Gottfried Wilhelm Leibniz Universität Hannover
Fakultät für Bauingenieurwesen und Geodäsie

Understanding the Requirements of Data Spaces in the Energy Sector

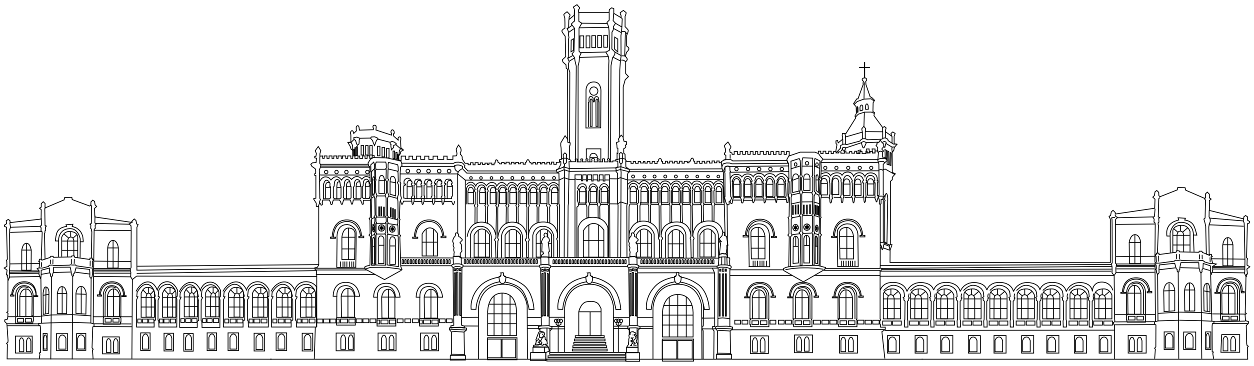
A thesis submitted in fulfillment of the requirements for the degree of
Master of Environmental Engineering

BY

Mazen Bechara
Matriculation number: 10040422
E-mail: mazen.bechara@stud.uni-hannover

First evaluator: Prof. Dr. Maria-Esther Vidal
Second evaluator: Dr.-Ing. Claudio Balzani
Supervisor: M.Sc. Philipp D. Rohde

March 7, 2024



Declaration of Authorship

I, Mazen Bechara, declare that this thesis titled, 'Understanding the Requirements of Data Spaces in the Energy Sector' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Mazen Bechara

Signature: ______________________________

Date: ______________________________

Acknowledgements

First, I would like to express my sincere gratitude to Prof. Dr. Maria-Esther Vidal for offering me the opportunity to pursue my thesis at TIB. Her expertise and guidance have been fundamental to my research. I am also grateful to Dr.-Ing. Claudio Balzani for his role in my thesis evaluation, which significantly enriched my research experience.

I am particularly grateful to my supervisor, Philipp D. Rohde, for his mentorship throughout this process. Special thanks to Mauricio Brunet for his crucial assistance with the extension implementation.

Finally, I would like to express my warmest thanks to my uncle, Suleiman Chammas, for being an idol and mentor in my life, and consistently offering guidance and inspiration. His unwavering support has shaped me in innumerable ways.

Abstract

Data Management (DM) is crucial for maintaining the transparency, integrity, and reproducibility of research findings by systematically organizing, storing, preserving, and sharing data throughout the lifecycle of research projects in various domains. This is particularly critical in data-intensive sectors like the energy sector. This sector faces unique challenges due to the complex nature of its data, ranging from sensor readings to policy assessments. DM is important not only for effective data handling, maintenance, and accessibility, but it also significantly enhances the reliability and trustworthiness of scientific research. By ensuring data is findable, accessible, interoperable, and reusable (FAIR), DM supports the credibility of outcomes and enhances data sharing practices, facilitating innovation and applied research in this rapidly evolving field.

In this thesis, we explored DM within the energy sector by identifying its requirements, assessing current practices, and understanding the perspectives of professionals in the field. Our research methodology began with a systematic literature review to collect foundational knowledge on the field's challenges and requirements. This was followed by a survey that focused mainly on the top 10 most mentioned DM requirements to understand the current state of DM in the energy sector. We discovered a strong emphasis on data quality for analytical purposes and the need for systems that are scalable and capable of integrating diverse data sources. Interestingly, while real-time data processing was not seen as a high priority by the majority of survey respondents, those with in-depth DM expertise highlighted its importance, indicating different perceptions based on DM knowledge. Additionally, our survey showed a preference for simulation tools over graphical visualization and highlighted a significant gap in familiarity with the FAIR principles among professionals, which pointed to limited external data sharing practices. To address one of these identified needs, we introduced the *ckanext-gitimport* extension as a proof of concept. This extension is designed to simplify the collection of metadata from external repositories. In summary, our work contributes to the understanding of DM in the energy sector by highlighting its current state, challenges, and areas for improvement. Through a combination of literature review, survey analysis, and the development of the extension, we lay the groundwork for future advancements in DM practices, essential for enabling data sharing in the energy sector.

Keywords: Data Management, Energy, Requirements

Contents

1 Introduction **1**
1.1 Motivating Example 3
1.2 Overview of the Document 4
1.3 Summary of the Chapter 4

2 Background **5**
2.1 Concepts, Tools, and Services 5
2.2 Data Management (DM) 8
2.3 FAIR Data Principles 8
2.4 Best Practices in DM 10
2.5 Challenges in DM 11
2.6 Exemplar Cases of DM 12
2.7 Summary of the Chapter 14

3 Related Work **15**
3.1 DM for Research 15
3.2 DM in the Medical Field 16
3.3 DM in the Energy Sector 17
3.4 Summary of the Chapter 18

4 Methodology **19**
4.1 Systematic Literature Review Process 19
4.2 Survey 30
4.3 DM Platform Selection 32
4.4 Summary of the Chapter 33

5 Data Manager **34**
5.1 CKAN 34
5.1.1 Architecture 35

5.1.2 Module Organization 37
5.1.3 Data Flow 38
5.1.4 Extensions and Code Organization 39
5.1.5 Key Concepts Used in the Implementation 40
5.2 Leibniz Data Manager 41
5.2.1 Architecture 42
5.2.2 Additional Features 42
5.3 Summary of the Chapter 42

6 Implementation **43**
6.1 Workflow of the Extension 44
6.2 The Scheming Extension 46
6.3 Additional Features 46
6.4 Unit Testing 48
6.5 CKAN Extension and LDM Customization 49
6.6 Summary of the Chapter 49

7 Results **50**
7.1 DM Expert Insights vs. Survey Findings 58
7.2 Summary of the Chapter 59

8 Conclusions and Future Work **60**
8.1 Conclusions 60
8.2 Limitations 61
8.3 Future Work 62
8.4 Summary of the Chapter 62

Bibliography **63**

List of Figures

1.1 Project Overview-Kennedy Energy Park Phase I 2

4.1 Methodology of the Thesis 19
4.2 Steps of the Literature Review 23
4.3 Geographic Distribution of Authors 27
4.4 Publication Year Distribution 28
4.5 Search Query Keyword Frequency 28
4.6 Distribution of Keywords Chosen by the Authors 29

5.1 Architecture and Context View of CKAN 36
5.2 Modules of CKAN 37
5.3 CKAN Data Flow 38
5.4 Code Structure of CKAN 39
5.5 LDM Overview 41

6.1 The Workflow of the Extension 45

7.1 Distribution of Affiliations Among Survey Respondents 51
7.2 Distribution of Respondents’ Job Positions 51
7.3 Respondents’ Primary Areas of Expertise in Energy 52
7.4 Distribution of Data Management Familiarity Responses 53
7.5 Distribution of Requirements Responses Among Survey Respondents 55
7.6 Distribution of Data Sharing Responses 56
7.7 Respondents Data Management Challenges 57
7.8 Respondents used Tools for Data Management Tasks 58

List of Tables

4.1 Portion of the Selected Articles 22
4.2 Portion of the Included Articles 26
4.3 Top 10 Requirements with Most Sources 31

6.1 GitHub Metadata Fields Defined in the YAML file 47

Acronyms

CKAN Comprehensive Knowledge Archive Network

CZ Critical Zone

CZO Critical Zone Observatory

DM Data Management

DOI Digital Object Identifier

DSO Distribution System Operator

GDPR General Data Protection Regulation

GPS Global Positioning System

GW Gigawatt

JSON JavaScript Object Notation

KG Knowledge Graph

LDM Leibniz Data Manager

MW Megawatt

MWh Megawatt Hour

NFDI Nationale Forschungsdateninfrastruktur

PLoS Public Library of Science

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PV Photovoltaic

RDM Reseach Data Management

RDO Research Digital Object

RFID Radio Frequency Identification

SoilTrEC Soil Transformations in European Catchments

UrhG Urheberrechtsgesetz

Chapter 1

Introduction

In the current scientific environment where data plays a crucial role, data management (DM) becomes increasingly important. DM includes a variety of tasks such as organizing, storing, preserving, and sharing data gathered during research. This practice is crucial not only for maintaining scientific rigor but also facilitates interdisciplinary collaborations. It is important to note that DM is not limited to data storage but extends to data curation, collaboration, and associated ethical considerations [55]. Particularly in the energy sector, there are dual challenges of rapid technological development and rising global energy demands. This sector is particularly rich in data, covering a range of complex datasets from sensor readings to policy assessments. Therefore, DM in this sector not only presents unique challenges but also opportunities for innovation and applied research [44]. A DM framework serves as the structured foundation for managing data efficiently across the entire project lifecycle, involving all project stakeholders. Such frameworks are instrumental in ensuring data integrity, facilitating collaboration among project participants, and enhancing the data's findability, accessibility, interoperability, and reusability (FAIR principles) [100].

Data ecosystems, which are infrastructures driven by data that enable stakeholders to exchange data [31], are an important component of DM. They should be designed to offer semantic interoperability, essential for integrating varied data sources across different domains. It needs to be able to provide practical guidelines for achieving this interoperability within modern data ecosystems. Furthermore, data spaces within the ecosystem are expected to support the ability of different data systems to work together seamlessly and autonomously, for specific use cases, facilitating secure and efficient collaboration and data sharing among various organizations. This approach ensures both trust and security in data transactions. The ecosystem should also empower individuals to control their personal data, enhance opportunities for data monetization, and amplify the socioeconomic impact of research data across domains and even countries. Additionally, establishing data repositories is crucial for improving government services and supporting evidence-based policy decisions [4] [34].

This work focuses on extracting the requirements necessary for DM in the energy sector through a comprehensive systematic literature review. It will be complemented by a survey filled out by energy experts, to validate the requirements and gain insights into them. As a proof of concept, this thesis presents an implementation of one of the requirements in the Leibniz Data Manager (LDM) [81], an existing framework for DM.

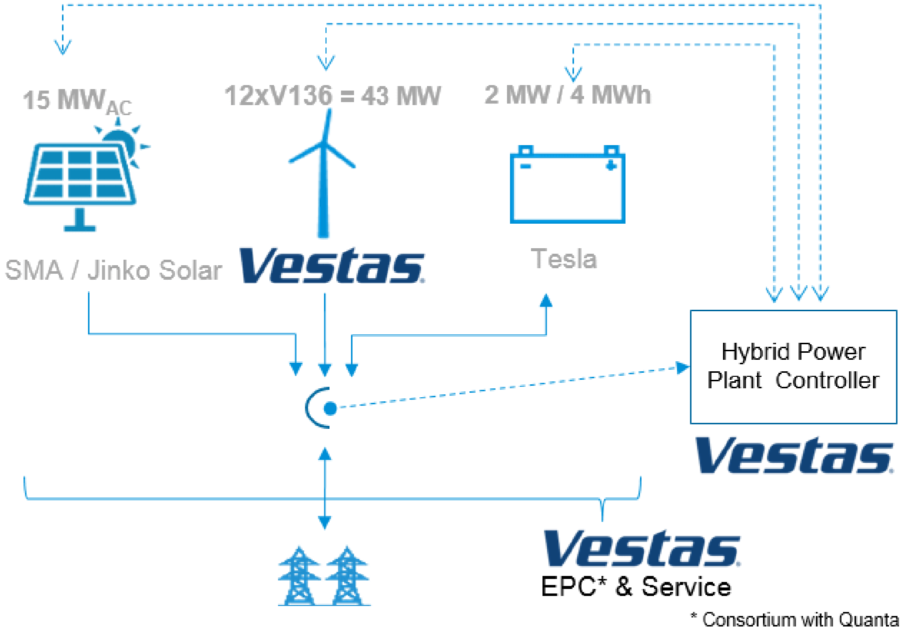


Figure 1.1: Project Overview-Kennedy Energy Park Phase I [69]

Several challenges are emerging in the energy sector. One of these challenges is the relatively new nature of the industry compared to others, which brings difficulties in collecting and handling data effectively. Another issue is the need for adaptability, especially in areas with diverse energy sources, requiring systems that can read data from different sources (RFID [83], GPS [93], camera, sensors, etc.) without major modifications. Additionally, managing various data inputs and standards from different providers is complex, and there is a need to find ways to efficiently use this data with minimal human intervention. Integration is also crucial, with the old methods of connecting systems no longer being feasible. Instead, a single, flexible system is needed to handle data from various sources while ensuring control and security. These challenges emphasize the importance of addressing DM issues in the energy sector [39].

1.1 Motivating Example

An example that motivates the challenges of integrating various data sources in energy data management is the Kennedy Energy Park, a hybrid renewable energy facility located in Hughenden, North Queensland Australia [7]. Developed in partnership between Windlab and Eurus, this facility integrates 15MW solar photovoltaic (PV), 43.2MW wind, and a 2MW/4MWh lithium-ion battery storage facility. Such a hybrid approach not only generates energy but also addresses one of the challenges faced by North Queensland's energy network. Traditionally, transporting energy from the southern parts of the state to the northern regions has proven costly and inefficient. The Kennedy Energy Park provides a solution by producing 60MW of renewable energy closer to the consumption point, reducing the pressure on long transmission lines. This project not only tests the combination of wind, solar, and storage, but it also sets the stage for the upcoming *Big Kennedy* phase, which is the second phase of the project that could further augment the renewable energy presence in North Queensland by providing 1.2GW of wind and solar energy [7].

The Kennedy Energy Park, with its integration of solar, wind, and battery storage, as can be seen in Figure 1.1, presents an example of the diverse challenges in DM within the renewable energy sector. This facility underscores the complexity involved in managing diverse data streams by combining various energy sources. Its use of advanced sensors and systems to optimize energy production highlights the necessity of sophisticated data integration and analysis techniques. Moreover, the scale and innovation of this project reflect the broader challenge of scaling DM frameworks in line with the evolving energy market. The variety of data sources, from energy meters in wind turbines to solar panels, captures invaluable data to ensure the efficient generation and storage of energy, making DM an important component in its management. These aspects of the Kennedy Energy Park provide an insightful and practical context for this thesis, demonstrating real-world applications and challenges in DM in this sector.

It is important to mention that inter-organizational data sharing can help realize the objectives of projects like the Kennedy Energy Park. The capability to share data effectively between varied entities is not only a technological requirement but a fundamental step in achieving innovation and generating value from data across different sectors [61].

1.2 Overview of the Document

The structure of this thesis is as follows: Chapter 2 provides a general background on DM, covering its practices, challenges, and relevant tools. Chapter 3 examines the contributions and findings in various sectors related to the field of DM. The methodology used in this research is detailed in Chapter 4, including a review of the literature, the extraction of requirements, a survey conducted for this study, and the selection of the DM platform. Chapter 5 discusses LDM, and its selection as the data manager for this thesis, followed by Chapter 6 which details the implementation of an extension of LDM. In Chapter 7, the thesis presents the results of the survey, discussing the outcomes and findings. The final chapter, Chapter 8, wraps up the thesis, offering conclusions, acknowledging the study's limitations, and suggesting avenues for future research.

1.3 Summary of the Chapter

This chapter outlines the significance of DM and mentions the essential features of a data ecosystem in the context of the energy sector. It highlights the critical needs and requirements for establishing an effective data ecosystem, emphasizing aspects like semantic interoperability and secure data exchange. The chapter also sets the objectives of this thesis, which includes conducting a survey based on certain identified requirements extracted from the literature review. It discusses the various challenges in energy data management, particularly focusing on data integration from diverse sources. An example that effectively illustrates this challenge is the Kennedy Energy Park, which demonstrates the practical implications and complexities of integrating multiple data streams in the energy sector.

Chapter 2

Background

This chapter discusses the main topics necessary for the comprehension of the development of the thesis. It explores data management (DM) and examines its roles in enhancing scientific research and practical applications. Key sections include DM, FAIR Data Principles, Best Practices and Challenges in DM, and Exemplar Cases Study. The discussion aims to highlight how effective DM contributes to scientific integrity, innovation, and sustainable development.

2.1 Concepts, Tools, and Services

The first section in this chapter defines some essential concepts, tools, and services crucial for understanding the following sections and chapters:

- **Data:** Refers to a collection of facts, measurements, or observations, often represented in the form of numbers, words, or images. It is the raw information from which conclusions can be drawn or analysis can be made. Data can be quantitative (numerical) like prices or weights, or qualitative (descriptive) like names or colors [13].

- **Metadata:** It is a detailed information that describes the data. Metadata can include details about how, when, and by whom data was collected, as well as what the data represents. Therefore, makes it easier to reproduce processes executed over the data [38].

- **Research Data:** Includes various forms of data that are collected or used to confirm and support the findings of the research. This can include numerical measurements, text, survey results, or observational notes, among other types.

Research data is diverse and reflects the wide range of methods and disciplines in academic research [95].

- **Digital Data:** Represents a broad array of information stored electronically. This includes but is not limited to experimental, simulated data; various software, codes and algorithms; textual content, audio files; and related metadata [94].

- **Big Data:** Refers to vast volumes of data that are too large or complex to be processed by traditional DM tools. It is characterized by the enormous scale (Volume), rapid flow (Velocity), and diverse types of data generated from various sources (Variety) [64].

- **Energy Data:** Energy data refers to specific types or categories of data related to the energy industry as defined by regulations, along with any additional data types specified by regulatory guidelines [53]. It is important to note that the definition of energy data can vary depending on the context and industry in which it is used.

- **Dataset:** It is essentially a collection of related information grouped together. It can include a variety of data, such as business details like names, salaries, and sales figures. In a database, a dataset could be all the data within it or specific groups of data, like the sales records of a particular department [88].

- **Data Service:** It typically includes a range of software solutions that assist in accessing, managing, and analyzing data. These services streamline how data is handled and processed in various computing applications [78].

- **Data Preservation:** Refers to ensuring that data remains accessible and usable beyond the duration of the research project for which it was originally created [94].

- **Data Ecosystem:** A data ecosystem is a comprehensive network that includes various tools, platforms, and processes, all connected to manage, process, and analyze data. It involves an array of elements like databases, data warehouses, integration tools, and analytics platforms, each playing a role in handling and making sense of data. The primary aim of a data ecosystem is to help organizations manage their data safely and efficiently, allowing them to derive meaningful insights and make well-informed decisions [31].

