Risky Business: Data-At-Risk in a Dynamic and Evolving Multidisciplinary Research Environment

LOUISE H. PATTERTON
THEO J. D. BOTHMA
MARTIE J. VAN DEVENTER

*Author affiliations can be found in the back matter of this article

RESEARCH PAPER

ubiquity press

ABSTRACT

At-risk data is an unfortunate research reality and can be present in all data formats in a range of research disciplines. This is defined as data that are at risk of loss due to various factors, including deterioration of the media, lack of accompanying documentation and data that exists in non-digital formats, which are often irreplaceable. Continued access to older data has a range of benefits. The factors that place valuable data at risk are therefore a cause for concern. This paper reports on a multi-method case study, comprising a survey and interviews. A web-based questionnaire was distributed to all research group leaders based at a leading South African research institute. This was followed by one-on-one interviews that were held with a sub-section of the same group of researchers. The combined findings of the two methods enabled a picture to be formed regarding factors that jeopardise research data, data rescue obstacles that the researchers encountered and the state of data rescue at the institute.

Several recommendations and strategies are put forward to address identified risk factors and challenges. Suggestions include the launch of a data rescue project, awareness training around data at risk, involving the institute's library and information services (LIS) section in data rescue and launching continued efforts to acquire a dedicated institutional data repository. It is also important to ensure that the scope of project risk management includes data considerations. The combined implementation of recommendations is anticipated to ensure the accessibility and usability of older at-risk data and reduce the chances of current and future data becoming compromised.

CORRESPONDING AUTHOR:
Louise H. Patterton
Department of Information Science, University of Pretoria, Pretoria, South Africa; Council for Scientific and Industrial Research, South Africa
LPatterton@csir.co.za

KEYWORDS:
data at risk; data rescue; historic data; legacy data; research data management

TO CITE THIS ARTICLE:
Patterton, LH, Bothma, TJD and van Deventer, MJ. 2024. Risky Business: Data-At-Risk in a Dynamic and Evolving Multidisciplinary Research Environment. *Data Science Journal*, 23: 11, pp. 1–21. DOI: https://doi.org/10.5334/dsj-2024-011

1. INTRODUCTION

The term 'data at risk' refers to endangered scientific data that are at risk of being lost (Murillo 2014) or the state of data when economic models and infrastructure are not in place to ensure access and preservation (Thompson, Davenport Robertson & Greenberg 2014). According to Griffin (2015), at-risk data is a blanket term for non-electronic data that are subject to a multitude of hazards of various types and origins, while Mayernik et al. (2020) refer to the absence of a dedicated plan to ensure that data is not at risk. At-risk data have become an unfortunate yet common phenomenon in research environments and should ideally be rescued. The rescue process should comprise the capture of data in a digital format for their long-term preservation and reuse (DataFirst 2022; World Meteorological Organization 2016). Within the context of this study, data rescue involves both analogue and digital data. The rescue of analogue data comprises the recovery, digitisation, description, sharing, and preservation of valuable historic data. The rescue of digital data, on the other hand, adds essential value (e.g., metadata, format information, and access) to digital data archives.

The benefits associated with rescuing valuable data-at-risk are numerous. Historical data are essential in many disciplines (Bradshaw, Rickards & Aarup 2015). Rescued data can extend the knowledge of a subject (Gallaher et al. 2015). Rescuing data makes science far more accurate (Hachileka 2015). Access to valuable data also saves time, effort, and money (Diviacco et al. 2015; Schmidt 2017).

While the rescue of data ensures its continued access, the nature of risk factors that result in data being at risk should still be cause for concern. Furthermore, data rescue efforts are typically costly in terms of skills, human resources, equipment, and time and are often not a viable venture. It is therefore crucial to not only investigate factors that play a role in data being at risk, but also the challenges that are experienced when addressing data at risk. This case study was conducted at a leading South African research institute and provides recommendations to reduce the effect of identified concerns. The aim is thus to minimise the accumulation of data that is at risk and facilitate the data rescue process.

2. LITERATURE REVIEW

This section contains a summary of published findings on factors that lead to data being at risk, as well as problematic issues that should be anticipated when one considers the rescue of compromised research data.

2.1 FACTORS LEADING TO DATA BEING AT RISK

The range of documented factors that may potentially result in data being at risk is portrayed in Table 1.

The information presented in Table 1 reveals that risk factors can be grouped into categories. The identified categories include researcher behaviour, institutional factors, format issues, and random events. Research behaviour that does not adhere to best practices is the category that is linked to the most risk factors. Lack of awareness, lack of metadata, displaced data, data that is saved in one location only, perceived data value and substandard data management practices all fall into this category. Overlap between categories can also be detected. An example is the issue of data deterioration, which can form part of both format issues and researcher behaviour.

2.2 DATA RESCUE OBSTACLES AND CHALLENGES

Studies that report on data rescue challenges are not commonplace, as published rescue studies tend to report on the success of the project. However, Downs (2015) has listed several problematic issues that research institutes need to consider when contemplating data rescue. It is important to mention the realistic rescue challenge of at-risk data that is in poor condition, incomplete, or partially curated. Moreover, Downs (2015) highlights the feasibility of a rescue project that depends on technical factors, including capabilities, resources and infrastructure. Training or the acquisition of experts is a possible data rescue project requirement. Additional data rescue challenges that need to be resolved prior to the start of a rescue project relate to

RISK FACTORS	DOCUMENTED INSTANCES
Deterioration of the record	Hachileka 2015; Arrouays et al. 2017
Catastrophic loss of records	Levitus 2012; Arrouays et al. 2017
Loss of human knowledge and skills	Wyborn et al. 2015; Mavraki et al. 2016
Outdated format and media	Muller 2015; Research Data Alliance 2019
Substandard quality	Brunet & Jones 2011; Mavraki et al. 2016
Missing or displaced data	Levitus 2012; Wyborn et al. 2015
Perceived data value	Nordling 2010; Griffin 2015
Lack of awareness	Griffin 2015; Schumacher & VandeCreek 2015
Changing priorities	Muller 2015; Arrouays et al. 2017
Metadata and documentation issues	Thompson, Davenport Robertson & Greenberg 2014; Wyborn et al. 2015; Mavraki et al. 2016
Government funding/administrative policy	Gaudin 2017; University of Nevada 2017
Non-digital formats	Levitus 2012; Thompson, Davenport Robertson & Greenberg 2014
Archiving/preservation policy	Downs & Chen 2017; Griffin 2015
Data in one location only	Murillo 2014; Gaudin 2017
Substandard data management practices	Levitus 2012; Muller 2015
Non-trustworthy copy	Guest Blogger 2017

Table 1 Documented factors leading to at-risk data.

decisions regarding validity and authenticity; whether the value of the data will justify the costs and whether the trade-off of rescue techniques might compromise the effort. Obstacles such as limited data quality information, incomplete data provenance, and delays in acquiring the required infrastructure or equipment should be considered and alleviated.

Published sources reporting on factors that lead to data being at risk and the challenges experienced by those considering the rescue of at-risk data sketch a bleak picture in this regard.

3. CHARACTERISTICS AND ATTRIBUTES OF THE SELECTED RESEARCH INSTITUTE

The research institute that formed part of this case study was established in 1945 and undertakes directed, multidisciplinary research and technological innovation to improve the quality of life of South Africans. Several major changes have occurred during the last seven decades. These include transformative events that emanated from management decisions, international trends, newly formed research disciplines, and a change in institutional strategic focus.

Adding to the factors mentioned above are universal interruptions and upheavals, such as researcher movement between research groups, resignations, retrenchments, and retirements. Frequent and dynamic transitions and transformations, while crucial for institutional progression, could potentially be accompanied by less-than-ideal data-related outcomes. Examples of potentially detrimental consequences include discarding data when a research group disbands, knowledge of older rescue projects and associated data context leaving the institute, and researchers who are familiar with older data formats, readers, and equipment no longer being available to the institute.

With the potential cumulative effects of factors that influence research continuity and the management of valuable data, it was crucial to obtain researchers' perspectives on data at risk, and their experiences when attempting to rescue vulnerable data.

This case study examined the following concerns to provide insight into the need to rescue at-risk data:

- the factors that have led to institutional data becoming 'at risk,'
- identified risk factors that are unique to the selected institute, or not commonly mentioned in published literature,
- the forces that may drive these risk factors,
- the challenges researchers experience when they conduct activities to rescue at-risk data, and
- the steps that should be implemented to mitigate and alleviate identified risk factors and challenges.

It is worth noting that the institute's data management implementation was still in its infancy when the data collection for this study was completed. A data management policy had been drafted, yet was not finally approved. Data management training had not yet been rolled out, and the institute was scoping a dedicated institutional data repository.

