UNIVERSITY OF OULU

Challenges and Requirements of Heterogenous Research Data Management in Environmental Sciences: A Qualitative Study

University of Oulu
Information Processing Science
Master’s Thesis
Muhammad Hamza Javed
2023

Abstract

The research focuses on the challenges and requirements of heterogeneous research data management in environmental sciences. Environmental research involves diverse data types, and effective management and integration of these data sets are crucial in managing heterogeneous environmental research data. The issue at hand is the lack of specific guidance on how to select and plan an appropriate data management practice to address the challenges of handling and integrating diverse data types in environmental research. The objective of the research is to identify the issues associated with the current data storage approach in research data management and determine the requirements for an appropriate system to address these challenges. The research adopts a qualitative approach, utilizing semi-structured interviews to collect data. Content analysis is employed to analyze the gathered data and identify relevant issues and requirements.

The study reveals various issues in the current data management process, including inconsistencies in data treatment, the risk of unintentional data deletion, loss of knowledge due to staff turnover, lack of guidelines, and data scattered across multiple locations. The requirements identified through interviews emphasize the need for a data management system that integrates automation, open access, centralized storage, online electronic lab notes, systematic data management, secure repositories, reduced hardware storage, and version control with metadata support. The research identifies the current challenges faced by researchers in heterogeneous data management and compiles a list of requirements for an effective solution. The findings contribute to existing knowledge on research-related problems and provide a foundation for developing tailored solutions to meet the specific needs of researchers in environmental sciences.

Keywords
Environmental research data, data storage, data tracking, cloud-based services, data management, hybrid data storage, and research challenges.

Supervisors
PhD, University Lecturer, Elina Annanperä
MSc, Doctoral Researcher, Henri Bomström

Foreword

I am grateful for the unwavering support and guidance I received throughout the thesis writing process. I extend special thanks to my advisors, Elina Annanperä, and Henri Bomström, who provided valuable assistance in formulating the initial thesis outline, enabling me to write with clarity. Their meticulous review of the draft and insightful comments greatly enhanced the quality of the thesis.

I would also like to express my heartfelt appreciation to my family and friends for their continuous encouragement and unwavering support, which played a vital role in helping me complete this thesis.

Additionally, I extend my thanks to all the authors whose works I have cited in this thesis. Your contributions serve as a beacon of wisdom, guiding me in the right direction.

Muhammad Hamza Javed
Oulu, June 1st, 2023

Contents

1. Introduction

Engineering sciences comprise numerous sub-disciplines with a wide variety of processes and methods that generate, process, and analyze diverse research data (Sandfeld et al., 2018). With advancements in positioning and sensor technologies, scientists now have increased access to large amounts of environmental data. Environmental data are heterogeneous in source and format and are typically collected at different spatiotemporal scales. To address the challenges posed by such diverse data, Research Data Management (RDM) is necessary. RDM refers to the planning, organization, storage, preservation, sharing, and dissemination of research data throughout the research lifecycle, from data collection to analysis and publication (Cox et al., 2017). Heterogeneous data refers to data that come from different sources, formats, structures, and characteristics, making integration and analysis difficult (Vetova, 2021). Proper storage of heterogeneous research data is critical as the amount of data generated by various fields continues to grow rapidly. The increasing complexity of data makes storage, processing, and analysis more challenging (Lima, 2020).

The main objective of implementing RDM practices and workflows highlighted by Wilkinson et al. (2016) is to ensure that data is organized and managed in a way that is easily discoverable, accessible, interoperable, and reusable, as outlined. Among the various stages involved in research data management, data storage is a crucial component, as highlighted by (Cox, 2017). It is essential to recognize the importance of effective data handling in research data management plans. Many research institutions such as the University of Oulu, provide only general recommendations called data management plans (DMPs) that do not provide specific guidance on how to choose a data management system that will meet there needs.Consequently,leaving researchers open to face challenges like identifying the specific requirements of their heterogeneous data and may struggle to manage it effectively over the long term.

Existing solutions for managing heterogeneous research data in environmental sciences have proven to be inadequate in addressing the specific challenges of data integration, organization, and accessibility (Nass et al., 2022). These solutions often lack the necessary features to handle diverse data types effectively, resulting in inconsistencies, scattered data, and limited interoperability (Nass et al., 2022). Solutions, for example, Zheng Yan (2019) Cloud Service Providers (CSPs), Da Silva et al. (2012) proposed the meta database extractor, and Pamart et al. (2022) introduced a metadata-enriched system in the context of a multi-modal digital imaging survey do not address all the issues rather focus on just on part of the problem like either they focus on just storage of data or on metadata which does not meet the requirements of interview participants.

It is important to note that this study does not aim to provide a ready-made solution for managing environmental research data. Instead, it focuses on specific aspects of research data management, specifically data handling. By addressing the research questions of this thesis, this thesis provides valuable insights and recommendations to improve research data management practices in environmental research. Ultimately, this can lead to more reliable and effective research outcomes. Further work may be needed to develop a comprehensive data management solution that can address all aspects of research data management for different types of research data.

Therefore, there is a critical need to identify best practices for managing research data and provide researchers with guidelines on how to implement these practices in a manner that is tailored to their unique needs and situations. This is particularly crucial for ensuring that research data is preserved, shared, and reused effectively, which can enhance scientific discovery and contribute to the advancement of knowledge. The purpose of this master's thesis is to find different challenges of data storage in RDM. The thesis aims to identify the issues associated with the current data storage approach and determine the requirements for desired solutions that researchers can utilize to address these issues to store their data, track their data, and share their data with the team based on data gathered in the interviews.

1.1 Research questions

The research questions for the thesis are as follows.

- RQ1: What are the key challenges that researchers face when dealing with environmental research data storage and data tracking?
- RQ2: How can the challenges in data storage and tracking of environmental research data be addressed effectively?

The first question seeks to identify the specific challenges that researchers encounter when dealing with environmental research data storage and data tracking. These challenges may include issues such as managing large data volumes, handling diverse data types and formats, ensuring data quality and integrity, addressing data security and privacy concerns, and coping with limited storage capacity. The second question aims to determine how challenges in environmental research data management can be addressed. The research aims to identify the needs and preferences of researchers, as expressed through the interview responses, to inform the development of a solution that effectively addresses the identified challenges in data storage and meets researchers' requirements. By answering these research questions, the thesis will provide insights and recommendations for improving research data management practices in the field of environmental research. The thesis employs a qualitative research methodology, utilizing semi-structured interviews as the primary data collection method. In addition, content analysis is applied to examine the gathered data, allowing for the identification of pertinent issues and requirements related to the desired solution.

The thesis is organized into six sections. The first section introduces the topic of the thesis. In the second section, background information and related work on research data management is provided. The second section also presents an overview of previous research in this area. The third section describes the research methods used in this study section four unveils the key findings obtained from the interviews and primarily focuses on the significant results. The section five, provides a discussion of the thesis, highlighting the main findings and their implication. Finally, section six concludes the thesis by summarizing the key points and providing recommendations for future research in this area.

2. Background

This section provides a concise overview of the university's DMP followed by an exploration of RDM, and related studies made in the realm of research data storage and organization.

2.1 Data management plan of the University of Oulu

The DMP provides details about what different steps researchers must do for their data but does not give any information on how such steps can be carried out which leads to different researchers storing data at different places and having no systematic way of managing data. The DMP serves as a comprehensive guide for effectively managing the data associated with the thesis project. The paper by Jones et al. (2018) emphasizes effective data management practices, including organization, backup, and documentation. The main findings highlight the importance of DMPs in ensuring consistent and transparent data management, compliance with funder requirements, and facilitating data sharing and preservation.

The importance of effective data management practices is emphasized, including the need for proper organization, backup, and documentation of research data. DMPs are identified as formal documents outlining data collection, storage, sharing, and preservation during research projects, ensuring consistency, transparency, and compliance with funder requirements (Jones et al., 2018). It encompasses several key aspects that contribute to maintaining control over the research data and ensuring its integrity (*LibGuides: Research Data Guide: Data Management Plan*, 2021).

- Description of the data: In this section, the DMP outlines the type of research data being collected. It includes details such as the nature of the data, whether it is original data or obtained from shared sources, and any specific file formats or specialized programs used for data processing. This description provides an overview of the data sources and formats involved in the research, enabling a better understanding of the data's characteristics and requirements.

- Ethical and other issues to be considered in the collection of data: Ethical considerations are crucial when handling research data, particularly if it contains personal or sensitive information. It contains information about agreements or collaborations with external parties, as it should be documented to ensure compliance and data security.

- Data documentation and metadata: This section focuses on how the research data is documented and organized. It outlines the principles for naming files, the existence of a list of abbreviations used in the data, version control measures, and timestamps indicating when the data was collected. It may also mention the location where the data was collected, and any specific equipment used for data collection. Anonymization procedures, if applied, should be described. Furthermore, the DMP assures the reader that there is a document or repository where detailed information about these aspects is recorded, ensuring transparency and traceability.

- Storage of research data: Data storage is a critical consideration to ensure the security and accessibility of research data. The DMP addresses the measures taken to store the data in a safe and secure location. It may mention the use of backup systems or data redundancy strategies to protect against data loss or corruption. By documenting data storage practices, researchers can demonstrate their commitment to data preservation and availability for future reference or analysis.

- Other matters related to the data: This section has information about access control and data-sharing policies. It clarifies who has authorized access to the data and under what conditions. This can include specifying whether the data is restricted to the research team, accessible to collaborators, or available to the public. Addressing data access ensures data confidentiality, intellectual property protection, and compliance with any legal or institutional requirements.

DMPs are aimed to provide a framework for researchers, facilitating organized and efficient data handling, storage, documentation, sharing, and preservation. The DMP ensures a clear understanding of the data, safeguards ethical considerations, facilitates documentation and metadata management, provides secure data storage, and outlines data access policies. Through careful attention to these aspects, researchers can enhance the overall quality, integrity, and usability of the research data, contributing to a more robust and transparent research process.

2.2 Research data management

The focus of RDM is on methods, best practices, and infrastructure implementations for archiving, publishing, and accessing research data (*Research Data Management*, 2023). Research data are valuable information assets that are generated or gathered by scientific methods. However, datasets are potentially fragile, susceptible to storage failure and obsolete technology, and may also be sensitive, which had private information for example, so it needs to be managed with the correct steps of security (Pinfield et al., 2014). The absence of activity for centralized data management and data storage within research institutions has the potential to make valuable data disappear. This situation can cause dark data to spread in the research institute. Dark data is data that cannot be indexed and carefully stored, so it makes data almost invisible to researcher and other potential users, because of that the data tend to be least utilized and disappeared at the end (Heidorn, 2008). That is why the activities of centralized research data management are essential to implement in each research institution.

According to Lau et al. (2021), RDM and DMP are two ideas in the field of research data management that are closely related. They are different in terms of their scope and purpose, but they all aim to manage research data properly. The thorough process of managing research data over the whole study lifecycle is known as research data management or RDM. Data management, storage, documenting, sharing, and preservation are just a few of the different tasks involved. RDM makes sure that research data is gathered, examined, saved, communicated, and archived in a structured and orderly manner. Through the course of the study process, it emphasizes the significance of maintaining data integrity, accessibility, and reusability (Lau et al., 2021).

Whereas a DMP is a detailed document that highlights the different steps using which the data will be handled by the researchers. DMPs are made at the start of any research project and may need revision in later stages if deemed necessary. DMP offers a thorough road map for managing research data, including details on data gathering techniques, backup and storage systems, data sharing arrangements, and data preservation tactics.

Researchers, funding organizations, and other stakeholders can use it as a reference guide to make sure that data management procedures adhere to best practices and any applicable laws or policies (Lau et al., 2021).

In essence, RDM represents the broader concept of managing research data, encompassing all aspects of data management, while a DMP is a tangible and specific document that outlines the strategies and procedures for data management within a research project. Both RDM and DMP are crucial components in ensuring the quality, accessibility, and long-term value of research data. By implementing effective RDM practices and creating well-structured DMPs, researchers can optimize their data management processes and contribute to the overall advancement of knowledge in their respective fields.

Surkis and Read (2015) research data lifecycle can be seen Figure. 1. The lifecycle consists of six phases. The first three phases in the lifecycle are creating data, gathering data, and processing data from a raw format into a useable form that can be analyzed, while the last three phases Preserving data, giving access to data, and reusing data focusing on data storage and findability of data.