- **Data Lifecycle:** Refers to all the stages that data goes through, from its initial creation to its eventual distribution and re-use. This lifecycle encompasses the entire journey of data, detailing how it is handled, modified, and utilized throughout its existence [38].

- **Linked Data:** Refers to a technique for sharing, connecting, and making data available on the Semantic Web, a framework designed to enable data to be interconnected and machine-readable across the internet. The primary purpose of Linked Data is to facilitate the discovery and utilization of data across various sources and fields by establishing interlinks between distinct datasets [97].

- **Data Repository:** A digital location where data is stored and preserved, often for sharing, keeping it secure, and making it easy for people to access it [38].

- **Data Sharing:** The practice of distributing data to other parties, aligning with the principles of open scientific communication and collaboration [38].

- **Data Security:** A series of strategies and protective measures designed to safeguard data from unauthorized intervention, misuse, or damage [38].

- **Research Digital Object (RDO):** A Digital Object (DO) can be defined as a sequence or set of bit sequences representing digital information, ranging from simple text files to complex data structures like Excel spreadsheets [11]. Building on this concept, a Research Digital Object (RDO) would be a specialized type of DO specifically designed for academic and scientific purposes.

- **Digital Object Identifier (DOI):** A distinctive digital identifier assigned to items irrespective of their nature (physical, digital, abstract). It provides a consistent way to locate and identify these items across various platforms and systems, ensuring their traceability and accessibility over time [29].

- **GitHub:** An online service that offers version control, enabling collaborative development of projects while maintaining the main code's integrity [37].

- **README:** A document, typically in a plain text format, that includes important information about other files found in the same directory or digital package. It frequently contains instructions or detailed context relevant to the data or software it accompanies [38].

- **Likert Scale:** It is a scale used in questionnaires to measure attitudes, opinions, or behaviors. It presents a statement to which respondents indicate their level of agreement or disagreement on a symmetric agree-disagree scale for a series of statements [74].

- **Empirical Data:** information obtained through evidence acquired from observation, experience, survey, or the use of scientifically calibrated instruments [91].

- **Knowledge Graph (KG):** Represents data through a network of interconnected entities and their relationships, allowing for the integration, querying, analysis, and comprehension of diverse and complex data sets. It enables semantic reasoning, which can improve data quality and utility. A KG can also be visualized using various tools that can show the structure and content of the graph, and it can have many applications and benefits for different domains and industries, such as energy, health, education, finance, etc. [41].

2.2 Data Management (DM)

DM includes the collection, storage, organization, and maintenance of data to support project operations and decision-making. DM functions span various disciplines, beginning with the development of a data architecture to guide the organization and deployment of data across systems. Databases play a central role in this process, serving both transactional and analytical purposes by organizing data for easy access, updates, and management. Key tasks include database administration, which involves setting up, monitoring, and tuning databases, as well as ensuring data security, backup, and recovery, alongside regular updates and maintenance to support operational and analytical needs. Effective DM is crucial for leveraging data as a strategic asset, enabling informed decisions, optimizing operations, and ensuring compliance with regulatory requirements [25].

2.3 FAIR Data Principles

The FAIR principles aim to enhance the **F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability of digital objects, emphasizing both human and machine usability. Established in 2016, these guidelines facilitate the ongoing discovery, integration, and citation of digital objects, contributing to scientific advancement. The principles are evaluated through metrics and maturity indicators, with tools like FAIRshake assessing compliance to ensure digital objects meet these standards, thus supporting broader knowledge integration and scientific discovery [47].

Enhancing Findability

The Findability principle necessitates that data should be easy to locate for both humans and machines. This is often achieved through the use of persistent identifiers like Digital Object Identifiers (DOIs) for datasets, ensuring that data can always be found even if its location changes. Metadata plays a critical role here, acting as a detailed index that makes data discoverable through search engines and specialized data repositories [100].

Accessibility for All

Accessibility goes beyond only being able to access data; it requires that once found, data can be accessed under well-defined conditions, respecting privacy and proprietary constraints where applicable. Accessibility implies that data and its metadata should remain retrievable even when the data itself is no longer available, ensuring that the knowledge it represents is not lost [100].

Promoting Interoperability

Interoperability involves ensuring that data can be integrated with other data, can be analyzed in various software environments, and can be used in conjunction with other datasets. This principle is crucial for multi-disciplinary research, where data from different fields must be combined to solve complex problems. Standards for metadata and data formats play a key role in achieving interoperability [100].

Reusability for Future Research

Finally, Reusability ensures that data can be reused in new research contexts, maximizing the value of the data long after the original research has concluded. This requires comprehensive metadata that provides context, clear data usage licenses, and documentation that outlines the data collection process, any transformations the data has undergone, and how it should be correctly cited [100].

DM and FAIRness

The connection between FAIR principles and DM is important to establish. Although the FAIR principles were initially designed for the context of research data, their applicability extends to any digital object, thereby directly informing and shaping effectively DM practices. By adhering to FAIR principles within DM, researchers and institutions can ensure that the data generated from research activities are managed in a way that maximizes their utility, impact, and contribution to ongoing scientific discovery and innovation [40].

2.4 Best Practices in DM

Efficient DM is crucial for enhancing reproducibility, facilitating data reuse, and recognizing data as a scholarly asset. The push towards data sharing for research verification and reuse highlights the need for robust DM. Effective DM practices enable researchers to quickly find, comprehend, and use their data throughout their project and beyond [10].

These practices streamline data analysis, visualization, and reporting, thereby easing the publication process. By adopting foundational DM practices, researchers can avoid common data handling mistakes, formalize their processes, and save significant time. This approach not only benefits individual projects but contributes to the broader scientific community, enhancing discovery speed, increasing the reliability of findings, fostering collaboration, and offering new educational data uses. Adopting these practices is a step towards better, more efficient research data handling across various disciplines [10].

According to [10], the best practices of DM can be summarized as follows:

- **Documentation:** Keeping detailed records of experiments and data is crucial. This includes the person who conducted the experiment, the procedures and materials used, and the conditions under which data was collected. Good documentation is important for data reuse and understanding, especially over time.

- **Organize Files and name them consistently:** Organizing files logically and naming them consistently helps in efficiently locating specific files. Creating a system that groups data by project, analysis type, or date, and sticking to it is essential.

- **Version Control:** Keeping different versions of a document as it evolves is crucial. This practice allows returning to earlier versions if needed and is particularly useful for procedures and data analysis.

- **Create a Security Plan:** For sensitive data, a clear security plan detailing who has access, data retention, and destruction policies is necessary. Security protocols should be regularly reviewed and updated.
- **Roles and Responsibilities:** Clearly defining and documenting the roles and responsibilities within a research team enhances efficiency and accountability.
- **Data Backup:** Following the 3-2-1 Rule (three copies of the data, two in different locations, and one on a different storage type) is a best practice to prevent data loss.
- **Tool Constraints:** The choice of tools for data collection, storage, and analysis should consider their limitations and compatibility with other components of the research workflow.
- **Project Closeout:** At the end of a project, it is important to identify and prepare key research files for long-term preservation, including creating master copies and snapshots of important files.
- **Using Data Repositories:** Depositing data in a repository ensures its long-term care and accessibility, making it easier for others to find and use.
- **Write these Conventions down (in a DM Plan):** A living document that describes DM conventions is important. It should be frequently referred to and updated as necessary.

2.5 Challenges in DM

Managing data has its own set of challenges. These need to be understood and addressed properly for effective DM. Here are some key challenges [68]:

1. **Copyright:** Figuring out who owns the copyright of data in research is tricky. It involves a lot of people like researchers, the people who collect data, data analysts, universities, and funding agencies. For instance, in Germany, the law states that the creator of a work owns the copyright and can decide how it is used by others (§31 of the German Copyright Act (UrhG) [12]). But in research, it is often a team effort to collect and analyze data. So, deciding who owns the copyright can be complex. DM needs clear rules about who owns the copyright in these cases.

2. **Data Licensing:** Deciding how to license data is another challenge. This involves setting rules on how the data can be used, shared, or changed by others. The type of license needed can vary depending on what kind of data it is – like text, numbers, or images.

3. **Wrong Interpretation of Data:** When data is shared without enough information on how it was collected or the conditions it was collected under, it can be misunderstood or misused. This can lead to wrong conclusions, even if the data is authentic. To reduce this risk, it is better to share data with a detailed explanation of how it was collected. This is often done in research papers or reports.

4. **Data Privacy:** In some studies, like those involving personal information about patients, data privacy is crucial. This kind of data is sensitive and needs to be handled very carefully to protect people's privacy. Tools and systems, like those offered by Dataverse [28], can assist in keeping this data private.

5. **Changing the Mindset:** One of the biggest challenges is changing how researchers, data owners, or providers think about sharing their data. Traditionally, researchers focus more on publishing papers than on sharing the data they used. But now, there's a growing understanding that sharing data is important for advancing research. Convincing researchers to share their data openly for others to use is a major shift in the research culture. In some countries, universities are starting to require researchers to publish some work before getting a PhD, but there is still no system for sharing the data from these studies. Steps need to be taken to encourage data sharing in research.

2.6 Exemplar Cases of DM

In every research project, data plays an important role. It is carefully collected, analyzed, organized, and used to conduct studies. Without accurate and reliable data, research in any field, including the energy sector, would not be possible. More and more, sharing data is becoming a common practice, sometimes even required by publishers and research institutions. For instance, the Public Library of Science (PLoS) demands that authors provide access to their research data when they submit articles for publication [68]. Sharing data has several benefits: it allows others to reuse data for different purposes, it gives credit to those who collect and analyze data, it increases trust in the data, and it brings transparency to the research process, saving time for researchers who can build upon existing data instead of starting from scratch [68].

DM is crucial in this context. It ensures the effective organization, preservation, and sharing of research data, whether it is measurements, survey results, or field notes. Proper DM leads to numerous benefits. It increases the visibility of research, often boosting the citation rate of associated publications. Funders and journals frequently require a comprehensive DM plan to ensure the data's accessibility and reusability, thereby maximizing the research's societal impact. DM also maintains research transparency, crucial for validating findings and motivating further studies. Additionally, it significantly reduces the risk of data loss through reliable storage solutions and sustainable file formats, ensuring long-term preservation and reusability of data. These practices are integral to good scientific practice, maintaining the integrity and credibility of research [36].

DM plays a critical role in both academic research and practical applications. A practical example of this importance can be seen in the Soil Transformations in European Catchments (SoilTrEC) project [33]. SoilTrEC, funded by the European Commission, was aimed at understanding soil processes within Earth's Critical Zone (CZ). The project was not only about collecting data but also about managing and modeling it effectively. The success of SoilTrEC was largely dependent on the delivery of key datasets and the development of tested modeling codes. One of the achievements of the project was the development of a new comprehensive soil model, which was designed to monitor soil function within the European Union. This model relied heavily on data collection, management, and modeling. It connected four EU field sites for in-depth soil process studies (e.g., the Damma Glacier Critical Zone Observatory (CZO) in Switzerland). These sites provided valuable datasets that were enhanced by targeted process studies to validate an integrated model of soil processes. SoilTrEC's research methodology shows a comprehensive plan that incorporated DM practices at many stages. This project was not confined to Europe alone; it included collaboration with partners from Asia and the United States, emphasizing the global nature of soil sustainability research [33, 59]. The application of DM in SoilTrEC covers several essential aspects [24]:

- **Data Collection and Organization:** The project involved extensive fieldwork and laboratory experiments across multiple sites, demanding systematic data collection and organization.
- **Data Storage and Backup:** Data was stored in secure databases or repositories, with regular backups to prevent data loss.

- **Data Sharing and Collaboration:** Given the collaborative nature of SoilTrEC, efficient data sharing among various partners was essential. DM facilitated this aspect, ensuring effective collaboration.
- **Metadata Creation:** The creation of standardized metadata was a critical part of the project, simplifying data understanding and reuse.
- **Long-Term Preservation:** For a lasting impact, SoilTrEC's DM strategies ensured the long-term preservation and accessibility of its research data.
- **Ethical Considerations:** Adhering to ethical standards, particularly in handling sensitive data, was a key component of the project's DM.

2.7 Summary of the Chapter

This chapter lays the foundation for understanding the fundamentals of DM. It discusses the significance of managing data effectively to enhance research integrity and facilitate practical applications in energy management and sustainable practices. The chapter provides insights into the benefits of data sharing and the implications of proper DM in increasing the visibility and credibility of research work. By exploring the Soil Transformations in European Catchments (SoilTrEC) project, it offers a practical example of how DM can be effectively implemented in a research project. Additionally, it addresses the challenges, practices, and essential concepts in DM, providing a comprehensive understanding of the role of data in research and its broader impact on scientific discovery and application.

Chapter 3

Related Work

This chapter reviews related work in the field of data management (DM), exploring the contributions and findings of other scholars and professionals. It is structured into three sections to provide a comprehensive understanding of DM's current state and its application in various domains. The first section presents work that offers a broad overview of DM for research, focusing on general practices, challenges faced, and recent advancements in the field. The second section examines work in DM within the medical sector, a field that, like the energy sector, deals with complex data management needs. This comparison helps to understand the problems and solutions in managing data in different sectors. Finally, the chapter concludes by concentrating on work related to DM in the energy sector, aiming to provide a clear picture of the current state and developments in this specific area.

3.1 DM for Research

The paper on *Research Data Management (RDM) practices and services* [6] explores the effective management of research data. It highlights that RDM is becoming increasingly important in the academic and research communities, particularly among researchers and academic libraries. However, its practice is still limited, especially in developing countries. It faces challenges that require the development of specific skills and active collaboration among various stakeholders, including university service departments. Policy formulation for RDM is often inadequate, especially in developing countries, often due to insufficient collaboration between higher education institutions, research boards, funding agencies, and higher education commissions. To address these challenges, the paper recommended that stakeholders collaborate to make it mandatory for researchers to deposit their data in institutional or subject repositories and publish in open access journals. Additionally, there should be procedures to address data ownership concerns and to ensure proper acknowledgment and citation of the work of previous researchers.

3.2 DM in the Medical Field

The first paper, titled *Functional Requirements for Medical Data Integration into Knowledge Management Environments* [50] addresses the challenge of isolated, non-interoperable data silos in patient care data management. It focuses on the need to transition from segregated data systems to unified information architectures. This transition is critical for effectively integrating medical data from various levels and sources. The authors employed a systematic two-step approach to determine DM requirements in the medical field. They began with a web-based systematic literature review using PRISMA [71] guidelines to identify relevant articles on medical data integration requirements. Next, they applied a document-based requirement elicitation approach to the identified literature, using reference management software for data extraction and conducting a thorough review. The final stage involved categorizing the requirements using the Data Lake Life Cycle model [46], covering aspects like data acquisition, processing, storage, and more. This process involved evaluating data, consulting with experts, and forming a unified set of requirements.

The second study, *Why It Takes a Village to Manage and Share Data* [9] offers a detailed look at data sharing in the biomedical field, especially in light of the United States National Institutes of Health's new data management and sharing policy. It emphasizes the need for collaboration among various stakeholders, including scientists and academic institutions, to build and sustain effective data sharing infrastructures. It discusses DM challenges for principal investigators, the impact of data nature on management practices, and the effect of knowledge infrastructures on community data sharing. The paper also highlights the roles of university leadership, computing, libraries, and departments in data management and calls for improved investment in data management infrastructure. Furthermore, it examines the stakeholder network, international privacy laws, and policies affecting data sharing. The study recommends a collective data sharing approach, urging shared responsibilities and continued support from funding agencies for data repositories, addressing biosciences' scaling data production challenges.

The third paper, *Research Data Management and Data Sharing for Reproducible Research—Results of a Community Survey of the German National Research Data Infrastructure Initiative Neuroscience* [51] investigates the challenges in neuroscience DM. It concentrates on the growing volume and complexity of data and the necessity for effective DM solutions. The main purpose of this work is a survey conducted by NFDI-Neuro [62] within the German neuroscience community to understand the current DM state and identify challenges, needs, and opinions.
The survey revealed significant gaps in data and metadata standards, provenance tracking methods, and secure data infrastructure. It also pointed out a lack of DM knowledge and skills among researchers and resource constraints. Despite these issues, there was a strong inclination within the community to share data and enhance DM capabilities. They also suggest a systematic approach to developing DM standards, tools, and infrastructure, stressing the importance of training, education, and additional resources. Continuous engagement with stakeholders, including policymakers, is crucial for a cultural shift in data management and sharing.

3.3 DM in the Energy Sector

The paper *Data Management in Energy Communities* [14] discusses the integration of energy communities into European law and their transposition into national laws. It highlights the lack of focus on data processing and protection within these communities. The paper emphasizes the need for frequent access to energy data for allocation and billing purposes, which raises privacy concerns. It also discusses the implementation of IT systems for energy allocation and billing, the responsibilities of the Distribution System Operator (DSO), and the need for compliance with the General Data Protection Regulation (GDPR). The paper concludes by noting the importance of considering data acquisition, usage, and privacy in the development and operation of energy communities.

Another work reviewed is the *DOE Policy for Digital Research Data Management* [94] that focuses on managing digital research data to support the U.S. Department of Energy (DOE)'s mission. This policy covers the entire lifecycle of data, including capture, analysis, sharing, and preservation, with a particular focus on data sharing and preservation. The policy applies to unclassified and otherwise unrestricted digital research data produced fully or partly with DOE funding. The policy outlines principles, roles, responsibilities, requirements, and guidance for various stakeholders, including DOE research offices, respondents to DOE research funding solicitations, and recipients of DOE research funding. Data management plans are central to this policy, detailing how data will be shared, preserved, and made publicly accessible. The policy also addresses the protection of confidentiality, personal privacy, and compliance with legal and policy requirements.

3.4 Summary of the Chapter

This chapter reviews the work done in the field of DM, exploring varied approaches and challenges across different sectors. It highlights the increasing recognition of DM's importance in academic and professional circles. The discussion points out the depth of research and literature available in the medical sector, where DM is extensively explored. Studies in this area address complex data integration challenges and emphasize the need for collaborative data management and robust data sharing infrastructures. The chapter also touches upon DM in the energy sector, where there is a noticeable exploration of how DM practices are applied and the specific challenges faced. The comparison suggests that while significant advancements have been made in DM, especially in medicine, there is a continuous need for research and development in other sectors like energy to fully leverage DM's potential.

Chapter 4

Methodology

This section outlines the methodology followed in this thesis, starting with a systematic literature review to identify key energy data management requirements. This is complemented by a survey conducted with experts in the energy field, aiming to validate preliminary findings from the literature with empirical data, which consists of firsthand observations and responses gathered directly from these experts. Following this, the thesis will focus on the practical application of these findings by introducing a proof of concept. This involves the development of an extension specifically designed to implement one of the requirements identified. This implementation is designed to help to reflect a data ecosystem, tailored to the needs of the energy sector. The goal of this approach is to validate the top 10 most frequently mentioned requirements collected from the literature by surveying domain experts, thereby contributing to the field of energy data management.