4. METHODOLOGY

This paper is a multi-method case study involving two data collection activities. A web-based questionnaire was distributed to all research group leaders (RGLs) based at the institute. One-on-one interviews were then held with a sub-section of the same group of researchers. The combined findings of the two methods enabled a picture to be formed regarding factors that jeopardise research data, as well as data rescue obstacles encountered by researchers. The research approach was qualitative, and both instruments used open-ended questions to extract the necessary information.

4.1 RESEARCH PARTICIPANTS

This case study collected information from the institute's RGLs. RGLs manage the institute's various research groups. They are experts within their research disciplines, and have knowledge of the data held by the research group, as well as the data activities performed by the group members.

The study made use of the two RGL sample groups:

- Sample A consisted of 49 RGLs and comprised the entire RGL population of the institute. Participants received invitations to complete the study's web-based questionnaire. Each of the institute's research groups is managed by a single RGL, and an RGL may manage more than one research group. The rationale behind the selection of Sample A was to have a group of experienced researchers who would:
    - provide a broad overview of the presence of data at risk in the institute;
    - provide a broad overview of the presence of data rescue activities in the institute; and
    - enable the selection of a sub-sample (Sample B), based on their responses.
- Sample B consisted of RGLs who had revealed in the web-based questionnaire that they either had at-risk data in their possession, or they had performed data rescue activities in the past. These respondents were invited to participate in one-on-one interviews. The invited sample comprised 18 RGLs. Of the 18 invited RGLs, eight agreed to be interviewed. It was unfortunate that 10 of the 18 invited RGLs, who had indicated in the questionnaire responses that they had at-risk data, declined to be interviewed.

4.2 WEB-BASED QUESTIONNAIRE

A first data collection stage involved the collection of information that would provide a good overall glimpse of the current state of data at risk and data rescue at the research institute. A web-based questionnaire presents several advantages, such as its capacity to reach remote participants (Adams & Lawrence 2015; Leedy & Ormrod 2005), its ability to reach many participants in a brief period (Menon & Muraleedharan 2020) and its capacity to export

questionnaire data to different formats, depending on the requirements or preferences of the interviewer (eSurv 2019). The researcher therefore decided to use this data collection method as a survey tool.

Regmi et al. (2016) list six vital factors to consider when designing an online or web-based questionnaire. These aspects influenced the design of the online survey:

- A progress indicator was used. It was possible to page back and forth between questions, and ample space was available for long responses.
- Participant selection was done very carefully (see Section 4.1), and sampling was purposive.
- The possibility of multiple responses by the same respondent was avoided as the online questionnaire was not anonymous. Identifying details were required to select Sample B.
- Data management was considered. As a web-based tool, eSurv could export questionnaire data to PDF, CSV, and Microsoft Excel formats. In addition, data were private, as the investigator's log-in details were required to access the data.
- The online questionnaire adhered to ethical requirements. All study components (including the questions asked, the wording of the cover letter, the wording of the consent form, and the use of data) were examined by the relevant Research Ethics Committees. Ethical clearance was obtained. An online indication of informed consent formed part of the web-based questionnaire.
- The web-based questionnaire was piloted before it was launched, and a link distributed to the RGLs.

After taking the above advantages and requirements into account, a short web-based questionnaire comprising eight questions on at-risk data and data rescue was distributed to the institute's 49 RGLs, also referred to as Sample A. The questions pertained to the existence of each group's at-risk data, the format/s of the data, a brief description of the data and the location of the at-risk data. Regarding data rescue, RGLs were asked whether they had previously performed any data rescue activities, whether the group had data rescue documentation or procedures in place, whether data documentation could be shared with the investigator, and whether participants were able to direct the study investigator to institutional parties who had performed data rescue activities. The concepts of 'at-risk data' and 'data rescue' were explained in the questionnaire. A copy of the web-based questionnaire's wording is attached as Appendix 1.

4.3 ONE-ON-ONE INTERVIEWS

Members of Sample B were invited to participate in one-on-one interviews. This sample comprised RGLs who had indicated in the web-based questionnaire that they either had at-risk data, or had performed data rescue activities in the past. The objective of the interviews was to obtain more detailed information pertaining to each research group's at-risk data. The interviewer also collected details about each research group's data rescue experiences. The main advantage of personal interviews is that the interviewer is able to clarify responses, probe when additional information is required, and even ask additional questions that do not form part of the interview schedule. These factors can improve the accuracy of responses (Adams & Lawrence 2015). Additional benefits include the high response rates (Akba Yrak 2000) and permitting long and complex questionnaires (Neuman 2014). Interviews conducted in this study were in-depth in nature, and made use of probing questions, as well as follow-up questions in the case of vague or ambiguous replies, and delved into data rescue-related topics that did not form part of the interview schedule.

Virtual one-on-one interviews took place with members of Sample B as this activity took place during the South African COVID-19 lockdown period. Virtual interviews offer the additional benefits of allowing for the collection of data over wide geographical areas (Jowett 2020), and they are often less expensive and more time-efficient than in-person interviews (Krouwel, Jolly & Greenfield 2019). Participants often feel more comfortable disclosing sensitive information by not being face to face with the interviewer, and by being able to participate in the interview from a familiar space such as their own home (Hanna 2012; Sipes, Roberts & Mulan 2019).

Interviews were semi-structured and were conducted by the principal investigator of the study (also the first author of the article). Verbal informed consent was given at the start of each interview. Interviews were recorded and subsequently transcribed by the interviewer. A copy of the outline of the interview schedule is attached as Appendix 2.

Transcribed interviews were subjected to thematic analysis, thereby enabling the identification of topics, ideas and patterns that came up repeatedly, such as the performance of data rescue activities, data risk factors and data rescue obstacles.

5. RESULTS

5.1 RESPONSES

As all of the institute's RGLs were invited to complete the questionnaire, it was important to examine both the conveyed information linked to data at risk and data rescue activities, and also the response rates, feedback pertaining to RGL group management, and the number of research groups involved in responses.

The responses to the web-based questionnaire revealed that:

- 23 of the 49 RGLs (51%) submitted completed or partially completed web-based questionnaires;
- two of the above-mentioned 23 RGLs managed more than one research group;
- research groups were never managed by more than one RGL;
- findings provided information on 25 research groups; and
- 18 RGLs met the interview invitation requirement by indicating that their group had at-risk data.

Eight of the 18 invited RGLs, including one proxy, agreed to be interviewed.

5.2 FACTORS PUTTING DATA AT RISK

In-depth interviews with eight RGLs resulted in the identification of 34 factors that have put data at risk (Patterton 2023b; Patterton 2023c). Each factor, together with its frequency, is shown in Figure 1. Factors mentioned by respondents are applicable to both analogue and digital data.

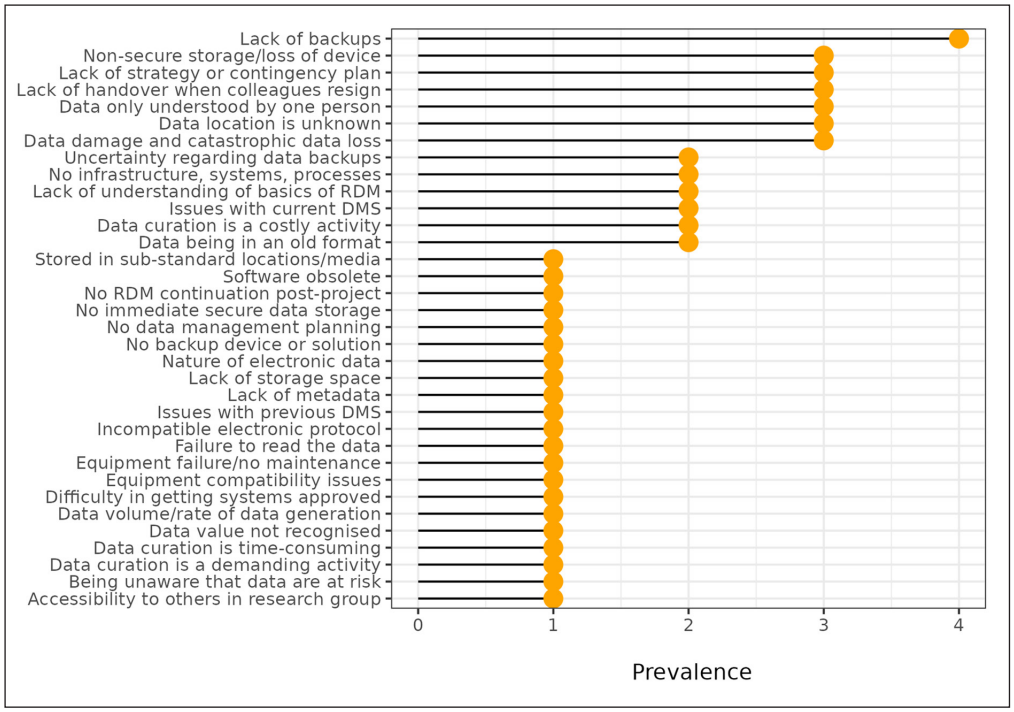

Figure 1 Factors that put data at risk.

