

Improving research data reusability through data conversations

Bridging gaps in metadata supply and demand

Sara Meie Op den Orth



Improving research data reusability through data conversations

Bridging gaps in metadata supply and demand

by

Sara Meie Op den Orth

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday February 6, 2025 at 11:00 AM.

Student number: 4610598
Project duration: December 21, 2023 – February 6, 2025
Thesis committee: Dr. C. Lofi, TU Delft, supervisor
Dr. T. Abeel, TU Delft

Cover: Created using OpenAI's DALL-E (Modified)

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

Abstract

Efficient and inclusive data reuse across research disciplines is based on high quality metadata that bridges the gap between data producers and consumers. This gap, referred to as the metadata gap, arises when the metadata provided by producers do not meet the needs of consumers. Through a comprehensive analysis of metadata supply and demand, this thesis identifies the motivations and barriers faced by producers in creating metadata, along with the challenges consumers face when reusing datasets. To address these issues, the thesis introduces context-bridging data conversations, a framework designed to make metadata creation a more collaborative and adaptive process. The proof-of-concept is built on four key mechanisms: involving consumers as co-creators, recognising and incorporating contextual metadata, leveraging real-time dialogue, and dynamically adapting metadata elicitation questions. Qualitative interviews were conducted to identify the factors that shape metadata practices, and AI-generated summaries were evaluated as a scalable tool to synthesise the insights of these conversations. The findings are applied to the data management plans of CropXR, an interdisciplinary research institute. This case study illustrates how the metadata gap analysis can identify specific areas for improvement in metadata practices and how the context-bridging data conversations framework can provide actionable recommendations to enhance data reusability. By analysing data reusability as a dynamic and context-dependent process, this thesis advances both practical methodologies and theoretical understanding of metadata management. These contributions offer actionable strategies to close the metadata gap, foster collaboration across scientific domains, and promote more efficient and inclusive research practices.

Preface

This thesis represents the result of months of work exploring approaches to improve research data reusability through context-bridging data conversations. I would like to acknowledge the support and contributions of the individuals and organisations that made this work possible.

My supervisor, Dr. Christoph Lofi, deserves special thanks for his guidance throughout this project. I especially appreciated his focus on pursuing meaningful research over unnecessary details, which helped keep the work (a little more) focused and manageable. His introduction to the CropXR¹ project also provided the starting point for this thesis.

I am also grateful to the CropXR project for providing access to presentations and materials on their private SharePoint, which informed the case study described in this thesis. Although this research was conducted independently and is not formally affiliated with CropXR, observing a real data management project in development provided much of the inspiration for this work. I extend my thanks to Auke Damstra, Technical Director of CropXR, for facilitating access to the project and offering opportunities to discuss plans and results. I also appreciate the insights shared by Sören Wacker, from the CropXR data management team at TU Delft, who allowed me to join meetings and discussed his plans and ideas for the project.

During the research stage, I used ChatGPT² primarily as a brainstorming partner to explore ideas and identify potential gaps in my argumentation. During the writing stage, it helped to streamline paragraphs and adjust tone, ensuring clarity and consistency. Although ChatGPT was an invaluable tool in these processes, all research and writing was carried out by me. The tool served as an additional resource to improve the quality of my work, much like any other tool in the toolkit of a researcher. My use of ChatGPT was done in consultation with my supervisor, who approved how it was integrated into my workflow. Additionally, I ensured that no sensitive personal data or confidential materials were uploaded during the process. I also used Writefull³ to grammar-check my entire thesis, further refining the language and ensuring correctness. Together, these tools helped me present my work effectively, without replacing or diminishing my own contributions.

Finally, on a personal note, I want to thank my partner and my mother for their patience and support throughout this process, my cat-loving friend for sharing in the thesis experience, and my brothers for helping me test and refine my ideas. Their encouragement and understanding played an important role in helping me complete this work.

*Sara Meie Op den Orth
Delft, January 2025*

¹<https://cropxr.org>

²<https://chatgpt.com>

³<https://writefull.com>

Contents

Abstract	i
Preface	iii
1 Introduction	1
2 Background	4
2.1 Importance of data reusability	4
2.1.1 Related data qualities influencing reusability	4
2.1.2 Benefits of reusability	5
2.2 The role of soft metadata management in data reusability	5
2.2.1 Understanding soft metadata	6
2.2.2 Current metadata management methods and tools	6
2.2.3 Data conversations	8
3 Methodology	10
3.1 Research design	10
3.1.1 Research methodology components	11
3.1.2 Participant selection and data anonymisation	12
3.2 Exploring data conversation potential through interviews	12
3.2.1 High-level data conversation design	12
3.2.2 Question-level interview design	13
3.2.3 Process of conducting interviews	17
3.2.4 Interview design limitations	17
3.3 Evaluating data conversation summaries through participant surveys	18
3.3.1 Process of creating transcript summaries	18
3.3.2 Summary evaluation design	20
3.3.3 Summary evaluation limitations	21
4 Results	23
4.1 Participant overview	23
4.1.1 Participant characteristics and insights	23
4.1.2 Participant pool limitations	25
4.2 Data producer and consumer interview results	25
4.2.1 Producers' metadata considerations and experiences	25
4.2.2 Consumers' challenges and solution approaches	28
4.2.3 Real-time dialogue and question adaptation in data conversations	31
4.2.4 Discussion of interview results	32
4.3 Survey-based evaluation results of the data conversation summary	32
4.3.1 Participant lasting impression of interviews	33
4.3.2 Summary usefulness in isolation	33
4.3.3 Summary usefulness compared to alternatives	34
4.3.4 Discussion of survey results	35
5 Findings	36
5.1 Mapping the metadata gap	36
5.1.1 Metadata supply: Motivations for creating high-quality metadata	37
5.1.2 Metadata supply: Barriers to creating high-quality metadata	40
5.1.3 Metadata demand: Barriers to efficient reuse of datasets	42
5.2 Bridging the metadata gap	44
5.2.1 Consumers as metadata co-creators to bridge the gap	46
5.2.2 Contextual metadata to bridge the gap	47