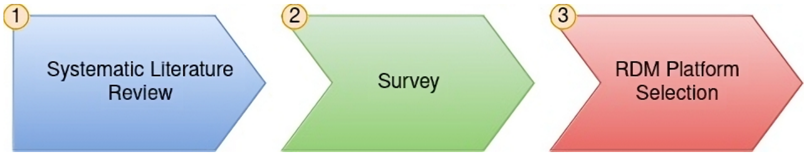


Figure 4.1: Methodology of the Thesis

4.1 Systematic Literature Review Process

This section explains the methodology used for conducting the systematic literature review. The process follows the following steps: a) formulating the research question, b) developing inclusion and exclusion criteria, c) conducting a comprehensive search, d) screening and selecting studies, e) extracting data and assessing quality, f) synthesizing and analyzing data, g) interpreting findings, and finally, h) reporting the outcomes. Each step is integral to the overarching objective of this review, which is to uncover the critical data management requirements within the energy sector, thereby contributing to its efficient and effective management [67].

Formulating Research Questions

The first step in the systematic literature review process is the articulation of the research question it seeks to answer. For this study, the question is formulated as follows: *What are the requirements of DM in the energy sector?*

Developing Inclusion and Exclusion Criteria

Prior to initiating the search, a set of inclusion and exclusion criteria was developed to systematically filter the literature. These criteria are important in ensuring that the search results are relevant and aligned with the objective of identifying the requirements of DM in the energy sector. These criteria will be applied to assess and filter the articles retrieved through the search query, ensuring a focused and relevant selection for review. The criteria are as follows:

- **Relevance to Energy Sector:** Articles should have a clear focus on the energy sector, such as wind energy or smart grids. Those where energy is mentioned but is not the main focus should be excluded, particularly those that focus on unrelated fields such as transportation or healthcare.

- **Focus on Big Data and Data Management:** Articles should emphasize big data and data management challenges and solutions relevant to the energy sector.

- **Avoidance of Narrow Scope:** Exclude articles with too specific or narrow scope that can not be generalized for the energy sector's broader DM needs, such as those focusing on AI or machine learning without direct relevance to energy.

Conducting a Comprehensive Search

For the literature collection, SCOPUS was chosen due to its comprehensive interdisciplinary coverage and extensive database (over 90M records) [75], aligning well with the cross-disciplinary nature of the thesis. It encompasses a broad range of fields, including the energy sector and data management, and is trusted for its reliable content, sourced from over 7,000 publishers and vetted by an independent Content Selection and Advisory Board [75]. Its global coverage and detailed citation metrics provide valuable insights into the impact and trends of research in these fields [52]. The search conducted on Scopus targeted articles related to *energy*, *requirements*, and *data management* from 2020 to 2023. Also, the articles needed to be in English with open access. Then, a specific filter for *Big Data* as a keyword was added to the query. This approach initially identified 102 articles, as determined by a search conducted on the 27th of November, 2023. The following query was used to conduct this search:

Search Query

```
(ALL (energy) AND ALL (requirements) AND ALL (”data management”)
AND PUBYEAR > 2019 AND PUBYEAR < 2024
AND (LIMIT-TO (DOCTYPE, ”ar”[*]))
AND (LIMIT-TO (LANGUAGE, ”English”))
AND (LIMIT-TO (OA[†], ”all”))
AND (LIMIT-TO (EXACTKEYWORD, ”Big Data”))
```

Screening and Selection of Studies

From the 102 identified articles, a screening and selection process started, initially focusing on identifying and removing duplicates and ensuring the basic relevance of the articles. This first level of screening was essential to ensure that each article was considered only once in the review process and to quickly ascertain the potential relevance based on titles and abstracts. After this initial screening, the articles underwent a more detailed review. At this stage, the inclusion and exclusion criteria previously developed were applied to check the content of each article's abstract. This phase was critical for assessing whether the articles clearly focused on the energy sector's data management challenges, emphasizing big data and solutions. Articles that did not align with these criteria were excluded, maintaining the review's relevance and quality.

A table named *decision_articles.xlsx*, containing detailed information on each article, is available in a Leibniz Data Manager [8] (LDM) entry. The table can be accessed via the following link [57]. It includes each article's title, along with information on whether it was included or excluded, and the reasons for each decision regarding inclusion or exclusion. Additionally, Table 4.1 provides an excerpt of that table for reference.

[*] DOCTYPE refers to the type of document; ’ar’ stands for article.

[†] OA denotes open access.

Table 4.1: Portion of the Selected Articles from *decision_ articles.xlsx*

Titles	**Status**	**Reason**
Big data challenges in overcoming China's water and air pollution: relevant data and indicators	Include	Focuses on environmental management in China, specifically water and air pollution, with the need of data management and reporting
Towards a Service-Oriented Architecture for the Energy Efficiency of Buildings: A Systematic Review	Include	Direct relevance to the energy sector, discussing energy efficiency in buildings and the application of big data in this context
Magnetic Force Classifier: A Novel Method for Big Data Classification	Exclude	Focuses on a machine learning classification method for big data, without specific application to DM in the energy sector
Efficient deadline-aware scheduling for the analysis of Big Data streams in public Cloud	Exclude	While addressing big data in cloud environments, the article does not specifically relate to the energy sector or its unique data management challenges
Advancing manufacturing systems with big-data analytics: A conceptual framework	Include	Concentrates on manufacturing systems, and the challenges of handling large amounts of complex data (big data)
Towards Energy-Efficient Framework for IoT Big Data Healthcare Solutions	Exclude	Focuses on healthcare solutions using IoT and big data, which is outside the scope of the energy sector's DM

After filtering the abstracts, 36 articles remained. An illustration of the selection process can be found in Figure 4.2, with n being the number of articles.

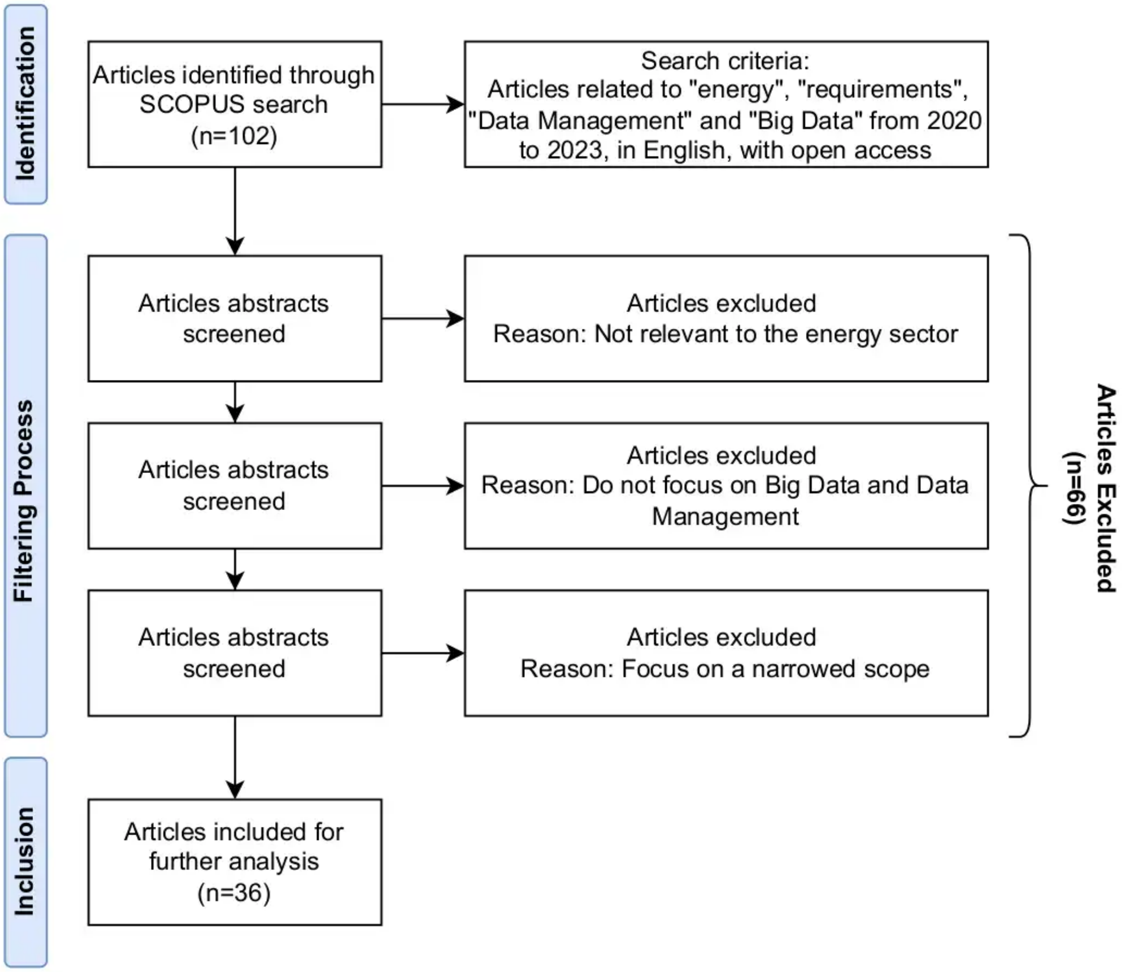

Identification
Filtering Process
Inclusion
Articles identified through SCOPUS search (n=102)
Search criteria: Articles related to "energy", "requirements", "Data Management" and "Big Data" from 2020 to 2023, in English, with open access
Articles abstracts screened
Articles excluded Reason: Not relevant to the energy sector
Articles abstracts screened
Articles excluded Reason: Do not focus on Big Data and Data Management
Articles abstracts screened
Articles excluded Reason: Focus on a narrowed scope
Articles Excluded (n=66)
Articles included for further analysis (n=36)

Figure 4.2: Steps of the Literature Review

Data Extraction and Quality Assessment

This step deals with the extraction of key DM requirements from the 36 selected articles, emphasizing the identification of challenges and solutions in energy data management. For quality assessment, each article was examined for its contribution to understanding the field, ensuring that the data extracted, such as main challenges, solutions, publication details, and keywords, met the research's intrusion criteria.

This examination guaranteed that only relevant and needed information was summarized and included in the analysis.

Data Synthesis and Analysis
Following extraction, the synthesis involved categorizing the extracted data to further organize and analyze these requirements. This step is important not just to facilitate the analysis of requirements, but also to help in managing the complex and multifaceted nature of energy data management, enabling a clearer and more structured analysis from initial data gathering to ensuring final security measures [84].

Building on this approach, this thesis categorizes the gathered requirements following the Data Lake Life Cycle model [46]. This model provides a clear and practical way to understand how data moves and changes in systems like data lakes and data warehouses. It looks at all the important parts of managing data throughout its life cycle, such as collecting data, processing it, analyzing it, storing it safely, managing (meta)data details, keeping track of data origins and changes (traceability and lineage), and keeping the data secure [50]:

- **Data Acquisition:** Involves the gathering, filtering, and cleaning of raw energy data, essential for the initial stages of data management.
- **Data Processing:** Focuses on the collection and manipulation of data to produce usable information.
- **Data Analysis:** Deals with exploring, mapping, and modeling data to extract relevant insights and information.
- **Data Storage:** Concerns the methods and practices of securely storing energy data (e.g., deletes, backups, etc.).
- **Data Lineage:** Tracks the origins, history, and evolution of the data throughout its life cycle.
- **Data Traceability:** Ensures the history, location, and application of data are verifiable and transparent.
- **Data Security:** Focuses on protecting data from unauthorized access, corruption, or breaches.
- **Metadata Management:** Involves the creation and management of metadata to enhance data retrieval and understanding.

In addition to the categories derived from the Data Lake Life Cycle model, two other categories were added, influenced by some specific requirements extracted from the literature:

- **Human Skills and Expertise in Data Management:** This category was introduced because the need for skilled professionals who can effectively manage and handle data within the energy sector was mentioned in four different articles.
- **Data Governance and Compliance:** This category was added because it fits better some specific requirements mentioned in several articles, such as the need for regular and impromptu audits to ensure compliance and quality in energy data practices, as well as strict penalties for any manipulation or falsification of energy data.

Interpretation of Findings

Following the detailed review and extraction of data from the 36 selected articles, key DM requirements in the energy sector were extracted. To facilitate the interpretation of these findings: A table was created to summarize the findings, listing the title of each article, the main DM challenges in the energy sector, main keywords, publication date, location, and article keywords. It is also available at this LDM entry [57] under the name *included_ articles.xlsx*. For quick reference and overview, a portion of this table is included in Table 4.2.

Table 4.2: Portion of the Included Articles from *included_ articles.xlsx*

(a) Titles, Challenges, and DM Requirements

Titles	**Challenges Identified**	**DM Requirements**
Big data challenges in overcoming China's water and air pollution: relevant data and indicators	- Scattered data across platforms - Lack of centralized data exchange - Insufficient EIA [43] disclosure - Data quality and transparency issues	- Centralizing data management (for data transparency, improve standards and security) - Electronic reporting of pollution data - Data sharing from trustworthy sources - Supervision and auditing (periodic and unscheduled) - Penalties for data falsification - Metadata inclusion for interpretation and reproducibility
Big Data Management in Smart Grids: Technologies and Challenges	- Sensors' data sometimes is updated and overwritten discarding the previous data - Data is generated with a precision of seconds resulting in Terabytes of data, and the analytical value per unit of data is low - The data communication requires the high-volume of data to be compressed before flow	- Integration of data as part of the operational process of smart grids is necessary - Robustness and a strong disaster recovery plan are needed for database management systems to handle the increasing data size and ensure data integrity - Efficient data compression methods - Effective storage and transfer mechanism needed to handle large amounts of data

(b) Keywords, Publication Date, Location and Article Keywords

Main Keywords	**Publication Date**	**Location**	**Article Keywords**
Data management requirements	8 March, 2021	China, USA	Indicators, Monitoring, Assessment, Data quality, Pollution management
Data management Requirements Energy	14 May, 2021	USA, Qatar	Apache spark, Big data, Data mining, Hadoop, Indexing, Management process, Smart grid, Stream mining

To further illustrate the findings from the table that examines the 36 selected articles, several plots have been produced. These plots provide different insights into various aspects of the articles:

1. **Geographic Distribution of the Authors (see Figure 4.3):** This histogram illustrates where the research contributions are originating from, highlighting the international nature of the energy data management discourse, with the USA at the forefront with 6 articles, followed by China and Germany with 5 articles each.

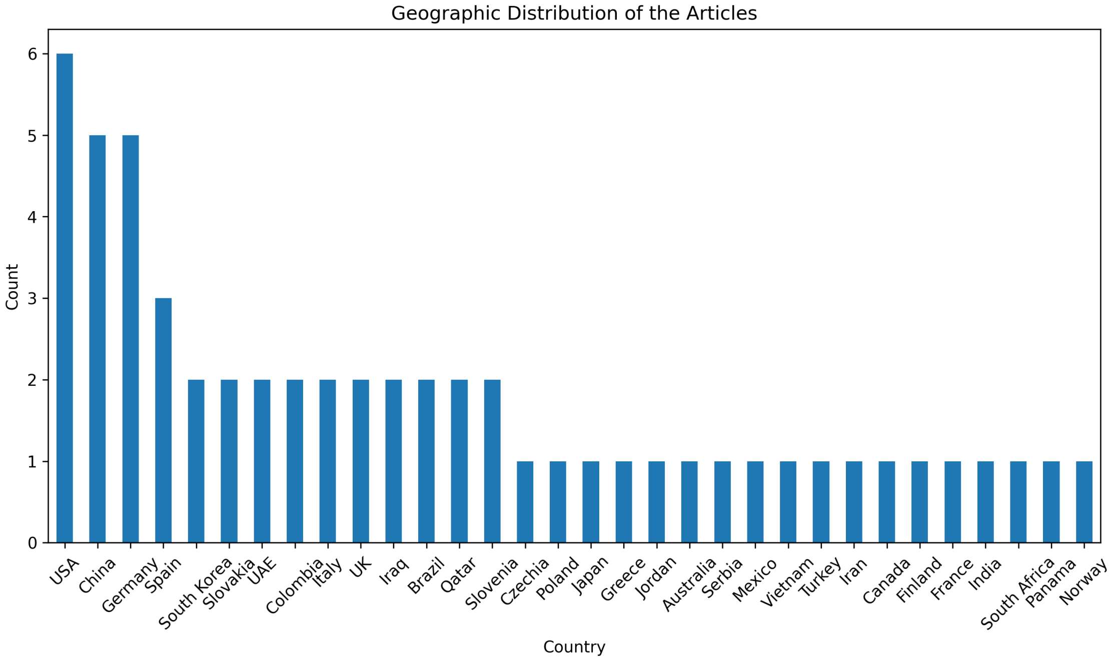


Figure 4.3: Geographic Distribution of Authors

2. **Publication Year Distribution (see Figure 4.4):** This histogram presents the number of articles published per year within the selected range, revealing that most of the chosen articles were published in the year 2020 with 14 articles.

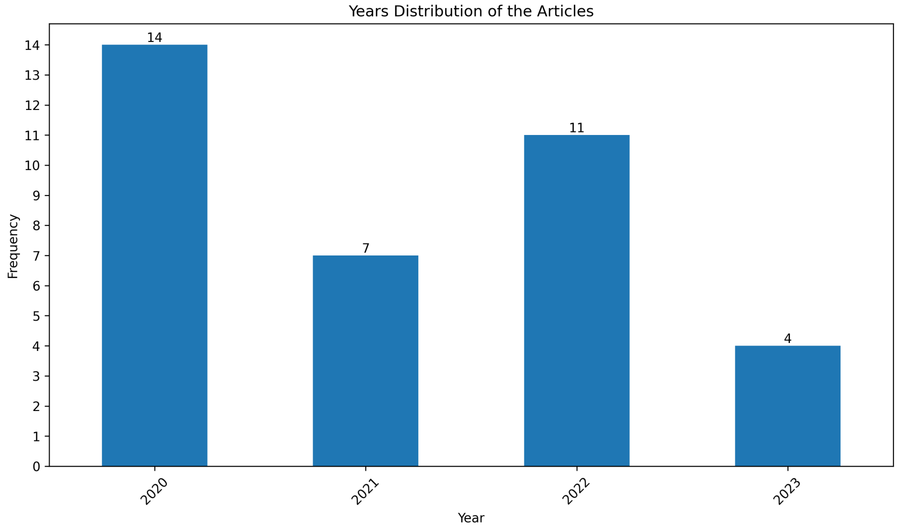


Figure 4.4: Publication Year Distribution

3. **Keyword Frequency (see Figure 4.5):** This histogram focuses on the occurrence of keywords that were added in the search query, it doesn't show a high difference between the different distributions with *Requirements* been mentioned in 31 articles, followed by *Data Management* in 30 and *Energy* in 28.