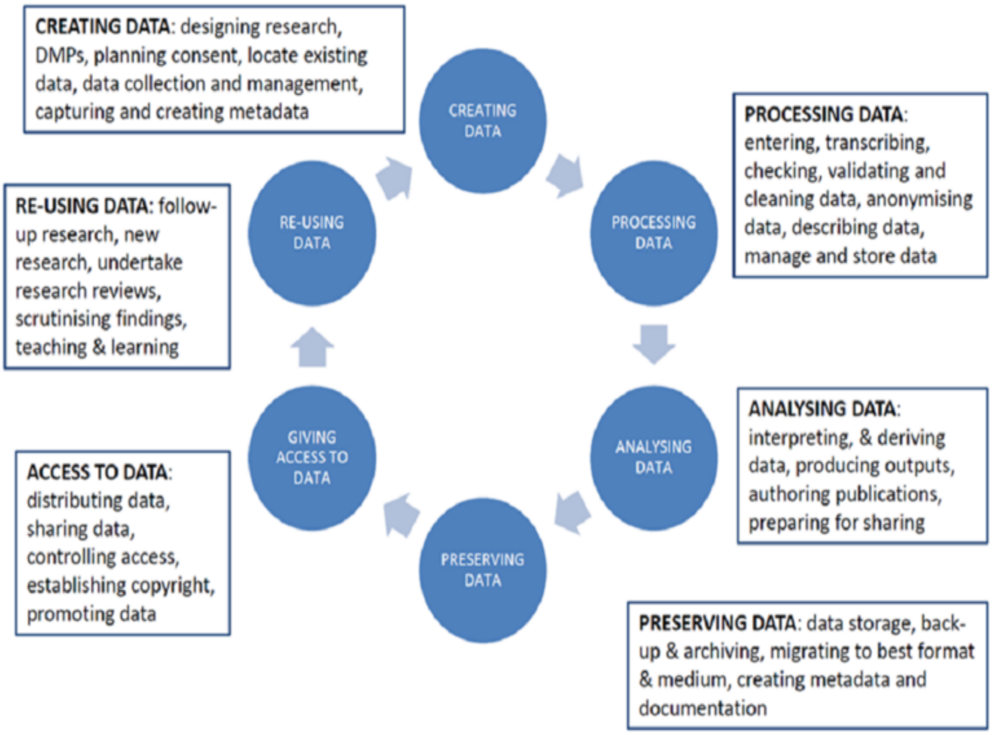
CREATING DATA: designing research, DMPs, planning consent, locate existing data, data collection and management, capturing and creating metadata
CREATING DATA
PROCESSING DATA: entering, transcribing, checking, validating and cleaning data, anonymising data, describing data, manage and store data
RE-USING DATA: follow-up research, new research, undertake research reviews, scrutinising findings, teaching & learning
RE-USING DATA
PROCESSING DATA
ANALYSING DATA: interpreting, & deriving data, producing outputs, authoring publications, preparing for sharing
GIVING ACCESS TO DATA
ANALYSING DATA
ACCESS TO DATA: distributing data, sharing data, controlling access, establishing copyright, promoting data
PRESERVING DATA
PRESERVING DATA: data storage, back-up & archiving, migrating to best format & medium, creating metadata and documentation

Figure 1. Research data life cycle source consisting of six-stage (Surkis & Read 2015).

If the data can be understood, it can be reused by another researcher that will use it for a validity test from the original results or analyze it again from the original data with a different method. Whereas the last three steps of the data lifecycle include maintaining data after research has been completed, giving access from data to another person, and finally reusing the data to do new research or to check reproducibility from original results.

Six-stage model for RDM was proposed by Surkis and Read (2015) to guide researchers in effectively managing their data throughout the research lifecycle. The model outlines the key stages involved in handling research data from planning to long-term preservation.

1. Planning: In this stage, researchers identify the types of data that will be collected during the research project. They determine how the data will be collected, organized, and documented. This includes defining data variables, measurement instruments, and data collection methods. Planning also involves outlining data ownership, access permissions, and data-sharing policies.
2. Collecting: The collecting stage involves the actual gathering of data according to the plan developed in the planning stage. This may involve conducting experiments, surveys, observations, or acquiring existing datasets. Researchers must ensure that the data collected is comprehensive, accurate, and properly documented to maintain data integrity.
3. Processing: During the processing stage, researchers clean and transform the collected data to ensure its quality and usability. This involves removing inconsistencies, errors, and outliers. Data may also be transformed or normalized to facilitate analysis. Researchers may use various techniques and tools for data cleaning and transformation, such as statistical software or scripting languages.
4. Storing: The storing stage focuses on securely and efficiently storing the data for easy retrieval and future use. Researchers should consider using appropriate storage solutions that meet their data storage needs, including security, accessibility, and scalability. This may involve using local servers, cloud storage, or institutional data repositories. Adequate backup and disaster recovery plans should also be in place to prevent data loss.
5. Sharing: The sharing stage involves making the research data available to others, which promotes transparency, reproducibility, and collaboration. Researchers may choose to share their data through various means, such as publishing it alongside their research findings or depositing it in data repositories. Proper documentation, data descriptions, and metadata should accompany the shared data to enhance its discoverability and understanding.
6. Preservation: The preservation stage focuses on ensuring the long-term accessibility and usability of the research data. As technology evolves, data formats and platforms may become obsolete. Therefore, researchers need to plan for data preservation by migrating data to newer formats or platforms, ensuring ongoing data integrity, and assigning persistent identifiers. Long-term preservation also involves considering legal and ethical aspects, such as data privacy and intellectual property rights.

The FAIR data principles, introduced by Wilkinson et al. (2016) and Mons et al. (2017), offer recommendations for producing data that are discoverable, accessible, interoperable, and reusable. These principles are designed to enhance the value and impact of research data by promoting effective data management practices.

The first principle, findability, emphasizes the importance of making data easily discoverable. To achieve this, researchers should assign persistent identifiers to their data, provide detailed metadata, and ensure that their data is indexed in appropriate repositories or catalogs. By following these practices, researchers enable others to locate and access their data with ease. The second principle, accessibility, focuses on ensuring that data is openly available for access and download. Researchers should establish clear and standardized access protocols, including open licenses or data usage agreements. By removing unnecessary restrictions, researchers facilitate broader access to their data, promoting transparency and enabling collaboration (Wilkinson et al., 2016; Mons et al., 2017).

The FAIR data principles are intended to offer recommendations for producing discoverable, accessible, interoperable, and reusable data (Wilkinson et al., 2016; Mons et al., 2017). FAIR data principles can be seen in the following Table 1.

Table 1. Summary of FAIR Data Principles (Wilkinson et al., 2016; Mons et al., 2017).

FAIR Principle	**For Infrastructure**	**For Researchers**
Findable	Both computers and people should be able to quickly find data and metadata. This means that in practice, data must be given distinct and persistent IDs, described using rich metadata, and registered or indexed in a searchable resource.	To ensure efficient data retrieval, research teams should implement standardized file organization practices, including defined file naming conventions. When sharing data publicly, it is advisable to upload it to a repository that provides persistent identifiers (e.g., DOIs, RRIDs) and utilizes standardized metadata to describe the datasets. Including comprehensive and high-quality metadata is essential for facilitating data discovery and establishing connections with related resources, such as article DOIs and author ORCIDs.
Accessible	The process of getting access to the data is well-defined. By employing a defined protocol that is open, free, and implementable by everyone, data should be retrievable by its identity. Even if the data is gone, the metadata should still be available.	A carefully defined process makes data available. Raw data, intermediary products, and other research resources ought to be accessible to the research team.

Table 1. continuation of Table 1

FAIR Principle	For Infrastructure	For Researchers
Interoperable	Data should be accessible to a variety of workflows and applications. For the representation of knowledge in data, formal, open, shareable, and widely applicable models should be used.	To make it simple to merge data sets that are similarly structured, data should be organized consistently. In practice, this entails putting into action a variety of procedures, such as characterizing and organizing data (e.g., applying the proper metadata, maintaining data dictionaries), and saving files in open or nonproprietary file formats.
Reusable	Data and metadata should be represented according to industry standards and have machine-readable licenses and clearly stated terms for reuse.	Data should be kept, arranged, and described with future reuse in mind. Future users could include a member of the research team who revisits the data after a hiatus of several months or years or another researcher who reuses the data for a different project.

The third principle, interoperability, underscores the need for data to be structured and formatted in a way that allows integration with other datasets and tools. This involves using common data formats, adopting standardized vocabularies, and adhering to established data models and ontologies. Interoperable data enhance data integration, comparison, and analysis across different research domains. The fourth principle, reusability, emphasizes the importance of well-documented and sufficiently described data. Researchers should provide comprehensive metadata, clear and consistent data documentation, and information on data provenance and quality. By ensuring the reusability of their data, researchers enable future researchers to effectively understand, interpret, and build upon existing data (Wilkinson et al., 2016; Mons et al., 2017).

2.3 Heterogenous data management

The cited studies in this section shed light on various aspects of research data management, including research data storage, data management challenges, cloud repositories, and polyglot data stores. Each study offers unique insights and findings. these studies contribute to the understanding of research data management by providing insights into different storage solutions, challenges, and best practices. They offer valuable recommendations for researchers to ensure data reliability, security, and accessibility while addressing the specific needs and requirements of their research projects.

Heterogeneous data refers to data that come from different sources, formats, structures, and characteristics, making integration and analysis difficult (Vetova, 2021). Several studies have been conducted in the field of data management, addressing different aspects and contexts. These studies contribute to the understanding of data storage, access control, and management, as well as the challenges associated with data heterogeneity and interoperability. The following studies are particularly relevant to the research topic of this thesis.

Yang et al. (2019) conducted a study focusing on heterogeneous data storage management across multiple Cloud Service Providers (CSPs). The study aimed to develop a scheme that delivers both access control and deduplication management simultaneously. By proposing a flexible approach, Yang et al. (2019) addressed the challenges associated with data storage in diverse environments, contributing to the understanding of efficient data management across CSPs.

Willems et al. (2014) discussed the use of open-source software for research data management infrastructure. The study emphasized the adoption of open-source and open standards, providing insights and case studies to simplify the process of uploading and sharing primary research data. Willems highlighted the importance of making data management simpler and more appealing for researchers. The incorporation of open-source technologies demonstrated the potential of these tools in facilitating geographic data management (Willems et al., 2014).

Da Silva et al., (2012) proposed the metadata base extractor, which aimed to promote interoperability between relational databases and other information sources. Their study focused on creating a display structure of relational database metadata alongside Dublin Core elements and additional elements. This approach allowed for improved access to relational databases through OAI-PMH requests, enhancing the sharing and harvesting of metadata. Da Silva et al.'s work contributed to the field by facilitating the integration of relational databases with other information sources (Da Silva et al., 2012).

Data management challenges in the context of the Internet of Things (IoT) were explored by (Selvi & Sasirakha, 2021). They identified several key issues, including the lack of a systematic approach to manage data, inconsistencies in data treatment and management, the risk of accidental data deletion, loss of knowledge due to personnel turnover, and difficulties in tracking data changes over time. To address these challenges, they proposed a four-layer architecture for data management in the IoT, considering data heterogeneity, security, and privacy. Their study shed light on the importance of developing frameworks for efficient data management in the IoT, providing valuable insights for researchers and practitioners.

Metadata is defined as data that provides information about other data. It is used to describe and provide context for data, making it easier to find, access, and use. There are many different definitions of metadata, with 96 separate ISO standards providing 46 different definitions of the term (Furner, 2020). The studies underscore the significance of metadata for data preservation, reuse, and interoperability. They offer distinct approaches and tools for extracting, transforming, and managing metadata from diverse data sources. Furthermore, domain-specific metadata schemes and metadata catalogs are highlighted as valuable resources for enhanced documentation, organization, and sharing of research data. The effectiveness and applicability of the proposed methods are demonstrated through real-world case studies, showcasing their contributions to specific domains such as digital imaging surveys, UML class diagrams, and experiment documentation.

Metadata management is important as it invloves the management of metadata, which serves as the underlying data for maintaining and managing the namespace, permissions, and address of file data blocks (Kong et al., 2013).Extensive work has been conducted in the context of metadata management using systems like large-scale distributed storage systems such as data lakes and distributed file systems (Kong et al., 2013).Additionally, metadata management has been explored where academic libraries and their staff are increasingly engaged in research data management practices and processes within universities (Wiley, 2014).

In the context of multi-modal digital imaging surveys, Pamart et al. (2022) introduce a metadata-enriched system. This system enables users to annotate and document metadata related to digital images, including acquisition parameters, processing methods, and scientific interpretations. It also includes a metadata catalog for convenient retrieval and sharing of metadata. For UML class diagrams, Di Felice et al. (2020) describes the design and implementation of a metadata repository specifically tailored to this domain. Their repository automatically extracts and stores information about the structure, attributes, and relationships of UML class diagrams. In the realm of experiment documentation, a method for semi-automatically extracting metadata from annotated documentation in the eLabFTW electronic laboratory notebook (Musyaffa et al., 2021). They utilize natural language processing techniques to identify key phrases and concepts, generating metadata tags for categorization and search purposes.

In the study by Bach et al. (2019), the main findings revolve around the design and implementation of a generic archive storage service for research data in Germany. The paper highlights its reliability, scalability, and cost-effectiveness. The methodology involved implementing a generic archive storage service. The paper discusses the use of hierarchical storage management and the integration of open-source software and existing research infrastructure. The authors argue for the value of such a solution in research data management and its potential for broader adoption beyond Germany (Bach et al.,2019).