5.2.3	Real-time dialogue to bridge the gap	48
5.2.4	Adaptive data conversation questions to bridge the gap	49
5.2.5	Transcript summarisation to bridge the gap	50
6	Discussion & Future Work	52
6.1	Interpretation of findings: bridging the metadata gap	52
6.2	The future of context-bridging data conversations	53
6.2.1	Moving beyond the proof-of-concept to a prototype	53
6.2.2	Managing the larger metadata haystack	54
6.2.3	Navigating the complexities of social interactions	54
7	Case study: Mapping and bridging metadata gaps in CropXR	56
7.1	The data management landscape of CropXR	56
7.1.1	Project overview	56
7.1.2	Current data management plan	57
7.1.3	The CropXR vision	58
7.2	Mapping the CropXR metadata gap	59
7.2.1	Connection through the centrally managed network	59
7.2.2	Knowledge fostering and dissemination through EduXR and DataXR	59
7.2.3	Metadata standards by the SAME group	59
7.2.4	A central long-term data hub through the Resilient Hub	60
7.2.5	Enhanced metadata elicitation and communication via Meta Buddy	60
7.3	Recommendations for Meta Buddy development	60
7.3.1	Recognise motivational needs of metadata producers	60
7.3.2	Involve data consumers as metadata co-creators	61
7.3.3	Capture and centralise (contextual) metadata	61
7.3.4	Plan for metadata iteration over time	62
7.3.5	Accommodate diverse use cases and users	62
7.3.6	Use real-time dialogue	63
8	Summary	64
	References	66
A	Informed consent forms	70
B	Interview protocol	76
B.1	Prologue statements and questions	76
B.2	Data producer questions	77
B.3	Data consumer questions	78
C	Survey protocol	79
D	Data conversation transcript summaries	81
D.1	Data producer transcript summaries	81
D.2	Data consumers transcript summaries	85
E	CropXR metadata gap	91

List of Tables

3.1	Interview producer themes	15
3.2	Interview consumer themes	16
3.3	Transcript summary evaluation questions	21
4.1	Research participant overview	24
4.2	Producer data conversation results: dataset details	26
4.3	Consumer data conversation results: dataset details	29
4.4	Transcript summary evaluation results: Likert scale questions	33
4.5	Transcript summary evaluation results: metadata elicitation ranking question	34
5.1	Metadata gap factors	37
5.2	Context-bridging data conversation mechanisms	45
5.3	Metadata gap linked to context-bridging data conversation mechanisms	45
B.1	Introductory interview questions	76
B.2	Producer interview questions	77
B.3	Consumer interview questions	78
C.1	Complete participant results to closed survey questions	80
E.1	CropXR metadata gap mapped	91

1

Introduction

Behind every research dataset lies a series of decisions made by producers of the data about what metadata to include to provide context to the raw data. In research, datasets are created to support the answer to a research question, serve as evidence for a hypothesis, or even constitute the answer itself. Metadata is primarily created with the producers' needs in mind, yet it also provides the essential context that allows future researchers to interpret and reuse the work, reducing redundancy, fostering innovation, and accelerating scientific discovery.

However, data reuse is rarely seamless. Data consumers interpret datasets through their own knowledge and adapt them for new purposes, often needing metadata that extends beyond what producers provided for their original research. Metadata serves as a bridge between data producers and consumers, providing the context needed to ensure that datasets are accessible and comprehensible across diverse disciplines and levels of expertise [1].

Despite its importance, creating metadata that facilitates efficient and inclusive reuse remains a complex task [2]. Data producers, primarily focused on presenting their findings clearly and efficiently, prioritise metadata that supports their own analyses. This approach can leave metadata underdeveloped for long-term reuse, as producers do not receive direct benefit from investing additional effort into metadata creation, especially for unknown future users [3]. As a result, many consumers are left to independently resolve ambiguities or gaps in metadata, creating inefficiencies and, in some cases, rendering datasets completely unusable [4], [5].

This disconnect between metadata supply and demand is referred to in this thesis as the metadata gap. A significant component of this gap lies in the handling of soft, also known as contextual, metadata [6]. This type of metadata is subjective and nuanced, including the reasoning behind key decisions, the consideration of alternative approaches, and the limitations of the data. Unlike hard metadata, such as file formats, variable names, or measurement units, soft metadata provides critical context for understanding the origins and appropriate use of a dataset.

Although some soft metadata may be included in research articles accompanying the dataset, it is often not recognised as metadata, even though it is essential for efficient reusability. Current metadata practices, such as metadata forms, which focus primarily on hard metadata, do not integrate contextual metadata well [7]. The subjective and complex nature of soft metadata also makes it more difficult to collect using overly standardised metadata methods [6], [8].

As datasets grow in size and complexity, and as open data initiatives expand, the challenges of metadata creation and reuse become increasingly significant. The push for more open data increases the demand for comprehensive high-quality metadata to ensure that datasets are not only accessible, but also usable across diverse disciplines and research contexts [9]. Moreover, modern research often relies on interdisciplinary and data-driven methods, which require metadata capable of bridging gaps between diverse fields and levels of expertise. Researchers who once managed small-scale data in spreadsheet editors now face the added challenge of creating metadata for larger, more complex datasets intended for a broad audience. Without comprehensive metadata, data consumers are forced to spend valuable time deciphering datasets instead of focussing on new discoveries [10], [11]. Improving metadata practices has the potential to overcome these inefficiencies, expanding opportunities for collaboration and innovation while enhancing the overall impact of scientific research [12]–[14].

