlights the urgency and source of the risk.

The risk matrix is not as helpful in prioritizing which data to save under capacity and urgency constraints. The risk matrix identifies the type and intensity of risk, but since all of the data is equally high-risk in this use case, the context of the data and its use case (evidence in a far-future legal case) are necessary for the following tasks of identifying, locating, and prioritizing data to save. This was done based on the best assessment and abilities of the available volunteers.

Ultimately, EDGI saved:

- Structured data from the Deer Park and nearby Lynchburg Ferry air quality monitors: saved with metadata to IPFS via Qri (qri.io) with script to keep pulling updates
- All of the PDF data (primarily directories of 20–100 links, typically to PDFs, including maps, images, narratives, and tables of data): saved to the Internet Archive as a full site snapshot

Assessing risks to rescued data

Following the data rescue operation, this risk matrix was used to assess ongoing risks to the repositories of rescued data: (1) the PDF data saved to the Internet Archive and (2) the structured data from air quality monitors saved to IPFS. The risk matrix was very effective for identification of vulnerabilities and potential next steps to better secure the data.

The full matrix (all of the categories and all of the risk factors) was applied twice: to PDF data saved to the Internet Archive, and to structured data saved on the decentralized web (IPFS). A scale of numeric values (1 (low) to 3 (high)) was used to rate the category versus the risk factor. For example, the risk factor of Media Deterioration was rated 3 (high) for Severity of Risk, but 1 (low) for Likelihood of Occurrence. This numeric rating was important to use of the full matrix— instead of removing columns as irrelevant, they could be down-rated where the risk was low.

Use of the matrix immediately highlighted the difference in risks important to the data stored on IPFS versus the Internet Archive. For example, data on the Internet Archive is well-governed and reasonably easy to find, but much more susceptible to natural disaster and hardware deterioration than the data on IPFS. IPFS is a new technology designed to store data across many physical locations— so it's very resilient to location-based risks, but its format may become obsolete as the technology develops.

The risk matrix is particularly useful when combined with spreadsheet tools. For example, a quick to-do list for EDGI as a data manager can be produced using a formula such as:

- Likelihood of occurrence > 1
- Resources for mitigation < 3
- Type of action: proactive
- Responsible party: EDGI
- Print mitigation action for rows where all of the above are true

Overall, the Risk Matrix outlined in this paper is a very useful tool for identifying risks to data and prioritizing next steps for mitigation— as long as the user has or can assume control over the data. However, in a data rescue use case, this risk matrix must be supplemented by additional context in order to prioritize which at-risk data should be saved when capacity is limited.

Conclusions and Lessons Learned

Risk assessments are instrumental for ensuring that existing data collections continue to be useful for scientific research and societal applications. Risk assessments are also an essential component of data rescue efforts in which interventions take place to prevent or minimize data loss. The data risk assessment framework presented in this paper provides a platform from which risk assessments can quickly begin.

To close out this paper, we discuss some observations and lessons learned in developing and applying the data risk assessment matrix. Data risk assessments can get significantly more in-depth and detailed than the basic template presented in **Table 3** and Appendix I. As one example, the US Geological Survey (USGS) has undertaken a substantive project to create risk calculations for USGS-held data collections based on a number of criteria (USGS 2019). The USGS process has involved the development of detailed formulas and weighting schemes to produce quantified assessments of data risk. The risk assessment matrix presented in this paper does not provide “out of the box” quantification measures or data risk prioritizations of the level of detail of the USGS project. The data risk matrix does, however, provide the foundations for an individual or organization to develop a more customized risk assessment rubric. The specifics of how risks were quantified or qualified, and how they were prioritized varied across the different uses of the matrix presented in the three case studies.

The three cases did demonstrate a common use pattern for the data risk matrix. The first step in each case was to review **Tables 1** and **2** to determine which risk factors and categorization methods were most relevant. Clearly not all of the risk factors are applicable to all cases, and some of the risk factors are closely related, such as the “lack of documentation & metadata” and “lack of provenance information.” Once the risk factors and categorization methods have been filtered down into a smaller matrix, the next step is to determine how to fill in the matrix cells for particular datasets or collections. It may not be obvious how this would work for some data collections. Our cases involved using a mix of quantitative, qualitative, and ordinal rankings (such as using “high, medium, and low” designations for particular cells). This step may take some trial and error by the matrix user(s) to determine ranking approaches that are the most useful.