5.3 DATA RESCUE ACTIVITIES

The results of the question 'Has your research group ever performed data rescue activities?' are portrayed in Figure 2.

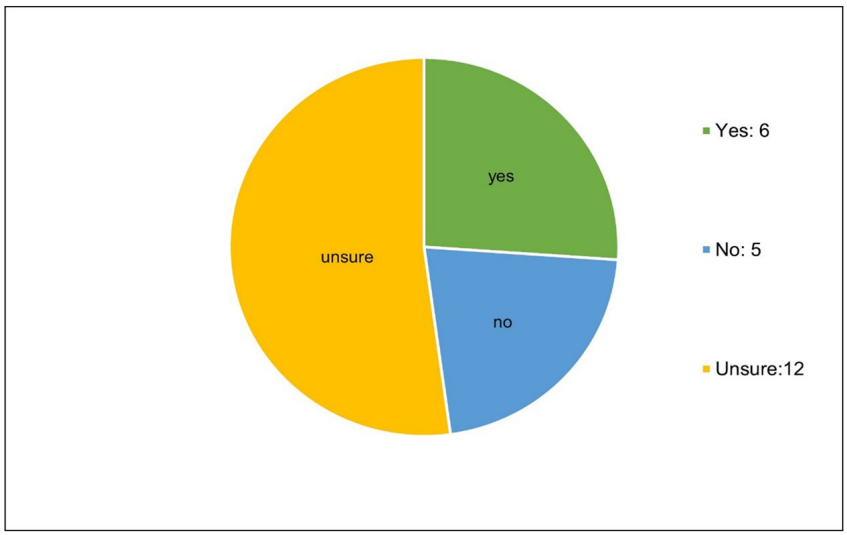

Figure 2 Occurrence of data rescue activities.

Most research groups had never performed any data rescue activities, with only a quarter of the respondents stating that some data rescue activities had taken place. The number of respondents that indicated that they were unsure whether data rescue had been done is troubling. However, it is not clear from the responses whether some of the RGLs were newly appointed managers (and unfamiliar with the group's activity history), or whether data management activities were not shared with all the members of the group, and with the managers.

These results were unexpected. Even though this researcher was not surprised by the dominance of data at risk in the institute, the overwhelming absence of data rescue projects and activities was not an anticipated outcome of the web survey. The expectation was that many of the research groups would somehow have attempted or organised data rescue initiatives, ranging between rudimentary to structured data rescue activities. The unanticipated nature of this finding is further exacerbated by the fact that most of the subsequently interviewed parties indicated that their at-risk data was of value and needed to be shared.

The interviews that followed the web-based questionnaires enabled deeper probing into data rescue activities that had been performed. Table 2 contains a summary of the nature and number of data rescue undertakings by the respective research groups.

DATA RESCUE ACTIVITIES EXECUTED	NUMBER
Uploaded data to a repository	n = 3
Added metadata to data	n = 3
Ensured data have backups	n = 2
Ensured digital storage space is sufficient	n = 1
Created and adhered to a data management plan	n = 1
Retrieved data after document management system failure	n = 1
Kept sensitive data offline	n = 1
Outsourced data rescue	n = 1
Retrieved data after data breach/ransomware incident	n = 1

Table 2 Nature and number of data rescue activities.

'n' refers to the number of interviewed RGLs who had stated during their interview that the group had performed the indicated data rescue activities.

As indicated, the two most common data rescue activities involved uploading data to a repository, and adding metadata. Other data rescue activities were mentioned by fewer than three of the eight respondents.

A single instance of a complete data rescue project was discovered. The rescue project involved early format digital audio data, with the main rescue steps resembling the typical rescue steps

described in published data rescue literature. The project also involved the outsourcing of many of the rescue steps to an external party, with the resultant product being data in a modern electronic format, accompanied by metadata that was made accessible to the public via a dedicated discipline repository.

Data rescue was a rare activity at the institute and many of the interviewed RGLs had experienced data loss. This either entailed the loss of data they had collected themselves, or the loss of data belonging to the research group. Activities that were not mentioned, but were expected to have been shared with the investigator, included the identification of valuable data, the identification of data at risk, and approaching the institute's research library for data rescue guidance. In addition, despite several RGLs having historic analogue data in their groups, reported data rescue activities only involved early and modern digital data.

5.4 DATA RESCUE OBSTACLES AND CHALLENGES

Respondents identified 26 data rescue obstacles. The RGLs each mentioned between two and eleven obstacles (Patterton 2023b). The most common data rescue challenges were found to be cost, data quantity/volume, time required to rescue data, and the fact that a data rescue venture would require skills, insight, and experience.

Figure 3 shows the data rescue obstacles mentioned by the RGLs when they were asked to list the obstacles they experienced when rescuing data or considering the rescue of data, or the obstacles they foresaw in a hypothetical data rescue scenario. The challenges and obstacles mentioned by the respondents are applicable to both analogue and digital data.

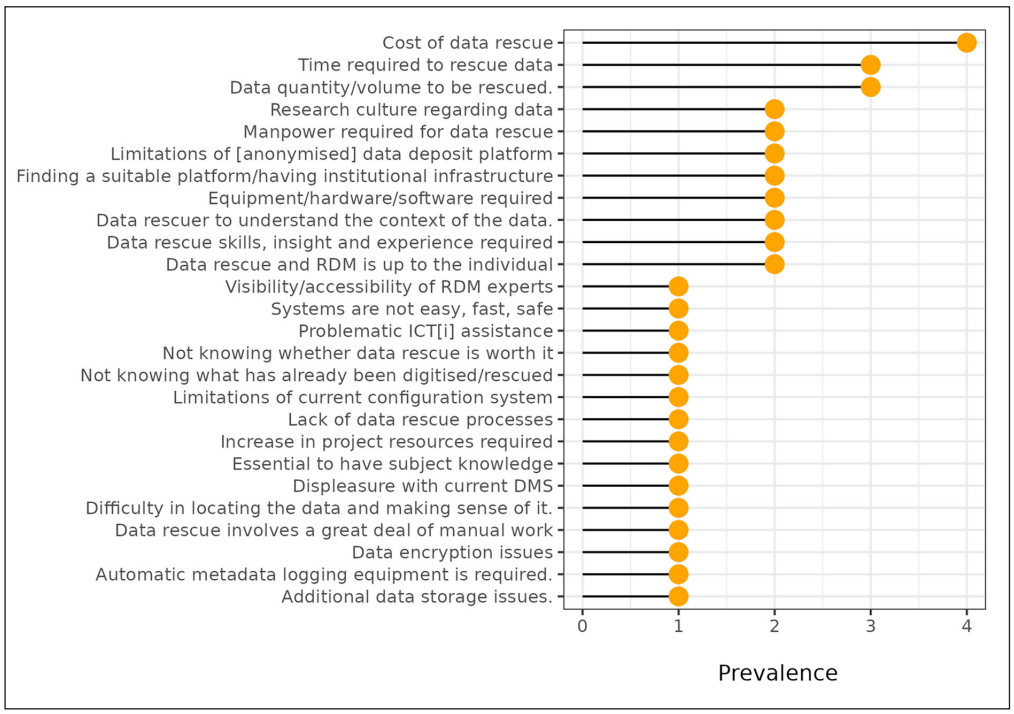

Figure 3 Data rescue challenges and obstacles.

6. DISCUSSION

This section describes, analyses, and interprets the findings presented in the previous section.

6.1 FACTORS THAT PUT DATA AT RISK

In-depth one-on-one interviews with eight RGLs resulted in the identification of 34 risk factors thought to lead to data being at risk. This number of risk factors is troublesome and greater than anticipated. Risk factors reflect on aspects involving human behaviour, information technology (IT) matters and institutional policy. With many of the risk factors sharing similarities in nature and root cause, it was possible to place the listed factors into broader categories. In addition, several of the listed risk factors fit into more than one of the broader categories. A great deal of overlap was detected. These broader divisions, containing instances of the respondents' mentioned data risk factors, are listed and expanded on below.