Crucial details about the data creation and reuse process are often lost, kept only for personal reference, shared informally among colleagues, or completely overlooked. Current systems lack mechanisms for capturing these insights, whether in the contextual metadata included in research articles or the metadata forms that producers complete when uploading data to repositories. This thesis not only examines the metadata gap, exploring the interplay between producer motivations and barriers, consumer needs, and the role of soft metadata in enabling reusability, but also proposes a solution: context-bridging data conversations. The proof-of-concept provides a framework to transform informal, water cooler discussions between producers and consumers into structured, scalable practices that capture the nuanced experiences of dataset creation and reuse. By focussing on consumer participation, contextual metadata, real-time communication, and adaptable questioning, this approach addresses key limitations of traditional metadata creation and communication methods, enhancing the efficiency and inclusivity of data reusability.

To demonstrate the practical relevance of these mechanisms, this thesis uses the CropXR¹ institute as a case study. CropXR, which conducts crop research and is establishing a data management platform, illustrates the real-world challenges of metadata creation and reuse. By involving researchers and stakeholders with varying levels of experience in data management, it provides a concrete example of how improved metadata practices can address the complexities of collaboration and dataset reuse.

The research questions of this thesis are:

- **RQ1: What motivations and barriers influence data producers in creating metadata for efficient dataset reusability?** This question examines the supply side of the metadata gap, focussing on the internal motivations, constraints, and limitations that influence data producers. The focus is on metadata that enhances data reusability, with the aim of making reuse more efficient and accessible to a broader audience. The literature review addressed RQ1 by exploring intrinsic motivations and identifying barriers. The interviews expanded on this by providing empirical evidence of the motivations and challenges affecting the producers.
- **RQ2: Which metadata-related barriers have the greatest impact on data consumers' ability to reuse datasets efficiently?** To understand what metadata is needed, this research question explores the demand side of the metadata gap. Specifically, it identifies the metadata-related barriers that data consumers face when attempting to reuse datasets, providing insights into the metadata needed for efficient data reuse. The literature review addressed RQ2 by exploring reusability challenges on the demand side of the metadata gap; however, since the literature is more focused on supply, most of the findings were derived from interviews, which provided empirical insight into consumer barriers.
- **RQ3: How can data conversations with data producers and consumers help bridge the metadata gap?** RQ1 and RQ2 establish the metadata gap, providing a foundation for investigating how data conversations can address this divide. The interviews addressed RQ3 by simulating context-bridging data conversations and examining four key mechanisms to improve the metadata management process: incorporating dialogue, involving consumers as metadata co-creators, emphasising contextual metadata, and adapting conversation questions dynamically. These simulations provided valuable information on how such conversations can bridge the metadata gap.
- **RQ3.1: To what extent can ChatGPT-generated summaries help efficiently distil meaningful metadata from data conversation transcripts?** This question explores the potential of AI-generated summaries for managing the large volume of metadata created during data conversations. Surveys were used to address RQ3.1 by evaluating the accuracy, clarity and usability of the ChatGPT-generated summaries, demonstrating their effectiveness in distilling key metadata.

The main contributions of this thesis are:

- **Comprehensive analysis of metadata supply and demand for reusability.** Identified key motivational factors and barriers that influence metadata producers in creating high-quality metadata for efficient reuse by data consumers. Highlighted the challenges faced by data consumers during reuse and developed a framework for data management projects to pinpoint problems and opportunities within metadata practices.

¹<https://cropxr.org>

- **Evaluation of context-bridging data conversations as a metadata elicitation approach.** Demonstrated how context-bridging data conversations address critical aspects of the metadata gap, using insights from scientific literature and data conversation simulations to evaluate their effectiveness. Showed that semi-structured interactions between data producers and consumers can significantly enhance metadata quality and usability.
- **Exploration of the role of context in data reusability.** Emphasised the importance of incorporating detailed contextual metadata and diverse perspectives, particularly those of data consumers, into metadata creation and maintenance processes. Advocated for a shift from informal, experiential learning to a structured approach to systematically document and communicate the reasoning behind decisions in data creation and reuse. Illustrated how this contextual information can be effectively captured and integrated into data management systems through a proof-of-concept.
- **Practical guidelines for CropXR implementation.** Developed actionable recommendations to improve the data management practices of an interdisciplinary collaborative research institute, using CropXR as a case study. Tailored solutions by analysing how the institute's current initiatives address the factors that make up the metadata gap. Provided specific guidance on incorporating elements of context-bridging data conversations into the data management plan to improve data reusability within the institute.

The remainder of this thesis is structured as follows. Chapter 2 explores the significance of data reusability in research, with a focus on the role and current challenges of soft metadata management. Chapter 3 outlines the research design, including the development of interview and survey protocols. Chapter 4 presents the key findings derived from the interviews and surveys. Chapter 5 synthesises these results to map the metadata gap, addressing RQs 1 and 2, and evaluates how data conversations can bridge this gap, addressing RQ 3. Chapter 6 discusses the implications of the findings, explores the scalability of the proposed solution, and identifies opportunities for future research. Chapter 7 applies the research insights to the data management challenges at the CropXR institute, offering practical recommendations to improve metadata practices in an interdisciplinary context. Finally, Chapter 8 summarises the thesis and its contributions.

2

Background

This chapter provides the groundwork for understanding how data conversations can improve research data reusability by providing a comprehensive overview of key concepts and practices. Section 2.1 begins with a discussion of the importance of data reusability, highlighting its role as a cornerstone of modern research and its integration with principles such as findability, accessibility, and interoperability. Next, Section 2.2 examines the role of soft metadata in data reusability. The chapter then reviews metadata management methods and tools, identifying current approaches and their limitations in handling soft metadata. Finally, the concept of data conversations is introduced, highlighting their potential to address gaps in metadata practices by fostering direct interactions between stakeholders. The goal of this chapter is to establish a robust theoretical foundation, connecting metadata practices with the innovative potential of data conversations to advance reusability.