In the paper by Waddington et al. (2015) the main findings center on the challenges researchers face in storing and managing data, with cloud repositories proposed as a potential solution. The methodology includes an analysis of the limitations of traditional data storage methods and the benefits offered by cloud repositories. The paper introduces the concept of cloud repositories as a scalable and flexible solution for research data management. The authors argue for the importance of data security, ethical and legal compliance, and effective data management practices when selecting a cloud repository (Waddington et al., 2015).

The paper by Kaur and Rani (2015) in the field of Computer Science, presents a smart polyglot solution aimed at effectively managing and analyzing big data in the healthcare industry. The proposed solution combines various technologies, including NoSQL databases, Hadoop, Apache Spark, and natural language processing (NLP) techniques. By leveraging these technologies, the solution addresses the challenges posed by the exponential growth of healthcare data. A key aspect of the solution is the integration of NLP techniques, which play a crucial role in extracting valuable information from unstructured data sources such as clinical notes, research articles, and patient forums. This integration enables the identification of medical terms, establishment of relationships between medical concepts, and facilitates advanced data analysis. Furthermore, the paper recognizes the importance of data privacy and security in healthcare. To address these concerns, the proposed solution incorporates privacy-preserving techniques such as data anonymization and access control mechanisms to safeguard sensitive patient information throughout the data management process (Kaur & Rani, 2015).

The paper by Oleksik et al. (2014) examines the design and practices surrounding electronic lab notebooks (ELNs) in a collaborative scientific setting. The authors conducted an empirical investigation during the 17th ACM conference on Computer supported cooperative work & social computing to explore the usage patterns and design features of ELNs (Oleksik et al., 2014). Through their study, the authors aimed to identify the emerging practices and design considerations that arise when using ELNs in collaborative scientific environments. They investigated the organization and structure of data within ELNs, the integration of multimedia content, and the collaborative features that facilitate knowledge sharing and collaboration among scientists (Oleksik et al., 2014). The paper highlights the challenges scientists face when adopting and implementing ELNs in their research workflows. These challenges include ensuring data security, addressing privacy concerns, and establishing interoperability between different ELN systems (Oleksik et al., 2014).

Research data storage focuses on the storage of research data, particularly for long-term preservation. This includes exploring different storage technologies, such as tape, disk, and cloud repositories, and addressing challenges related to scalability, reliability, and cost-effectiveness (Bach et al., 2019). Data management challenges highlight the challenges faced by researchers in storing and managing data. These challenges include limitations in storage capacity, access, security, data treatment inconsistencies, potential data deletion risks, knowledge loss due to staff turnover, and scattered data across multiple locations (Waddington et al., 2015).

Heterogeneous data management is one of the issues noted in the literature. Yang et al. (2019) heterogeneous data management strategy was developed for Cloud Services 2019. However, the study does not go into great length about the drawbacks and difficulties of the approach they offer. The lack of information on the research methodology used in the study also contributes to credibility and dependability issues. The study by Willmes et al. (2014) uses open-source software and the author supplies data management infrastructure, but the study's scope and potential are not specified. These studies show that to improve the interpretation and real-world implementation of the findings, more thorough discussions on constraints, potential biases, and research methodology are required.

In terms of approaches to data management, the literature reveals divergent viewpoints and proposed solutions. Yang et al. (2019) focus on developing a scheme tailored to managing heterogeneous data storage within CSPs, while (Selvi & Sasirakha, 2021) propose a four-layer architecture specifically designed for data management challenges in the Internet of Things (IoT) context. These differing approaches reflect variations in authors' perspectives on the most effective strategies for data management. Additionally, the emphasis on open-source software, as highlighted by Willmes et al. (2014), suggests the authors' dedication to exploring the benefits and feasibility of open-source solutions in research data management infrastructure.

The importance of metadata management in effective data storage and tracking. Metadata plays a critical role in data preservation, reuse, and interoperability. Various studies propose systems and methods for metadata management. Delgado et al. (2021) introduce a co-development approach between metadata experts and researchers to create customized metadata schemes, while Hasan and Abu Bakar (2021) focus on extracting and transforming metadata from diverse data sources using the R programming language. These studies offer valuable insights into the potential approaches for metadata management, but they differ in terms of methodology, tools, and specific domains of application.

Furthermore, the reviewed studies highlight the significance of data storage solutions in research data management. Bach et al. (2019) propose a generic archive storage service for research data in Germany, emphasizing its reliability, scalability, and cost-effectiveness. On the other hand, Waddington et al. (2015) discuss the challenges faced by researchers in storing and managing data and propose cloud repositories as a potential solution. Kaur and Rani (2015) explore the concept of polyglot data stores as an alternative approach to traditional monolithic data stores. These studies provide insights into different storage technologies, their strengths, weaknesses, and potential benefits.

The conflicting perspectives in the literature reflect the diverse range of data storage solutions for research data. Bach et al. (2019) propose a centralized archival system tailored for research data in Germany, focusing on factors such as data integrity, preservation, and long-term accessibility. Conversely, Waddington et al. (2015) argue against traditional storage methods and advocate for the use of cloud repositories, highlighting scalability, flexibility, and data security as key considerations. These differing perspectives underscore the need for customizable and adaptable storage solutions that can integrate with existing research systems and workflows while ensuring data security and compliance.

The findings from the literature review also suggest the importance of effective data management practices. Jones et al. (2018) emphasize the significance of DMPs in ensuring proper data organization, documentation, and preservation. The study by (Khurshid et al., 2020) highlights the role of data governance frameworks in promoting standardized data management practices, ensuring data quality and compliance with regulatory requirements. These studies underscore the need for comprehensive data management strategies and frameworks that encompass data organization, documentation, quality control, and governance.

2.4 Research data management issues

Irawan et al. (2019) conducted a study that addresses the crucial issue of RDM and its implications within the Indonesian context. The authors emphasize the significance of RDM in promoting transparency, reproducibility, and collaboration in scientific endeavors. They identify challenges specific to Indonesia, such as limited infrastructure, inadequate data documentation practices, lack of awareness and training, limited funding, and policy gaps. To tackle these challenges, the authors propose a comprehensive framework for research data management in Indonesia, providing practical recommendations for each stage of the data lifecycle. They also emphasize the importance of institutional support and collaborative efforts in enhancing data management practices (Irawan et al., 2019). This study focuses on research data management challenges and opportunities within the Indonesian context. It highlights the significance of data management practices in promoting transparency, reproducibility, and collaboration in scientific endeavors. The authors identify challenges such as limited infrastructure, inadequate documentation practices, lack of awareness and training, limited funding, and policy gaps. This study provides valuable insights into the specific challenges faced in the Indonesian research landscape, which may resonate with similar contexts. It offers a comprehensive framework for research data management in Indonesia, with practical recommendations for each stage. This framework can be a useful reference to inform the development of strategies to overcome challenges in your research context (Irawan et al., 2019).

Birkbeck et al. (2022) conducted a literature analysis to identify challenges and issues related to RDM practices. They highlight the importance of effective RDM practices in advancing scientific inquiry, collaboration, and innovation. The challenges identified include the lack of clear policies and guidelines, inadequate infrastructure and resources, limited awareness and training, ethical and legal considerations, and managing diverse data formats and types. The paper emphasizes the need to address these challenges through the establishment of policies, resource provision, training programs, and appropriate data management strategies. This study conducts a comprehensive literature analysis to identify challenges in research data management practices. It emphasizes the importance of effective RDM practices in ensuring the accessibility, transparency, and reusability of research data. The authors identify challenges such as a lack of clear policies and guidelines, inadequate infrastructure and resources, limited awareness and training, ethical and legal considerations, and diversity of data formats and types. This study provides a broader understanding of challenges in research data management beyond the Indonesian context. It highlights the common challenges faced by researchers and institutions worldwide, which can help contextualize your research objectives. The identified challenges can be used to inform the development of strategies and interventions to address similar issues in your research context (Birkbeck et al., 2022).

Buys and Shaw (2015) present the findings of a survey conducted to assess data management practices within an institution. Their focus is specifically on data storage practices. The paper highlights various storage solutions adopted by the institution, including on-premises servers, cloud-based storage, NAS, and external hard drives. They identify challenges such as limited capacity, security concerns, difficulties in retrieval and sharing, and the need for backup and recovery strategies. The authors offer recommendations to enhance data storage practices, including scalable storage solutions, encryption, and access control measures, backup and recovery mechanisms, and awareness and training programs. This study focuses on data storage practices within an institution. It highlights different storage solutions employed, challenges faced, and recommendations for enhancing data storage practices. While this study has a narrower focus on data storage, it contributes to the broader understanding of research data management (Buys & Shaw, 2015).

Mancilla et al. (2019) explore research data management practices at Delft University of Technology in the Netherlands. Their findings highlight a lack of awareness and understanding of data storage and preservation practices among researchers, leading to a risk of data loss. Insufficient documentation practices are also identified as a risk to data integrity and reproducibility. The paper emphasizes the need for cultural change and the adoption of best practices in research data management, including the implementation of policies, training programs, and integration into the research workflow. This study explores research data management practices at a specific university, highlighting the need for cultural change and best practices. It emphasizes on institutional policies and guidelines, training and support for researchers, and integration of data management into the research workflow. This study emphasizes the importance of the adoption of best practices in research data management. It provides insights into the importance of institutional support and the integration of data management practices within the research workflow (Mancilla et al., 2019).

The literature highlights several common challenges faced in research data management. Firstly, limited infrastructure and resources hinder effective data management, impacting storage and accessibility (Birkbeck et al., 2022). Secondly, inadequate documentation practices pose challenges in understanding, reproducing, and reusing research data, emphasizing the need for improved documentation standards (Irawan et al., 2019).

Thirdly, researchers and institutions often lack awareness and training in RDM practices, underscoring the importance of knowledge dissemination and training programs (Birkbeck et al., 2022). Fourthly, policy gaps and limited funding contribute to challenges in RDM, necessitating the development of comprehensive policies and increased funding support (Irawan et al., 2019). Lastly, ethical, and legal considerations, such as data privacy and compliance, present challenges in managing research data (Birkbeck et al., 2022).

Scholars have proposed various frameworks and recommendations. Comprehensive frameworks provide a structured approach to RDM, encompassing stages such as data creation, documentation, storage, sharing, and reuse (Irawan et al., 2019). Additionally, practical recommendations tailored to specific contexts offer guidance on overcoming the constraints and needs of RDM practices (Irawan et al., 2019). Furthermore, institutional support and collaboration are crucial for effective RDM, emphasizing the integration of data management policies, training programs, and collaborative efforts among stakeholders (Irawan et al., 2019). Effective data storage practices require attention to different storage solutions based on specific requirements. These solutions include on-premises servers, cloud-based storage, and external hard drives (Buys & Shaw, 2015). However, challenges such as limited capacity, security concerns, difficulties in retrieval and sharing, and the need for backup and recovery strategies should be addressed (Buys & Shaw, 2015). Recommendations include implementing scalable storage solutions, encryption and access control measures, and awareness and training programs to enhance data storage practices (Buys & Shaw, 2015).

Promoting cultural change and adopting best practices play a crucial role in effective RDM. The cultural change recognizes the need to adopt best practices, emphasizing the importance of institutional policies and guidelines to promote responsible data management (Mancilla et al., 2019). Adequate training and support for researchers are vital in enhancing data management capabilities, including proper storage, preservation, and accessibility (Mancilla et al., 2019). Integrating data management practices into the research workflow ensures the seamless implementation of RDM principles and processes (Mancilla et al., 2019).

3. Research methods

This section presents the research methods used in this thesis. A qualitative research method is used, and the motivation for using qualitative research methods is characterized by its emphasis on exploring complex phenomena, generating in-depth and rich descriptions of the subject matter, and developing theories or frameworks that capture the complexity of the phenomenon under investigation (Creswell, 2014). First, qualitative research is described through existing literature. Next, participant selection and data collection processes are explained. Last, the analysis techniques for the collected data are explained. Figure. 2 shows the research method of the thesis.