6.1.1 Factors related to institutional policies and procedures

This category can be viewed as one that encompasses many, if not most of the root causes of factors that lead to at-risk data. Institutional policies and procedures (or the lack thereof) can indirectly influence the categories of researcher behaviour, recommended, or implemented equipment and systems, and IT-related services. These three categories have also been identified as broader categories of data being at risk and appear later in this section.

Examples of identified risk factors that fall within the category of institutional policies and procedures are the lack of data processes (this includes data management processes and processes that affect research activities, and data being at risk), problems in getting systems approved (approval procedures and the accompanying red tape result in delays in acquiring, implementing and using the required equipment and systems), displeasure expressed by respondents with both the previous and current document management system (DMS), and the absence of an effective strategy.

6.1.2 Factors related to IT

Factors that fall into this category include substandard storage, insufficient storage space, loss of storage, protocols not being compatible with equipment, the nature of the electronic data, lack of infrastructure, DMS issues, data damage, and no available backup device.

Reasons for the preponderance of factors in this category include insufficient funding (funding affects the storage infrastructure, devices and backup equipment that can be purchased), and the purchase, implementation and recommendation of systems by the institute without considering the possible needs of all researchers and their data. This can lead to hesitancy in, or failure of uptake.

In the absence of data management topics on risk management plans (e.g., storage, backup, data volume), it is likely that IT-related issues will contribute to data being at risk. Furthermore, the absence of data management plans (DMPs) will result in the increasing contribution of IT-related issues to data being at risk. DMPs typically state the data volume, data storage and backup locations, infrastructure required, and the long-term preservation activities linked to a planned research project.

6.1.3 Factors related to researcher behaviour

This category is two-pronged in nature. On the one hand, it refers to behaviour patterns based on informed non-compliance. On the other, it includes activities or practices that are unwittingly performed due to the absence of awareness of best practices. Examples of the former are poor handover upon resignation or retirement, only one person in the group understanding the data, or storing data on sub-standard devices. By contrast, failure to create a DMP or trusting the institutional DMS to be secure are examples of practices that could be addressed via data management training.

It is also probable that random and unforeseen behavioural events can result in data being at risk. Sudden resignations or departures are examples of typical events that can lead to a situation where research group members are unable to access the data or understand it.

There may be a link between the ease and convenience of data handling or data storing behaviour and the resultant risk of data loss. This aspect must be investigated further. Examples of convenient data handling include storing electronic data only on a laptop, making use of a USB storage medium that happens to be nearby, or stashing paper-based data in an unmarked cardboard box in an unused office.

The line between inferior data management practices due to deliberate non-compliance and practices that are caused by ignorance is often blurred. For this reason, the importance of data management awareness and training, and ensuring that institutional data policies and procedures are visible, clear, concise and viable, is discussed in the recommendations section.

6.1.4 Factors related to equipment, systems, and tools

Stated risk factors include software obsolescence, equipment incompatibility, data damage, equipment failure, substandard storage, insufficient storage space, and failure to read the data.

A major reason behind this category of risk factors, which features a multitude of factors, is likely inadequate funding, as cost can be a vital determinant when deciding whether to make use of secure off-site paper data storage options, or to digitise paper, or simply to store the paper in an unused office nearby. Policy regarding recommended platforms and systems may also contribute to these practices being a risk factor. The institutional recommendations regarding the previous DMS, where substantial volumes of data were later found to be inaccessible or corrupted, is a case in point. Another example is the institutional recommendation that the new DMS (implemented from 2019 onwards) be used for data storage. The interviewed RGLs indicated their displeasure with the system and expressed their doubts regarding its speed and user-friendliness.

The absence of DMPs in earlier years of research can be linked to the use of equipment that has turned out to be sub-standard, faulty, or not ideal for data storage. Drafting DMPs and having them reviewed by other parties (e.g., funders, research group members, or external collaborators) could potentially have resulted in the detection of the planned use of non-ideal equipment, systems, and platforms.

It is noticeable that many of the equipment-related factors are also linked to IT-related factors. This emphasises the importance of clear and open communication channels with the institute's IT department, and prudent reporting and feedback when errors are detected or problems experienced. Minimal input from the institute's research sector prior to the institutional purchase and rollout of equipment can also be regarded as a risk factor.

The institute has a dedicated asset management division, which performs regular asset verification and asset disposal tasks. This is an auditing requirement and ensures that asset devaluation is up-to-date, and insurance coverage is correct. However, it is vital that older and seemingly 'outdated' equipment is discarded with caution to ensure that rare equipment, that will potentially be required to access older, valuable data at a later stage, is not discarded.

6.1.5 Project-specific dilemmas

Factors in this category tend to be sporadic and unusual in their nature and occurrence. They are deemed to be project specific as they often go hand in hand with the unique circumstances surrounding each project and may dissipate as a risk factor once the project is concluded. Examples of risk factors include the inability to immediately store data in a secure location during fieldwork, the inability to make immediate backups of data during fieldwork, or the rate of data generation (including data volumes) connected with the project.

A major reason behind the project-specific 'random' risk factors entails funding, which affects access to secure mobile storage during fieldwork, or being able to store vast amounts and velocities of data. Insufficient training can also be viewed as a reason. During the interviews, RGLs stated that training field workers to upload data to a cloud-based environment while still based in the field can reduce the risk of data loss.

In certain instances, the non-mitigating nature of risk factors should be seen as a research reality. Steps to reduce the chances of data loss include making backups of data as soon as possible, prioritising data that are irreplaceable, and ensuring that available off-site data management options have been considered. Even so, catastrophic data loss or criminal events can occur, but the chances are minimised through data management best practices.

6.1.6 Scholarly comparison and implications

On the whole, data at risk factors identified via the literature review tend to mirror many of the data at risk factors listed by the respondents involved in this study. The risks that were linked to institutional factors such as policies (Thompson, Davenport Robertson & Greenberg 2014), changing priorities (Muller 2015), and downsizing (Muller 2015) were also factors mentioned by the interviewees. Interview respondents revealed a higher occurrence of IT-related risk factors than was established via the literature review. This can be attributed to respondents placing great emphasis on digital data during the interviews. Of particular concern were the statements by respondents that the institutional lack of infrastructure, systems and processes, loss of devices, and an inferior DMS were placing their digital data at risk.

Broadening the scope of IT factors to include technology issues such as software obsolescence, incompatible readers, sub-standard media, failure to read the data, and data being in an old format shows the similarities between the concerns of RGLs and the factors stated by Levitus (2012), Murillo (2014), Hachileka (2015), Muller (2015), and Wyborn et al. (2015).

Similarities in data risk factors linked to researcher behaviour were detected, with the interviewed RGLs echoing the published risk factors of the loss of skills (Wyborn et al. 2015), non-secure storage (Murillo 2015), inferior data management practices (Levitus 2012), not being aware of the value of data (Griffin 2015; Nordling 2010), and lack of metadata (Murillo 2014; Thompson, Davenport Robertson & Greenberg 2014; Wyborn et al. 2015). However, respondents also stated lack of backups, lack of strategy, and data not being accessible to others as being detrimental to the survival of their data.

Other factors uncovered through the interviews and not widely published in scholarly literature are mostly institute-specific and include the massive data volumes generated in certain research groups, insufficient data storage space or storage media, and the general acceptance of non-secure data storage options. One of the respondents mentioned the danger of running out of storage media for valuable, sensitive physical samples during the hard COVID-19 lockdown and mentioned the resultant financial implications should the media have been exhausted. Other risk factors linked to the specific institute include not having a backup device while working remotely, incompatible research protocols, and equipment displaying non-compatibility issues.

Several respondents stated that lack of time, lack of funding, and insufficient human resources can result in data being at risk. This appears to be a factor of particular concern at this institute and is most likely related to the institute's recent restructuring activities and the freezing of non-vital positions.

Another institute-specific risk factor (which may also fall under ICT-related factors): The lack of a formal data management platform or a dedicated data repository, pointed to inferior data storage infrastructure at the time of data collection. Institutional-wide data management planning templates were also not yet available. Equipment malfunction and the potential resultant loss of data were also mentioned as risk factors, as were problematic issues with the institute's DMS. The previous DMS (instituted before 2019) was linked to the corruption of records, an inferior search facility and the inaccessibility of many items. Respondents described the current DMS (implemented since 2019, and in use at the time of this study's data collection) as being slow and not being user-friendly.