2.1. Importance of data reusability

Metadata plays a crucial role in ensuring data reusability, which is a cornerstone of the FAIR data principles. These principles emphasise making data findable, accessible, interoperable, and reusable [1]. Reusability, as defined by the FAIR principles, refers to the quality of data resources, tools, vocabularies, and infrastructures that allows their reuse by third parties. Achieving this requires (meta)data to be richly described with accurate and relevant attributes, accompanied by clear and accessible usage licences, detailed provenance, and adherence to domain-relevant community standards. Such characteristics ensure that research objects are well documented, legally and ethically reusable, and compatible with existing workflows, facilitating integration, analysis, and reuse by both researchers and computational agents.

The ability to combine and analyse existing datasets exemplifies the transformative potential of reusability. It allows researchers to uncover new patterns, insights, and relationships that would not be possible with isolated datasets, significantly advancing the research landscape.

2.1.1. Related data qualities influencing reusability

Reusability does not exist in isolation. Its interconnectedness with other foundational principles is already clear from the FAIR data principles, [1]. These principles collectively ensure that datasets are not only reusable, but also well-integrated into the broader research ecosystem. This section highlights some of these interdependent qualities to provide a broader context for understanding reusability, namely findability, accessibility, replicability, reproducibility, transparency, traceability, and interoperability. Although these qualities are distinct and can be addressed independently, they are deeply interconnected and collectively influence reusability by enhancing the usability, accessibility, and overall impact of datasets. This list is not exhaustive, but highlights the range of factors considered integral to reusability in this thesis, even though they are conceptually independent.

- **Findability.** Ensures that datasets are discoverable through appropriate metadata, unique identifiers, and indexing systems, making it possible for researchers to locate relevant data efficiently [1].
- **Accessibility.** Ensures that data and associated tools are openly available and easily retrievable, serving as a prerequisite for reusability [1].
- **Replicability.** Involves independently repeating an entire study, using new data and following the same methods, to verify the consistency and robustness of the results [15].
- **Reproducibility.** Requires that others can use the original data and the associated code to replicate all the numerical findings of the study, ensuring credibility and transparency [15].
- **Transparency.** Involves openly sharing the methods, decisions, and processes underlying the generation and analysis of data, enabling others to understand and scrutinise the research workflow.
- **Traceability.** Complements transparency by focussing on the ability to track the origins, transformations, and history of data or methods, ensuring that all steps in the research process are well documented and verifiable.
- **Interoperability.** Refers to the ability of data, tools, and systems to work seamlessly together across different platforms and disciplines, enabling datasets to be integrated and analysed in various workflows [1].

2.1.2. Benefits of reusability

Reusability is critical to maximise the value of research efforts, promote transparency throughout the research lifecycle, and encourage interdisciplinary collaboration. These benefits are also influenced by related qualities such as accessibility, transparency, and interoperability, reflecting the interconnected nature of these principles.

- **Efficiency and resource savings.** Reusability reduces the need for duplicate efforts in data collection and processing, saving both time and resources [12]. By allowing researchers to build on existing datasets, it streamlines workflows and accelerates the pace of scientific discovery. This efficiency is further supported by accessibility, which ensures that datasets are easily retrievable, and interoperability, which allows seamless integration into new workflows.
- **Improved reproducibility.** Simplifying reproducibility is another significant benefit. By facilitating replication of previous studies, reusability improves the integrity and accountability of research [10], [11]. Reproducible data can help identify errors, discourage fraudulent practices, and strengthen the reliability of scientific findings [13]. These results also depend on transparency and traceability, which ensure that the methods and data history are well documented and verifiable.
- **Broadened accessibility.** Reusability also makes data more accessible to a wider range of users. Easily interpretable datasets allow students, early career researchers, and those with limited technical expertise to engage in scientific inquiry [13]. Accessibility and interoperability play key roles in democratising data use as well, ensuring that datasets can be found and utilised effectively in diverse research contexts.
- **Maximized impact.** Finally, reusability amplifies the impact of research data. Collecting and processing data can be costly, and ensuring its effective reuse helps justify these investments [12]–[14]. The long-term utility of data is based on interoperability and adherence to domain standards, which facilitate integration into new research and applications.

2.2. The role of soft metadata management in data reusability

Metadata is essential to improve data reusability, as emphasised by the FAIR data principles [1]. By providing a contextual framework that transforms raw data points into meaningful, interpretable datasets, metadata enables researchers to locate, understand, and use data effectively.

However, the concept of metadata itself is not well defined, as evidenced by the diversity of definitions presented in the literature [16, p. 34]. For example, [17, p. 26] describes metadata as “a potentially informative object that describes another potentially informative object,” while [18, p. 2] defines it as “the sum total of what one can say at a given moment about any information object at any

level of aggregation.” Similarly, [19, p. 491] emphasise the structured and encoded nature of metadata as a tool for discovery, assessment, and management, while [20, p. 1] offers a simpler perspective: “the information we create, store, and share to describe things.”

In this paper, we adopt a broad definition of metadata as any information that provides context to raw data points, essentially, “data about data,” aligning with the term’s literal meaning. This broad framework also sets the foundation for distinguishing between two key categories of metadata: hard metadata and soft (or contextual) metadata [8].

2.2.1. Understanding soft metadata

Metadata is traditionally associated with hard metadata - objective and measurable attributes such as temperature, recording time, or stimulus parameters. These attributes are typically objective, quantifiable and collection can often be done automatically [8].

In contrast, soft metadata, also referred to as contextual metadata in this thesis, provides the descriptive and subjective details necessary for a deeper understanding of a dataset. Examples include the rationale behind research methodologies, explanations of failed experiments, or insights from researchers about domain-specific conventions [3], [12]. Soft metadata is critical for “meaning making,” as it offers the contextual information needed to interpret data accurately [21, p. 156]. However, capturing soft metadata is inherently more challenging because it is difficult to standardise, is often undocumented, and often requires manual input [8], [22]. Furthermore, the subjective nature of soft metadata means that what is included often depends on the individual researcher’s experience and perspective. Even when documented, it often lacks the depth and richness that can be achieved through direct discussions between data producers and consumers [8].