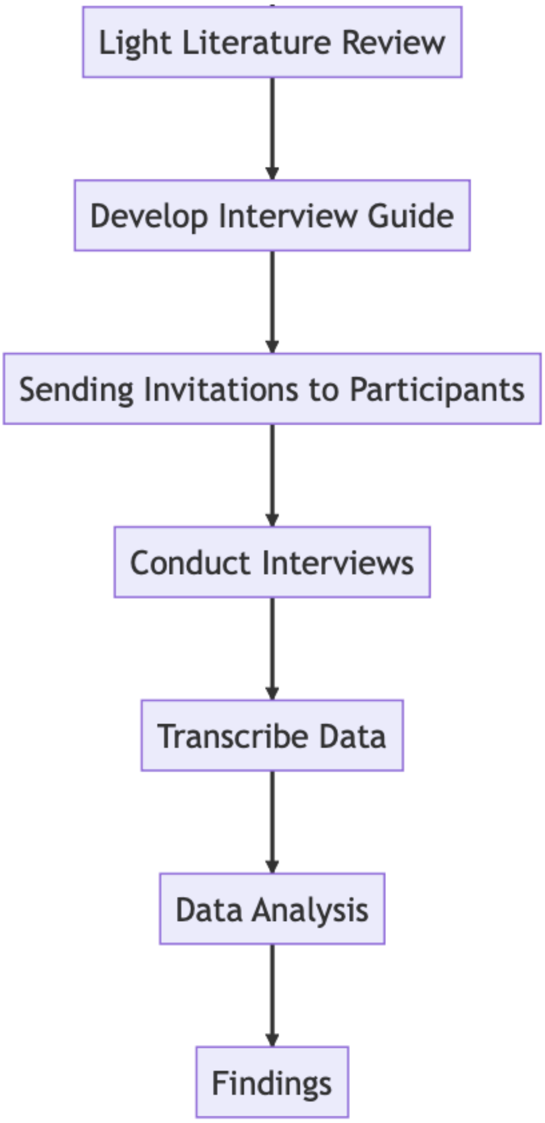

Figure 2. Eight-step research method for the thesis

The research method consists of four main components starting with a literature review, moving into data collection through semi-structured interviews, analyzing the collected data, and finally presenting the findings to the Interview participants. Firstly, a review of existing literature relevant to the research question was conducted. The study identified relevant sources and reviews and summarizes the literature to gain an understanding of the topic. Databases such as Google Scholar, ACM Digital Library, Research Rabbit, and Scopus were searched to retrieve articles and studies. The literature search covered a period from 2000 to 2022 to ensure the inclusion of recent and comprehensive sources and works before that were not used because of the technology has advanced and not many suitable studies were found. The second component is the semi-structured interviews phase. In this phase, semi-structured interviews are used as the data collection method. It begins with developing an interview guide, which includes carefully balancing thematic and follow-up questions, avoiding potential biases in participant responses, and considering ethical considerations. Then the invitations are to participants and conducts the interviews. The data from the interviews are transcribed for further analysis.

The third component is the Analysis phase. In this phase, the transcribed data is organized for subsequent analysis. Content analysis was used on the data to explore patterns and themes within it. The fourth and final component is the Present Findings phase. Here, the findings derived from the data analysis are interpreted. The interpretations are then presented to the participants of the interviews for validation and further discussion. This phase serves to ensure the accuracy and credibility of the findings before concluding.

3.1 Qualitative research and semi-structured interviews

The motivation for choosing qualitative research methods instead of quantitative research methods for this study stems from the aspiration to understand the current workflow of the participants and get more information about the research data management practices that they currently have in place. Qualitative research in RDM typically involves the use of in-depth interviews, focus groups, and document analysis to explore researchers' experiences and perspectives regarding data management. These methods can provide insights into the complex social and institutional factors that influence researchers' attitudes and behaviors toward RDM. By using qualitative research methods, this thesis aims to explore the subjective experiences and perspectives of users with a focus on understanding how these systems are integrated into organizational workflows and how they impact decision-making processes.

Qualitative research has multiple definitions and meanings to different people (Garcia & Quek,1997). Qualitative research is a type of research that aims to understand the meaning and interpretation of social phenomena by examining subjective experiences and perspectives of individuals and groups (Denzin, 2011; Creswell, 2014). Unlike quantitative researchers, the qualitative researcher attempts to keep the question layout as open as possible and to avoid presumptions (Camic et al., 2003).This enables avoiding premature closure of potential research areas. On the other hand, quantitative researcher evolves with a theory-based hypothesis which is then tested. A qualitative researcher often bases her research on inductive logic, unlike a quantitative researcher who bases her research on deductive logic (Camic et al., 2003). A comparison between qualitative and quantitative can be seen in Table 2.

Table 2. Summary of qualitative and quantitative research differences (Camic et al., 2003).

Qualitative	**Quantitative**
Depth	Breadth
Non-numerical	Numerical
Small and Purposeful	Large and Random
Subjectivity	Objectivity

Before proceeding to interviews, the interview guide was created. An interview guide is a structured set of questions or prompts used to guide an interview. It is designed to ensure that all relevant topics are covered, and that the data collected is rigorous and reliable (Kallio et al., 2016). Kallio et al. (2016) five-phased framework, which can be seen in Figure 2, was followed. The initial stage of the research involved formulating preliminary research questions, which served as a foundation for the study. Semi-structured interviews were chosen to answer the research questions. After conducting the light literature review the next phase was the creation of a preliminary interview guide.

Semi-structured interviews mean collecting qualitative data that allows for flexibility and in-depth exploration of a topic. Semi-structured interviews have been highlighted as a 2016preferred method for research due to various motivations identified in several studies (McIntosh & Morse, 2015; DeJonckheere & Vaughn, 2019; Kakilla, 2021;). One significant advantage of semi-structured interviews is their ability to facilitate in-depth exploration of a topic by utilizing open-ended questions. This approach allows researchers to delve deeply into the subject matter, uncovering nuanced and comprehensive insights. Another strength of semi-structured interviews lies in their flexibility, enabling interviewers to adapt their line of questioning based on the interviewee's responses. This adaptability ensures that the interview remains relevant and productive. Additionally, semi-structured interviews provide a platform to explore individual perspectives and experiences, allowing researchers to gain a comprehensive understanding of diverse viewpoints. These interviews also prove valuable in investigating motivations, attitudes, and the identification of enabling factors or barriers related to a specific topic, behavior, or intervention. Overall, the use of semi-structured interviews offers researchers a versatile and robust methodology for gathering rich and nuanced data in various research contexts (McIntosh & Morse, 2015; DeJonckheere & Vaughn, 2019).

The semi-structured interviews were used as a data collection method. One of the paramount factors in receiving plausible results from qualitative semi-structured interviews is to have a well-prepared interview guide. Kallio et al. (2016) present a framework where the creation of the interview guide can be divided into five phases and can be seen in Figure 2. The model has five phases, and each phase provides useful information that is needed to proceed to the next phase.

The first phase involves assessing the suitability of a semi-structured interview as a data collection method concerning the research question. Three types of interviews Structured Interview, Semi-Structured Interview, Unstructured Interview and were considered, and after consulting with two senior researchers, the semi-structured interview style was selected over structured and unstructured interviews.

In the second phase, a literature review was conducted to gather previous knowledge on the topic using various sources, such as Google Scholar, ACM Digital Library, Research Rabbit, and Scopus, to find existing papers and studies related to heterogeneous RDM. The purpose of the literature review was to gain an understanding of the topic and identify previous research conducted in this area. The review was kept brief to maintain an open-minded approach for the interviews.

The third phase involves the development of a preliminary interview guide, which includes carefully balancing thematic and follow-up questions, avoiding potential biases in participant responses, and considering ethical considerations. The interview guide was created before conducting the interviews and can be found in Appendix A, following the framework depicted in Figure. 3. The preliminary research questions were also formulated, and detailed information about their selection can be found in Section 3.1.1.

The fourth phase consists of pilot testing the preliminary interview guide through internal testing, expert assessment, and field testing. Feedback is collected, and no modifications are made to the guide before proceeding to the next phase (Kallio et al., 2016). Initially, the questions were reviewed by two senior researchers, and then two pilot interviews were conducted with volunteers who were not experts in the field but were university students. The aim was to refine the interview timeline and ensure the proper sequencing of discussions. The fifth and final phase involves presenting the complete interview guide in the study paper, providing a polished and logical guide for data collection.

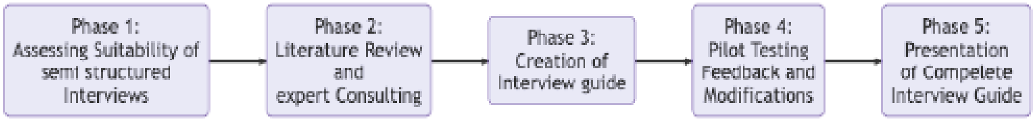
Phase 1: Assessing Suitability of semi structured Interviews
Phase 2: Literature Review and expert Consulting
Phase 3: Creation of Interview guide
Phase 4: Pilot Testing Feedback and Modifications
Phase 5: Presentation of Compelete Interview Guide

Figure 3. Semi-structured interview guide development (Kallio et al., 2016).

The five participants of this study were volunteers selected by a team's contact person who reached out to participants and asked about their willingness to participate in the interviews on the thesis topic over Microsoft Teams meetings or in-person interviews. The participants were very much interested in data management processes for their research data and how they can store their data more effectively. The invites that explained the purpose of the interviews were sent to participants through emails. Five different people were selected who were working on different projects.

The research team consisted of members involved in various projects, contributing to the diversity of data collected for the thesis. One team member had expertise in the field of water, energy, and environmental engineering, specifically leading a thematic group on water quality. Their research encompassed areas such as water engineering, hydrology, land use change, and riverine system processes, and utilized various tools. Data management was emphasized as a crucial aspect of their work. Another interviewee was a senior researcher in the same research unit, specializing in applied and environmental microbiology. They oversaw a team engaged in different types of work, generating a wide range of data. The third interviewee worked with data in remote sensing, with a focus on utilizing drones for monitoring headlands in Finland. Lastly, a member of the research team in the water, energy, and environmental engineering unit was responsible for conducting research related to water supply, stormwater management, wastewater engineering, groundwater modeling, and water resource management. Their specific research interests included water purification and the comprehensive management of the water system.

3.1.1 Description of questions

The participants' job experience and the nature of their work were the subjects of the first three interview questions. The purpose of questions 4 and 5 was to learn more about the participants' perceptions of the present workflow and their thoughts on any potential problems with the current work paradigm. Eight questions of the interview sought their perspective on environmental data and what made it unique compared to other sorts of data. Inquiries 9 and 10 sought their opinions on the ideal course of action that would resolve the problems raised earlier and found during the interview. The final query gave participants the chance to share any other pertinent information they felt was important during the last question.

1. Tell us about yourself and what you do for work.
2. What kind of projects have you worked on where you collected data? What scale?
3. What type of data have you recently worked with, and for what purpose?
4. Give us some details on the process of Data management from start to end that you use right now?
5. What kind of issues do you have with data management right now?
6. What is the most common use of data?
7. What kind of tools do you use?
8. Can you share your opinion on what makes Environmental data different from other data?
9. What in your mind would make managing your data easier?
10. What do you think will be the ideal way of solving the issues you currently face?
11. Do you have anything else you want to tell me?

The interviews were performed in the year 2023 between February to March. The language used was English, and the participants were from Finland. The average interview lasted 45 minutes, the shortest was 40 minutes, and the longest was 58 minutes. The interviews were mainly scheduled during office hours but in free time according to the participant's availability. The interviews were video calls on Microsoft Teams (Microsoft Teams, 2023) and they were recorded with a recording option available in the software. The recordings were transcribed using the built-in functionality of Teams software and securely stored on a personal laptop and Google Derive.

3.2 Content analysis

Content analysis is a qualitative research method that involves analyzing and interpreting the content of textual, audio, or visual data to identify patterns, themes, and meanings. Content analysis is *"a research technique for making replicable and valid inferences from texts to the contexts of their use"* (Krippendorff, 2013, p. 18). Content analysis can be used to explore a wide range of research questions in various fields such as communication, psychology, sociology, and education. Elo and Kyngäs (2008) noted that content analysis *"is a versatile method that can be applied to a variety of data sources and research questions"* (p. 1093).

There are two main types of content analysis: manifest content analysis and latent content analysis. Manifest content analysis involves analyzing the surface-level content of the data, such as words and phrases. On the other hand, latent content analysis involves analyzing the underlying meaning and context of the data. The manifest content analysis technique was used in this research. According to Mayring (2014), the process of content analysis involves several steps that can be seen in Figure. 4. Content analysis can be used as a standalone method or in combination with other qualitative or quantitative methods. As noted by Hsieh and Shannon (2005) *"content analysis can complement and enrich other qualitative research methods, such as in-depth interviews, participant observation, and focus groups"* (p. 1285).

In this thesis content analysis is selected as a method because of its ability to analyze large amounts of data in a systematic and structured way. Content analysis can be used by researchers for different reasons. One reason is that it allows objective and reliable analysis process of text data. Researchers can identify patterns, themes, and trends in the data that might be hidden in other forms of data analysis (Neuendorf, 2016). Content analysis can help fill in studying a variety of topics such as media representations, social media conversations, political discourse, and organizational culture (Krippendorff, 2019).

Content analysis flexibility makes it compatible with any language across all cultures (Neuendorf, 2016). The method can be adapted to the needs of research questions and the context of the work. Researchers have the option to choose from different types of content analysis like deductive or inductive and can customize analysis techniques as they see fit (Krippendorff, 2019). Researchers may choose content analysis for objective results for a large amount of data and answer a wide range of questions and its fixability allows it to be adopted across cultures and cross languages.