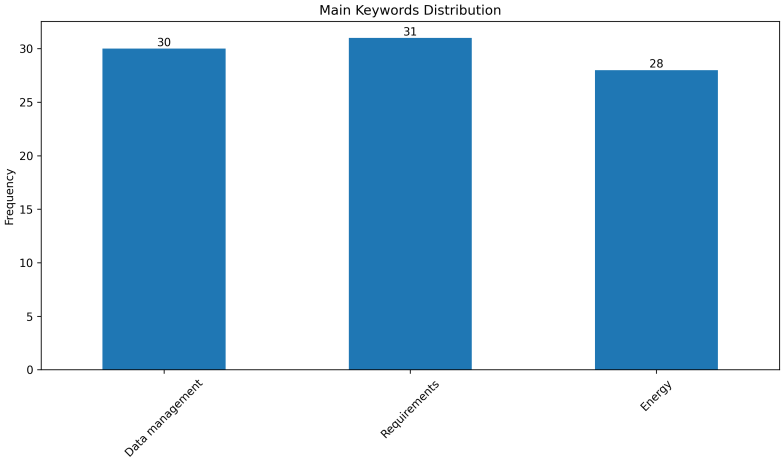


Figure 4.5: Search Query Keyword Frequency

4. **Distribution of Keywords Chosen by the Authors (see Figure 4.6):** The word cloud offers a visual representation of the keywords selected by the authors, with the size of each word proportional to its frequency. This visualization helps to quickly identify the core topics and terms that define the research landscape.

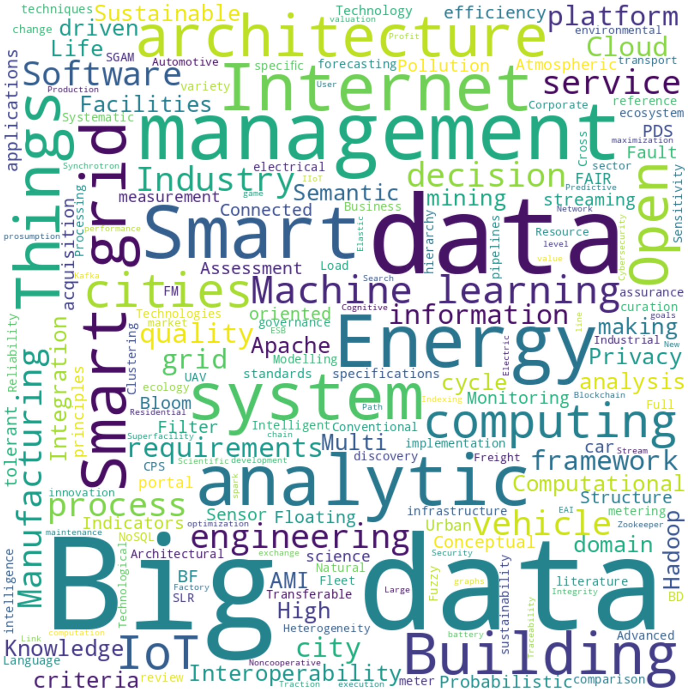


Figure 4.6: Distribution of Keywords Chosen by the Authors

A total number of 103 requirements were extracted at the end of the literature review process. However, upon closer examination, it was apparent that many of these requirements overlapped or were semantically similar. Therefore, they were merged into a refined set of 47 distinct requirements.

With these additions, the thesis categorizes the 47 refined requirements into 10 distinct categories. To gain a complete overview, detailed documentation of these categories, including lists of the requirements, their respective categories, the articles in which they are mentioned, and the frequency of their mentions, can be accessed via the following link [57].

Reporting
The systematic literature review results in the identification and categorization of key data management requirements in the energy sector. Through the examination of 36 selected articles, this process extracted 47 unique requirements organized into 10 categories, offering a structured insight into the sector's challenges and needs. The primary findings are documented in files available via the LDM entry mentioned above. These resources detail the identified requirements, their categorization, and the supporting literature. Following this, the study narrows its focus to the top 10 most frequently mentioned requirements. This selection serves as the foundation for a subsequent survey aimed at engaging experts within the energy sector. The survey seeks to validate these key requirements through expert opinions, ensuring the relevance and applicability of the findings in this sector.

4.2 Survey

The survey forms an important element of this work, specifically designed to build upon and validate the findings from the systematic literature review. Its purpose is to collect empirical data from professionals in the energy sector, with the aim of validating and enhancing the findings from the literature review. Designed to target energy sector professionals with any level of data interaction, IT specialists in energy companies, and researchers involved in energy-related projects. It aims to gather insights and validate the identified requirements through the perspectives of professionals in the field. The survey is structured into three sections, it delves into various aspects of DM in the energy sector:

1. **Professional Background and Familiarity with DM:** This section consists of five questions. It gathers information about the respondents' affiliations, roles, and areas of expertise in energy. Additionally, it assesses their familiarity with data management, establishing a baseline for their understanding and experience in this field.

2. **Importance of DM Requirements:** Consisting of ten questions, this section focuses on evaluating the importance of the top 10 requirements most frequently mentioned in the literature (refer to Table 4.3). Using a Likert scale, respondents rate the importance of each requirement within their professional roles, offering insights into the practical applicability of these requirements.

3. **DM Tools, Standards, and Challenges:** The final section, with four questions, explores the specific tools and standards used in data management tasks,

the use of public repositories, data sharing practices, and challenges faced in managing data.

Table 4.3: Top 10 Requirements with Most Sources (all the articles, with their enumerations, can be found in the following LDM entry [57])

Requirement	Articles	Count
Implementing security and privacy measurements to safeguard sensitive energy information	7, 8, 9, 12, 14, 15, 17, 21, 25, 26, 31, 32, 34, 36	14
Ensuring immediate (high data rates) and simultaneous processing of energy data for timely and efficient decision-making	articles 8, 12, 13, 15, 16, 19, 20, 21, 22, 23, 24, 25	12
Integrating various types of energy data from different sources	3, 5, 6, 7, 8, 9, 10, 11, 12, 13	10
Processing a large volume of data from different energy systems	5, 9, 10, 16, 20, 21, 23, 26, 27	9
Adapting and growing efficiently to meet increasing demands in energy data systems (Scalability)	5, 7, 11, 15, 17, 27, 29, 34, 35	9
Comprehensive Metadata for interpretation and reproducibility	2, 5, 18, 19, 30, 33, 34	7
Developing tools for clear visualization and interpretation of energy data	4, 15, 16, 18, 31, 32	6
Monitoring the quantity and quality of data	3, 4, 16, 18, 26, 33	6
Standardizing data models and incorporating data analysis to ensure consistency and relevance of data	7, 8, 11, 16, 26	5
Identifying and dealing with anomalies (e.g., outliers, duplicates, etc.) in preprocessing phase	4, 5, 13, 24, 25	5

To facilitate a structured and quantitative analysis, the survey predominantly employs multiple-choice questions and Likert scales, allowing respondents to quantify their opinions and provide measurable insights. It was conducted over the span of one month in February 2024 and is estimated to take 10-15 minutes to complete. Google Forms was selected as the platform due to its ease of use and its capability to automatically analyze data. As a free tool, it facilitated a quick setup of the survey and direct analysis of results [82]. The survey was distributed to experts in various energy fields such as solar, geothermal, and wind energy, as well as IT specialists in energy companies, and researchers involved in energy-related projects. Ethical considerations were a priority, therefore participants were informed about the academic purpose of the survey and assured of confidentiality and anonymity. The survey responses were continuously monitored to ensure diverse representation from different areas of expertise within the energy sector.

4.3 DM Platform Selection

After extracting the requirements from the literature review and validating a subset of them through the survey, the next step involves developing a proof of concept that implements a crucial requirement identified: *collecting metadata from different repositories*. The goal of this step is to extend a data manager to address a specific requirement identified for the energy sector. The efficacy of DM practices is significantly influenced by the choice of an appropriate DM platform. As the energy sector continues to evolve, the need for platforms capable of handling complex and diverse data becomes crucial.

In this context, CKAN (Comprehensive Knowledge Archive Network) emerges as a strong candidate, due to its widespread use, large developer community, and adaptability to meet the evolving needs of the energy sector. CKAN's capabilities in handling diverse data types, its support for non-standard metadata, features like faceted search and dataset versioning, and its modular extendable design make it a good choice for customizing a data manager for specific sector requirements [2].

Building on CKAN's foundation, the Leibniz Data Manager (LDM) [8] offers a more tailored solution. Designed to address the broader challenges in DM, LDM enhances the lifecycle management of data and research data. It adheres to the FAIR principles [100], supports various types of research digital objects (RDOs) [81], including datasets in different formats, data services demonstrated as live code using Jupyter notebooks [49], and data visualizations. The platform allows researchers to manage, analyze, and cite RDOs effectively, providing options to generate Digital Object Identifiers (DOIs) [29] for each uploaded RDO. LDM enhances CKAN's functionalities by importing RDOs from other repositories, maintaining updated metadata through synchronization, and creating a comprehensive knowledge graph. For a more detailed exploration of CKAN and LDM, including their architecture, features, and specific capabilities, refer to Chapter 5 of this thesis.

4.4 Summary of the Chapter

This chapter presents a structured methodology for the thesis, it begins with a systematic literature review, focusing on identifying key DM requirements in the energy sector. This step is complemented by a survey among energy field experts, validating the literature's findings with empirical data. The next step involves applying one of these requirements in the development of a proof of concept, implementation of an LDM extension. The chapter details each methodological step, underscoring their significance in the broader research context.

Chapter 5

Data Manager

This chapter discusses the chosen data manager for this thesis, the Leibniz Data Manager (LDM). To understand it properly, it is important to first introduce the Comprehensive Knowledge Archive Network (CKAN). Since LDM is built on top of CKAN, a comprehension of CKAN's architecture and features is crucial for understanding the LDM extension that is implemented in this thesis (discussed in Chapter 6). CKAN is the foundational framework for LDM and serves as its interface. However, LDM benefits from CKAN's ability to be extended and its flexibility to upgrade from a data manager into a knowledge-driven data manager [81]. Therefore, the following sections will explore CKAN's architecture, data organization, and flow mechanisms, followed by an explanation of LDM's features and capabilities for data management.

5.1 CKAN

CKAN is an open-source data management system. It is distinguished by its large community of over 240 contributors [16], more than 267 registered extensions [21], and over 2000 active CKAN instances [19], which significantly enhance its capabilities. These extensions allow for a tailored experience to meet specific needs, including advanced visualization tools, document previews, custom themes, various storage options, efficient link management, and advanced metadata management. It is globally recognized for its adaptability, powering over 30 major government sites worldwide with a 100% open-source codebase on GitHub [18]. Some examples of governments that use CKAN for their DM needs include the Singapore Government, employing it as an open data portal; the Australian Government, using it to manage data from over 800 organizations; and the Governments of Canada and the United

States, which have adopted CKAN for their open data catalogs. Another interesting instance is Open Africa, a platform aiming to be the largest repository of data on the African continent [18]. These instances highlight CKAN's versatility and its ability to efficiently manage diverse data requirements. Additionally, these portals manage petabytes of government data and provide access to millions of datasets, with over 15 million datasets accessible through these platforms [18].

The platform's ease of use and large community contribute to its popularity, making it a preferred choice for open data initiatives. It offers a rich set of features, including but not limited to:

- **Data Publishing and Sharing:** Facilitates easy publication and sharing of datasets [26].
- **User Experience Enhancements:** Offers keyword autocomplete, translation functions, and tagging for datasets [17].
- **Security and Governance:** Implements authority management through access controls and license management [26].
- **Dataset Harvesting:** Enhances dataset discovery and collection from multiple CKAN instances [26].
- **Federated Network:** Supports the establishment of a network of data portals that interconnect and share data [20].
- **File Storage:** Allows for the uploading of media and image files, with storage options on the server or in the cloud through extensions [17].
- **Geospatial Capabilities:** Provides advanced features for geospatial data, including preview, search, and discovery functionalities [26].

5.1.1 Architecture

CKAN's design is modular, extensible, and scalable [30]. Its scalability is not only a technical advantage but also aligns with the crucial requirements of the energy sector, highlighted in the literature review. It facilitates the customization or expansion of features through extensions, aligning with the dynamic needs of DM in this field.
It fundamentally employs Flask [65] as the primary web framework, and the back-end is primarily written in Python [73], which handles data processing, authentication, and other server-side tasks. The user interface is created using JavaScript [45], which

enhances the user experience with dynamic interactions and real-time updates. The data, organized in PostgreSQL [70] databases, allows for structured storage and efficient retrieval. CKAN's data model includes various tables, such as those for datasets, resources, organizations, and users, facilitating complex queries when users search for specific datasets.

The search functionality is powered by Apache Solr [5], an open-source search platform built on Apache Lucene that offers powerful full-text search capabilities, faceted navigation, and efficient indexing. Additionally, CKAN installations provide Web APIs [1] for external applications and services to interact with CKAN programmatically. These APIs offer endpoints for querying, creating, updating, and deleting datasets, resources, and other CKAN entities. Developers can use these APIs to integrate data from CKAN into their own applications, build custom dashboards, or automate DM tasks [3, 26]. Figure 5.1 provides a detailed visualization that illustrates this architecture further, highlighting both the structural design and the practical application scenarios of CKAN.

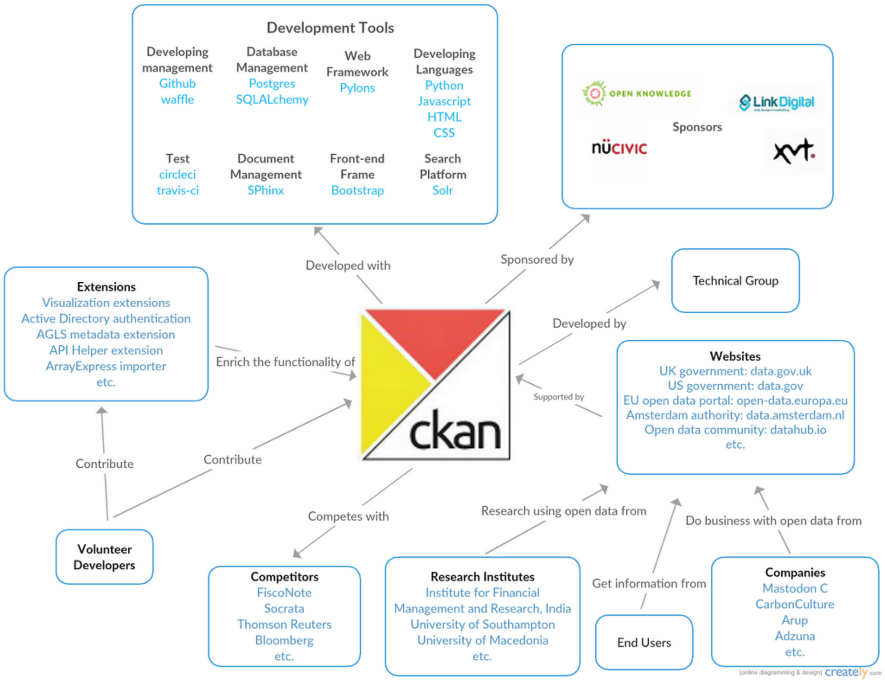


Figure 5.1: Architecture and Context View of CKAN [3]

5.1.2 Module Organization

To better understand how CKAN works, it is important to examine it at its module structure. CKAN's internal structure follows the Pylons web framework, which uses an architectural design pattern called Model-View-Controller (MVC) [92]. The framework is divided into four main layers [26]:

1. **Routes:** These define the connection between URLs and the corresponding Views within the system, directing requests from specific URLs to the appropriate Views that will handle and respond to these requests.

2. **Views:** They are responsible for receiving requests and generating responses. They use the *Action* function for reading and updating data, to manage the request-response cycle. Views also render the *Jinja2* [72] template for responses and employ the *Template Helper* for frequently reused or too complex to be included in the template itself.

3. **Logic:** This layer contains critical functions such as *Actions* and *Auth*, which are responsible for CKAN's internal operations. These functions also govern core functionalities and are accessible externally through API URLs with the same names as the *Action* functions.

4. **Models:** They handle the storage and retrieval of data objects in the*PostgreSQL* database using Object-Relational Mapping (ORM) [42] with *SQLAlchemy* [86] for efficient database interaction.

Fig 5.2 provides a more detailed view of the main layers of CKAN.

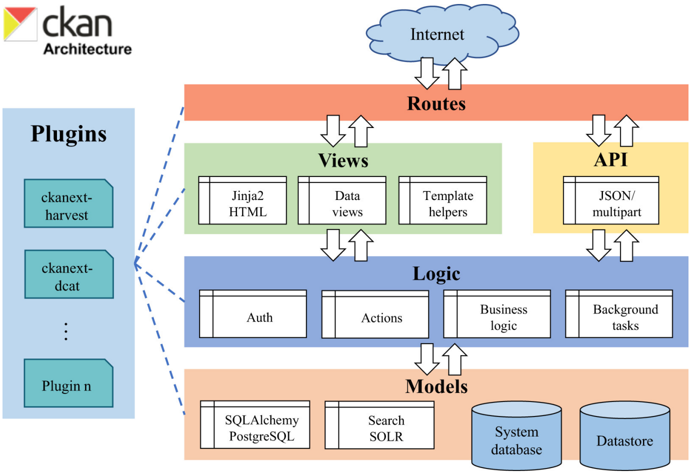


Figure 5.2: Modules of CKAN [26]

5.1.3 Data Flow

The data model utilized by CKAN consists of three primary entities: organizations, datasets, and resources. Organizations act as the highest level of categorization and allow for the grouping of datasets. Each dataset, representing a collection of related information, may contain multiple resources. These resources are either actual data files or links to data stored externally. CKAN uses the pull-push data flow mechanism. In this system, data producers (e.g., the creators or contributors of information) upload or "push" datasets into the platform, organizing them under relevant organizations and as part of specific datasets. On the other side of this mechanism, data consumers engage in a "pull" process. They utilize CKAN's interface to search for specific datasets, downloading them directly or accessing the data via CKAN's API. This dynamic between pushing new or updated datasets by producers and pulling required datasets by consumers ensures a continual flow and update of information within CKAN [3]. Figure 5.3 shows the complete data flow in CKAN.