The third step is then to use the cell values in the matrix to guide conversations and decisions about risk mitigation priorities. In this sense, the matrix exercise can provide a high level overview of data collections, risks they may face, and the relative urgency and challenges that those risks present to the data stewards. The matrix can serve as a common reference point for discussions of resource allocations and stewardship priorities. However, as exemplified in the EDGI use case, prioritization in real-time, as would be required during catastrophic events such as disasters or wars particularly where there may be political interference,

is difficult if not impossible. As such, preventing or minimizing data loss requires pre-planning at a scale rarely available.

The goal in creating this data risk assessment matrix has been to provide a light-weight way for data collections to be reviewed, documented, and evaluated against a set of known data risk factors. As the understanding of the value that scientific data have for research and societal uses increases, many initiatives recognize that “old data is the new data” (NIWA 2019). Risk assessments are critical to ensure that “old data” can become “new data,” and are also critical to ensure that new data can continue to be newly useful into the future.

Appendix I – Data Risk Assessment Template

RISK FACTORS	<i>Categorization Methods</i>									
	<i>Severity of risk</i>	<i>Likelihood of occurrence</i>	<i>Length of recovery</i>	<i>Impact on user</i>	<i>Who is responsible</i>	<i>Cause of problem</i>	<i>Degree of control</i>	<i>Proactive vs reactive response</i>	<i>Nature of mitigation</i>	<i>Resources req'd for mitigation</i>
Lack of use										
Loss of funding for archive										
Loss of funding for specific datasets										
Loss of knowledge										
Lack of docs & metadata										
Data mislabeling										
Catastrophes										
Poor data governance										
Legal status for ownership and use										
Media deterioration										
Missing files										
Dependence on service provider										
Accidental deletion										
Lack of planning										
Cybersecurity breach										
Over-abundance										

(Contd.)

RISK FACTORS	Categorization Methods									
	Severity of risk	Likelihood of occurrence	Length of recovery	Impact on user	Who is responsible	Cause of problem	Degree of control	Proactive vs reactive response	Nature of mitigation	Resources req'd for mitigation
Political interference										
Lack of provenance information										
File format obsolescence										
Storage hardware breakdown										
Bit rot and data corruption										

Acknowledgements

This project was organized and supported by the Data Stewardship Committee within the Earth Science Information Partners (ESIP). We thank ESIP and the committee participants for feedback on the project at numerous points in the past few years.

The work of Robert R. Downs was supported by the National Aeronautics and Space Administration under Contract 80GSFC18C0111 for the Socioeconomic Data and Applications Distributed Active Archive Center (DAAC).

Alexis Garretson acknowledges the support of the Environmental Data Initiative Summer Fellowship program and the Earth Science Information Partners Community Fellows Program. Alexis also acknowledges Mohonk Preserve staff, particularly the staff of the Daniel Smiley Research Center: Elizabeth C. Long, Megan Napoli, and Natalie Feldsine. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1842191. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The work of Chung-Yi (Sophie) Hou was supported by the National Center for Atmospheric Research.

The National Center for Atmospheric Research is sponsored by the U.S. National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of NCAR or the NSF.

Competing Interests

The authors have no competing interests to declare.

References

- Allen, L, Stewart, C and Wright, S.** 2017. Strategic open data preservation. *College & Research Libraries News*, 78(9). <https://crln.acrl.org/index.php/crlnews/article/view/16771/18312>. DOI: <https://doi.org/10.5860/crln.78.9.482>
- Anderson, WL, Faundeen, JL, Greenberg, J and Taylor, F.** 2011. Metadata for data rescue and data at risk. In: *Conference on Ensuring Long-Term Preservation in Adding Value to Scientific and Technical Data*. <http://hdl.handle.net/2152/20056>
- Aven, T.** 2016. Risk assessment and risk management: Review of recent advances on their foundation. *European Journal of Operational Research*, 253(1): 1–13. DOI: <https://doi.org/10.1016/j.ejor.2015.12.023>
- Becker, C, Maemura, E and Moles, N.** 2020. The design and use of assessment frameworks in digital curation. *Journal of the Association for Information Science and Technology (JASIST)*, 71(1): 55–68. DOI: <https://doi.org/10.1002/asi.24209>
- Cervone, HF.** 2006. Project risk management. *OCLC Systems & Services: International Digital Library Perspectives*, 22(4): 256–62. DOI: <https://doi.org/10.1108/10650750610706970>