In general, metadata management, especially when dealing with soft metadata, can be viewed as a form of knowledge organisation. As [23] argues, knowledge systems embed data into actionable frameworks, enabling organisational learning, adaptability, and resilience. Similarly, metadata acts as an organisational artefact, transforming raw data into usable knowledge by capturing and structuring contextual information. This perspective suggests that metadata practices have the potential to go beyond documentation, functioning instead as tools for generating, disseminating, and applying knowledge across research communities. Integrating soft metadata into metadata management systems could transform them into more comprehensive knowledge systems. These systems would not only capture the information necessary for researchers to support their findings but also document the nuanced practices and conventions involved in generating the data. By preserving this contextual knowledge, such systems would enable both the data and the underlying processes to be effectively reused in future research.

A particularly interesting type of soft metadata is the actions and rationale behind the actions researchers take when creating or reusing datasets. The literature around organisational routines distinguishes between ostensive routines (structured, documented processes) and performative routines (improvisational, real-world practices) [24]. In the current context, performative routines refer to the actual steps and decisions that researchers make during the creation or reuse of a dataset. These routines are rarely collected or discussed, yet they offer crucial insights that go beyond the simplified, ostensive routines typically described in the methodology of a dataset. Documenting performative routines through metadata provides valuable contextual information that enhances reusability and minimises the need for repeated workarounds [25].

2.2.2. Current metadata management methods and tools

Metadata management encompasses a wide range of practices aimed at making datasets interpretable, reusable, and interoperable. These practices can be broadly categorised into three primary functionalities: capturing metadata, standardising metadata, and collaborating on metadata creation. Although each approach addresses specific challenges, significant gaps remain in effective capture of soft metadata, which requires context, nuance, and adaptability.

A recent preprint, [26], underscores that the distance between data creators and reusers remains a key challenge for metadata management. The authors identify six dimensions—domain, methods, collaboration, curation, purposes, and time—that hinder knowledge exchange. This work highlights the ongoing need for metadata practices that provide nuanced, context-rich information to bridge these metadata supply and demand gaps.

Capturing metadata

Efforts to capture metadata are often producer-led, relying on researchers who know the dataset best to document its attributes. This approach is logical, as metadata is traditionally created to support research findings and ensure the usability of the dataset. However, creating metadata to meet immediate research needs may not fully align with the requirements of metadata for long-term reuse. Producers face competing demands, limited time, and a lack of dedicated tools, which frequently result in incomplete or inconsistent metadata, particularly for soft metadata. As a result, producer-led efforts often prioritise hard metadata, objective measurable parameters such as temperature or recording time, while neglecting the subjective descriptive information needed for broader interpretability.

In addition to producer-led efforts, dedicated roles like data stewards can support researchers in metadata management [27]. These professionals provide guidance and expertise to help researchers create high-quality metadata without requiring them to become data management experts. Although effective, this approach relies on institutional support and is resource intensive.

Automation offers an alternative to metadata capture, particularly through provenance tracking systems that document data lineage and transformations. These tools are widely used in machine learning research and other data-intensive research fields, where reproducibility and traceability are critical. Although automation reduces manual workload and ensures consistency, it is poorly suited for capturing soft metadata, as much of the necessary contextual information is never formally documented.

Standardising metadata

Standardisation plays a central role in creating consistent and structured metadata. Common methods include metadata templates, controlled vocabularies, taxonomies, and ontologies, all of which aim to facilitate machine and human readability. However, these approaches face a fundamental challenge: balancing the need for structure with the flexibility required to capture diverse and nuanced data. Standardisation is often easier to implement within a single domain or for a specific type of dataset, where shared conventions and practices provide a common foundation. However, when dealing with interdisciplinary datasets or diverse user groups, predefined structures may not accommodate the variability and complexity of soft metadata.

Highly structured metadata can simplify integration and improve searchability, but it often imposes rigid frameworks that make it difficult for producers to fully capture the complexities of their datasets. For example, metadata templates and controlled vocabularies are built on the assumption of a predefined understanding of user needs, which may not align with the evolving requirements of data consumer needs [7].

Recognising these limitations, some approaches let go more of rigid standardisation in favour of more inclusive and adaptable frameworks. For example, wiki-based systems, such as semantic wikis, allow users to collaboratively negotiate meanings and definitions, fostering interoperability and adaptability [7], [28]. Although these systems encourage adaptability, they leave unresolved the critical issue of efficiently capturing metadata.

Collaborating on metadata

To capture diverse perspectives on metadata and strike a better balance between standardisation and adaptability, several collaborative approaches have been proposed. These methods leverage community involvement to create higher quality metadata by incorporating multiple points of view and spreading the workload [20].

Tagging systems allow users to apply descriptive tags, enhancing searchability and supporting context-specific metadata generation. Collaborative tagging, in particular, distributes the burden of metadata creation across a community while incorporating diverse perspectives [29]. However, researchers often have highly personal ways of organising their work, and existing tools for semantic annotation and tagging lack the customisation and personalization needed to meet individual preferences [6].

Network-based systems, such as metadata association networks and linked open data, take collaboration a step further by linking related metadata through semantic annotation and cross-matching. These systems improve metadata reusability by providing richer context and facilitating integration across datasets [20]. Despite their advantages, these approaches remain heavily reliant on machine-readable formats and often fail to address the human-centred needs of soft metadata, such as subjective insights and nuanced context.

Although these collaborative approaches incorporate more diverse perspectives, soft metadata is still not given sufficient attention. Direct collaboration through data conversations provides a powerful, albeit resource-intensive, method for capturing nuanced and contextual information. By allowing producers and consumers to exchange information directly, these conversations can bridge gaps that formalised or automated approaches cannot address. The strengths and challenges of data conversations will be explored in detail in the next section.

Innovations in collaborative data management tools

A variety of features have been developed to facilitate collaborative data management, offering innovative features to enhance metadata practices. This section highlights two notable examples, GitHub Discussions and OpenMetadata, to explore their relevance in the context of this research.