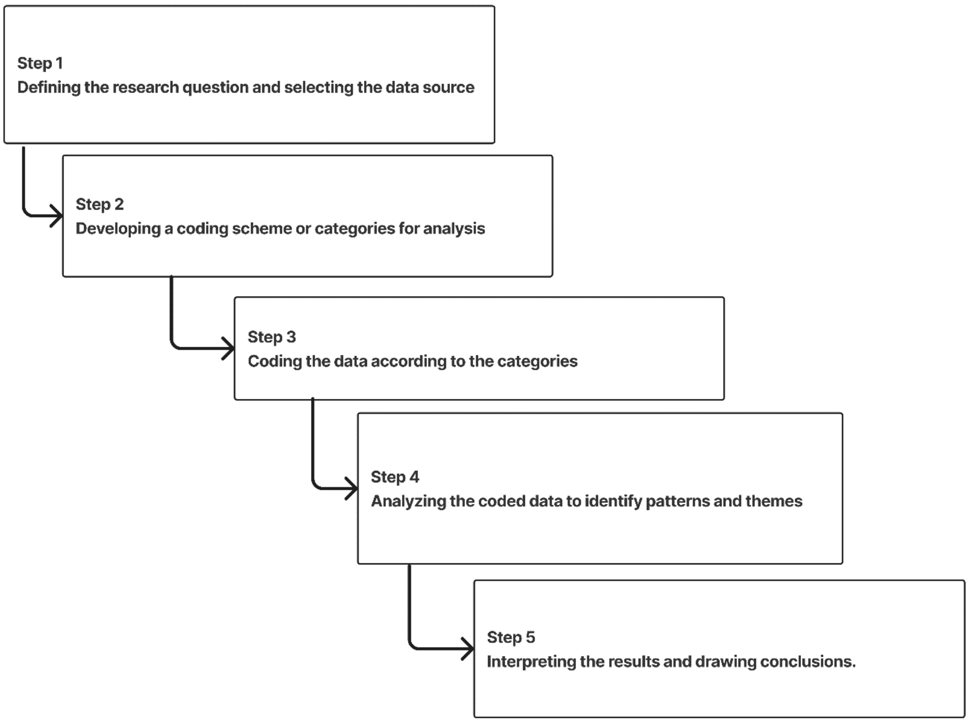
Step 1
Defining the research question and selecting the data source
Step 2
Developing a coding scheme or categories for analysis
Step 3
Coding the data according to the categories
Step 4
Analyzing the coded data to identify patterns and themes
Step 5
Interpreting the results and drawing conclusions.

Figure 4. Five-phased framework for the process of content analysis by (Mayring, 2014).

Mayring (2014) five steps for conducting content analysis involve defining the research question and selecting the data source, creating a coding scheme, applying the coding scheme to the data, analyzing the data, and interpreting the results. These provide a structured and systematic approach to content analysis that can help ensure the reliability and validity of the findings. How these steps were performed in this thesis can be seen in section 3.3.1.

Content analysis involves several phases that are crucial for conducting a systematic analysis of data. The first phase involves defining the research question and selecting the material to be analyzed (Mayring, 2014). This step entails identifying the purpose of the analysis and choosing the most relevant data source that aligns with the research question. After reading through the literature and conducting interviews the research questions were defined. The second phase focuses on creating a coding scheme. This step entails developing a set of codes or categories that will be used to analyze the data in a structured and systematic manner. The coding scheme should be relevant to the research question and provide a framework for organizing the data. The third phase involves applying the coding scheme to the data. This step requires reading and analyzing the data carefully, assigning each segment of text to one or more codes or categories according to the coding scheme developed in the previous phase. This systematic approach ensures consistency in the analysis process. In the fourth phase, the data is analyzed by examining patterns and themes that emerge from the coded data. This step involves counting the frequency of codes, identifying relationships between codes, and searching for patterns or trends in the data. Through this analysis, researchers gain insights into the underlying patterns and themes within the data. The final phase of content analysis is interpreting the results. In this phase, researchers analyze and interpret the findings concerning the research question. They consider the limitations of the analysis and discuss the implications of the findings for theory or practice. This phase aims to draw meaningful conclusions from the analysis and provide insights that contribute to the existing body of knowledge (Mayring, 2014).

3.3 Analysis process

Mayring (2014) five-phased framework for the content analysis process was followed with little modifications throughout the analysis process. Figure. 5 provides a complete illustration of the process.

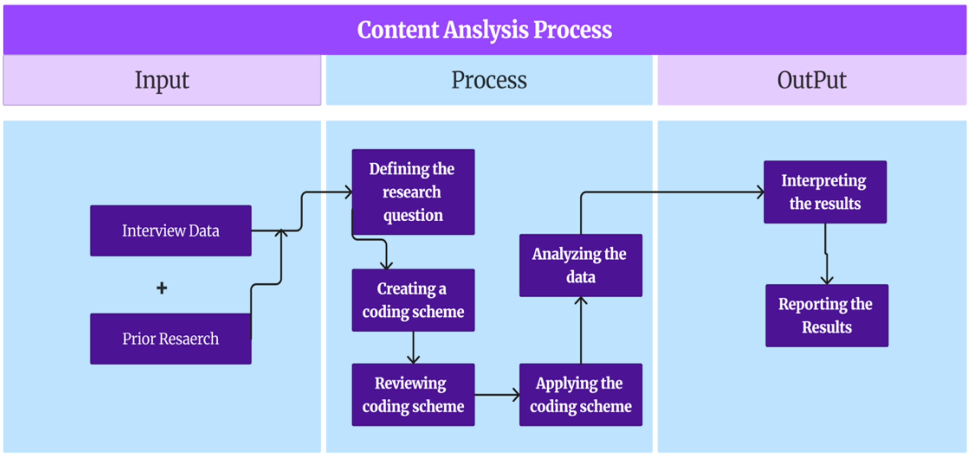

Figure 5. The content analysis process of the thesis

At the end of phase two, 12 different code schemes were created from the data set and four of which were selected named data collection, current data management process, data management issues, and Ideal Solution. The list of codes is provided in Table 1. Then code Schemes were applied to the data in the third phase and phase four the data was analyzed to gather the requirements and highlight issues of the current system in the final stage the results were presented. The implementation of the analysis process can be seen in section 3.3.1.

3.3.1 Analysis process

In the first phase of the five-phased framework for content analysis (Mayring, 2014), a dataset consisting of five interviews was selected. In the second step, study sought to develop codes by identifying repeated terms and phrases in the data in the context of the research question. The terms and phrases were selected if they were repeated more than two times. A total of 12 codes were identified to aggregate comparable information from various participants for efficient analysis. In the end, however, only four of these codes data collection, current data management process, data management issues, and ideal solution were chosen based on their frequency in relation to the topic of this thesis and the research objectives presented in sections 1.1.

The codes generated because of questions that asked participants about their introduction and work expertise were not included because they did not provide any vital information for the thesis topic. The codes that were created from data that only occurred once and no other participant talked about were eliminated. In the third stage, the codes produced in the second step were applied by rereading the transcript and highlighting the passages or data segments that corresponded to each code. Once the data from all the interviews had been highlighted and categorized into related categories, phase four commenced, in which evaluation of the data began. This may involve recognizing subjects, concepts, or ideas that appear repeatedly in the data.

Afterward, was started find correlations between codes and searched for patterns. Pattern identification refers to numerous references of characteristics or concerns, repeated complaints or compliments, or consistent feedback from a specific set of researchers. In the final phase, the results were analyzed considering the research questions and developing conclusions pertinent to the study subject. The analysis was done by interpreting the meaning of repeated patterns that were identified in the previous stage in the context of the information provided in the interviews. The interview results were presented to the interview participants as part of the thesis methodology to establish the dependability of interview findings.

4. Findings

The current data management process of interview participants is fraught with challenges that require resolution. The challenges like the absence of a systematic approach to data management, inconsistent treatment, and handling of data among researchers, the risk of inadvertent data deletion, potential loss of valuable knowledge during personnel changes, lack of clear guidelines for data management, data dispersed across multiple locations, and difficulties in tracking data changes over time. By addressing these challenges and implementing robust RDM practices, interview participants can enhance their efficiency in managing data, ensure data integrity, and facilitate the reproducibility of research findings. Detailed explanations of these issues can be found in sections 4.3 and 4.4.

The requirements identified through the interviews emphasize the necessity for a comprehensive data management system that incorporates automation, open access, centralized storage, online electronic lab notes, systematic data management, secure repositories, reduced reliance on hardware storage, and version control with metadata assistance. The specific requirements are elaborated upon in section 4.4. Quotations from participants for different codes are provided in Tables 4.

Table 4. The resulting four codes of content analysis

Code name	Description	Quotations
Data collection	Many projects collect original data rather than using publicly available data or reusing data from other sources.	*"All projects involve original data collection. "* *"Laboratory work involves taking notes in a notebook, transferring data to an Excel table or Word document, and sharing it on OneDrive. "* *"Projects worked on together involve shared folders for all involved researchers to access. "*
Current Data Management Process	There are various methods of data storage, including using the wrong OneDrive or a shared folder for collaborative projects. The issue of having multiple copies of files and the lack of version control may also be a common theme.	*"People have them on the wrong OneDrive."* *"Depending on what the type of work is."* *"First you will write in some notebook you take like some notes like what you have done on the day."* *"At some point, we'll get transferred to let's say Excel table or Word document."* *"This is then shared, that's this is then like put to but that's for say wrong OneDrive."* *"Usually have these shared folders that all people that are involved in the project can access."* *"Stored right away on some repositories."*

Table 4. continuation of Table 4

Code name	Description	Quotations
Data management issues	Participants mentioned the need for electronic lab notebooks and consistent naming conventions to improve data management practices. The issue of data hygiene was also raised, which relates to keeping track of the different types of data and minimizing the number of files.	*"We don't have a systematic way to do it."* *"Depending on each researcher, might have a little bit different way to do it."* *"Small risk that it might influence the data quality."* *"Storage of the data."* *"In theory, there's a possibility that one researcher could delete everything."* *"Our data is not yet Open Access."* *"People change quite often, so certain kind of knowledge is escaping with that as well."*
Ideal solutions	The participants discussed potential solutions to the data management issues, including creating a protocol for data management and developing conventions for labeling files, and providing additional information about the data.	*"Good guidelines that show how to do data management it."* *"Put it on the on like a repository where it's like a very safe, safely located and stored for."* *"Get it Open Access."* *"Stored right away on some repositories."* *"Automatic data handling."* *"Very systematic way which everyone follows."* *"Good guidelines that how to do it."* *" Inconsistencies in data treatment and management between researchers"*

This section provides a detailed explanation of the findings of content analysis performed on the interview data and explains different RDM issues that were identified in the current approach. Firstly, this section talks about the data collection practices of interviewees. In the second part, current RDM is explained, and in the third section, the issues of RDM are highlighted.

4.1 Data collection

The data collection code aims to investigate the current data collection practices employed by participants during interviews. It seeks to identify the various data collection methods utilized by research team members at the institute. Effective data collection is crucial for supporting research questions or hypotheses by gathering and recording relevant information or data. The interviewees provided different aspects of data collection in various research contexts.

The interviewee's statements highlight some important aspects of data collection and research practices. By emphasizing the importance of original data collection, good data management practices, and collaboration, researchers can ensure that their work is rigorous, accurate, and meaningful.

One of the interviewees also noted that *"Projects worked on together involve shared folders for all involved researchers to access."* highlighting the importance of collaboration and communication in data collection. The interviewee emphasized that when multiple researchers are involved in a project, it is essential to have a system in place to share data and facilitate communication. Shared folders or other collaborative tools can ensure that all researchers have access to the data and can contribute to its analysis and interpretation.

According to an interviewee, *"All projects involve original data collection"* highlighting that any research project requires the collection of data that has not been previously collected. The type of data collected can vary depending on the research question, the research design, and the methodology used, and data collection methods may include surveys, interviews, observations, or experiments.

Regarding laboratory work, the interviewee stated that *"laboratory work involves taking notes in a notebook, transferring data to an Excel table or Word document, and sharing it on OneDrive."* emphasizing the importance of proper data management in laboratory settings. AN interviewee stated that data is often collected through experiments or other scientific procedures, and the researcher must ensure that the data is recorded accurately and clearly in a notebook or other documentation. Once the data is collected, it may be transferred to electronic formats such as Excel or Word to facilitate analysis, and sharing data through cloud-based platforms like OneDrive can ensure that the data is accessible to all involved researchers and can promote collaboration.

4.2 Current data management process

This code provides a context of the current RDM that different participants use in different fields and different projects. The interviewees offered different perspectives on data management processes, that different teams are currently using based on their requirements. They highlighted both the importance of proper data storage and the potential challenges that can arise.

During our interview, one of the participants provided insights on data management processes that are important for research projects. One issue they highlighted was the risk of storing data on the wrong OneDrive, which can lead to confusion and difficulties in accessing the data. According to the interviewee, this underscores the importance of using the correct storage location and ensuring that all involved researchers have access to it.

The need for tailoring data management processes emphasizes the specific needs of each project as mentioned by one of the participants *"Depending on what the type of work is."* The type of data collected, the size of the dataset, and the number of people involved in the project are all factors that can influence the most effective data management approach. For example, they explained that in laboratory work, *"First, you will write to some notebook you take like some notes like what you have done on the day."* this highlights the importance of keeping detailed records during data collection. Proper documentation ensures that data is accurately recorded and can be properly analyzed later.