- Charifson, DM, Huth, PC, Thompson, JE, Angyal, RK, Flaherty, MJ and Richardson, DC.** 2015. History of Fish Presence and Absence Following Lake Acidification and Recovery in Lake Minnewaska, Shawangunk Ridge, NY. *Northeastern Naturalist*, 22: 762–781. DOI: <https://doi.org/10.1656/045.022.0411>
- Chodacki, J.** 2018. Data Mirror: Complementing data producers. *Against the Grain*, 29: 35. <https://escholarship.org/uc/item/2n1715ff>. DOI: <https://doi.org/10.7771/2380-176X.7877>
- CoreTrustSeal Data Repository Certification.** 2018. <https://www.coretrustseal.org/>.
- Cornelius, KB and Pasquetto, IV.** 2018. “What data?” Records and data policy coordination during presidential transitions. In: *Transforming Digital Worlds*, 155–163. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-78105-1_20
- Cook, BI, Cook, ER, Anchukaitis, KJ, Huth, PC, Thompson, JE and Smiley, SF.** 2009. A Homogeneous Record (1896–2006) of Daily Weather and Climate at Mohonk Lake, New York. *Journal of Applied Meteorology and Climatology*, 49: 544–555. DOI: <https://doi.org/10.1175/2009JAMC2221.1>
- Cook, BI, Cook, ER, Huth, PC, Thompson, JE, Forster, A and Smiley, D.** 2008. A cross-taxa phenological dataset from Mohonk Lake, NY and its relationship to climate. *International Journal of Climatology*, 28: 1369–1383. DOI: <https://doi.org/10.1002/joc.1629>
- Data Rescue Interest Group.** 2018. Research Data Alliance. <https://rd-alliance.org/groups/data-rescue.html>.
- Dennis, B.** 2016. Scientists are frantically copying U.S. climate data, fearing it might vanish under Trump. *The Washington Post*, Dec. 13, 2016. <https://www.washingtonpost.com/news/energy-environment/wp/2016/12/13/scientists-are-frantically-copying-u-s-climate-data-fearing-it-might-vanish-under-trump/>.
- Donaldson, DR, Dillo, I, Downs, R and Ramdeen, S.** 2017. The perceived value of acquiring Data Seals of Approval. *International Journal of Digital Curation*, 12(1). DOI: <https://doi.org/10.2218/ijdc.v12i1.481>
- Downs, RR and Chen, RS.** 2017. Curation of scientific data at risk of loss: Data rescue and dissemination. In: Johnston, L (ed.), *Curating Research Data. Volume One, Practical Strategies for Your Digital Repository*. Association of College and Research Libraries. DOI: <https://doi.org/10.7916/D8W09BMQ>
- Faundeen, J.** 2017. Developing criteria to establish trusted digital repositories. *Data Science Journal*, 16: 22. DOI: <https://doi.org/10.5334/dsj-2017-022>
- Gallaher, D, Campbell, GG, Meier, W, Moses, J and Wingo, D.** 2015. The process of bringing dark data to light: The rescue of the early Nimbus satellite data. *GeoResJ*, 6: 124–134. DOI: <https://doi.org/10.1016/j.grj.2015.02.013>
- Graf, R, Ryan, HM, Houzanme, T and Gordea, S.** 2017. A decision support system to facilitate file format selection for digital preservation. *Libellarium: Journal for the Research of Writing, Books, and Cultural Heritage Institutions*, 9(2). DOI: <https://doi.org/10.15291/libellarium.v9i2.274>
- Griffin, RE.** 2015. When are old data new data? *GeoResJ*, 6: 92–97. DOI: <https://doi.org/10.1016/j.grj.2015.02.004>
- Guidelines to the Rescue of Data At Risk.** 2017. Research Data Alliance. <https://www.rd-alliance.org/guidelines-rescue-data-risk>.
- Hsu, L, Lehnert, KA, Goodwillie, A, Delano, JW, Gill, JB, Tivey, MA, Ferrini, VL, Carbotte, SM and Arko, RA.** 2015. Rescue of long-tail data from the ocean bottom to the Moon: IEDA Data Rescue Mini-Awards. *GeoResJ*, 6: 108–114. DOI: <https://doi.org/10.1016/j.grj.2015.02.012>
- [ISO] International Organization for Standardization.** 2012a. ISO 14721:2012 (CCSDS 650.0-M-2): Space data and information transfer systems – Open Archival Information System (OAIS) – Reference model. <https://www.iso.org/standard/57284.html>.
- [ISO] International Organization for Standardization.** 2012b. ISO 16363:2012 (CCSDS 652.0-R-1): Space data and information transfer systems – Audit and certification of trustworthy digital repositories. <https://www.iso.org/standard/56510.html>.
- Janz, M.** 2018. Maintaining access to public data: Lessons from Data Refuge. *Against the Grain*, 29: 30–33. DOI: <https://doi.org/10.31229/osf.io/yavzh>
- Knapp, KR, Bates, JJ and Barkstrom, B.** 2007. Scientific data stewardship: Lessons learned from a satellite–data rescue effort. *Bulletin of the American Meteorological Society*, 88(9): 1359–1362. DOI: <https://doi.org/10.1175/BAMS-88-9-1359>
- Lamdan, S.** 2018. Lessons from Datarescue: The limitations of grassroots climate change data preservation and the need for federal records law reform. *University of Pennsylvania Law Review Online*, 166(1): Article 12. https://scholarship.law.upenn.edu/penn_law_review_online/vol166/iss1/12.