GitHub Discussions¹, launched in 2020, is a feature in GitHub repositories designed to create a dedicated space for project-related conversations. Unlike Issues or Pull Requests, which focus on specific tasks or code changes, Discussions provide a platform for broader dialogues, such as conceptual questions, general feedback, and troubleshooting. Although primarily geared toward software development, GitHub Discussions illustrates the potential benefits of fostering collaborative environments for knowledge exchange. As [30] found, the feature played a crucial role in advancing project development, demonstrating how open dialogue can improve teamwork and problem solving.

OpenMetadata,² launched in 2021, is a platform explicitly designed for collaborative metadata management. It includes features such as activity feeds, glossary creation, and discussions directly linked to datasets, fostering collaboration through interactive tools such as conversation threads, mentions, and emoji reactions. In addition to facilitating team engagement, OpenMetadata supports automated metadata generation, including data profiling and lineage tracking, which streamlines the documentation of data processes. Its “Knowledge Center” centralises long-form contextual insights and links them to specific datasets, enhancing the richness and usability of metadata. This integration makes OpenMetadata a compelling tool for managing both hard and soft metadata, bridging gaps in traditional metadata practices by combining automation with collaboration.

Since both GitHub Discussions and OpenMetadata are relatively new, more research is needed to evaluate their long-term impact, effectiveness, and potential applications, particularly to bridge the gaps between hard and soft metadata and extend beyond their primary contexts. Nonetheless, their development demonstrates a growing interest in an expanded view of what information is considered relevant, emphasising the importance of collaboration and contextual knowledge in modern data practices.

2.2.3. Data conversations

Data conversations refer to the exchange of information about datasets, often focussing on metadata, methodologies, and usage implications. They encompass interactions between data producers, users, and other stakeholders. A familiar example is informal, in-person dialogue between researchers, such as when a principal investigator discusses dataset specifics with successors. These discussions often focus on the reasoning behind data collection decisions or contextual nuances, highlighting aspects of soft metadata.

Defining data conversations: perspectives from the literature

The term “data conversations” encompasses a variety of interactions centred around datasets, from informal dialogues to structured professional frameworks.

[31] define data conversations as “informal, lunchtime talks with time for discussion” [31, p. 79] focussing on using researchers’ enthusiasm to explore the human aspects of research data management (RDM). This definition emphasises informality and personal engagement, positioning these conversations as a space to share experiences, discuss challenges, and connect on a human level about data practices.

[32] approach data conversations from a more formalised perspective. In their work on data literacy coaching, they highlight how data conversations can facilitate professional development. Here, data conversations are not just informal exchanges, but organised, intentional dialogues designed to build data literacy capacity and enhance understanding among participants. This definition expands the

¹<https://github.com/features/discussions>

²<https://open-metadata.org>

concept beyond spontaneous discussion to include methodical efforts aimed at improving practical data skills.

[33] provide yet another lens, emphasising the fluid and multifaceted nature of data conversations. Their research discusses interactions ranging from real-time collaborations to asynchronous exchanges, such as commenting on datasets or providing feedback on public repositories. This view highlights the diversity of contexts in which data conversations occur, suggesting that these exchanges are not confined to specific settings or methods but can encompass a wide array of participatory and observational interactions.

Finally, [34] underscore the value of data conversations in understanding user needs and tailoring services to researchers. In their study, data conversations emerge as tools for service design, enabling researchers and support staff to co-create solutions for data management challenges. This definition reinforces the idea that such interactions are instrumental not just for individuals but also for institutional development, providing a foundation for responsive and user-centred systems.

These varying perspectives on data conversations illustrate a continuum ranging from informal exploratory discussions to structured, goal-orientated frameworks. While [31] stress the importance of human connection and informality, [32] and [34] emphasise the role of formalised approaches in achieving specific professional or institutional outcomes. [33] position data conversations as enablers of collaboration, addressing practical needs such as feedback, quality assurance, and building trust. Together, these perspectives reveal the adaptability of data conversations, demonstrating their utility in diverse research, professional, and institutional contexts.

Benefits and challenges of data conversations

Data conversations provide detailed context that other metadata elicitation methods often miss, improving the comprehension and usability of datasets [8], [33]. By fostering dialogue between data producers and users, these interactions bridge disciplinary gaps, encourage interdisciplinary understanding, and promote the reuse of datasets in innovative ways.

However, despite these benefits, data conversations face notable challenges. Scalability remains a significant issue, as these interactions often lack structure or documentation, making it difficult to extend their benefits to larger teams or organisations [32], [34]. Furthermore, the ad hoc nature of many conversations limits their ability to provide durable and systematic records, reducing their long-term utility [31], [34]. Inclusivity can also be a barrier, as not all collaborators have equal access to participate, particularly in distributed or asynchronous research environments where participation opportunities are unevenly distributed [33].

This chapter highlights the pivotal role of metadata in ensuring data reusability and demonstrates how interdependent qualities like findability, transparency, and interoperability enhance the wider research ecosystem. Soft metadata emerges as a key enabler for deeper dataset understanding, though its capture and management remain significant challenges due to its subjective and contextual nature. Current metadata practices are useful, but do not adequately address soft metadata management, leaving a metadata gap. Data conversations offer a solution by facilitating direct exchanges between data producers and users, bridging gaps left by formalised and automated approaches. However, challenges related to scalability and documentation persist. These insights underscore the need for innovative strategies to integrate soft metadata effectively and harness data conversations for broader, more impactful improvements in research data reusability.

3

Methodology

The methodology of this thesis consists of three main components: a literature review, interviews, and surveys. The literature review provided the foundation for designing the interview structure, ensuring alignment with the research objectives.

The interviews captured researchers' real-life experiences with creating and reusing datasets. More importantly, they served to simulate four key mechanisms of context-bridging data conversations with both producers and consumers, conducted within a controlled interview environment. The interview setting did not replicate the complexities of integrating these conversations into a data management system. These simulations act as proof-of-concept to test and demonstrate the potential of context-bridging data conversations.