With several institute-specific factors not featuring strongly within the literature reviewed for this research, it is prudent to examine the reasons for this institute's researchers mentioning these risk factors. It is likely that the common denominators in many of the risk factors amount to the institute's lack of suitable infrastructure, systems, platforms and servers. Lack of management support and buy-in are suspected to be related to the stated issues. The absence of the required data management foundations, coupled with insufficiencies in systems and storage options, have potential dire consequences for valuable data, and are addressed in the recommendations section.

Implications emanating from this institute's at-risk data factors are numerous. The most important ramification is that the factors leading to data being susceptible to loss should be acknowledged and addressed. Most of the factors causing data to be in a threatened state can potentially be reduced or even eliminated. These factors comprise a lack of a data management or rescue strategy, a lack of data management or rescue processes, or making use of sub-standard data storage. The stated factors should be attended to, with possible actions entailing the inclusion of 'at-risk data' in data management training sessions, and ensuring that 'at-risk data' forms part of data management awareness material and guidelines.

The link between an institutional data management policy, data management procedures and the resultant data being at risk is obvious. Additionally, many of the behavioural factors that lead to data being at risk are inversely proportional to good research practices and ethical research behaviour. Of particular importance is the creation of a DMP at the start of every research project. A DMP is expected to stipulate data management best practices and detail the envisaged data storage locations, data backup practices, access arrangements, metadata activities, and long-term preservation strategy. Adhering to the DMP in conjunction with considering the project's risk management plan are vital steps in minimising future at-risk data.

The mitigating activities listed above emphasise the role and need of a data manager (either institutional or within the research group) and data management awareness training for research group members.

The fact that IT-related factors such as participation, systems, platforms, and infrastructure were frequently mentioned as data risk factors demonstrates the need for close and clear communications with the institute's IT department. The involvement of and collaboration with the IT department during the mitigation of at-risk data factors cannot be disregarded.

Despite all aspects listed here, it is part of the research reality that risk factors will always be present. Limited human resources, limited skills, budget cuts, equipment obsolescence, and equipment failure are as much a part of the research environment as are best research practices and approved procedures.

6.2 DATA RESCUE OBSTACLES AND CHALLENGES

The study identified 26 data rescue challenges or obstacles. The myriad stated challenges provides an insightful glimpse into the problems envisaged by researchers when asked to list challenges experienced during past rescue efforts or to envisage challenges that would be encountered during a hypothetical data rescue project.

The number and range of listed challenges are worrying, but not unexpected. Previous research into the institute's data management practices revealed low adherence to data management best practices, an absence of vital data management platforms or tools, and data management procedures not yet in place (Patterton 2014; Patterton 2016; Patterton, Bothma & Van Deventer 2018). In addition, many of the stated data rescue challenges show similarity with factors that lead to data being at risk. Examples are the high incidence of IT-related factors, lack of processes, and problems with the current and available institutional systems. While the findings emanate from data collected at a single research institute, it is suspected that many of the obstacles are universal in nature and will be prevalent and cause for concern at other research institutes in a similar stage of data management maturity.

The 26 data rescue obstacles and challenges listed by respondents can be clustered into four broad categories, with several obstacles fitting into more than one category.

6.2.1 Institute-related challenges

Obstacles in this category emanate from the nature of institutional policies, procedures, and guidelines in place, as well as the lack of relevant policies, procedures, and guidelines. Examples of the stated challenges are problematic IT assistance, lack of data-related infrastructure, and lack of data rescue processes.

Infrastructure and IT-related challenges, in particular, are cause for concern, as the accessibility and security of modern data depend on infrastructure, IT-related systems, and IT support. With modern data mostly being in an electronic format, saved on institutional servers, backed up to similar devices, and relying on systems to provide the required access requirements and restrictions, these challenges are issues that cannot be ignored. Put simply, the status quo is not advantageous to working with either historic or modern data.

6.2.2 Equipment-related challenges

Equipment-related challenges are obstacles that emanate from the equipment available to the research group. One could also state that this category is linked to the previous one, as institutional policies often determine the equipment to be purchased, or that is permissible or recommended. Examples of obstacles are the stated lack of applicable equipment, slow systems that are not user-friendly, and the configuration system not meeting the group's envisaged needs. In addition, the desire for equipment to automatically log metadata, and troublesome storage issues also fall into this category.

6.2.3 Resource-related challenges

While linked to 'institute-related challenges' and even 'equipment-related challenges,' resource-related challenges refer to a research group's lack of the means to successfully perform data rescue. Included in this category are statements related to a lack of human

resources, funding, time, skills, and data understanding should data rescue be required or considered. Skills and means to deal with big data (either in volumes held or rate of generation) also fall under this heading.

6.2.4 Challenges related to sub-standard data management

Sub-standard data management can also be described as a behaviour-related challenge, and can even be ascribed to the institute's research culture. Examples of challenges in this category include data management and data rescue being up to the individual, the inability to assess data value, and not knowing whether paper-based data have already been digitised.

Many of the behavioural obstacles can be mitigated through applicable training. An understanding of data management best practices and insight into data assessment steps are crucial requirements. Conversely, a preponderance of behavioural factors exists due to the current research environment, research culture, or economic climate. Factors that contribute to sub-standard practices may include human resources constraints, funding difficulties, or contractual obligations. Funder mandates, which often drive institutional data management (Digital Curation Centre 2023), did not appear to be a factor in detected research data management activities at the selected institute.

6.2.5 Scholarly comparison and implications

The investigation of data rescue challenges and obstacles did not form a big part of this study's literature review. Moreover, it was not a common feature of data rescue studies, as the published outputs were mostly the result of successful data rescue efforts. By implication, published ventures rarely contain details of insurmountable data rescue obstacles that result in aborted rescue projects. However, the challenges reported by RGLs show some similarities with the problematic data rescue issues listed by Downs (2015). Similarities in concerns regarding capabilities, resources and infrastructure were detected. Two examples illustrated the need for data rescue experts and specialised infrastructure. It is also anticipated that concerns regarding data quality and data provenance will surface once organised data rescue projects are rolled out at the selected institute. Schmidt (2017) mentioned that data rescue can be labour intensive. This sentiment was echoed at the selected institute. Thompson, Davenport Robertson and Greenberg (2014) mentioned the funding required for data curation. This is an issue identified by half of the interviewed RGLs as a data rescue challenge.

The following implications, which result from the findings of data rescue obstacles at the selected institute, were identified:

- The uptake of non-approved systems and tools due to dissatisfaction with the current options is a troubling, yet likely scenario.
- Data rescue activities and data rescue projects will not be ready for implementation at the selected research institute until certain data rescue challenges have been addressed and mitigated.
- The difficulty of getting a data rescue project off the ground is regarded as a likely scenario.
- The absence of data rescue projects can be ascribed to a combination of obstacles or challenges, and not the result of one single factor. Similarly, only addressing a single obstacle is unlikely to result in the sudden update of data rescue activities.

7. LIMITATIONS OF THE STUDY

Research limitations are the theoretical or practical shortcomings of a study and are often outside of the researcher's control (Viera 2024). Addressing the limitations of this study, particularly the issues affecting the rigour of the research or the robustness of the results, ensures the study's transparency (Dissertation Team 2024; Munch 2020).

The following study limitations, linked to the samples and their responses, should be mentioned:

- It is possible that some of RGLs who did not submit an online questionnaire also had data at risk. Survey fatigue is an unfortunate reality at the selected institute.

- It is possible that some of RGLs who submitted an online questionnaire were not fully aware of the at-risk data held by the group. This would particularly be the case where a research group had a newly appointed RGL.
- RGLs who did not accept interview invitations had at-risk data or had performed data rescue, as was revealed via their web-based questionnaire responses.
- It is possible that the institute has data at risk that is not accounted for, due to research groups often being disbanded.

It is also likely that the unique characteristics of the selected institute, particularly its research disciplines, resource availability and research policies, might affect the universality of study findings linked to at-risk data and data rescue activities.

8. RECOMMENDATIONS AND THE WAY FORWARD

Based on the findings of the study, several recommendations are made to address the state of at-risk data and data rescue at the selected institute, but also for other research institutions. While suggestions are institute-specific, and take the institutional systems, tools, and staff complement into consideration, the broader gist of each recommendation is applicable to other research institutes as well.

8.1 ADDRESS AND PROMOTE AWARENESS OF AT-RISK DATA

Lack of insight into the value of older data and the correct way of handling at-risk data are likely to result in the current volumes of at-risk data increasing daily. Failure to address this scenario may lead to a state where only a small percentage of at-risk data rescue is deemed feasible. It is vital for researchers to be aware of the following:

- The potential value of historic data
- The correct ways of handling older data
- The important role of metadata and data documentation
- The enormous costs associated with data rescue
- Parties to contact should at-risk data be located and identified

Awareness creation can take the form of an intranet article, posts on the blog of the institute's Library and Information Services (LIS) division, visits of the LIS division's representatives to research groups, and a possible institute-wide webinar.