The interviewee stated that *"At some point, we'll get transferred to let's say Excel table or Word document."* suggesting the importance of ensuring that data is recorded in a format that facilitates analysis. Electronic formats like Excel or Word can help organize and structure data, making it easier to analyze and interpret. This highlights the importance of accurate documentation and proper transfer of data to ensure that it is reliable and useful for analysis Another important aspect of data management highlighted by the interviewee was the use of shared folders to facilitate collaboration and communication among researchers. The interviewee mentioned, *" Usually We have these shared folders that all people that are involved in the project can access."* They explained that shared folder is crucial for promoting collaboration and ensuring that everyone has access to the most up-to-date data. Finally, the interviewee stated *that "Data Stored right away on some repositories."* emphasizing the importance of storing data in repositories right away. This can help ensure that the data is safe and accessible.

4.3 Data management issues

Effective data management is essential for research teams to ensure the reliability, reproducibility, and discoverability of their research data. The current data management process of many research teams is not systematic and has several issues that need to be addressed, including inconsistencies in data treatment, the risk of accidental deletion of data, loss of knowledge due to personnel turnover, lack of guidelines, and data stored in different locations. Research teams need to develop a comprehensive data management plan that addresses these issues and includes clear guidelines for data collection, storage, sharing, and preservation. With a systematic and organized approach to data management, research teams can ensure the quality and integrity of their data and facilitate the sharing and reuse of research findings.

The current data management process of research teams is not systematic, and several issues need to be addressed. An interviewee's quotes shed light on some of these issues. Another interviewee stated, *"We don't have a systematic way to do it."* highlighting the lack of a structured approach to data management. Another interviewee mentioned that *"Depending on each researcher, [they] might have a little bit different way to do it."* indicating inconsistencies in data treatment and management among researchers. These variations in approach can lead to challenges in maintaining data integrity and consistency. One of the interviewees also expressed concern about the risk of accidental deletion of data, stating, *"In theory, there's a possibility that one researcher could delete everything."* emphasizing the need for robust data backup and preservation practices. Additionally, the issue of data storage may be stored in different locations, including personal computers and cloud storage, making it difficult to manage and access effectively. Multiple participants also noted, *"Our data is not yet Open Access."* indicating a need for greater accessibility and sharing of research data. Finally, an interviewee mentioned the challenge of knowledge loss due to personnel turnover, stating, *"People change quite often, so certain kind of knowledge is escaping with that as well."* underscoring the importance of capturing and retaining institutional knowledge related to data management practices. The list of issues identified is as follows.

- Lack of a systematic way to manage data.
- Inconsistencies in data treatment and management between researchers
- Risk of accidental deletion of data
- Loss of knowledge due to turnover of personnel
- Lack of guidelines for data management
- Data is stored in different locations, including personal computers and cloud storage.
- Difficulty in tracking changes made to data over time.

In conclusion, the current data management process within research teams exhibits several challenges and areas for improvement. The issues highlighted by the interviewees emphasized the need for a systematic approach to data management, addressing inconsistencies, implementing robust backup and preservation measures, enhancing data accessibility, and sharing, capturing institutional knowledge, establishing guidelines, and centralizing data storage. By addressing these issues, research teams can enhance the integrity, accessibility, and long-term usability of their data, ultimately contributing to more efficient and reliable scientific research.

The issues highlighted by interviewees underscore the need for clear protocols, guidelines, and backup systems to ensure the quality and integrity of research data. Research data management, encompassing the organization, storage, and sharing of data, is a critical aspect of scientific research. However, several challenges were identified that can impede effective data management. Firstly, the lack of a systematic approach to managing data results in wasted time, resources, and inconsistencies in data analysis, compromising the accuracy and reliability of research findings. Inconsistencies in data treatment and management between researchers further contribute to discrepancies in results, hindering scientific progress and impeding collaboration and study replication. Additionally, the risk of accidental data deletion poses a significant concern, necessitating the implementation of proper backup systems and safeguards to prevent the loss of valuable research data. The absence of clear guidelines for data management creates confusion and hampers collaboration and data reuse, while the turnover of personnel leads to knowledge loss and disruptions in data handling. To address these issues, it is crucial to establish standardized practices, promote collaboration, and provide training and resources for data management, ensuring the efficiency and effectiveness of research data management processes.

4.3.1 Lack of a systematic way to manage data

The lack of a systematic way to manage data can have significant consequences for research projects. As one researcher explains, *"We don't have a systematic way to do it. So, part of the data is stored in the cloud, and part of the data is stored in the researcher's computer. Some of the data is in joint."* This haphazard approach to data storage can make it difficult for researchers to access and share data.

To address this issue, some researchers have proposed a more systematic approach to data management. *"Idea that when the new data comes, everyone, every researcher will update it in a certain format all right away, but this is already one step this kind of."* says another researcher. However, even with this step, there is still a lot of manual work involved. To improve the situation, some researchers suggest putting all measurements directly into the cloud in code and using automatic data handling.

> *"To improve the situation, but still it's a lot of manual work so of course in the future it will be optimal if all the measurements would put in code directly in the cloud and then there would be some automatic data handling."*

One of the major problems with the lack of systematic data management is the risk of inconsistencies in how data is treated. *"One is that since it's not systematic. Like uh, treated the data, so depending on each researcher might have a little bit different way to do it."* explains a researcher. This can lead to issues with data quality, as there is always a small risk that inconsistencies may influence the results. Another issue that arises from the lack of systematic data management is the loss of knowledge when researchers leave the project or university notes a researcher.

> *"There is that we as in the university we. People change quite often, so people come to make a Ph.D. and then they leave, and so certain kind of knowledge is escaping with that as well."*

> *"From the data management point of view. stored right away on some repositories where it's like a very safe, safely located and stored for so that's. Guidelines are part of the data management as well."*

To address these issues, guidelines for data management should be established. As one participant mentioned, By establishing guidelines for data management, researchers can ensure that data is stored, treated, and shared consistently and effectively.

4.3.2 Inconsistencies in data treatment

The management of research data can vary widely among different research teams. One researcher might store their data on their personal computer, while another might store it in the cloud or a shared database. For instance, some databases are set up where the researchers have access to a joint OneDrive. However, this lack of standardization can lead to inconsistencies in data management practices, as each researcher might have their approach. As one researcher noted,

> *"Since it's not systematically treated, the data, each researcher might have a little bit different way to do it. There's always a small risk that it might influence the data quality. It's not that systematic yet".*

Moreover, this lack of standardization can pose risks to the security of the data, as well as the possibility of data loss. One researcher highlighted the potential issue of accidental deletion, stating that.

> *"If someone accidentally deleted everything, this is not a good situation, and we would need to have a good recipe where it would be Open Access as well."*

To mitigate these risks, it would be beneficial to establish a more systematic approach to data management that all team members can follow. As another researcher suggested, *"It would be better to have a very systematic way which everyone follows, and it will be easier to guide the new persons as well."*

From a data management perspective, proper analysis and storage of the data are crucial steps. As one researcher pointed out,

> *"Of course, from the data management point of view, data analysis is part of it. To make it smooth, we need to get it Open Access, stored right away on some repositories where it's very safely located and stored."*

However, even with the best intentions, researchers may still encounter challenges, such as having data stored on the wrong OneDrive. As one researcher noted, *"People have data on the wrong OneDrive, depending on what the type of work is."* Therefore, it is essential to establish clear guidelines and practices for data management to ensure that all research data is handled appropriately and accurately.

4.3.3 Risk of accidental deletion of data

Valuable research data can be lost due to the risk of accidental deletion of data which can lead to significant issue in research data management.Accidental deletion can occure due to different reasons like hardware failure, software errors or human errors as the one of interviewee stated that *"there's in theory, there's a possibility that one researcher could delete everything by accident. And then, of course, this is not a good, good situation."*This highlights the importance of having a reliable backup system in place to ensure that data is not lost or destroyed due to accidents or other unforeseen events.

Loss of data can also occur if proper backup and recovery systems are not in place that can help preventing such losses as a participant mentioned that *"We need to implement safeguards to prevent accidental deletion of data, such as requiring confirmation before deleting files."*This can help in preventing the loss of valuable research data and ensuring the continuity of research projects.

4.3.4 Lack of guidelines for data management

One of the challenges in research data management is the lack of guidelines for data management. Researchers may not have clear guidelines or standards for organizing, storing, and sharing data, leading to inconsistencies and difficulties in collaboration and data reuse.

The lack of guidelines for data management can also be problematic, as it can lead to inconsistent practices and uncertainty among researchers, as participants mentioned in the interview. When asked about it, one participant stated, *"Without clear guidelines for data management, employees may not know how to handle data properly."* The participants also highlighted that since the data is not systematically treated, each researcher might have a slightly different way to do it. They expressed concern about the small risk that this might influence the data quality, especially when people save data in different locations, including personal laptops and hard drives. Additionally, difficulty in tracking changes made to data over time can make it challenging to ensure the accuracy and validity of research findings. As one participant explained,

> *"Another issue is that people change quite often in the university, so certain knowledge is escaping with that as well. We need a very systematic way that everyone follows, and it will be easier to guide the new persons as well."*

These challenges highlight the importance of implementing a systematic approach to ensure the security and quality of data.

4.3.5 Data stored in different locations.

The storage of data in different locations could lead to difficulties in accessing and managing the data as one interviewee stated that " Some data is stored in the cloud, some on researchers' computers, some in hard drives, and some in joint folders. " It is essential to establish clear procedures for data storage and access, including protocols for data backup and recovery. Physical hard drives and personal computers may not provide adequate security or redundancy, making data vulnerable to lose or damage.

4.3.6 Loss of knowledge due to turnover of personnel

According to the interviewee, turnover of personnel can result in the loss of critical knowledge, as they stated: *"People change quite often, so certain kind of knowledge is escaping with that as well."* This issue is a common challenge in research data management and can have significant consequences. Another participant mentioned that *“When employees leave, they take their knowledge of the data with them, which can make it difficult for new employees to pick up where they left off."* When personnel leave an organization, they take with them the knowledge they have accumulated during their tenure, including knowledge about the data they have managed. This loss of knowledge can lead to difficulties in accessing and interpreting the data, as well as in maintaining its integrity over time.

To mitigate this problem, it is crucial for organizations to have standardized documentation procedures and to encourage knowledge sharing between employees. One way of solving the issue was pointed out by participants as they stated, *“We need to develop processes for documenting data and making sure it is accessible to others, even after employees leave."* Therefore, establishing standardized procedures for documenting and sharing data can help prevent the loss of valuable data knowledge, even when personnel turnover occurs. By doing so, organizations can ensure that their research data is well-managed, accessible, and can be effectively utilized for future research.

4.4 Requirements for the Solution

In terms of improving the data management process, the interviewee suggested that an ideal solution would involve automation and open access. An interviewee stated,

> *"From the data management point of view, of course, this data analysis is part of it. So, to make it smooth, but then of course like mention before to get it Open Access, get it to and stored right away on some repositories where it where it's like a very safe, safely located and stored for so that's. Of course, if that chain could be done automatically, that would be optimal."*

This indicates that an automated system that automatically stores and manages data in a secure and accessible way would be the most effective solution. The ideal solution would involve automating the data management process and making the data open access. This would require a more systematic approach to data management, standardized formats for data treatment, and secure repositories for data storage. Some quotes on ideas that came out as ideal solutions are as follows.

- *"Centralization data storage system."*
- *"Online electronic lab notes to minimize lab data loss."*

- *"System that helps to make data accessible and findable."*
- *"System that reduces hardware storage."*
- *"System should have version control and help with metadata."*

In conclusion, research data management is a critical aspect of scientific research, and the issues highlighted in this article underscore the need for clear protocols, guidelines, and backup systems to ensure the quality and integrity of research data. A lack of a systematic way to manage data can lead to wasted time, resources, and errors in data analysis. Inconsistencies in data treatment and management between researchers can lead to discrepancies in results and hinder scientific progress. The risk of accidental deletion of data can result in the loss of valuable research data. Finally, the lack of guidelines for data management can create difficulties in collaboration and data reuse and can lead to inconsistent practices and uncertainty among researchers. It is crucial to establish clear data management protocols, backup and recovery systems, and guidelines for data management to ensure the reliability and validity of research data.

This section presents the system requirements derived from interviews in which research team members explained what kind of functionalities they want from the system. These requirements aim to address the challenges identified in the current data management process and provide recommendations for the development of a system that supports effective research data management. The following requirements were gathered.

4.4.1 Hybrid database system for storage

The system should automate the storage and management of data to ensure a smooth data analysis process. Centralized data storage system should be included that will allow secure storage of data. By automating these procedures, the system reduces the possibility of mistakes or data loss that could happen while maintaining data by hand. Additionally, by giving researchers rapid and simple access to the relevant data, it streamlines the data analysis workflow and increases productivity while allowing researchers to concentrate more on data analysis and interpretation.