Furthermore, the study tested how well AI could summarise the transcripts of these data conversations, with the aim of making the communication of metadata generated during the conversations easier. The surveys evaluated the accuracy and clarity of these AI-generated summaries, both independently and in comparison to full interview transcripts. This evaluation provided insights into the usefulness of AI summaries in improving metadata reusability.

The surveys evaluated the accuracy and clarity of these AI-generated summaries, both independently and in comparison to full interview transcripts. This evaluation provided insights into the usefulness of the summaries in improving metadata reusability.

Figure 3.1 presents an overview of how these components interconnect and support the findings discussed in Section 5. An overview of the research design is given in Section 3.1. For further details on the interviews and surveys, refer to Sections 3.2 and 3.3, respectively.

3.1. Research design

This section describes the approach of the study to investigating the metadata gap using mixed methods. The methodology includes a literature review, interviews, and surveys, in conjunction with participant recruitment and anonymisation procedures, to ensure robust and ethical research practices.

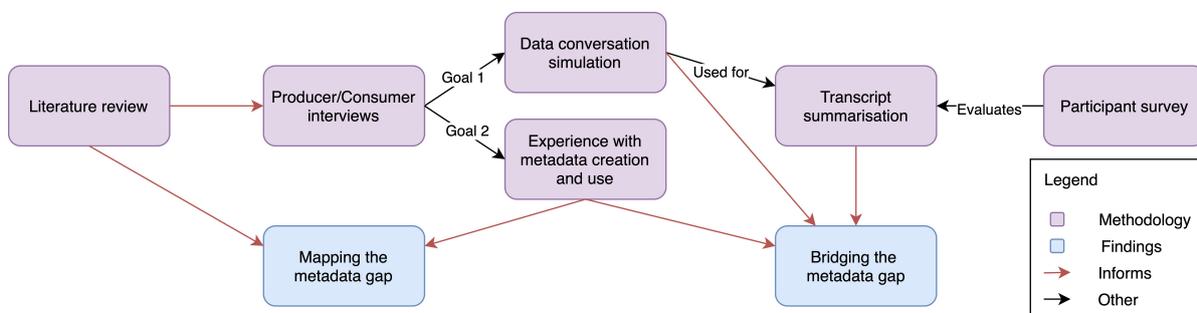


Figure 3.1: Overview of the interconnections between methodology components and their support for the findings discussed in Chapter 5. The figure illustrates the relationships and dependencies among the components, highlighting their roles in the study's conclusions.

3.1.1. Research methodology components

The study used three methods: a literature review to examine existing metadata practices, qualitative interviews to explore participant experiences and challenges, and surveys to evaluate data conversation outcomes and AI-generated summaries. The literature review and interviews addressed RQ1 and RQ2, focussing on motivations, barriers, and challenges related to metadata for producers and consumers. Interviews and surveys explored RQ3, investigating how data conversations and AI-generated summaries (RQ3.1) could help bridge the supply-demand gap.

Examining metadata practices through literature review

The literature review explored the gap between metadata standards and real-world practices, focussing on contextual metadata from both supply and demand perspectives. The analysis explored data reusability and how contextual, or soft metadata, is considered on both the supply and demand sides of the gap, as well as existing methods to capture this information (see Chapter 2). Rather than specifying the exact types of metadata that should be included for reusability, the review acknowledged that this varies between domains and is best determined by domain experts. The review considered research on collaborative data stewardship, exploring how collaborative approaches could enhance metadata practices with the aim of improving data reusability.

RQ1 was addressed by exploring the motivations and barriers that researchers had identified to provide high-quality metadata. Although no existing theoretical framework perfectly fit the study's focus, the Self-Determination Theory by [35] was used to structure the motivational aspects. This theory aligned well with the review's findings in regards to the the importance of fostering intrinsic motivation for metadata creation. The demand side of the metadata gap was also explored by looking at the reusability challenges faced by data consumers when metadata is incomplete or insufficient, informing RQ2. To bridge these gaps, the literature review explored potential collaborative solutions, emphasising data conversations as a promising approach to improve metadata creation for reusability, informing RQ3.

Exploring participant experiences with interviews

The interviews were designed as a qualitative study to investigate metadata practices, motivations, and barriers among participants, based on and expanding the findings of the literature review. The interview questions, shaped by the literature review, were developed to achieve two main goals.

Firstly, the interviews were designed to capture the metadata practices and challenges of the participants, providing empirical evidence to address RQ1 and RQ2. This part of the study focused on mapping the metadata gap by examining the needs and behaviours of data producers and consumers, highlighting discrepancies between the metadata currently available and what is required for effective dataset reuse. This research focusses on the internal motivations of producers to create high-quality metadata, as there are not many external incentives to prioritise metadata for reusability (see Section 2.2). Unlike data consumers, who need metadata to complete their work and can switch datasets if necessary, producers do not face an absolute requirement to provide high-quality metadata. Their decision to do so is driven by intrinsic factors, making their motivations crucial to understand. Motivations identified during interviews are categorised according to the three fundamental human needs for intrinsic motivation described by: competence, autonomy, and relatedness. These categories are part of self-determination theory (SDT) [35]. In this study, we focus primarily on relatedness and competence, while addressing autonomy only peripherally, as research activities are typically already highly autonomous. More information on each category can be found in Section 5.1.1.

Motivations are closely tied to the barriers that can hinder or shape them. This research categorises these barriers into constraints, external factors such as time or resource shortages, and limitations, internal factors such as lack of knowledge or skills. Producers face barriers such as time constraints, lack of recognition, or uncertainty about data usage, which can hinder their motivation to provide high-quality metadata. In contrast, data consumers rely on metadata to complete their work; their motivation to use metadata is inherent because reusing datasets is often their only option. Creating their own dataset would require significantly more resources. For this reason, we focus only on barriers to reuse for consumers, as their motivation is already built-in.