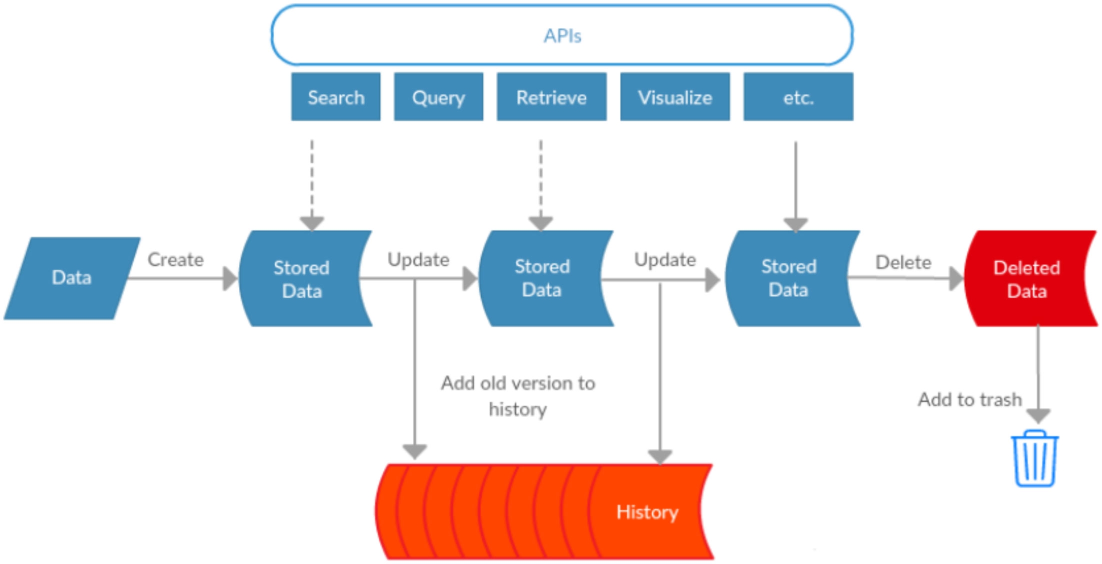


Figure 5.3: CKAN Data Flow [3]

5.1.4 Extensions and Code Organization

As mentioned earlier, the ability to extend and customize CKAN is a key feature of its design. In CKAN, extensions are used to modify or enhance the platform's default functionalities, following the unique needs of various projects or organizations. This adaptability is especially useful in scenarios that require custom dataset fields or integration with external databases. Understanding the structure of CKAN's source code is important for developing extensions. The main directory, labeled `ckan`, houses crucial components such as models, views, controllers, as well as modules for data migration and testing. Extensions implemented by users are stored in a separate directory named `ckanext`. This directory, along with others containing configuration files and binaries, forms the backbone of CKAN's customizable framework [3].

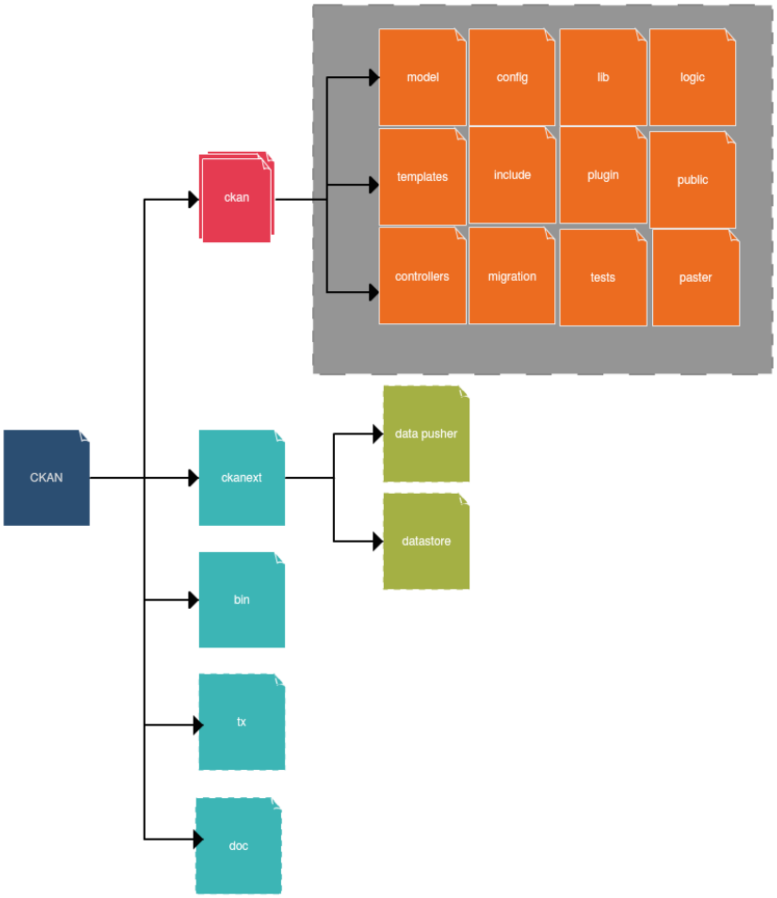


Figure 5.4: Code Structure of CKAN [3]

Maintaining the integrity and functionality of the code through testing is crucial. CKAN enforces rigorous testing protocols to ensure the reliability of both its core code and any new extensions. New or modified code, including that within extensions, must pass a set of predefined tests before integration into the main codebase [17].

5.1.5 Key Concepts Used in the Implementation

This part focuses on crucial concepts for understanding the implementation aspects of CKAN, particularly its extension mechanism and integration capabilities [17].

- **Plugin Interfaces:** These are instrumental in modifying its standard functionality. Extensions use these interfaces to interact with the core system, allowing for the creation of new web pages or altering existing functionalities. For instance, the IRoutes interface is utilized to change CKAN's routing, enabling the development of custom web pages.

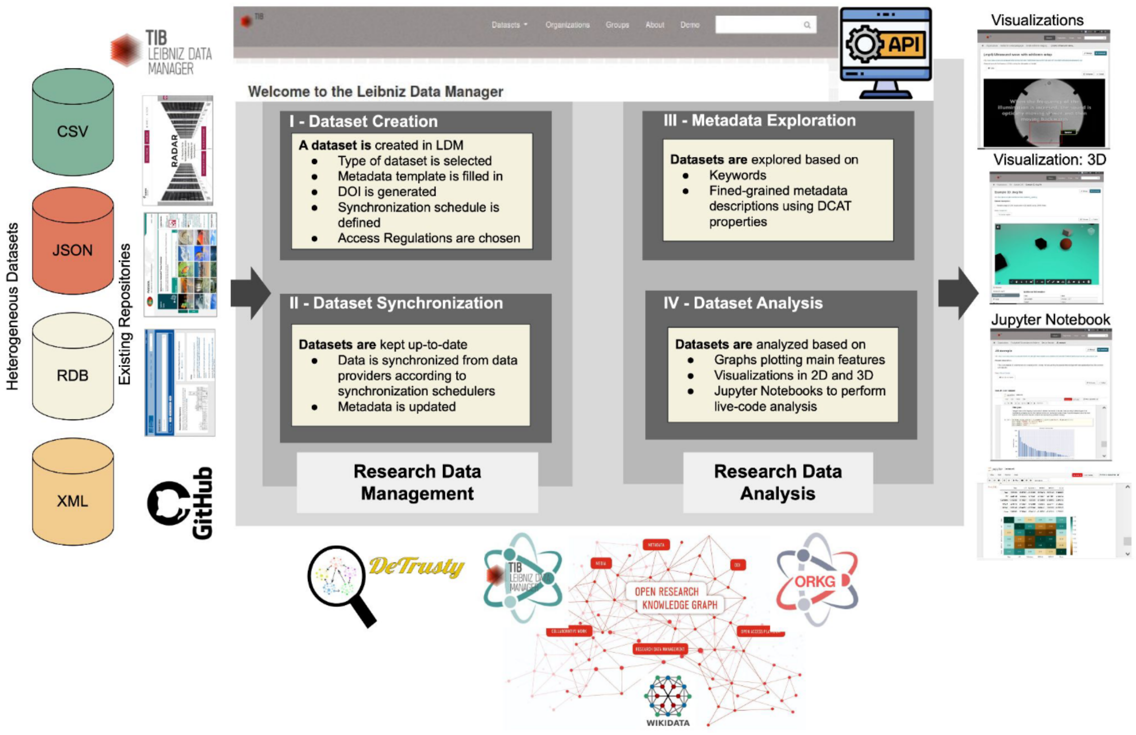
- **Plugin Toolkit:** This plugin is a flexible Python module, that provides developers direct access to a range of methods, classes, and exceptions from the core system. It includes a `get_action()` method, which extensions can use to invoke internal methods from CKAN's Action API. This API exposes CKAN's core operations for use by both clients and extensions, ensuring compatibility and safe usage across different versions of CKAN.

- **IBlueprint Interface:** It works with Flask, CKAN's underlying micro web framework, the IBlueprint interface allows for the creation of modular components or Blueprints. These Blueprints enable developers to add new functionalities, such as web pages or API endpoints. This not only enhances CKAN's modularity but also its scalability, allowing it to adapt to various project requirements.

- **Configuration and Common Functionality:** The `ckan.common module` and `config.get()` methods play crucial roles in managing configuration settings and accessing shared functionalities within a CKAN instance. This includes managing global variables and configuration parameters to request objects, thereby ensuring a seamless and customizable experience across the CKAN platform.

5.2 Leibniz Data Manager

LDM [8] is a data management system developed in alignment with the FAIR data principles, emphasizing the importance of machine-processable metadata for efficient data discovery, accessibility, interoperability, and reuse. Built upon Semantic Web technologies, LDM provides comprehensive support to researchers in documenting, analyzing, and sharing research datasets. It stands out for its ability to integrate datasets from diverse repositories and solve interoperability challenges, utilizing well-established vocabularies like DCAT [96] and DataCite [27] for effective metadata presentation. Additionally, LDM integrates Jupyter notebooks as data services, enabling the execution of live code for interactive data analysis [8]. Figure 5.5 provides an overview of LDM.



A dataset is created in LDM
- Type of dataset is selected
- Metadata template is filled in
- DOI is generated
- Synchronization schedule is defined
- Access Regulations are chosen

Datasets are kept up-to-date
- Data is synchronized from data providers according to synchronization schedulers
- Metadata is updated

Datasets are explored based on
- Keywords
- Fined-grained metadata descriptions using DCAT properties

Datasets are analyzed based on
- Graphs plotting main features
- Visualizations in 2D and 3D
- Jupyter Notebooks to perform live-code analysis

Figure 5.5: LDM Overview [81]

5.2.1 Architecture

LDM's architecture is designed to facilitate the entire lifecycle of DM, including data planning, collection, processing, analysis, publishing, preservation, and reuse. The system is capable of handling data in various formats and integrates datasets from different repositories (e.g., Leibniz University Hannover [54], PANGAEA [66], or RADAR [76]). In adherence to Linked Data and FAIR principles, LDM employs DCAT and DataCite for metadata description. Each dataset within LDM is uniquely identified with a Digital Object Identifier (DOI), enhancing traceability and citation. The platform offers robust functionalities for dataset exploration and analysis, including keyword-based searches, various RDF [77] serializations of metadata, and the utilization of multiple plots and visualizations [8].

5.2.2 Additional Features

As an open-source extension of the CKAN data repository system, LDM includes several advanced features that significantly enhance its DM capabilities. Alongside harvesting datasets from other repositories, it features a synchronization scheduler, an important tool for ensuring that the imported metadata remains updated and accurate. In addition, LDM generates a knowledge graph from the metadata, whether they are published by LDM users or imported. This knowledge graph provides detailed insights into the properties of DCAT resources, licensing information, and links to scientific publications. The knowledge graph also connects to Wikidata [99] resources, thus offering extensive and detailed descriptions of these objects. Another significant feature is the SPARQL [98] endpoint accessibility. This functionality allows users to access the knowledge graph via a SPARQL endpoint, facilitating sophisticated queries and data retrieval processes. Furthermore, the federated query engine DeTrusty [79] is integrated into LDM. This tool enables the execution of queries over the LDM KG, ORKG [87], and Wikidata [99] knowledge graphs, enhancing the platform's capabilities for in-depth data analysis [80, 81]. Collectively, these features underscore LDM's role as a versatile and powerful tool for DM, offering a wide array of functionalities to meet the diverse needs of researchers and data managers.

5.3 Summary of the Chapter

This chapter examines LDM, a DM system developed on top of CKAN. It covers the key aspects of both LDM and CKAN, including their architectural designs and main features. Additionally, the chapter explores the unique capabilities and functionalities of LDM that extend beyond the core features of CKAN, highlighting its role as a comprehensive tool in the field of DM.

Chapter 6

Implementation

As already mentioned in Chapter 5, CKAN is an extensible platform that provides multiple entry points for extension development through its programming interfaces. These interfaces play an important role in managing the lifecycle of datasets and resources within CKAN, which includes the creation, updating, and deletion of datasets. While the `IResourceController` interface, integral to CKAN's Python-based framework, is primarily focused on resources, managing datasets often involves interactions with interfaces like `IDatasetForm` or `IPackageController`. These interfaces provide functions to manage key events in the dataset lifecycle [17].

In the context of CKAN, a *package* typically represents a dataset and includes information about it, such as various attributes and a list of associated resources. These resources usually represent external files, each linked via a unique URL [17]. The `ckanext-gitimport` extension, developed in this thesis, enhances the dataset creation process by automating the population of dataset fields with metadata retrieved from GitHub repositories. Furthermore, it incorporates the README file from the corresponding repository as a resource within the dataset. This functionality serves as a proof of concept and directly addresses the need for efficient, metadata collection and management (see Chapter 7), a DM tool highlighted as important by survey participants in the energy sector.

6.1 Workflow of the Extension

The workflow of the `ckanext-gitimport` extension initiates when a user inputs a GitHub repository name into the designated template interface and activates the metadata retrieval by clicking the *Fetch Metadata* button. This action starts the extension's automated process for populating dataset fields with metadata from the GitHub repository, streamlining dataset creation within the CKAN framework. The extension is designed to handle sensitive operations, such as token retrieval and secure metadata retrieval, through server-side Python scripting. Furthermore, it employs JavaScript to facilitate responsive client-side interactions. The extension has three main modules: views.py, plugin.py, and gitimport.js, each serving a distinct role in the workflow.

1. **views.py - Server-Side Logic:** This module acts as the 'bridge' between the GitHub API and the user interface, by keeping sensitive data safe during transmission. It is responsible for setting up a blueprint using Flask [65], which defines the endpoints for interaction within the CKAN instances. It also manages the secure retrieval of GitHub access tokens from the ckan.ini configuration file, a critical step for authenticated API requests. In the event of an API call, this module extracts the repository name from the incoming parameters and fetches the corresponding metadata from GitHub. Once the metadata is obtained, views.py compiles it into a JSON [48] format and serves it back to the client for further use.

2. **plugin.py - CKAN Configuration and Plugin Registration:** This module manages the extension's configuration and its registration within the CKAN framework. It integrates the extension with CKAN by implementing plugin interfaces that augment the platform's capabilities, including the registration of the blueprint established in views.py. Additionally, it declares the necessary resources such as templates for the user interface and directories for static files. To ensure the extension's seamless operation within the CKAN ecosystem, plugin.py modifies the platform's configuration to include the settings specific to the extension.

3. **gitimport.js - Client-Side Interaction:** It manages the dynamic aspects of the extension's interaction with users. It captures the GitHub repository name as user input and retrieves metadata in JSON format from the server-side endpoint defined in views.py. Following the metadata retrieval, this module processes and presents the metadata, which includes key details for a GitHub repository such as owner, license, description, contributors, etc. This process

enhances the user experience by automating the population of dataset fields with relevant GitHub metadata. In addition, the gitimport.js module fetches the README file URL from the GitHub repository, ensuring its automatic integration into the appropriate resource field on the subsequent page within the CKAN user interface. This inclusion is crucial for the resource aspect of a dataset in CKAN, as resources are a mandatory element, and the README file serves as a good option due to its usual comprehensive documentation about the dataset.

Figure 6.1 provides a visual representation of the extension workflow.

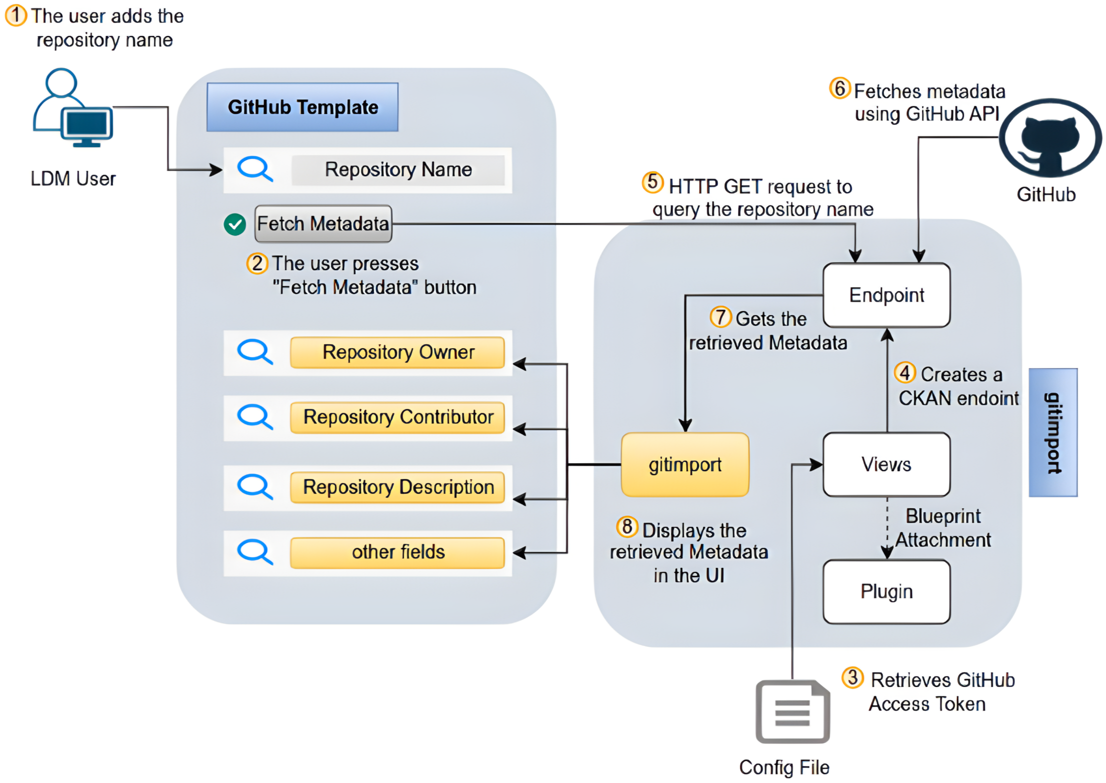

① The user adds the repository name
LDM User
GitHub Template
Repository Name
Fetch Metadata
② The user presses "Fetch Metadata" button
Repository Owner
Repository Contributor
Repository Description
other fields
⑥ Fetches metadata using GitHub API
GitHub
⑤ HTTP GET request to query the repository name
Endpoint
⑦ Gets the retrieved Metadata
④ Creates a CKAN endoint
gitimport
gitimport
Views
⑧ Displays the retrieved Metadata in the UI
Blueprint Attachment
Plugin
③ Retrieves GitHub Access Token
Config File

Figure 6.1: The Workflow of the Extension

6.2 The Scheming Extension

The `ckanext-scheming` extension [22] is a popular CKAN extension, that enables the configuration and customization of metadata schemas using YAML or JSON files. This extension supports custom validation and template snippets for both editing and display purposes, providing a flexible way to tailor metadata schemas to specific needs.