Promotional activities that are focused on at-risk data will go hand in hand with an understanding of data rescue (see Section 8.2).

8.2 PROMOTE AWARENESS OF DATA RESCUE

In-depth interviews with RGLs have shown that data rescue activities are not commonly performed. Knowledge of the concept of data rescue is also limited. It is doubtful whether the skills set required for the rescue of data is currently present within the research sector to which the institute belongs, as the only example of a complete data rescue project found comprised the outsourcing of the project to an external party.

To deal with these gaps in knowledge, data rescue should be promoted as both a concept and an activity, with the following concepts forming part of knowledge-sharing sessions:

- Conveyance of the benefits of data rescue
- Explanation of the typical data rescue workflow
- Demonstration of a data rescue workflow model
- Indication of the anticipated involvement of various institutional sectors

As will be the case for the advancement of institutional knowledge pertaining to at-risk data, data rescue awareness at research institutes can make use of the institutional LIS blog or website, an intraweb article, personal visits by the designated LIS rescue trainer to the research groups and an institute-wide data rescue webinar. It is anticipated that the two concepts—at-risk data and data rescue—will form part of the same training session, webinar, or personal visit.

It is important that awareness sessions also explain that not all data can be rescued, and that not all data should be rescued. The importance of data assessment to determine the value of data, and of evaluating data rescue resources should form part of the preliminary stages of each prospective research project.

8.3 LAUNCH A DATA RESCUE PILOT PROJECT

Following the activities described in Section 8.1 and Section 8.2, the launch of a data rescue pilot project is recommended. This pilot project will be the first of its kind at the selected institute and entails testing the rescue activities and the required outputs, as well as the feasibility and efficacy of the involvement of different institutional sections (e.g., researchers, members of the LIS division and the IT department). Even though the data rescue project will adhere to the data rescue workflow model created by the institutional data librarian (Patterton 2023a), the pilot project experience is anticipated to result in several changes and adaptations to the recommended model.

The main recommended stages of a single generic data rescue project, with its crucial activities, are indicated in Table 3. While this table contains the steps for a rescue project involving paper-based data or data in an early digital format, several of the rescue steps (e.g., data description or data sharing) can also be applied to digital data rescue.

Table 3 Data rescue stages and accompanying activities.

DATA RESCUE STAGE	ACTIVITIES
Preparing for data rescue	• Locate at-risk data • Create data inventory • Assess at-risk data • Assess data rescue resources • Make data rescue decision
Planning for data rescue	• Select data rescue team • Draft data rescue project plan • Draft data management plan for the data
Storing and preserving the data	• Clean the at-risk data • Prepare and clean the storage area • Lable and store the at-risk data accordingly
Digitising/converting the data	• Create digitisation inventory • Create predetermined files and folders • Image, scan, key, or convert the data • Quality control and validate
Describing the data	• Select applicable repository, and examine its metadata standard • Draft metadata template • Complete metadata document • Complete data documentation document
Sharing the data	• Upload data to selected repository • Ensure metadata and data documentation accompany the data • Add DOI information to all relevant documents
Archiving the data	• Select secure and stable preservation location • Ensure data are in preservation format • Upload data • Ensure regular monitoring takes place
Project closure	• Thank all parties involved • Create awareness of rescued data • Share data rescue learnings • Implement changes to data rescue model if required • Draft and distribute final project report

Elaborative details regarding the activities linked to each stage form part of the data rescue workflow model created by the institute's data librarian (Patterton 2023a).

8.4 INCREASE DATA MANAGEMENT AWARENESS AND COMPLIANCE

Findings have shown that the presence of sub-standard data management practices is both a risk factor for data and an obstacle during data rescue activities. Examples of non-ideal practices and their linked detrimental effects include the mention of a certain 'research

culture,' researchers adhering to their 'own processes,' the absence of data management planning and the lack of metadata linked to datasets. It is interesting to note that while funder mandates drive data management in many countries (Digital Curation Centre 2023), this had not yet manifested at the research institute in question. To deal with these issues, the recommendation is made that all researchers be aware of good data management practices, with certain practices being subjected to a monitoring and compliance process.

The approval of an institutional data management procedure, coupled with institute-wide data management training, had not yet commenced during this study's data collection stages, but followed soon afterwards. The attendance of the mandatory sessions was relatively poor. It is therefore recommended that future training sessions be complemented by online self-training tools, personal visits of the data librarian to research groups, and a slow roll-out of data management compliance monitoring.

In addition to data management adhering to best practices, researchers should also be familiar with an institute's data management policy and data management procedure. Although following best practices regarding data management will not undo the threatened state of an institute's historic data, it can ensure that data currently being generated, and to be generated in the future, do not succumb to the same predicament.

8.5 BROADEN THE SCOPE OF THE RISK MANAGEMENT PLAN

A risk management plan currently forms part of the project registration requirements at the selected institute. It is suggested that potential risk concerns pertaining to data be added to the risk document templates used by this institute and elsewhere, and that potential threats to a project's data accessibility, understanding and usability be anticipated and documented.

8.6 INVOLVE THE LIS SECTOR IN DATA RESCUE

The institute's LIS division currently performs many tasks, similar to the rescue activities mentioned in Sections 8.2. and 8.3. In addition to familiarity with data rescue activities, the study findings have indicated that researchers experience data rescue constraints with regard to time, costs, human resources, and skills. Moreover, the RGLs also referred to the dangers of researchers having their own research practices, researchers' data management activities displaying sub-standard qualities, and research groups not having a strategy when it comes to the rescue of at-risk data. To overcome many of these obstacles, it is recommended that an institute's LIS section be involved with data rescue projects. The roles and responsibilities listed by Patterton (2023a) can be used for guidance. Based on the suggested roles, an institute's LIS section may potentially be involved in all data rescue stages, with the section's participation including data rescue supervisory activities and contributing towards data rescue training material.

The involvement of an institute's LIS section is important. Suggestions for the section's participation in data rescue include the following:

- **Training of the LIS sector:** For the LIS sector (both at the selected institute and wider) to be optimally involved with data rescue, it is crucial to make provision for on-the-job training, to attend data rescue and data curation workshops, to engage in self-training by means of published data rescue manuals and guidelines, and for active LIS professionals to share their data rescue knowledge, skills, and experience with the LIS community. Examples of published manuals or guidelines that are available include the paper-based climate data rescue manual of the World Meteorological Organization (WMO) (2016) and the elaborate digital marine species data rescue manual of Kennedy (2017).
- **Institute-wide view:** Researchers focus on the project, and perhaps on its contribution to a discipline. LIS specialists support and enable a variety of these research projects across the organisation. This makes it possible to share best practice and knowledge, and assists with the identification of assets across the organisation. LIS specialists are also regularly in contact with a variety of staff members. They have someone to talk to about problems and can express concerns regarding neglected collections. These collections could be evaluated for their value in rescuing data.

- **LIS metadata expertise:** The survey results reported by Thompson, Davenport Robertson, and Greenberg (2014) indicate that there is a general institutional expectation that a librarian, archivist, or information professional should be responsible for the generation of metadata. Their findings also indicate that LIS professionals are indeed the parties that add metadata to data in the overwhelming majority of instances. The current study's RGL feedback and scholarly publications emphasise the significance of metadata. Coupling this with the metadata expertise and experience of LIS professionals indicates the importance of including the LIS sector in data rescue activities, especially during the stage concerned with metadata creation.
- **LIS curriculum change:** In North America, many universities have developed graduate programmes to prepare information professionals for data curation (Keralis 2021). Including data rescue and data curation modules in the LIS curriculum is an important step in equipping the LIS workforce. These curriculum inclusions, ideally occurring at postgraduate level where students already have some degree of data familiarity, will contribute to a better-equipped LIS workforce and LIS professionals who are able to participate in data rescue upon entering the workforce.
- **Sharing of LIS data rescue expertise:** Adding to the point about LIS data rescue training is the recommendation for sharing data rescue expertise. LIS workshops, conferences, and symposia, particularly those involving academics, researchers, and special libraries, should feature sessions on at-risk data and data rescue. Handover activities, such as the imparting of data rescue knowledge and skills preceding retirement and resignation, is another form of expertise sharing.

The involvement of library and information services in data rescue can involve all members of the sector: undergraduate technicians or assistants, semi-professional staff members, interns, research library professionals, and high-level managerial staff. Additionally, LIS professionals with data management experience can contribute to the areas of training, the creation of guidelines, the adaptation of the model, data rescue project management, and data management plan training and quality control.