- Levitus, S.** 2012. The UNESCO-IOC-IODE “Global Oceanographic Data Archeology and Rescue” (GODAR) Project and “World Ocean Database” Project. *Data Science Journal*, 11: 46–71. DOI: <https://doi.org/10.2481/dsj.012-014>
- Maemura, E, Moles, N and Becker, C.** 2017. Organizational assessment frameworks for digital preservation: A literature review and mapping. *Journal of the Association for Information Science and Technology*, 68(7): 1619–1637. DOI: <https://doi.org/10.1002/asi.23807>
- Mayernik, MS, Downs, RR, Duerr, R, Hou, C-Y, Meyers, N, Ritchey, N, Thomer, A and Yarmey, L.** 2017. Stronger together: The case for cross-sector collaboration in identifying and preserving at-risk data. Figshare. DOI: <https://doi.org/10.6084/m9.figshare.4816474.v1>
- Mayernik, MS, Huddle, J, Hou, C-Y and Phillips, J.** 2018. Modernizing library metadata for historical weather and climate data collections. *Journal of Library Metadata*, 17(3/4): 219–239. DOI: <https://doi.org/10.1080/19386389.2018.1440927>
- McGovern, NY.** 2017. Data rescue. *ACM SIGCAS Computers and Society*, 47(2): 19–26. DOI: <https://doi.org/10.1145/3112644.3112648>
- Michener, WK, et al.** 1997. Nongeospatial metadata for the ecological sciences. *Ecological Applications*, 7(1): 330–342. DOI: [https://doi.org/10.1890/1051-0761\(1997\)007\[0330:NMFTES\]2.0.CO;2](https://doi.org/10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2)
- Mohonk Preserve, Belardo, C, Feldsine, N, Forester, A, Huth, P, Long, E, Morgan, V, Napoli, M, Pierce, E, Richardson, D, Smiley, D, Smiley, S and Thompson, J.** 2018a. History of Acid Precipitation on the Shawangunk Ridge: Mohonk Preserve Precipitation Depths and pH, 1976 to Present. *Environmental Data Initiative*. DOI: <https://doi.org/10.6073/pasta/734ea90749e78613452eacec489f419c>
- Mohonk Preserve, Forester, A, Huth, P, Long, E, Morgan, V, Napoli, M, Pierce, E, Smiley, D, Smiley, S and Thompson, J.** 2018b. Mohonk Preserve Ground Water Springs Data, 1991 to Present. *Environmental Data Initiative*. DOI: <https://doi.org/10.6073/pasta/928feed7ee748509ab065de7e3791966>
- Mohonk Preserve, Feldsine, N, Forester, A, Garretson, A, Huth, P, Long, E, Napoli, M, Pierce, E, Smiley, D, Smiley, S and Thompson, J.** 2019. Mohonk Preserve Amphibian and Water Quality Monitoring Dataset at 11 Vernal Pools from 1931–Present. *Environmental Data Initiative*. DOI: <https://doi.org/10.6073/pasta/864aea25998b73c5d1a5b5f36cb6583e>
- National Research Council.** 1983. Risk Assessment in the Federal Government: Managing the Process. Washington, DC: The National Academies Press. DOI: <https://doi.org/10.17226/366>
- NIWA.** 2019. The week it snowed everywhere. *NIWA Media Release*, Nov. 21, 2019. <https://niwa.co.nz/news/the-week-it-snowed-everywhere>.
- Peng, G.** 2018. The state of assessing data stewardship maturity – An overview. *Data Science Journal*, 17: 7. DOI: <https://doi.org/10.5334/dsj-2018-007>
- Peng, G, Milan, A, Ritchey, NA, Partee, RP, II, Zinn, S, McQuinn, E, Casey, KS, Lemieux, P, III, Ionin, R, Jones, P, Jakositz, A and Collins, D.** 2019. Practical application of a data stewardship maturity matrix for the NOAA OneStop project. *Data Science Journal*, 18: 41. DOI: <https://doi.org/10.5334/dsj-2019-041>
- Peng, G, Privette, JL, Kearns, EJ, Ritchey, NA and Ansari, S.** 2015. A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, 13: 231–253. DOI: <https://doi.org/10.2481/dsj.14-049>
- Pienta, AM and Lyle, J.** 2018. Retirement in the 1950s: Rebuilding a longitudinal research database. *IASSIST Quarterly*, 42(1). DOI: <https://doi.org/10.29173/iq19>
- Poli, P, Dee, DP, Saunders, R, John, VO, Rayer, P, Schulz, J, Bojinski, S, et al.** 2017. Recent advances in satellite data rescue. *Bulletin of the American Meteorological Society*, 98(7): 1471–1484. DOI: <https://doi.org/10.1175/BAMS-D-15-00194.1>
- Ramapriyan, HK.** 2017. NASA’s EOSDIS: Trust and Certification. Presented at: *2017 ESIP Summer Meeting*, Bloomington, IN. Figshare. DOI: <https://doi.org/10.6084/m9.figshare.5258047.v1>
- Richardson, DC, Charifson, DM, Stanson, VJ, Stern, EM, Thompson, JE and Townley, LA.** 2016. Reconstructing a trophic cascade following unintentional introduction of golden shiner to Lake Minnewaska, New York, USA. *Inland Waters*, 6: 29–33. DOI: <https://doi.org/10.5268/IW-6.1.915>
- Ryan, H.** 2014. Occam’s razor and file format endangerment factors. In: *Proceedings of the 11th International Conference on Digital Preservation (iPres)*, October 6–10, 2014, Melbourne, Australia, 179–188. https://www.nla.gov.au/sites/default/files/ipres2014-proceedings-version_1.pdf.
- Slovic, P.** 1999. Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield. *Risk Analysis*, 19(4): 689–701. DOI: <https://doi.org/10.1023/A:1007041821623>
- Thompson, CA, Robertson, WD and Greenberg, J.** 2014. Where have all the scientific data gone? LIS perspective on the data-at-risk predicament. *College & Research Libraries*, 75(6): 842–861. DOI: <https://doi.org/10.5860/crl.75.6.842>