This extension was used to create the GitHub template, created within this YAML file, that includes a variety of fields customized to GitHub metadata. Each field is designed to capture specific aspects of a GitHub repository, ensuring that the created dataset is rich with relevant and useful information. These fields include but are not limited to, repository details such as the name, description, contributor and their name, topics, and other metadata elements like stars, forks, and license information. For a detailed explanation of each field, refer to Table 6.1. This table outlines every field defined in the GitHub template, providing a quick definition and its relevance in the context of GitHub. By integrating these fields into the CKAN dataset schema, the extension significantly enhances the dataset's metadata, making it more informative and useful for users who rely on CKAN for DM and discovery. This integration exemplifies the flexibility and power of the ckanext-scheming extension in customizing CKAN to suit specific metadata requirements.

6.3 Additional Features

The extension incorporates several additional features to ensure the accurate incorporation of metadata into the CKAN database. These functionalities specifically address potential user input errors and changes within the dataset creation form, focusing on user interactions with the repository name field (`github_repo`) and the dynamic nature of the metadata fields.

A key feature is the ability to reset fields when there is a partial change in the repository name. After metadata is fetched by pressing the *Fetch Metadata* button, any modification to the `github_repo` field by the user triggers a reset of all metadata fields to a default status of *no value* or an empty string. This preventive measure ensures that the metadata remains accurate and relevant to the dataset's information, avoiding incorrect association with a different repository. If the `github_repo` field is altered after the initial fetch, users need to press the fetch button again to retrieve metadata for the new repository name.

Table 6.1: GitHub Metadata Fields Defined in the YAML file

Field Name	**Description**
github_repo	Name of the GitHub repository.
name	URL-friendly slug for the dataset must be unique.
github_owner	Owner of the GitHub repository (user or organization).
github_contributor	Username of a contributor to the GitHub repository.
extra_contribs	Additional contributors' usernames, supports multiple entries.
github_author	Full name of the main author or maintainer of the repository.
extra_authors	Repeating field for additional authors' full names.
owner_org	Organization hosting the dataset.
github_description	Description of the GitHub repository.
license	License information for the dataset.
pb_doi *	Digital Object Identifier (DOI) for the dataset.
repository_topics	Topics or tags related to the GitHub repository.
extra_topics	Repeating field for additional topics or tags.
repository_stars	Number of stars received by the repository on GitHub.
repository_forks	Number of forks the repository has on GitHub.
programming_language	Primary programming language of the repository.
Resource Fields	
url	URL of the resource (by default the README file).
name	Unique name of the resource.
description *	Description of the resource.
format *	Format of the resource.

Note: Fields marked with star (*) are optional and not automatically populated.

Another feature involves reloading the entire page if the `github_repo` field is cleared. This page reload serves a dual purpose: it resets the form to its initial state based on the default GitHub template and removes any dynamically added fields like those for contributors, authors, and topics. This ensures that the form reverts to its original, unaltered state, maintaining the form's integrity and relevance, especially when users remove the repository name after having populated fields with metadata.

Additionally, the extension handles scenarios where a user might paste a new text into the `github_repo` field over an existing entry. In such cases, if the existing repository name is partially highlighted and new text is pasted over it, the extension resets all metadata fields to their default *no value* status. This feature is another safeguard to ensure that metadata fields reflect the current repository name accurately and prevent erroneous associations.

6.4 Unit Testing

Unit testing is a fundamental aspect of software development, focusing on independently testing individual units or components to ensure their proper functionality. In the `ckanext-gitimport` extension, the unit testing suite employs jsdom [32], Sinon [85], and Chai [15] to create a simulated browser-like environment within Node.js [63]. This setup is essential for testing JavaScript functions that interact with the Document Object Model (DOM) [58], replicating browser behavior in the Node.js runtime. The suite uses jsdom to create a simulated DOM environment, crucial for testing functions that manipulate web page elements. Sinon is used for creating stubs, spies, and mocks, with its primary role in these tests being to stub the fetch function. This strategy is key for testing network request scenarios, allowing the simulation of various network conditions without actual network calls. Chai's *expect* syntax provides a clear and expressive means for writing assertions, ensuring effective validation of test outcomes. Also, Mocha [60], a versatile JavaScript test framework, was used to run these tests in Node.js, it is suitable for both server-side and client-side JavaScript. It works together with Chai, facilitating various assertions about the JavaScript code [23].

The testing process covers several key functions of the extension and has eight unit tests including `clearFieldsById`, `clearDynamicFields`, `resetFields`, and `fetchGitHubMetadata`. These functions are crucial for manipulating form fields and fetching data using the fetch API. For instance, the `resetFields` function is crucial for resetting all GitHub metadata fields to their default state. Its effectiveness is tested by creating a mock DOM structure, filled with input elements, and then executing the `resetFields` function. The test checks that the values of these inputs are correctly reset, typically to empty strings, ensuring that the function reliably clears all relevant fields in the user interface. Similarly, the `fetchGitHubMetadata` function fetches metadata from GitHub and populates the fields in the extension's form. Testing this function is more complex due to its use of network requests. Using Sinon to stub the global fetch function, the test suite simulates different response scenarios. This includes successful fetches where the stub resolves with mock data, resulting in the correct population of DOM elements, and fetch failures where the stub rejects with an error, prompting `resetFields` to reset the form fields. These scenarios ensure that the function can handle various network conditions and respond appropriately. The results of these tests are indicative of the functions' reliability and effectiveness within the extension. All eight tests pass, underscoring the robustness of the extension's operation. More information about these tests is available in the public GitHub repository of the extension, discussed in section 6.5.

6.5 CKAN Extension and LDM Customization

One of the goals of implementing the `ckanext-gitImport` extension is to contribute to the CKAN community. Therefore, the extension is available in a public GitHub repository, accessible to the wider community for use and contribution. The repository can be found at the following URL [56]. The public availability of this extension not only benefits the CKAN community but also invites further development and enhancements from other developers.

Customization for LDM
In addition to creating a general CKAN extension, this thesis also customizes the extension to specifically work with LDM. The main logic and server-side work of the extension remain consistent across both the CKAN instances and the LDM instance. However, certain customizations were necessary to adapt the extension to the unique interface of the LDM instance. The primary changes in the LDM-specific version of the extension include modifications to the HTML file that generates the header and the addition of a new JavaScript script named get_datasets. It adds a new item to the dropdown list in the user interface, providing a seamless integration with LDM's existing layout and navigation structure. In contrast, for standard CKAN instances, the `ckanext-gitimport` functionality is introduced as a new button in the navigation bar, making it readily accessible and easy to use for CKAN users.

In addition, some minor modifications were also necessary in the `gitimport.js` script to account for differences in field ID names and other instance-specific characteristics between the CKAN and LDM systems. These tweaks ensure that the extension functions correctly in both environments, catering to the specific needs and configurations of each system.

The dual development approach for both general CKAN and LDM instances highlights the extension's versatility and adaptability, hopefully making it a tool for a wide range of users and instances within the CKAN community.

6.6 Summary of the Chapter

In this chapter, the development process of a new extension called `ckanext-gitImport` is discussed, including an exploration of the range of tools, workflow, and technical decisions employed in its implementation. The extension is designed to automate the importation of metadata from GitHub repositories into CKAN datasets, enhancing the user experience by minimizing manual input and reducing potential errors in metadata entry.

Chapter 7

Results

The survey was designed to mainly address the following research questions:
RQ1) What are the key requirements for DM in the energy sector, and **RQ2)** What is the current state of DM within the energy community? Therefore, this chapter presents the analysis of the survey responses conducted to answer these questions. The survey gathered responses from a diverse group of professionals with varying roles and expertise in the energy sector.

Survey Participants
The survey captured insights from 14 participants, revealing an adequate representation between the academic and private sectors within the energy field. From these participants, eight of the respondents are associated with research institutions, such as Fraunhofer IWES [35], TIB [89], and TU Clausthal [90], indicating a strong representation from academic and research-focused organizations. The remaining participants are from the private sector, enriching the data with varied perspectives (see Figures 7.1 and 7.2). The survey included a variety of roles, with six participants holding Doctor of Engineering titles and two being PhD students. The rest of the participants are energy experts working in different energy sectors, with roles such as project developer and planning engineer. A diverse set of expertise areas is represented, with wind energy being the most commonly reported, accounting for 40.9% of the survey participants. This is followed by other areas such as solar energy at 22.7%, geothermal energy at 13.6%, and hydrogen energy at 9.1% (see Figure 7.3).

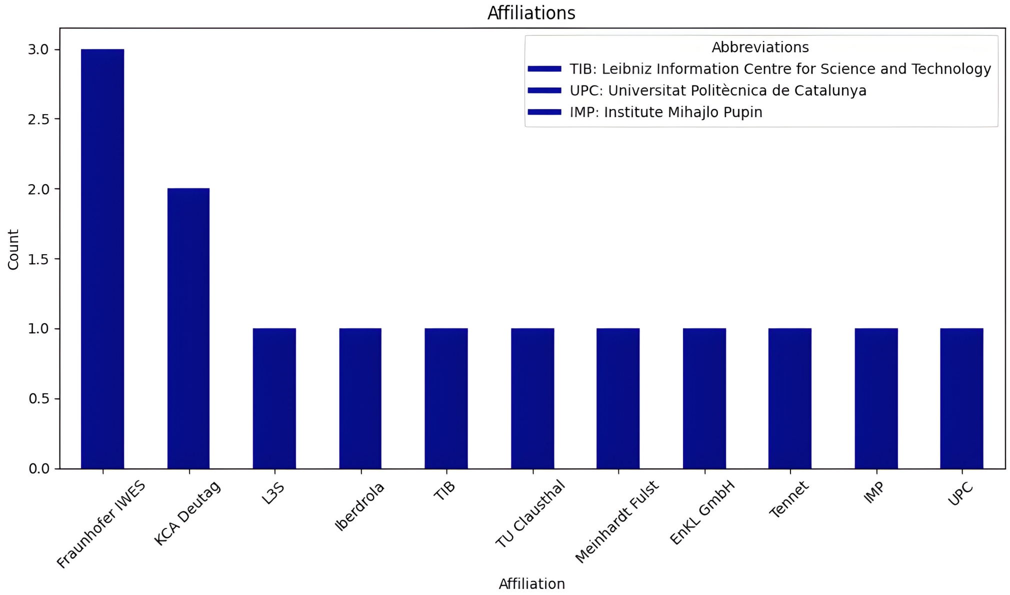


Figure 7.1: Distribution of Affiliations Among Survey Respondents

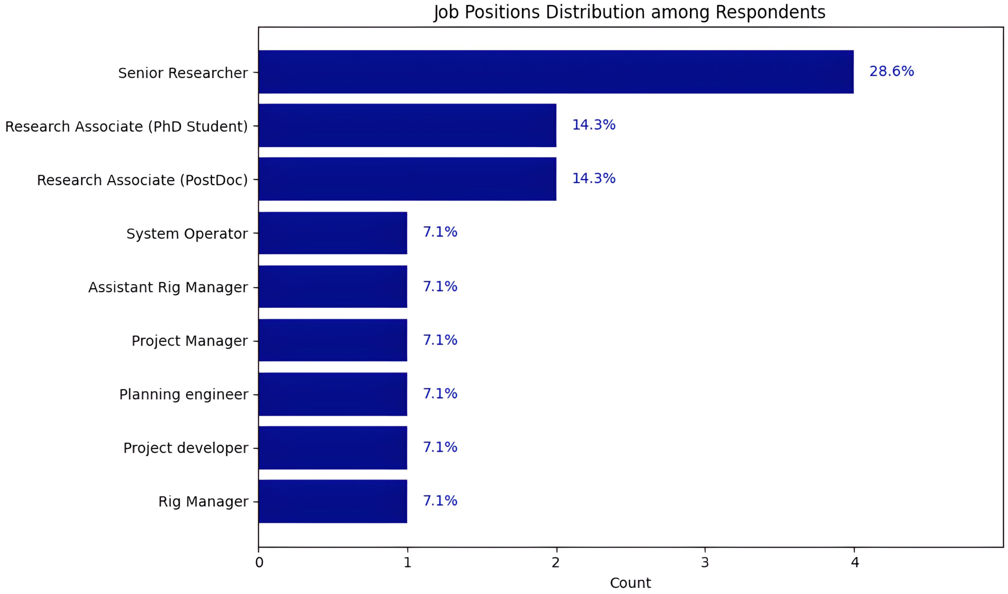


Figure 7.2: Distribution of Respondents' Job Positions

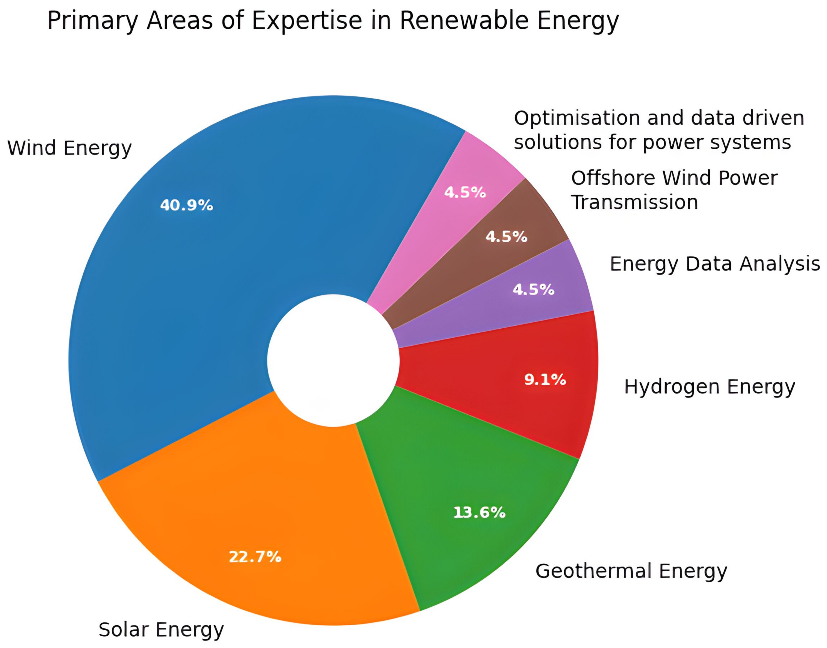


Figure 7.3: Respondents' Primary Areas of Expertise in Energy

Familiarity with DM

As can be seen from Figure 7.4, the survey results indicate that the majority of the respondents, at 64.3%, are somewhat familiar with DM, possessing a general understanding but with limited practical experience. A further 21.4% have heard of it but do not understand it well, and only 14.3% are very familiar, having extensive knowledge and experience. This distribution of familiarity levels underscores a potential opportunity for educational initiatives to enhance the practical skills required for DM in the energy sector.

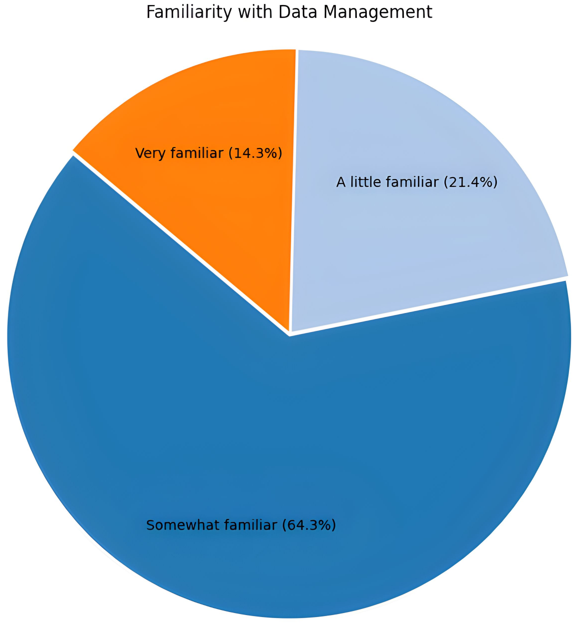


Figure 7.4: Distribution of Data Management Familiarity Responses

Given that two of the respondents (i.e., 14.3%), are energy experts with extensive knowledge and experience in DM, their answers will be compared with the overall findings of the survey in Section 7.1. This comparison will elucidate how the views of these experts on energy DM align with or diverge from the broader survey results, thereby providing a deeper understanding of the significance of specific DM aspects within the energy sector.

Key Requirements for DM in the Energy Sector

The survey highlights the top 10 mentioned requirements in the articles, extracted from the literature review (see Chapter 4), for DM in the energy sector, providing a clear indication of the priorities and challenges faced by professionals in this field. As can be seen from Figure 7.5, 71.4% of respondents place high importance on the quality of data in analytical tasks, indicating a primary focus on the precision and dependability of data for analysis purposes. Security measures and privacy practices are also a top concern, with half of the participants rating them as highly important.

This shows a clear need for robust data protection systems within the sector.

When also considering moderate importance, integration of various types of data for interoperability and compatibility is chosen by 50% of respondents, with an additional 42.9% viewing it as highly important. This underscores the sector's requirement for diverse systems to work together effectively and for different data types to function without conflicts. Processing large volumes of data is similarly critical, cited as moderately important by 50% and highly important by 35.7% of respondents. This reflects the challenge of managing increasing data volumes and diversity in the energy sector.

Surprisingly, real-time data processing and metadata management are seen as less critical, according to the survey participants. Real-time data processing is considered of low importance by 28.6% and not important by 14.3% of respondents. This could suggest that in some fields, such as geothermal energy or some research-oriented projects, reliance on immediate sensor data or other real-time inputs may not be important. Similarly, metadata management for interpretation and reproducibility is regarded as of low importance by 42.9% and not important by 7.1%. This might be more common among professionals in private companies, where the focus is often on the direct usage of data for operational purposes, rather than investing resources in comprehensive metadata management, which might be more critical for long-term research and development projects.

In conclusion, the survey results reveal that the energy sector values data quality for analysis purposes and the security and privacy of data above all in DM. At the same time, the capacity to handle diverse and large datasets efficiently is also essential. These insights should inform the development of DM strategies that align with these identified priorities.