Ideally, data rescue should become a dedicated LIS responsibility at research institutes, with the research sector associating assistance regarding at-risk data and data rescue with the institute's LIS sector.

8.7 CONSIDER COLLABORATIVE RESEARCH PROJECTS

RGLs have commented on the fact that data rescue can be costly, time-consuming, labour-intensive, and an activity requiring specialised skills. While a previous recommendation (see Section 8.6) has referred to the beneficial involvement of the LIS sector, it is also suggested that data rescue projects consider collaborating with additional partners, including retired research experts, citizen scientists, and postgraduate students. Moreover, data rescue partnerships with entities, including science councils and departments in tertiary institutes, are additional collaboration opportunities.

Data rescue projects ideally suited to collaborative ventures are those where the envisaged project is associated with challenges regarding rescue funding, the available human resources, the available rescue skills, available insight into the research discipline, and equipment to read the data.

8.8 INVESTIGATE DEDICATED INSTITUTIONAL DATA REPOSITORY OPTIONS

Institutes without a dedicated institutional data repository are likely to find themselves in a position where research data are stored in a range of locations, on different devices, and not necessarily accompanied by metadata or data documentation. The recommendation is therefore made that these research institutes investigate the possibility of acquiring a dedicated institutional data repository to be used for the preservation and sharing of institutional research data by default.

Considerations linked to this suggestion involve the data repository meeting an institute's unique requirements, such as varying data access levels, the ability to accommodate a range of data formats, a handle assigned to the data for sharing and citation purposes, or integration with established institutional systems.

While older at-risk data are unlikely to end up in a data repository, unless it forms part of a rescue project, the future access and use of current and future data will undoubtedly be enhanced by a dedicated data repository. The availability of a dedicated institutional data repository may coincide with beneficial outcomes, including knowing where data are stored, the creation of metadata and data documentation, the assurance of data backups, easier handover, limiting issues with the current DMS, easier compliance with the data management procedure, and alleviating the use of sub-standard and non-secure storage options.

8.9 ENSURE HANDOVER CONSIDERS RELEVANT DATA

The research findings have shown that RGLs are concerned about data not being accessible to members of the research group, poor data management continuation, and group members being unaware that data are at risk. To deal with these factors, it is recommended that institutional handover activities consider the potentially detrimental effects any transitions will have on any valuable research data and data knowledge. Institutional handover activities and change management strategies also need to ensure that there is a record of the affected data, that several parties are aware of the data, that the data can be accessed, and that the data can be understood and used.

Changes in strategic outlook and the restructuring of research groups, coupled with resignations and retirements, form part of the research work environment. When not managed well, with detrimental outcomes not foreseen, these events can affect the future access and usability of institutional data.

9. CONCLUSION

The overriding goal of the study was to gain insight into the nature of 'at-risk data' and data rescue. The research intended to identify factors leading to data being at risk, identify risk factors unique to the institute, identify forces driving the risk factors, identify data rescue challenges, and propose steps to alleviate risk factors and challenges.

A range of factors has led to the selected institute's data becoming at risk. These factors are often identical to the risk factors found via scholarly publications. Despite similarities, the study revealed additional novel risk factors (not found in scholarly literature), including the lack of backups, inferior infrastructure (and systems and processes), troubling data volumes, and difficulty in getting systems approved. Although research groups had analogue at-risk data, the emphasis on factors putting digital data at risk reveals the current priorities and concerns of the interviewed RGLs.

Data rescue activities were not common at the selected institute, and it is likely that a similar trend will manifest at institutes that reveal similar stages of data management maturity. Data rescue challenges experienced by researchers are plentiful and diverse in nature, with the interviewed RGLs revealing the obstacles that emanate from institutional challenges, equipment challenges, resource challenges, and challenges related to sub-standard data management practices.

Addressing the occurrence of at-risk data and data rescue obstacles in the research environment is a complicated task that requires a multi-faceted approach. While not all risk factors and challenges can be minimised, this study has suggested a combined strategy to deal with the topic at a research institute. It therefore recommends measures to alleviate this challenge, such as awareness training, the monitoring of data management, the roll-out of a data rescue pilot project, the significant involvement of an institute's LIS section in data rescue, and expanding handover activities to include data concerns.

Notwithstanding the fact that recommendations are based on institute-specific findings, it is anticipated that research institutes and tertiary entities elsewhere may benefit from learning about researcher responses to at-risk data and data rescue and the study's suggestions for dealing with the indicated risk factors and obstacles. Recommendations are expected to be important in environments where the occurrence of at-risk data is suspected but not yet addressed via mitigation measures.

The study succeeded in addressing all of the posed research challenges. The brief takeaway message stemming from this study is that data rescue is a rare research phenomenon. Data

risk factors are real, universal and troubling, and data rescue challenges are worthy of attention. The good news is that relevant and directed interventions can limit factors that put data at risk, and minimise the obstacles experienced by researchers when rescuing data.

DATA ACCESSIBILITY STATEMENT

The data presented in this publication are published in the Figshare repository (Patterton LH, 2023b and Patterton, LH, 2023c).

ADDITIONAL FILES

The additional files for this article can be found as follows:

- **Appendix 1.** Web Questionnaire Questions. DOI: https://doi.org/10.5334/dsj-2024-011.s1
- **Appendix 2.** Interview Schedule. DOI: https://doi.org/10.5334/dsj-2024-011.s2

ACKNOWLEDGEMENTS

We would like to thank all research participants for their time and for the open and frank discussions.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

The corresponding author made substantial contributions to the paper as the paper is an output from her PhD studies. The research was supervised by the co-authors. The authors collaborated closely throughout this study and agreed to be on the author list.

AUTHOR AFFILIATIONS

Louise H. Patterton orcid.org/0000-0002-8067-8545
Department of Information Science, University of Pretoria, Pretoria, South Africa; Council for Scientific and Industrial Research, South Africa

Theo J. D. Bothma orcid.org/0000-0001-7850-3263
Department of Information Science, University of Pretoria, Pretoria, South Africa

Martie J. van Deventer orcid.org/0000-0002-9776-1177
Department of Information Science, University of Pretoria, Pretoria, South Africa

REFERENCES

Adams, KA and **Lawrence, EK.** 2015. *Research Methods: Statistics, and Applications*. Thousand Oaks, CA: Sage.

Akba Yrak, B. 2000. A comparison of two data collecting methods: Interviews and questionnaires. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 18: 1–10. Available at http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/1051-published.pdf [Last accessed 8 February 2024].

Arrouays, D, et al. 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. *GeoResJ*, 14: 119. DOI: https://doi.org/10.1016/j.grj.2017.06.001

Bradshaw, E, Rickards, L and **Aarup, T.** 2015. Sea level data archaeology and the Global Sea Level Observing System (GLOSS). *GeoResJ*, 6: 9–16. DOI: https://doi.org/10.1016/j.grj.2015.02.005

Brunet, M and **Jones, P.** 2011. Data rescue initiatives: Bringing historical climate data into the 21st century. *Climate Research*, 47(1–2): 29–40. DOI: https://doi.org/10.3354/cr00960

DataFirst. 2022. *DataFirst and the rescue of data at risk: Partnering to enable research access to historical African data*. Available at https://www.datafirst.uct.ac.za/services/data-rescue?highlight=WyJyZXNjdWVkIl0= [Last accessed 17 November 2023].

Digital Curation Centre. 2023. *Funders' data plan requirements*. Available at https://www.dcc.ac.uk/resources/data-management-plans/funders-requirements [Last accessed 12 November 2023].

Dissertation Team. 2024. *Limitations of a Study: The Complete Guide*. Available at https://us.dissertationteam.com/blog/limitations-of-a-study/ [Last accessed 20 November 2024].

Diviacco, P, et al. 2015. Data rescue to extend the value of vintage seismic data: The OGS-SNAP experience. *GeoResJ*, 6: 4452. DOI: https://doi.org/10.1016/j.grj.2015.01.006

Downs, RR. 2015. Data rescue at a scientific data center. In: *Best Practices Exchange 2015*, Harrisburg, Pennsylvania, 20 October 2015. Available at https://bpexchange.files.wordpress.com/2017/01/bpe2015_20151020_downs_datarescuescidatactr.pdf [Last accessed 18 November 2023].

Downs, RR and **Chen, RS.** 2017. Curation of scientific data at risk of loss. Data rescue and dissemination. In: Johnston, LR (ed.), *Curating Research Data: Practical Strategies for your Digital Repository*, pp. 263–277. Available at https://academiccommons.columbia.edu/doi/10.7916/D8W09BMQ [Last accessed 18 November 2023].