4.4.2 Open Access for Data

The system should provide open access to research data, enabling easy retrieval and utilization by users. It should incorporate an online platform with features such as online electronic lab notes to facilitate the recording and preservation of experimental procedures, observations, and results. Requiring open access to data can promote transparency, collaboration, and the advancement of scientific knowledge. It can allow for the verification and replication of research, encourages interdisciplinary collaboration, and facilitates discoveries.

4.4.3 Metadata management for research projects

Metadata management for research projects refers to the process of systematically collecting, organizing and maintaining descriptive information about the research data. It involves capturing and documenting important details about the data, such as its characteristics, origin, structure, and relationships with other data elements. Metadata provides context and meaning to the research data, facilitating its discovery, interpretation, and effective use by researchers.

The metadata management system should encompass features such as metadata capture, organization, and categorization. It should offer the flexibility to include descriptive information, such as keywords, abstracts, and annotations, to enhance the context of the research data. Ensuring interoperability with relevant metadata standards and formats is vital to facilitate seamless integration with other systems and tools. Establishing relationships and linkages between different data elements within the system can promote data exploration and discovery.

In addition, the system should provide versioning and tracking capabilities to monitor changes made to metadata over time and ensure data integrity. It should also incorporate access control mechanisms to safeguard the confidentiality and privacy of the metadata, granting appropriate permissions for authorized individuals or teams to access and modify the information. Robust search functionality will be essential for researchers to efficiently search and retrieve relevant research data based on metadata attributes.

4.4.4 Secure Repositories for Data Storage

The system should provide secure repositories that ensure the safety and integrity of research data. It should incorporate efficient data storage mechanisms to reduce the reliance on personal computers and cloud storage, which can pose risks to data security and accessibility. By providing secure repositories, the system safeguards research data from unauthorized access, accidental deletion, or loss. Additionally, it should have built-in version control and metadata assistance to track changes, and revisions, and facilitate accurate interpretation and understanding of the data. This ensures the traceability of data modifications, supports data integrity, and enhances data reproducibility.

4.4.5 Data Accessibility and Findability

The system should provide secure repositories that ensure the safety and integrity of research data. It should incorporate efficient data storage mechanisms to reduce the reliance on personal computers and cloud storage, which can pose risks to data security and accessibility. By providing secure repositories, the system safeguards research data from unauthorized access, accidental deletion, or loss. Additionally, it should have built-in version control and metadata assistance to track changes, and revisions, and facilitate accurate interpretation and understanding of the data. This ensures the traceability of data modifications, supports data integrity, and enhances data reproducibility.

4.4.6 Electronic Lab Notebook

An Electronic Lab Notebook (ELN) is a digital platform or software application that serves as a modern replacement for traditional paper lab notebooks. Develop and implement a ELN application with comprehensive features to enable users to efficiently record, organize, and manage experimental procedures, observations, and results in a digital format. The ELN should include advanced functionalities such as data organization, collaboration, data security, accessibility across devices and platforms, powerful search capabilities, and an intuitive user interface. The implementation should prioritize user experience while ensuring data integrity, confidentiality, and seamless integration with other laboratory instruments and data analysis tools.

5. Discussion

In this section, thesis will go through the answers to the research questions and discuss the themes and other topics that emerged in the interviews, as well as what previous studies have revealed about those subjects. The section also examines possible limitations of the study and ideas on how it could have been improved.

This study contributes to the existing body of knowledge in the field of environmental research by providing insights into the challenges of data storage and tracking. The identification of patterns, relationships, and trends enhances our understanding of the factors influencing data management practices in this domain. The findings emphasize the need for improved data management strategies, infrastructure, and support services, as well as the potential benefits of cloud-based storage solutions. Moreover, this research brings attention to the complex relationship between data complexity and storage challenges. By highlighting the specific challenges associated with different data types from different fields, the study guides researchers, and institutions in developing targeted solutions for data management.

5.1 Answer to research question one

One of the research questions of the thesis was "What are the key challenges that researchers face when dealing with environmental research data storage and data tracking?" The major findings of this study point out a number of significant problems with research data management. Inconsistencies and inefficiencies result from the lack of a systematic method for handling data, which is the first problem. Second, inconsistent data management and treatment practices among researchers can jeopardize the accuracy and reproducibility of research findings. Thirdly, the possibility of data erasure by mistake emphasizes the necessity of appropriate backup systems and preservation mechanisms. Fourth, staff turnover increases the risk of knowledge loss and makes it more difficult to ensure data continuity. Furthermore, efficient techniques for organization and storage are hampered by the absence of explicit data management principles. Moreover, issues with data accessibility and integrity are exacerbated by data stored in several places and the difficulty of tracking changes over time. To increase research integrity, reproducibility, and collaboration, it is imperative to address these results.

The findings presented indicate various issues associated with data management in the research context, which are aligned with the works (Buys & Shaw, 2015; Irawan et al., 2019; Mancilla et al., 2019; Birkbeck et al., 2022). One of the key findings from the current study highlights the lack of a systematic way to manage data. This aligns with the research conducted by Irawan et al. (2019) in the Indonesian context, where they identified limited infrastructure, inadequate documentation practices, and a lack of awareness and training as major challenges. Birkbeck et al. (2022) noted the absence of clear policies and guidelines for research data management further contributes to the inconsistent practices observed in data treatment and management between researchers.

The risk of accidental deletion of data and the loss of knowledge due to personnel turnover are important concerns identified in the current study. These challenges are also recognized by Irawan et al. (2019) who emphasize the need for targeted interventions and the establishment of data management policies and training programs. Additionally, the difficulties in tracking changes made to data over time, as mentioned in the findings, can impede data transparency and reproducibility. The storage and location of data emerge as critical aspects of research data management. Buys and Shaw (2015) reveal that different storage solutions are employed, including on-premises servers, cloud-based storage, network-attached storage, and external hard drives. The challenges related to limited storage capacity, data security concerns, and difficulties in data retrieval and sharing align with the current study's findings. These challenges underscore the importance of implementing scalable storage solutions, encryption, access control measures, and effective backup and recovery strategies, (Buys & Shaw, 2015).

The personal changes that occur within the research unit can significantly impact the integrity of data. When personnel change or leave the research team, there is a possibility of data loss or misinterpretation of the context in which the data was used. This can lead to challenges in reproducing or understanding research findings, hindering the overall integrity of the data. Furthermore, the findability of data can be compromised as well. If data is scattered across different locations, including personal computers and cloud storage, it becomes difficult to locate and access the required data when needed. This issue, specifically the impact of personnel changes and data storage locations, has not been extensively discussed in the context of research data management.

The interviews conducted with research team members revealed that data stored in different locations was a significant concern. This newly identified issue highlights the importance of centralizing data storage to maintain data integrity and accessibility. When data is stored in various locations, it becomes challenging to ensure consistent data management practices and data security measures. By addressing this issue, the system can provide a centralized data storage system that ensures data integrity, reduces the risk of data loss, and promotes efficient data management. It is worth noting that these specific challenges related to personal changes and data storage locations have not been extensively addressed in the existing literature on research data management.

This thesis's findings specifically address the challenges and issues of RDM in a particular context, whereas the referenced studies examine research data management practices in different contexts such as Indonesia, institutions within a country, and selected universities in Iraq. While the thesis's findings mention inconsistencies in data treatment between researchers and the turnover of personnel as challenges, the referenced studies discuss additional challenges such as limited infrastructure, inadequate documentation practices, lack of awareness and training, limited funding, policy gaps, and diversity of data formats and types. The studies provide specific recommendations and strategies to address the challenges they identified.

5.2 Answer to research question 2

One of the research questions of the thesis was" How can the challenges in data storage and tracking of environmental research data be addressed effectively? The collection requirements can greatly contribute to the storage and metadata aspects of a DMP. The findings of this study have several implications and contribute to the existing body of knowledge in the field of data storage and management for environmental research.

Regarding the hybrid database system for storage (4.4.1) implementing a centralized data storage system, researchers can securely store and manage their research data, reducing manual effort and minimizing the risk of errors or data loss. By consolidating data into a single database, it becomes easier to manage, organize, and store research data effectively. This storage ensures data integrity, reduces redundancy, and eliminates inconsistencies that may arise from multiple copies of the same data. Researchers can easily store and retrieve data, facilitating efficient data storage and retrieval for the DMP (Petersen et al., 2008; Curdt et al., 2017).

In terms of open access for data (4.4.2) the study highlights the importance of providing easy access to research data. Implementing an online platform with open-access features facilitates the sharing and retrieval of data, promoting transparency and collaboration among researchers. The findings align with the objective of creating an open-access system that maximizes the impact of research findings, fosters innovation, and encourages interdisciplinary research (IFPRI: International Food Policy Research Institute, 2022).

The study's exploration of metadata management for research projects (4.4.3) emphasizes the significance of capturing and documenting descriptive information about research data. Metadata management is a key aspect discussed in (Pamart et al., 2018; Di Felice et al., 2020; Musyaffa et al., 2021; Hasan & Abu Bakar, 2021). Metadata plays a crucial role in describing and providing context to the underlying data. This thesis also recognizes the importance of metadata management as an essential functionality. It emphasizes the need to extract and transform metadata effectively to enhance data organization, searchability, and understanding.

Regarding secure repositories for data storage (4.4.4) the study highlights the need for secure mechanisms to store and protect research data. By providing secure repositories with built-in version control and metadata assistance, the system ensures the safety, integrity, and traceability of research data. The findings contribute to the objective of developing a system that safeguards research data from unauthorized access, accidental deletion, or loss, enhancing data security and reproducibility (Lakshmanan et al., 2016).

The study's focus on data accessibility and findability (4.4.5) by implementing collected requirements can help make research data easily accessible and discoverable, the system will researchers to efficiently locate and retrieve specific datasets or information. This promotes data reuse, interdisciplinary collaborations, and diverse explorations of research data. Lastly, the study's exploration of ELNs (4.5.6) can help record, organize, and manage experimental procedures and results (Oleksik et al., 2014). By adopting electronic lab notebooks, researchers can enhance data organization, collaboration, and accessibility.

While the findings of this study support the relationship between data complexity and storage challenges, unexpected or contradictory results were also observed. For example, a subset of participants reported lower levels of difficulty in managing highly complex data types. Researchers with specialized knowledge or advanced skills in data management may have developed effective strategies to mitigate the challenges associated with complex data.

Furthermore, alternative explanations for the preference for cloud-based storage solutions could be considered. While the scalability, accessibility, and data security features of cloud repositories were emphasized in the findings, other factors such as cost-effectiveness, ease of use, and familiarity with cloud technologies might influence participants' choices (Yang et al., 2019) Future research could explore these alternative explanations through qualitative interviews or surveys to gain a more comprehensive understanding of researchers' decision-making processes.

In summary, this study has examined various aspects of data storage and management in environmental research. The findings underscore the importance of automating data storage, providing open access to research data, implementing metadata management, ensuring secure repositories, improving data accessibility and findability, and incorporating electronic lab notebooks. These findings contribute to the existing knowledge by emphasizing the significance of efficient data management practices, enhancing research productivity, fostering collaboration, and maximizing the impact of research findings. The implications of this study are crucial for the field of environmental research as they provide insights and recommendations for developing systems that streamline data analysis processes, promote transparency and collaboration, enhance data discoverability and usability, ensure data security, and facilitate effective documentation. Implementing these findings will ultimately contribute to advancing environmental research and addressing pressing environmental challenges.

5.3 Implications

The study's practical implications focus on improving data management practices in research. For academics, organizations, and politicians to create and put into practice improved data management policies, it offers useful insights. The quality and integrity of scientific research can be strengthened by addressing the stated difficulties by enhancing the organization, documentation, storage, and exchange of research data. Standardized practices and protocols are emphasized to enhance research reproducibility and boost the reliability and credibility of research findings. The risks of data loss due to accidental deletion or personnel turnover highlight the importance of proper data backup, recovery, and knowledge transfer mechanisms.

The theoretical implications contribute to the understanding of research integrity by promoting transparency, accountability, and reproducibility in scientific research. The inconsistencies in data treatment emphasize the need for further research and discussions to develop comprehensive guidelines, best practices, and standardized frameworks for research data management. The current RDM's lack of adherence to FAIR principles presents challenges in data findability and accessibility. The suggested steps outlined in the thesis are crucial for improved data management that aligns with FAIR principles.

In conclusion, this study identifies key challenges in research data management, emphasizing the need for standardized guidelines, improved infrastructure, and training programs to enhance research integrity, reproducibility, and collaboration within the scientific community.

5.4 Limitations and future work

Every study has limitations that should be acknowledged to gain a comprehensive understanding of its scope and impact. The sample size may be one restriction, which may have an impact on how broadly the results can be applied. To better understand the difficulties and implications of managing research data, future studies should consider larger and more diverse samples. Contextual specificity is another limitation, as findings may not apply to other institutions or countries. Exploring data management practices across various contexts can overcome this limitation. Exploring data management practices across various contexts can overcome this limitation. The scope and generalizability of the research may be limited to specific contexts and scopes, highlighting the need for exploration across different domains and contexts.