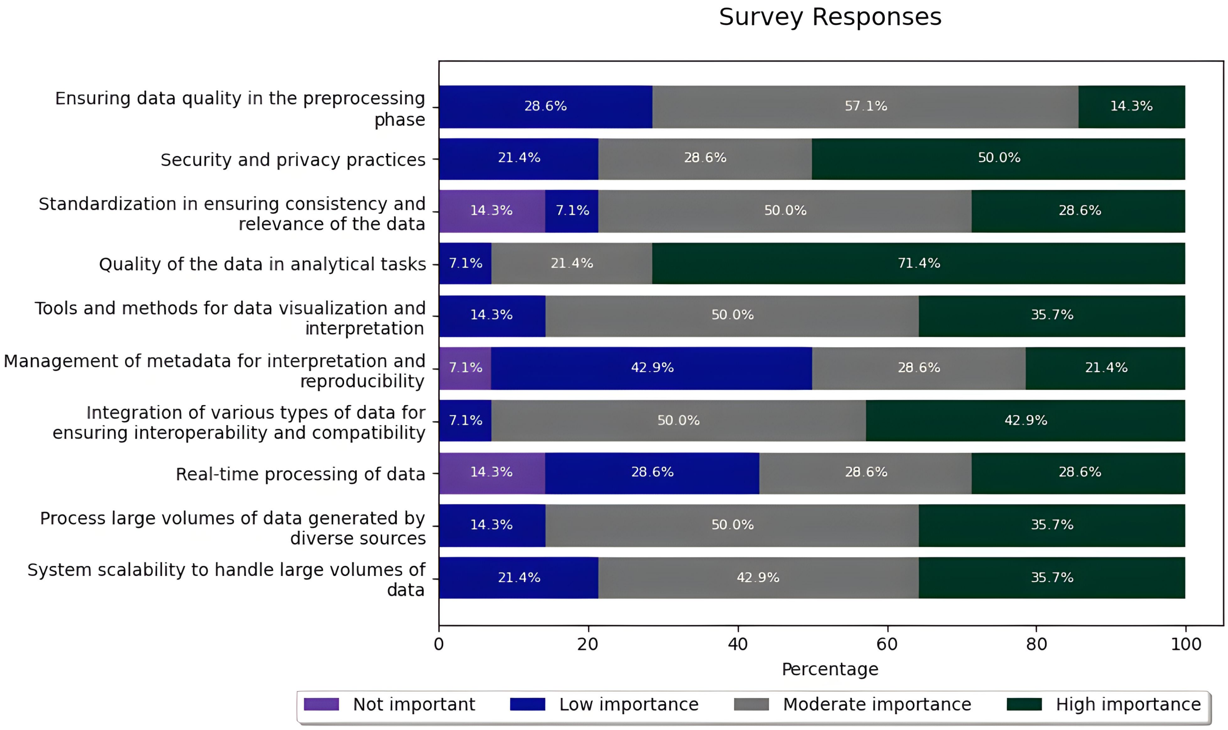


Figure 7.5: Distribution of Requirements Responses Among Survey Respondents

Data Sharing Practices

Data sharing practices within the energy community are predominantly oriented towards internal collaboration, with 10 out of 14 respondents indicating that they share data primarily with internal collaborators within their institution. A smaller proportion of the survey participants, only 28.6%, utilize public repositories or platforms for data sharing, highlighting a tendency towards closed-loop data exchange within organizational boundaries (see Figure 7.6). However, it is important to note that the survey format restricted respondents to select only one option for data sharing, which could have influenced the results. For instance, although no respondents indicated sharing data with external partners, this might not accurately represent the complete picture. One respondent commented that they would have also chosen external partners if multiple responses were permitted. This suggests that, if given a second option, some participants might engage in broader data sharing practices, including external collaborations.

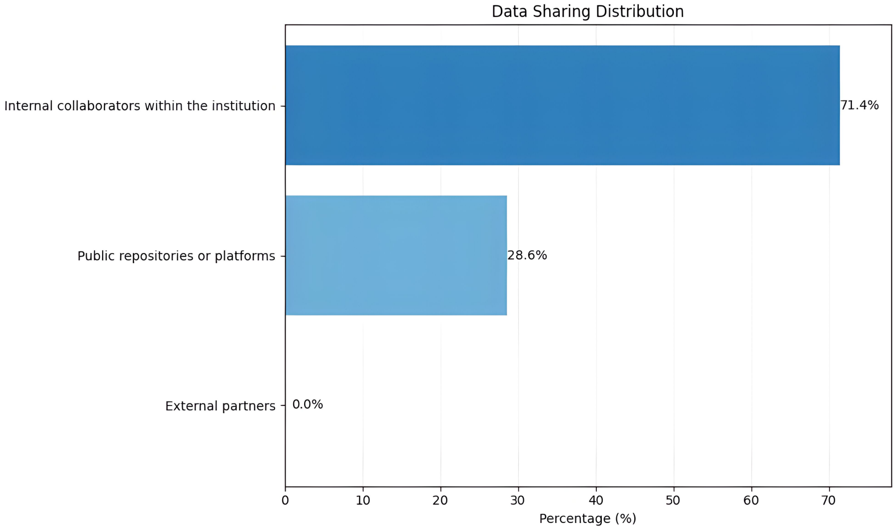


Figure 7.6: Distribution of Data Sharing Responses

Challenges and Tools in DM

The survey also addressed the challenges in DM, with respondents identifying three primary concerns. As can be seen from Figure 7.7, the two most common challenges are data ownership and intellectual property, and technical difficulties with data uploading both at 26.9%, the third concern is the lack of expertise in DM at 19.2%. Concerns over data ownership and intellectual property, chosen by respondents seven times, reflect a concern about the legal and proprietary aspects of DM. This highlights the importance of establishing clear data governance policies. The technical difficulties with data uploading, such as file size limitations, format compatibility, and data integrity issues, was also chosen seven times. This indicates that the current DM infrastructures may struggle to handle the volume and variety of data efficiently. Additionally, the lack of expertise or resources for DM, mentioned five times, suggests a need that could be met by increased investment in training and infrastructure.

In the context of tools used for DM tasks, Figure 7.8 presents a range of tools and standards employed by professionals in the energy sector. Foremost, data analysis tools emerge as the most important, with 19.3% of survey participants incorporating them into their DM routines. This underlines the crucial role of data analysis in driving insights and decisions within the sector. Simulation and modeling tools also receive significant attention, with 14% of the participants reflecting the preference for predictive and scenario-based analyses in understanding and navigating complex energy systems. Metadata collection and management tools, along with data format tools, are utilized by 12.3% of respondents. This underscores the sector's need for efficient organization and use of collected metadata.

Conversely, data visualization and data exchange with collaborators are less common in the responses, marked at 1.8% and 5.3% respectively. The modest reliance on data visualization tools could suggest a preference for simulations over graphical representations in the current DM plans and best practices. The relatively low interest in tools for data exchange with collaborators points towards a more inward-looking approach to data sharing. This also validates the findings of Section 7, which suggest a preference for keeping data within the organization

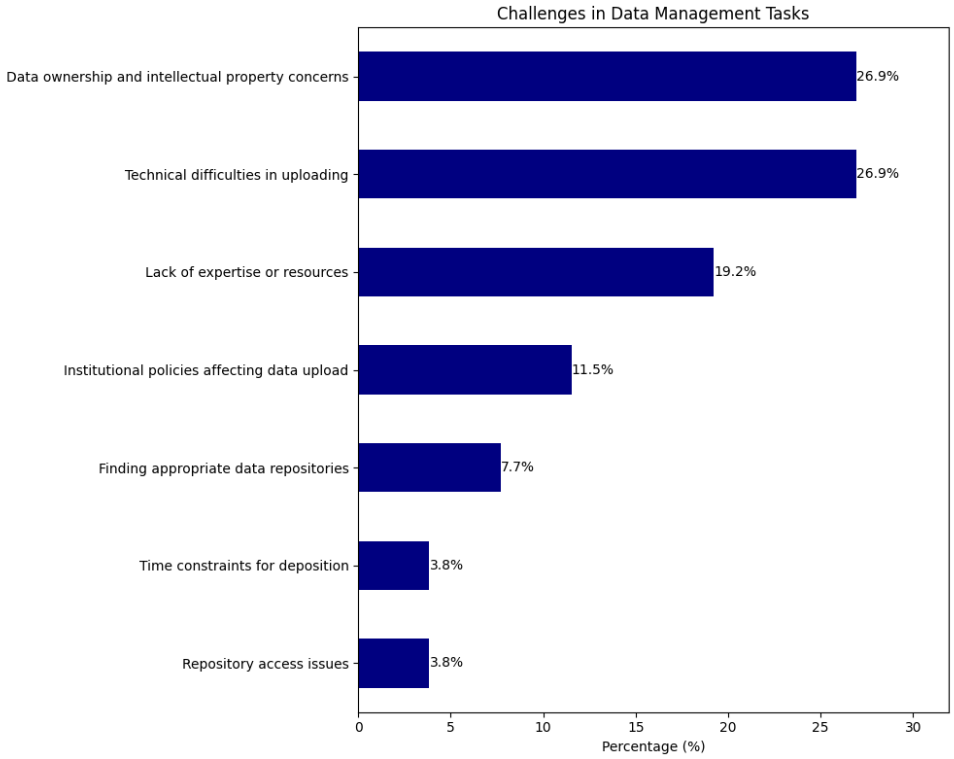


Figure 7.7: Respondents Data Management Challenges

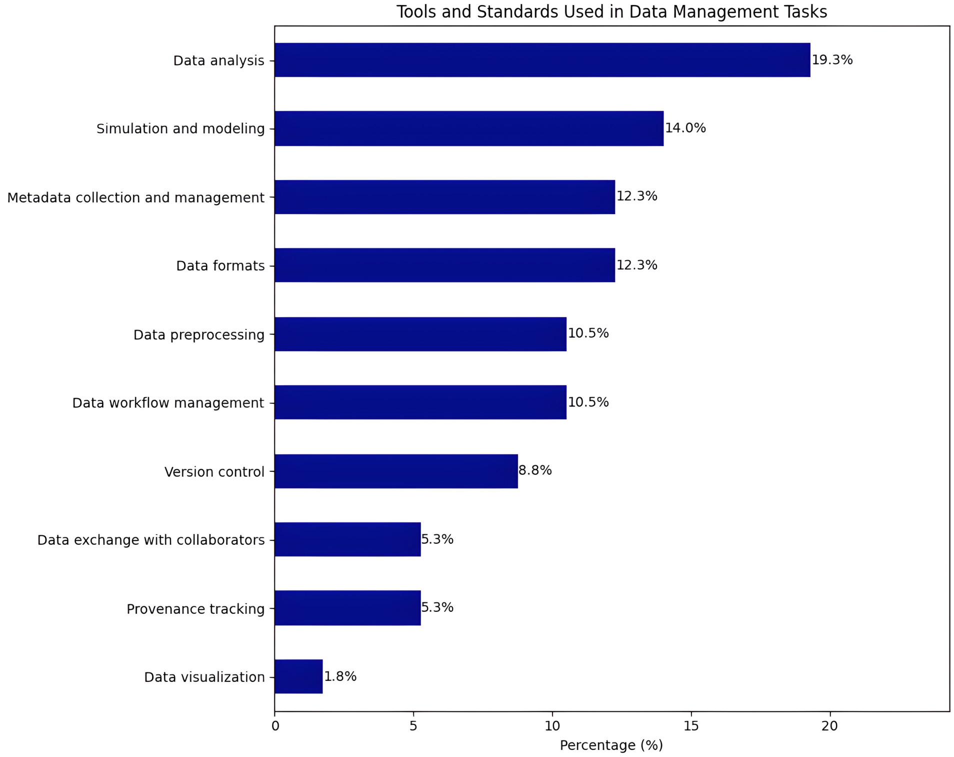


Figure 7.8: Respondents used Tools for Data Management Tasks

7.1 DM Expert Insights vs. Survey Findings

In the survey, the majority of participants had a low level of familiarity with DM, with only two respondents indicating extensive knowledge and experience. This section compares the findings of these two experts, both wind and solar energy professionals, with the broader findings to understand how expert opinions align with or differ from general perceptions.

In terms of DM requirements, both professionals highlighted the importance of the scalability of systems and data processing capabilities, which aligns with the survey's emphasis on the efficiency of handling large volumes of data. However, they also identified the importance of real-time data processing, a viewpoint that was not as strongly reflected in the overall survey responses, where the need for real-time processing was not considered of moderate or high importance by 50% of the respondents. The need for interoperability and compatibility was strongly advocated by the experts, echoing the survey results that stressed the importance of enabling different systems and data types to function cohesively. Their views on data quality for analytical purposes were consistent with the survey results. However, there was a difference in the importance of metadata management, which was considered highly significant by one expert and less so by another, reflecting the divided opinions of the survey participants on this matter. The challenges related to data sharing described by the experts, such as issues of data ownership, technical barriers, and resource limitations, reflect the obstacles identified in the survey, indicating these are widespread concerns within the industry.

7.2 Summary of the Chapter

The survey results present a picture of the energy sector's engagement with DM. Data analysis emerges not just as a critical tool, with 19.3% of professionals reporting its use in their DM tasks, but is also considered as the most important DM requirement by 71.4% of respondents.

Also, the survey indicates that security and privacy practices are a top priority for energy professionals, reflecting the need to safeguard energy data. This strong focus on security measures resonates with the sector's need for protecting sensitive and proprietary information. However, the survey also shows less concern for real-time data processing and metadata management for interpretation and reproducibility within the current DM practices. While these aspects are recognized, they do not hold the same significance as data analysis and security, with each being considered of low to no importance by almost 50% of the responses. Additionally, data sharing does not appear to be a priority, with a significant number of respondents favoring internal over external or public data exchanges. This preference could be influenced by proprietary data concerns or a focus on internal data utilization strategies.

The survey also reveals that while there is some familiarity with DM concepts, there is a potential need for enhanced expertise. This is indicated by the fact that 64.3% of participants have some understanding of DM but may lack comprehensive practical experience, suggesting that further education and training in DM in general could benefit the sector.

These findings should not be seen as a complete representation of the entire energy sector but can offer valuable insights into the priorities and potential areas for the development of DM practices as perceived by the survey participants.

Chapter 8

Conclusions and Future Work

This thesis explored DM within the energy sector, focusing on its requirements, the current state, and practices among professionals. Through a comprehensive literature review and a targeted survey, this study aimed to discover key requirements of DM that are critical for the advancement and efficiency of research and operations in the energy sector. This chapter summarizes the main findings, acknowledges the limitations of the study, and proposes directions for future research.

8.1 Conclusions

This study began with the review of the literature, examining 36 articles to set the stage for a comprehensive survey. The survey was developed to gather insights from professionals in the energy sector, uncovering several important findings. It was found that professionals within the energy sector place a high priority on data quality for analytical tasks, alongside the need for systems that can scale and integrate various data types for effective DM. This emphasis reflects the sector's need to ensure data reliability for analysis and the necessity for systems to work together without issues.

Another interesting observation from the survey is that real-time data processing is not considered of high importance by half of the respondents. In contrast, experts who are very familiar with DM view it as crucial, indicating differing opinions on its significance based on the respondent's level of expertise in DM. Additionally, the survey showed a stronger preference for using simulation tools rather than tools for graphical visualization. This suggests a greater interest in using predictive models and scenario analysis over visual representation of data. Furthermore, the survey uncovered a significant lack of understanding and implementation of the FAIR principles among the respondents, with many indicating they only share data within their own organizations. This reveals a broader challenge of limited external data sharing and a general unfamiliarity with these key principles.

The survey also highlighted a pressing need for better DM expertise and skills within the sector, pointing out the importance of further training and development in this area. Also, this thesis presented the *gitimport* extension, developed as a proof of concept, representing a tool type highly valued by survey participants: metadata collection. This extension, is specifically designed for gathering metadata from external repositories, in this case GitHub. This work demonstrates the potential for tools to streamline and improve the process of metadata acquisition, aligning with the needs highlighted by professionals in the energy sector.

8.2 Limitations

Given that the vast majority of the respondents surveyed are located and work in Germany, with only one exception, this geographical concentration represents a limitation of this work. Such a constraint could potentially affect the generalizability of the findings across different national and cultural contexts, which may face unique DM challenges and practices. This situation indicates a need for future research to include a broader and more geographically diverse participant base.

Another limitation was the survey's relatively limited sample size and scope, which might have impacted the findings. This underscores the need for future research to engage a larger and more diverse participant base. Additionally, the survey's design, which mainly focused on renewable energy experts, may require refinement to encompass a broader and more varied set of respondents. The initial survey targeted energy experts across a spectrum of DM experience levels, with the majority having very little familiarity with the subject. Future surveys present an opportunity to specifically target energy experts who possess a significant background in DM, aiming for a more informed and comprehensive understanding of DM's current state within the sector.

8.3 Future Work

Future research needs to extend the survey to specifically target energy experts with extensive experience in DM. This focused approach is likely to provide deeper insights into the real-world challenges and applications of DM in the energy field. The next steps should involve the development and testing of new tools or extensions that meet these specific needs, like data analytics and simulations.

An important direction for subsequent work is also to raise awareness among professionals in the energy sector about the importance of the FAIR principles. Educating these individuals on the significance of making data Findable, Accessible, Interoperable, and Reusable is crucial because it can enhance data usability across different platforms and projects, leading to more collaborative and efficient research outcomes. Changing the mindset around data sharing and improving knowledge in DM are essential steps towards building a more open and cooperative research environment within the energy sector.

8.4 Summary of the Chapter

In summary, this thesis contributes to the ongoing discourse on DM in the energy sector. By highlighting the current priorities, challenges, and gaps in DM practices, this work lays the foundation for future research and development aimed at enhancing the sector's data management capabilities. Through targeted surveys, proof-of-concept tool development, and the identification of key areas for improvement, this study seeks to pave the way for more efficient, secure, and collaborative research and operational practices in the energy field.

Bibliography

[1] Amazon. *What Is An API (Application Programming Interface)?* Accessed on February 19, 2024. URL: `https://aws.amazon.com/what-is/api/`.

[2] Ricardo Amorim, João Aguiar Castro, João Rocha, and Cristina Ribeiro. "A Comparative Study of Platforms for Research Data Management: Interoperability, Metadata Capabilities and Integration Potential". In: *Advances in Intelligent Systems and Computing* 353 (Jan. 2015), pp. 101–111. DOI: `10.1007/978-3-319-16486-1_10`.

[3] Andy Chiu, Boyang Tang, Jihong Ju, Bo Wang. *CKAN: The open source data portal.* Accessed on February 19, 2024. URL: `https://delftswa.gitbooks.io/desosa2016/content/ckan/chapter.html`.

[4] Anil Turkmayali. *Semantic interoperability: a common language for data sharing.* September 14, 2023. URL: `https://internationaldataspaces.org/semantic-interoperability-a-common-language-for-data-sharing/`.

[5] Apache Solr. *PSolr is the popular, blazing-fast, open source enterprise search platform built on Apache Lucene.* Accessed on February 19, 2024. URL: `https://solr.apache.org/`.