Esurv. 2019. *Make online surveys for free!* Available at https://esurv.org/ [Last accessed 8 December 2019].

Gallaher, D, et al. 2015. The process of bringing dark data to light: The rescue of the early Nimbus satellite data. *GeoResJ*, 6: 124134. DOI: https://doi.org/10.1016/j.grj.2015.02.013

Gaudin, S. 2017. *Coders and librarians team up to save scientific data*. ComputerWorld, 20 March. Available at https://www.computerworld.com/article/3182384/data-storage/coders-and-librarians-team-up-to-save-scientific-data.html [Last accessed 17 November 2023].

Griffin, RE. 2015. When are old data new data? *GeoResJ*, 6: 92–97. DOI: https://doi.org/10.1016/j.grj.2015.02.004

Guest Blogger. 2017. *How data refuge works, and how YOU can help save federal open data.* Sunlight Foundation, 6 February. Available at https://sunlightfoundation.com/%0b2017/02/06/how-data-refuge-works-and-how-you-can-help-save-federal-open-data/ [Last accessed 17 November 2023].

Hachileka, E. 2015. *Why we need to save Africa's historical climate data.* United Nations Development Programme, 14 October. Available at https://undp-cirda.blogspot.com/2015/11/why-we-need-to-save-africas-historical.html [Last accessed 16 November 2023].

Hanna, P. 2012. Using internet technologies (such as Skype) as a research medium: A research note. *Qualitative Research*, 12(2): 239–242. DOI: https://doi.org/10.1177/1468794111426607

Jowett, A. 2020. *Carrying out qualitative research under lockdown – Practical and ethical considerations*. London School of Economics and Political Science, 20 April. Available at https://blogs.lse.ac.uk/impactofsocialsciences/2020/04/20/carrying-out-qualitative-researchunder-lockdown-practical-and-ethical-considerations/ [Last accessed 16 November 2023].

Kennedy, M. 2017. *Guidelines for marine species occurrence data rescue – The OBIS Canada Cookbook*. COINAtlantic. Available at https://www.coinatlantic.ca/_files/ugd/cf2ff9_82e7008749294b83b31c4a8a9ecd99cf.pdf [Last accessed 17 November 2023].

Keralis, SDC. 2021. Data curation education: A snapshot. In: *The Problem of Data,* pp. 32–43. Available at https://www.clir.org/wp-content/uploads/sites/6/pub154.pdf [Last accessed 7 February 2024].

Krouwel, MJ, Jolly, K and **Greenfield, S.** 2019. Comparing Skype (video calling) and in-person qualitative interview modes in a study of people with irritable bowel syndrome – an exploratory comparative analysis. *BMC Medical Research Methodology*, 19(1): 1–9. DOI: https://doi.org/10.1186/s12874-019-0867-9

Leedy, PD and **Ormrod, JE.** 2005. *Practical Research: Planning and Design*. Upper Saddle River, NJ: Pearson.

Levitus, S. 2012. The UNESCO-IOC-IODE "Global Oceanographic Data Archaeology and Rescue" (GODAR) Project and "World Ocean Database" Project. *Data Science Journal*, 11: 46–71. DOI: https://doi.org/10.2481/dsj.012-014

Mavraki, D, et al. 2016. Rescuing biogeographic legacy data: The "Thor" expedition, a historical oceanographic expedition to the Mediterranean Sea. *Biodiversity Data Journal*, 4: e11054. DOI: https://doi.org/10.3897/BDJ.4.e11054

Mayernik, MS, et al. 2020. Risk assessment for scientific data. *Data Science Journal*, 19(1): 15. DOI: https://doi.org/10.5334/dsj-2020-010

Menon, V and **Muraleedharan, A.** 2020. Internet-based surveys: Relevance, methodological considerations and troubleshooting strategies. *General Psychiatry*, 33: e100264. DOI: https://doi.org/10.1136/gpsych-2020-100264

Muller, C. 2015. Rescuing early digital assets and preserving data rescue capabilities. In: B*est Practices Exchange, Pennsylvania State Archives*, Harrisburg, USA, 19–21 October 2015. SlideShare. Available at https://www.slideshare.net/ctm0608/data-rescue-and-preserving-dr-capabilities?qid=460c012e-76e2-4e1f-b142-7292775eb3b4&v=&b=&from_search=1 [Last accessed 17 November 2023].

Munch, J. 2020. *How to discuss your study's limitations effectively*. Available at https://www3.mdanderson.org/library/about/pdf/write-stuff/spring-2020.pdf [Last accessed 20 February 2024].

Murillo, AP. 2014. Data at risk initiative: Examining and facilitating the scientific process in relation to endangered data. *Data Science Journal*, 12: 207219. DOI: https://doi.org/10.2481/dsj.12-048

Neuman, WL. 2014. *Social Research Methods: Qualitative and Quantitative Approaches*. Essex: Pearson Education.

Nordling, L. 2010. Researchers launch hunt for endangered data. *Nature*, 468: 17. DOI: https://doi.org/10.1038/468017a

Broadening the scope of IT factors to include technology issues such as software obsolescence, incompatible readers, sub-standard media, failure to read the data, and data being in an old format shows the similarities between the concerns of RGLs and the factors stated by Levitus (2012), Murillo (2014), Hachileka (2015), Muller (2015), and Wyborn et al. (2015).

Similarities in data risk factors linked to researcher behaviour were detected, with the interviewed RGLs echoing the published risk factors of the loss of skills (Wyborn et al. 2015), non-secure storage (Murillo 2015), inferior data management practices (Levitus 2012), not being aware of the value of data (Griffin 2015; Nordling 2010), and lack of metadata (Murillo 2014; Thompson, Davenport Robertson & Greenberg 2014; Wyborn et al. 2015). However, respondents also stated lack of backups, lack of strategy, and data not being accessible to others as being detrimental to the survival of their data.

Other factors uncovered through the interviews and not widely published in scholarly literature are mostly institute-specific and include the massive data volumes generated in certain research groups, insufficient data storage space or storage media, and the general acceptance of non-secure data storage options. One of the respondents mentioned the danger of running out of storage media for valuable, sensitive physical samples during the hard COVID-19 lockdown and mentioned the resultant financial implications should the media have been exhausted. Other risk factors linked to the specific institute include not having a backup device while working remotely, incompatible research protocols, and equipment displaying non-compatibility issues.

Several respondents stated that lack of time, lack of funding, and insufficient human resources can result in data being at risk. This appears to be a factor of particular concern at this institute and is most likely related to the institute's recent restructuring activities and the freezing of non-vital positions.

Another institute-specific risk factor (which may also fall under ICT-related factors): The lack of a formal data management platform or a dedicated data repository, pointed to inferior data storage infrastructure at the time of data collection. Institutional-wide data management planning templates were also not yet available. Equipment malfunction and the potential resultant loss of data were also mentioned as risk factors, as were problematic issues with the institute's DMS. The previous DMS (instituted before 2019) was linked to the corruption of records, an inferior search facility and the inaccessibility of many items. Respondents described the current DMS (implemented since 2019, and in use at the time of this study's data collection) as being slow and not being user-friendly.

With several institute-specific factors not featuring strongly within the literature reviewed for this research, it is prudent to examine the reasons for this institute's researchers mentioning these risk factors. It is likely that the common denominators in many of the risk factors amount to the institute's lack of suitable infrastructure, systems, platforms and servers. Lack of management support and buy-in are suspected to be related to the stated issues. The absence of the required data management foundations, coupled with insufficiencies in systems and storage options, have potential dire consequences for valuable data, and are addressed in the recommendations section.

Implications emanating from this institute's at-risk data factors are numerous. The most important ramification is that the factors leading to data being susceptible to loss should be acknowledged and addressed. Most of the factors causing data to be in a threatened state can potentially be reduced or even eliminated. These factors comprise a lack of a data management or rescue strategy, a lack of data management or rescue processes, or making use of sub-standard data storage. The stated factors should be attended to, with possible actions entailing the inclusion of 'at-risk data' in data management training sessions, and ensuring that 'at-risk data' forms part of data management awareness material and guidelines.

The link between an institutional data management policy, data management procedures and the resultant data being at risk is obvious. Additionally, many of the behavioural factors that lead to data being at risk are inversely proportional to good research practices and ethical research behaviour. Of particular importance is the creation of a DMP at the start of every research project. A DMP is expected to stipulate data management best practices and detail the envisaged data storage locations, data backup practices, access arrangements, metadata activities, and long-term preservation strategy. Adhering to the DMP in conjunction with considering the project's risk management plan are vital steps in minimising future at-risk data.