Implementing the suggested steps for improved data management may present challenges due to resource constraints, technical expertise, and organizational barriers

Another limitation of the study is the potential lack of generalizability to other institutions or countries. The challenges and requirements identified in your research may be influenced by specific contextual factors, such as organizational structures, funding resources, or data management practices unique to the participating institutions. To overcome this limitation, it would be beneficial for future studies to explore data management practices across a broader range of contexts. This could involve conducting interviews or surveys with researchers from various institutions or countries to capture a more comprehensive understanding of the challenges and requirements associated with heterogeneous research data management.

Many possible directions for further research might be suggested based on the restrictions and gaps found in this study. Future studies could also examine how well other data management solutions, such as training courses or the use of specialized data management software, work to solve the problems that have been discovered. Comparative studies evaluating the outcomes of various interventions would contribute to the development of evidence-based best practices for data management in environmental research.

6. Conclusion

This study has examined various aspects of data storage and management in environmental research. The findings underscore the importance of automating data storage, providing open access to research data, implementing metadata management, ensuring secure repositories, improving data accessibility and findability, and incorporating electronic lab notebooks. These findings contribute to the existing knowledge by emphasizing the significance of efficient data management practices, enhancing research productivity, fostering collaboration, and maximizing the impact of research findings. The implications of this study provide insights and recommendations for developing systems that streamline data analysis processes, promote transparency and collaboration, enhance data discoverability and usability, ensure data security, and facilitate effective documentation. Implementing these findings will ultimately contribute to advancing environmental research and addressing pressing environmental challenges.

To address these challenges, the thesis proposes several requirements for an effective data storage solution. These requirements include implementing a database system for storage, emphasizing metadata management, promoting open access policies, and integrating electronic lab notebooks. By implementing these steps, researchers can ensure centralized storage, efficient organization, accurate metadata management, and secure data storage, thereby facilitating data sharing, and collaboration. It is important to acknowledge the limitations of this thesis. The sample size used in the interviews may limit the generalizability of the findings. Additionally, the specific context in which the research was conducted may not fully reflect the challenges and requirements in other institutions or countries. Future studies should consider larger and more diverse samples and explore data management practices across various contexts to overcome these limitations.

In summary, this thesis has shed light on the challenges of research data management in the environmental research field and provided practical recommendations for addressing these challenges. By implementing the proposed requirements, researchers can enhance data storage, organization, and tracking, leading to improved research integrity, reproducibility, and collaboration. The findings and implications of this research contribute to the field of research data management and have the potential to impact various research domains and sectors reliant on effective data management.

References

Bach, F., Schembera, B., & van Wezel, J. (2019). Design and Implementation of the First Generic Archive Storage Service for Research Data in Germany. Datenbank-Spektrum, 18(3), 161-171. doi: 10.1007/s13222-018-0298-7

Birkbeck, Gail & Nagle, Tadhg & Sammon, David. (2022). Challenges in research data management practices: a literature analysis. Journal of Decision Systems. 31. 1-15. 10.1080/12460125.2022.2074653.

Buys, C., & Shaw, C. (2015). Data Management Practices Across an Institution: Survey and Report. 3(2), 1225–1225. doi: 10.7710/2162-3309.1225

Cox, A. M. (2017). Developments in research data management in academic libraries: towards an understanding of research data service maturity. Journal of the Association for Information Science and Technology, 68(9), 2182-2200.

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.

Curdt, C., Hoffmeister, D., Waldhoff, G., Jekel, C., & Bareth, G. (2012). DEVELOPMENT OF A METADATA MANAGEMENT SYSTEM FOR AN INTERDISCIPLINARY RESEARCH PROJECT. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 7-12.

Da Silva et al. 2012. Extracting and Exposing Relational Database Metadata on the Web. IADIS International Conference WWW/Internet 2012. pp 35-42

Dasapta Erwin Irawan, Cahyo Darujati, Santirianingrum Soebandhi, Fierly Hayati, & Sari. (2019). How to Extend your Data Lifetime: Research Data Management in Indonesia's Context. doi:10.2991/iclick-18.2019.33

Denzin, N. K , & Lincoln, Y. S. (2018). The SAGE handbook of qualitative research (5th ed.). SAGE Publications.

DeJonckheere, M., & Vaughn, L. M. (2019). Semistructured Interviewing in Primary Care research: a Balance of Relationship and Rigour. Family Medicine and Community Health, 7(2). doi:10.1136/fmch-2018-000057.

Di Felice, P., Paolone, G., Paesani, & Marinelli, M. (2020). Design and Implementation of a Metadata Repository about UML Class Diagrams. A Software Tool Supporting the Automatic Feeding of the Repository. Journal of Information Science Theory and Practice, 8(2), 26-41.

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. Journal of Advanced Nursing, 62(1), 107-115. doi: 10.1111/j.1365-2648.2007. 04569.x

Furner, J. (2020). Definitions of "Metadata": A Brief Survey of International Standards. Journal of the Association for Information Science and Technology, 71.

FPRI: International Food Policy Research Institute. (2022). Research Data Management and Open Access (RDMOA) Policy. Retrieved May 5, 2023, from https://www.ifpri.org/research-data-management-and-open-access-rdmoa-policy

Garcia, L., Quek, F. (1997). Qualitative Research in Information Systems: Time to be Subjective?. In: Lee, A.S., Liebenau, J., DeGross, J.I. (eds) Information Systems and Qualitative Research. IFIP — The International Federation for Information Processing. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-35309-8_22

Gerard Oleksik, Natasa Milic-Frayling, and Rachel Jones. 2014. Study of electronic lab notebook design and practices that emerged in a collaborative scientific environment. In Proceedings of the 17th ACM conference on Computer supported cooperative work &amp; social computing (CSCW '14). Association for Computing Machinery, New York, NY, USA, 120–133. doi: 10.1145/2531602.2531709

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative Health Research, 15(9), 1277-1288. doi: 10.1177/1049732305276687

Hasan, F. F., & Abu Bakar, M. S. (2022). An approach for metadata extraction and transformation for various data sources using R programming language. Indonesian Journal of Electrical Engineering and Computer Science, 26(3), 1520-1529. doi:10.11591/ijeecs.v26.i3.pp1520-1529

Hui Keng Lau, Ser Yee Lee, & Ali, Y. (2021). Effectiveness of data auditing as a tool to reinforce good research data management (RDM) practice: a Singapore study. 22(1). doi:10.1186/s12910-021-00662-y.

Jones, S., Pryor, G., & Whyte, A. (2018). Data Management Planning: Principles and Practice. Facet Publishing.

Kallio, H., Anna-Maija Pietilä, Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. 72(12), 2954–2965. doi:10.1111/jan.13031

Krippendorff, K. (2019). Content analysis: An introduction to its methodology. Sage publications.

Krippendorff, K. (2013). Content analysis: An introduction to its methodology (3rd ed.). Thousand Oaks, CA: Sage Publications.

Kong, J.S., Kim, M.J., Lee, W.Y., & Ko, Y.W. (2013). Two-Level Metadata Management for Data Deduplication System.

Kaur, K., & Rani, R. (2015). A Smart Polyglot Solution for Big Data in Healthcare. IT Professional, 17, 48-55. doi: 10.1109/MITP.2015.111

LibGuides: Research Data Guide: Data management plan. (2021). Oulu.fi. Retrieved April 2, 2023, from https://libguides.oulu.fi/Researchdata/Data_Management_Plan

Lima, M.S. (2011). Visual Complexity: Mapping Patterns of Information.

Mayring, P. (2014). Qualitative content analysis: Theoretical foundation, basic procedures and software solution. Klagenfurt, Austria:

Microsoft Teams. (2023). Microsoft.com. Retrieved June 1, 2023, from https://teams.microsoft.com/_#/conversations/19:meeting_ZTkyODg1OTAtYmMzNy00OTZhLWEyMzctZjM3Y2NkZDc2ZTVm@thread.v2?ctx=chat

McIntosh MJ, Morse JM. Situating and Constructing Diversity in Semi-Structured Interviews. Global Qualitative Nursing Research. 2015;2. doi:10.1177/2333393615597674

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O., Wilkinson, M. D., ... & Bongcam-Rudloff, E. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use, 37(1), 49-56.

Musyaffa, F. A., Rapp, K., & Gohlke, H. (2021). LISTER: Semi-automatic metadata extraction from annotated experiment documentation in eLabFTW. Journal of Information Science Theory and Practice, 9(4), 4-15.

Mancilla, H. A., Teperek, M., van Dijck, J., den Heijer, K., Eggermont, R., Plomp, E., Turkyilmaz-van der Velden, Y., & Kurapati, S. (2019). On a Quest for Cultural Change - Surveying Research Data Management Practices at Delft University of Technology. LIBER Quarterly: The Journal of the Association of European Research Libraries, 29(1), 1–27. doi:10.18352/lq.10287

Neuendorf, K. A. (2016). The content analysis guidebook. Sage publications.

Nass, A., Mühlbauer, M., Heinen, T., Böck, M., Munteanu, R., d'Amore, M., Riedlinger, T., Roatsch, T., Strunz, G., & Helbert, J. (2022). Approach towards a Holistic Management of Research Data in Planetary Science - Use Case Study Based on Remote Sensing Data. Remote. Sens., 14, 1598.

Pamart, A., Livio De Luca, & Philippe Véron. (2022). metadata enriched system for the documentation of multi-modal digital imaging surveys. 6(1), 1–24. doi:10.14434/sdh.v6i1.33767

Petersen, M. D., Fleischer, C. C., Agger, R., & Hokland, M. (2008). A Database Solution for Laboratory Information Management. doi:10.1111/j.0300-9475.2004.01423bc.x

Pinfield, S., Cox, A. M., & Smith, J. (2014) Research data management and libraries: Relationships, activities, drivers and influences. PLoS ONE, 9(12), 1–29. doi:10.1371/journal.pone.0114734

P. Bryan Heidorn. (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends, 57(2), 280–299. doi:10.1353/lib.0.0036

Rabenhorst, S. D., & Schäfer, M. S. (2019). Overstating the climate consensus: A study in the persuasive function of online comments. Public Understanding of Science, 28(4), 435-453.

Research Data Management. (2023). OCLC; OCLC. Retrieved May 5, 2023, from https://www.oclc.org/research/areas/research-collections/rdm.html

Surkis, A., & Read, K. (2015). OF INTEREST * Research data management, 103(July), 154–156. doi:10.3163/1536-5050.103.3.011

Sandfeld, S., Dahmen, T., Fischer, F. O.R., Eberl, C., Klein, S., Selzer, M., Nestler, B., Möller, J., Mücklich, F., Engstler, M., Diebels, S., Tschuncky, R., Prakash, A., Steinberger, D., Kübel, C., Herrmann, H.-G., Schubotz, R. (2018). Strategiepapier Digitale Transformation in der Materialwissenschaft und Werkstofftechnik. Retrieved May 5, 2023, from https://edocs.tib.eu/files/e01fn18/1028913559.pdf

T.Kalai Selvi & Dr.S.Sasirakha, (2021). Data management issues and study on heterogeneous data storage in the Internet of Things. International Journal of Engineering Research and Technology, 14(4), 247-252.

Vetova, S. (2021). Big heterogeneous data integration and analysis. doi:10.1063/5.0043627

Wiley, C. (2014). Metadata use in research data management.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018.

Waddington, S., Zhang, J., Knight, G., Jensen, J., Downing, R., & Ketley,C. (2015). Cloud repositories for research data – addressing the needs of researchers. Journal of Cloud Computing, 4(1), 1-11. doi: 10.1186/s13677-015-0030-8

Willmes, Christian & Kürner, Daniel & Bareth, Georg. (2013). Building Research Data Management Infrastructure using Open-Source Software. Transactions in GIS. 18. n/a-n/a. 10.1111/tgis.12060.

Yan, Z., Zhang, L., Ding, W., & Zheng, Q. (2017). Heterogeneous Data Storage Management with Deduplication in Cloud Computing. IEEE Transactions on Big Data. doi:10.1109/TBDATA.2017.2701352.

7. Appendix A.

Interview Guide

Introduction:

• Begin by introducing yourself and your role in the interview process.

• Explain the purpose of the interview, which is to gather information for the participant's thesis.

• Request the participant's consent to record the interview for accurate notetaking and referencing.

Background Questions:

• Provide a brief background about yourself.

• Share your relevant work experience.

• Discuss your understanding of the current state of research on the thesis topic.

Thesis-related Questions:

• Provide a summary of the research problem you aim to address.

• Explain the methodology or approach you used for your research.

• Discuss any challenges or limitations you faced in your research.

Bonus Questions:

• Share any additional insights or experiences related to your earlier responses.

• Feel free to share any personal anecdotes or experiences that relate to your thesis topic.

Next Stage of Thesis and Conclusion:

• Thank you for your time and valuable insights shared during the interview.

• Let's confirm the next steps of the thesis process, such as potential follow-up interviews or data analysis.

• If you have any questions or need clarification about the thesis or interview process, please ask.

Note: Remembered to actively listen to the participant's responses, took clear notes, and engaged in a conversational manner to encourage openness and fruitful discussion.