- Usability.gov.** 2019. Card Sorting. US Department of Health & Human Services. <https://www.usability.gov/how-to-and-tools/methods/card-sorting.html>.
- USGS.** 2019. USGS Data at Risk: Expanding Legacy Data Inventory and Preservation Strategies. US Geological Survey. <https://www.sciencebase.gov/catalog/item/58b5ddc3e4b01ccd54fde3fa>.
- Varinsky, D.** 2017. Scientists across the US are scrambling to save government research in 'Data Rescue' events. *Business Insider*, Feb. 11, 2017. <http://www.businessinsider.com/data-rescue-government-data-preservation-efforts-2017-2>.
- Yakel, E, Faniel, I, Kriesberg, A and Yoon, A.** 2013. Trust in digital repositories. *International Journal of Digital Curation*, 8(1). DOI: <https://doi.org/10.2218/ijdc.v8i1.251>
- Yoon, A.** 2017. Data reusers' trust development. *Journal of the Association for Information Science and Technology*, 68(4): 946–956. DOI: <https://doi.org/10.1002/asi.23730>
- Zimmerman, DE and Akerelrea, C.** 2002. A group card sorting methodology for developing informational web sites. In: *Proceedings IEEE International Professional Communication Conference*, 437–445. IEEE. DOI: <https://doi.org/10.1109/IPCC.2002.1049127>

How to cite this article: Mayernik, MS, Breseman, K, Downs, RR, Duerr, R, Garretson, A, Hou, C-Y, EDGI and ESIP Data Stewardship Committee. 2020. Risk Assessment for Scientific Data. *Data Science Journal*, 19: 10, pp.1–15. DOI: <https://doi.org/10.5334/dsj-2020-010>

Submitted: 19 December 2019

Accepted: 02 February 2020

Published: 12 March 2020

Copyright: © 2020 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 