[6] Murtaza Ashiq, Muhammad Usmani, and Muhammad Naeem. "A systematic literature review on research data management practices and services". In: *Global Knowledge, Memory and Communication* ahead-of-print (Dec. 2020). DOI: `10.1108/GKMC-07-2020-0103`.

[7] Australian Renewable Energy Agency. *Kennedy Energy Park.* Last updated 24 November 2020. URL: `https://arena.gov.au/projects/kennedy-energy-park/`.

[8] Anna Beer, Mauricio Brunet, Vibhav Srivastava, and Maria-Esther Vidal. "Leibniz Data Manager – A Research Data Management System". In: *The Semantic Web: ESWC 2022 Satellite Events.* Springer, 2022, pp. 73–77. DOI: `10.1007/978-3-031-11609-4\_14`.

[9] Christine L. Borgman and Philip E. Bourne. *Why it takes a village to manage and share data.* 2022. arXiv: `2109.01694 [cs.DL]`.

[10] Kristin Briney, Heather Coates, and Abigail Goben. "Foundational Practices of Research Data Management". In: *Research Ideas and Outcomes* 6 (July 2020). DOI: `10.3897/rio.6.e56508`.

[11] Bundesdruckerei. *Was sind FAIR Digital Objects?* Accessed on February 14, 2024. URL: `https://www.bundesdruckerei.de/de/innovation-hub/was-sind-fair-digital-objects`.

[12] Bundesministerium der Justiz. *Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz).* Accessed on February 12, 2024. URL: https://www.gesetze-im-internet.de/urhg/__31.html.

[13] Byron Galbraith. *Data vs. Information: What's the Difference?* JUN 05, 2023. URL: https://bloomfire.com/blog/data-vs-information/.

[14] Stephan Cejka. "Data management in energy communities". In: Nov. 2021. URL: https://www.researchgate.net/publication/356987239_Data_management_in_energy_communities.

[15] Chai Assertion Library. *Chai is a BDD / TDD assertion library for node and the browser.* Accessed on February 25, 2024. URL: https://www.chaijs.com.

[16] CKAN. *A community of open data enthusiasts.* Accessed on February 20, 2024. URL: https://ckan.org/community.

[17] CKAN. *CKAN documentation Release 2.11.0a0.* Accessed on February 12, 2024. Jul 25, 2023. URL: https://docs.ckan.org/_/downloads/en/latest/pdf/.

[18] CKAN. *CKAN for Government.* Accessed on February 20, 2024. URL: https://ckan.org/government.

[19] CKAN. *Essential questions are overarching or topical questions about CKAN project.* Accessed on February 20, 2024. URL: https://ckan.org/faq/essential.

[20] CKAN. *Federate.* Published December 22, 2023. URL: https://ckan.org/features/federate.

[21] CKAN. *Find CKAN Extensions.* Accessed on February 20, 2024. URL: https://extensions.ckan.org/.

[22] ckanext-scheming's contributors. *ckanext-scheming public repository.* Accessed on March 2, 2024. URL: https://github.com/ckan/ckanext-scheming.

[23] Corinne Ling. *How to Get Better at Unit Testing.* 04-29-20, Accessed on February 25, 2024. URL: https://sparkbox.com/foundry/improve_unit_testing_with_mocha_chai_jsdom.

[24] Andrew Cox and Eddy Verbaan. "Case Study of RDM in an Environmental Engineering Science Project". In: *Exploring Research Data Management.* Facet, 2018, pp. 33–40.

[25] Craig Stedman. *What is data management and why is it important?* Accessed on March 2, 2024. URL: https://www.techtarget.com/searchdatamanagement/definition/data-management.

[26] Dasol Kim, Myeong-Seon Gil, Minh Chau Nguyen, Heesun Won, Yang-Sae Moon. "Comprehensive Knowledge Archive Network harvester improvement for efficient open-data collection and management". In: *ETRI Journal* (June 2021). DOI: 10.4218/etrij.2020-0298.

[27] DataCite. *Connecting Research, Advancing Knowledge.* Accessed on February 22, 2024. URL: https://datacite.org/.

[28] Dataverse. *The Dataverse Project: Open source research data repository software.* Accessed on February 12, 2024. URL: https://dataverse.org/.

[29] DOI Foundation. *DOI: Homepage.* Accessed on February 12, 2024. URL: https://www.doi.org/.

[30] Dragan Avramovic & Yoana Popova. *Adapting CKAN to a New Era of Search Engines.* Published December 22, 2023. URL: https://ckan.org/blog/adapting-ckan-new-search-engines.

[31] Dremio. *Deliverable D2.4 The PLATOON Unified Knowledge Base Creation.* Accessed on February 14, 2024. URL: https://platoon-project.eu/wp-content/uploads/2023/02/D2.4-Unified-knowledge-base-creation-v1.pdf.

[32] Eric Boersma. *JSDOM: How to Get Started.* Accessed on February 25, 2024. URL: https://www.testim.io/blog/jsdom-a-guide-to-how-to-get-started-and-what-you-can-do/.

[33] European Soil Data Centre (ESDAC). *SoilTrEC : Soil Transformations in European Catchments.* Accessed on February 11, 2024. URL: https://esdac.jrc.ec.europa.eu/projects/soiltrec.

[34] fairsfair. *FAIR Semantics, Interoperability, and Services.* October 2023. URL: https://www.fairsfair.eu/fair-semantics-interoperability-and-services-0.

[35] Fraunhofer IWES. *Fraunhofer-Institut für Windenergiesysteme.* Accessed on February 25, 2024. URL: https://www.iwes.fraunhofer.de/.

[36] Friedrich-Alexander-Universität (FAU) Universitätsbibliothek. *Advantages of research data management.* Accessed on February 11, 2024. URL: https://ub.fau.de/en/research/data-and-software-in-research/advantages-of-research-data-management/.

[37] GitHub. *GitHub Docs.* Accessed on February 14, 2024. URL: https://docs.github.com/en.

[38] Harvard Medical School. *Data Management Terminology.* Accessed on February 13, 2024. URL: https://datamanagement.hms.harvard.edu/about/data-management-terminology.

[39] Hicham Zinalabdin. *Top 8 Data Management Challenges For The Green Energy Industry.* July 3, 2013. URL: https://blog.ze.com/the-zema-solution/top-8-data-management-challenges-for-the-green-energy-industry/.

[40] Rosie Higman, Daniel Bangert, and Sarah Jones. "Three camps, one destination: the intersections of research data management, FAIR and Open". In: *Insights: the UKSG journal* (May 2019). DOI: 10.1629/uksg.468.

[41] IBM. *What is a knowledge graph?* Accessed on February 14, 2024. URL: https://www.ibm.com/topics/knowledge-graph.

[42] Ihechikara Vincent Abba. *What is an ORM – The Meaning of Object Relational Mapping Database Tools.* Accessed on February 20, 2024. URL: https://www.freecodecamp.org/news/what-is-an-orm-the-meaning-of-object-relational-mapping-database-tools/.

[43] International Institute for Sustainable Development (IISD). *EIA: What? Why? When?* Accessed 12 February 2024. URL: https://www.iisd.org/learning/eia/eia-essentials/what-why-when/.

[44] Valentina Janev, Maria-Esther Vidal, Kemele Endris, and Dea Pujic. "Managing Knowledge in Energy Data Spaces". In: *Companion Proceedings of the Web Conference 2021*. New York: ACM, 2021, pp. 7–15. ISBN: 978-1-4503-8313-4.

[45] JavaScript. *Ready to try JavaScript?* Accessed on February 19, 2024. URL: `https://www.javascript.com/`.

[46] Tomcy John and Pankaj Misra. *Data Lake for Enterprises*. Packt Publishing, 2017. ISBN: 9781787281349. URL: `https://www.oreilly.com/library/view/data-lake-for/9781787281349/`.

[47] John Cheadle. *The Introduction to FAIR Principles*. Accessed on February 12, 2024. URL: `https://docs.nih-cfde.org/en/latest/the-fair-cookbook/content/Introduction/fair-principles/`.

[48] JSON. *Introducing JSON*. Accessed on February 25, 2024. URL: `https://www.json.org/json-en.html`.

[49] Jupyter. *Free software, open standards, and web services for interactive computing across all programming languages*. Accessed on February 12, 2024. URL: `https://jupyter.org/`.

[50] Benjamin Kinast, Hannes Ulrich, Björn Bergh, and Björn Schreiweis. "Functional Requirements for Medical Data Integration into Knowledge Management Environments: Requirements Elicitation Approach Based on Systematic Literature Analysis". In: *J Med Internet Res* 25 (Feb. 2023). DOI: `10.2196/41344`.

[51] Carsten M. Klingner, Michael Denker, Sonja Grün, Michael Hanke, Steffen Oeltze-Jafra, Frank W. Ohl, Janina Radny, Stefan Rotter, Hansjörg Scherberger, Alexandra Stein, Thomas Wachtler, Otto W. Witte, and Petra Ritter. "Research Data Management and Data Sharing for Reproducible Research—Results of a Community Survey of the German National Research Data Infrastructure Initiative Neuroscience". In: *eNeuro* 10.2 (2023). DOI: `10.1523/ENEURO.0215-22.2023`.

[52] Kotranaresh. *Navigating Academic Research with Scopus Cited Journals: A Comprehensive Guide*. Oct 21, 2023. URL: `https://medium.com/@kotranaresh5/navigating-academic-research-with-scopus-cited-journals-a-comprehensive-guide-bec2b016bf2a`.

[53] Law Insider. *energy data definition*. Accessed on February 14, 2024. URL: `https://www.lawinsider.com/dictionary/energy-data`.

[54] Leibniz Universität Hannover. *Leibniz Universität Hannover*. Accessed on February 22, 2024. URL: `https://www.uni-hannover.de/de/`.

[55] Maria-Esther Vidal. *Responsible Data Management*. BIOMEDAS-Programm (joint initiative of Leibniz Universität Hannover and Hannover Medical School. Sommer 2021. URL: `https://www.tib.eu/de/forschung-entwicklung/forschungsgruppen-und-labs/scientific-data-management/lehre/verantwortungsvolles-datenmanagement`.

[56] Mazen Bechara. *ckanext-gitimport*. Accessed on February 25, 2024. URL: `https://github.com/Mazen-B/ckanext-gitimport`.

[57] Mazen Bechara. *Understanding the Requirements of Data Spaces in the Energy Sector- MA Thesis Material*. DOI: `10.57702/0sgry643`.

[58] MDN contributors. *Document Object Model (DOM)*. Accessed on February 25, 2024. URL: https://developer.mozilla.org/en-US/docs/Web/API/Document_Object_Model.

[59] Manoj Menon, Svetla Rousseva, Nikolaos Nikolaidis, Pauline Gaans, Panos Panagos, Danielle Maia de Souza, Kristin Ragnarsdottir, Georg Lair, Liping Weng, Jaap Bloem, Pavel Kram, Martin Novák, Brynhildur Davidsdottir, Guðrún Gísladóttir, David Robinson, Brendan Reynolds, T. White, Lars Lundin, Bin Zhang, and Steven Banwart. "SoilTrEC: A global initiative on critical zone research and integration". In: *Environmental science and pollution research international* 21 (Dec. 2013), pp. 3191–3195. DOI: 10.1007/s11356-013-2346-x.

[60] mochajs. *simple, flexible, fun.* Accessed on February 25, 2024. URL: https://mochajs.org/.

[61] Maryam R. Nezami, Mark L. C. de Bruijne, Marcel J. C. M. Hertogh, and Hans L. M. Bakker. "Collaboration and Data Sharing in Inter-Organizational Infrastructure Construction Projects". In: *Sustainability* 14.24 (2022). ISSN: 2071-1050. DOI: 10.3390/su142416835. URL: https://www.mdpi.com/2071-1050/14/24/16835.

[62] NFDI Neuroscience. *Sharing Data // Connecting Reseach // Pooling Brains.* Accessed on February 18, 2024. URL: https://nfdi-neuro.de/.

[63] nodejs. *Node.js® is an open-source, cross-platform JavaScript runtime environment.* Accessed on February 25, 2024. URL: https://nodejs.org/en.

[64] Oracle. *What is Big Data?* Accessed on March 2, 2024. URL: https://www.oracle.com/big-data/what-is-big-data/#:~:text=is%20Big%20Data%3F-,Big%20data%20defined,especially%20from%20new%20data%20sources..

[65] Pallets. *FLASK.* Accessed on February 25, 2024. URL: https://flask.palletsprojects.com/en/3.0.x/.

[66] PANGAEA. *Data Publisher for Earth and Environmental Science.* Accessed on February 22, 2024. URL: https://www.pangaea.de/.

[67] Paperpile. *How to write a systematic literature review [9 steps].* Accessed 3 March 2024. URL: https://paperpile.com/g/systematic-literature-review/.

[68] Dimple Patel. "Research data management: a conceptual framework". In: *Library Review* 65 (July 2016), pp. 226–241. DOI: 10.1108/LR-01-2016-0001.

[69] Lennart Petersen, Bo Hesselbaek, Antonio Martinez, Roberto Borsotti-Andruszkiewicz, Nathan Steggel, Dave Osmond, and Germán Tarnowski. "Vestas Power Plant Solutions Integrating Wind, Solar PV and Energy Storage". In: *3rd International Hybrid Power Systems Workshop.* May 2018.

[70] PostgreSQL. *PostgreSQL: The World's Most Advanced Open Source Relational Database.* Accessed on February 19, 2024. URL: https://www.postgresql.org/.

[71] PRISMA. *TRANSPARENT REPORTING of SYSTEMATIC REVIEWS and META-ANALYSES.* Accessed on February 18, 2024. URL: http://www.prisma-statement.org/.

[72] PyPi. *Jinja2 3.1.2.* Accessed on February 20, 2024. URL: https://pypi.org/project/Jinja2/.

[73] Python. *Python is a programming language that lets you work quickly and integrate systems more effectively.* Accessed on February 19, 2024. URL: https://www.python.org/.

[74] Qualtrics. *Likert scales: definition, benefits & how to use them.* Accessed on March 2, 2024. URL: https://www.qualtrics.com/uk/experience-management/research/likert-scales/?rid=ip&prevsite=en&newsite=uk&geo=DE&geomatch=uk.

[75] Rachel McCullough. *The Scopus Content Coverage Guide: A complete overview of the content coverage in Scopus and corresponding policies.* March 5, 2023. URL: https://blog.scopus.com/posts/the-scopus-content-coverage-guide-a-complete-overview-of-the-content-coverage-in-scopus-and.

[76] RADAR. *Research Data Archive. Share. Publish.* Accessed on February 22, 2024. URL: https://www.radar-service.eu/radar/en/home.

[77] RDF Working Group. *Resource Description Framework (RDF).* Accessed on February 22, 2024. 2014-02-25. URL: https://www.w3.org/RDF/.

[78] Red Hat. *Understanding data services.* Published December 16, 2022. URL: https://www.redhat.com/en/topics/data-services.

[79] Philipp D. Rohde, Mazen Bechara, and Avellino. *DeTrusty.* 2024. DOI: 10.5281/zenodo.5997121. URL: https://doi.org/10.5281/zenodo.5997121.

[80] Philipp D. Rohde, Enrique Iglesias, and Maria-Esther Vidal. *FedORKG: Accessing Federations of Open Research Knowledge Graphs.* Jan. 2024. DOI: 10.5281/zenodo.1059144. URL: https://doi.org/10.5281/zenodo.1059144.

[81] Philipp D. Rohde, Ahmad Sakor, Mauricio Brunet, Enrique Iglesias, Mazen Bechara, Susanne Arndt, Mathias Begoin, Angelina Kraft, and Maria-Esther Vidal. *The Leibniz Data Manager: Supporting Researchers in the Lifecycle of Research Data.* Feb. 2024. DOI: 10.5281/zenodo.10610501. URL: https://doi.org/10.5281/zenodo.10610501.

[82] Sandra Melo. *Advantages and disadvantages of Google forms.* June 15, 2018. URL: https://datascope.io/en/blog/advantages-and-disadvantages-of-google-forms/.

[83] Sarah Amsler and Sharon Shea. *RFID (radio frequency identification).* Accessed on February 12, 2024. URL: https://www.techtarget.com/iotagenda/definition/RFID-radio-frequency-identification.

[84] Saylor Academy. *Documenting and Managing Requirements.* Accessed 8 February 2024. URL: https://learn.saylor.org/mod/book/view.php?id=66773&chapterid=60452.

[85] sinonjs. *Standalone test spies, stubs and mocks for JavaScript.* Accessed on February 25, 2024. URL: https://sinonjs.org/.

[86] SQLAlchemy. *The Python SQL Toolkit and Object Relational Mapper.* Accessed on February 20, 2024. URL: https://www.sqlalchemy.org/.

[87] Technische Informationsbibliothek (TIB). *Open Research Knowledge Graph.* Accessed on February 22, 2024. URL: https://orkg.org/.

[88] TechTarget Contributor. *data set.* Accessed on February 14, 2024. URL: https://www.techtarget.com/whatis/definition/data-set.

[89] TIB. *TIB – Leibniz Information Centre for Science and Technology.* Accessed on February 25, 2024. URL: https://www.tib.eu/en/.

[90] TU Clausthal. *Technische Universität Clausthal.* Accessed on February 25, 2024. URL: https://www.tu-clausthal.de/.

[91] Tulsa Community College (TCC) Library. *What are empirical sources or data?* Accessed on March 06, 2024. URL: https://askus.library.tulsacc.edu/faq/387787.

[92] Tutorials Point. *MVC Framework - Introduction.* Accessed on February 20, 2024. URL: https://www.tutorialspoint.com/mvc_framework/mvc_framework_introduction.htm.

[93] U.S. government. *GPS: The Global Positioning System.* Accessed on February 12, 2024. URL: https://www.gps.gov/.

[94] United States Department of Energy. *DOE Policy for Digital Research Data Management.* Accessed on February 18, 2024. URL: https://www.energy.gov/datamanagement/doe-policy-digital-research-data-management.

[95] University of Leeds. *What is research data?* Accessed on February 13, 2024. URL: https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained.

[96] W3C. *Data Catalog Vocabulary (DCAT) - Version 3.* Accessed on February 22, 2024. Mar 07, 2023. URL: https://www.w3.org/TR/vocab-dcat-3/.

[97] W3C. *LinkedData.* Accessed on February 14, 2024. URL: https://www.w3.org/wiki/LinkedData.

[98] W3C. *SPARQL 1.1 Query Language.* W3C Recommendation 21 March 2013, Accessed on February 22, 2024. URL: https://www.w3.org/TR/sparql11-query/.

[99] Wikidata. *Wikidata:Introduction.* Last edited on 30 January 2024, Accessed on February 22, 2024. URL: https://www.wikidata.org/wiki/Wikidata:Introduction.

[100] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3 (2016). DOI: 10.1038/sdata.2016.18.