Second, the interviews included simulations of the proposed context-bridging data conversation framework. These simulations tested four key mechanisms to address Research Question 3 (RQ3) by improving the extraction of contextual metadata from both producers and consumers, helping to bridge

the metadata gap. The data conversations were designed to simulate a dialogue with a data steward, providing a proof-of-concept to examine the types of metadata participants create, reference, or require during dataset production or reuse. A data steward is a professional responsible for ensuring that project data is well organised, documented, and compliant with standards, enabling efficient reuse and interoperability. In these simulations, the interviewer acted as a proxy for the data steward, identifying gaps in metadata and prompting participants to reflect on their practices and needs. By approximating the role of a data steward, the conversations highlighted the challenges participants face in creating or using metadata, while exploring ways to enhance dataset reusability. The focus was particularly on contextual or (soft) metadata that participants encountered during dataset creation or reuse. Producers were asked about the considerations they made to improve the reusability of their datasets, while consumers discussed the types of metadata they relied on to effectively locate and reuse data. Although these simulations were not part of a larger data management plan, they demonstrated the potential of data conversations as a tool for metadata extraction.

Evaluating data conversation summaries with surveys

Each of the four context-bridging data conversation mechanisms was designed to increase the amount of metadata generated for a dataset. To be able to still communicate this information efficiently, we tested whether ChatGPT-generated transcript summaries could preserve key information while improving readability (addressing RQ3.1). To evaluate these summaries, participants were sent a survey asking them to rate various qualities and compare the summary with the full transcript. A survey was chosen because it provided a low barrier and a consistent way for participants to give feedback.

3.1.2. Participant selection and data anonymisation

Participants in this study were selected through the author's social and professional network, including outreach through the CropXR Slack channel. We aimed to gather individuals with various levels of research experience and from a wide range of academic and professional domains. The study intentionally focused on individuals with a STEM-focused academic background (Science, Technology, Engineering, and Mathematics) SDT whose primary expertise did not include data management. The focus was on interviewing people with little to no formal data management education, as they represent a significant portion of the research community and have the most potential for improvement in metadata practices. This focus also aligned with the CropXR project, where many biologists have developed their data management skills through experience rather than formal training. Although this emphasis may limit the generalisability of the findings to non-STEM fields, it provides an opportunity to identify ways to improve metadata practices for those who frequently work with data but lack formal training. Furthermore, their reliance on large datasets underscores the importance of effective metadata management in their work.

Ethical research practices included informed consent (see Appendix A), voluntary participation, and pseudo-anonymisation of data by removing identifiable details (e.g., names, organisations). Despite these precautions, complete anonymisation remains challenging, and participants were informed of their rights to review and request transcript corrections.

3.2. Exploring data conversation potential through interviews

The interviews aimed to explore the needs, practices, and challenges faced by both data producers and consumers regarding metadata, with a focus on improving dataset reusability. They also incorporated a simulation of the proposed context-bridging data conversation framework. Section 3.2.1 details the four key mechanisms tested during the data conversation simulation. Section 3.2.2 provides an overview of the interview questions, while Section 3.2.3 describes the process of conducting the interviews. Lastly, Section 3.2.4 addresses the limitations of this part of the methodology.

3.2.1. High-level data conversation design

This research focused on testing four key mechanisms that were expected to have a significant impact on improving metadata extraction compared to traditional methods. These mechanisms were selected because they could be practically tested within the interview setting, each addressing different aspects of the metadata management process. Although the interview was subdivided into a data conversation segment and a more general section focussing on experience, this distinction was made primarily to

test the effectiveness of AI-generated summaries (see Section 3.3). However, the mechanisms were evaluated throughout the interview process.

- **Consumers as co-creators.** To address the limited perspective and low incentives that producers face when providing high-quality metadata, we involved data consumers in the metadata creation process. By including consumers, we hypothesise that metadata can be better aligned with metadata needs of the consumer, thus improving its relevance and usability.
- **Contextual metadata.** The inclusion of contextual metadata, such as decision rationales and detailed background information, was chosen to address its absence in current practices despite its importance for reusability. As highlighted in the literature review (see Section 2.2), contextual metadata provides critical details that enhance the interpretability and application of datasets.
- **Real-time dialogue.** Dialogue, defined here as a live and interactive exchange of information, introduces a dynamic approach to metadata creation by allowing participants to clarify and elaborate in real time. Unlike written metadata forms, real-time dialogue can reduce barriers by making the process faster and more engaging. Additionally, it captures detailed and nuanced information that improves the quality and depth of metadata.
- **Question adaptation.** This mechanism is designed to refine the interview questions during and between sessions to maximise the flexibility of the interview format. Adapting questions allows the metadata elicitation process to adjust to the evolving needs of data consumers, simulating a dynamic and responsive data conversation system.

The following explains how each mechanism was incorporated into the interview design. The next section covers the details of the interview structure for each question.

- **Consumers as co-creators.** To test the usefulness of involving consumers as metadata co-creators, the interviews examined whether producers' awareness of potential consumers influenced their metadata decisions and explored consumers' actual needs to better inform these decisions.
- **Consumers as co-creators.** This mechanism was incorporated by examining whether producers' awareness of potential consumers influenced their metadata decisions. Additionally, the interviews explored the specific needs of consumers, providing insight into how their involvement could benefit metadata management practices.
- **Contextual metadata.** To gain insights into current perspectives and experiences with contextual metadata from both producers and consumers, we examined how each group interacted with and used this type of information. For producers, the focus was on whether they had considered including contextual metadata in their practices. For consumers, we investigated whether they frequently identified this type of metadata as useful when reusing datasets.
- **Contextual metadata.** To understand current perspectives and experiences with contextual metadata from both producers and consumers, we explored how each group engaged with this type of information. For producers, we examined whether they had ever considered including this type of metadata. For consumers, we investigated whether this was the kind of information they often found missing when reusing datasets.
- **Real-time dialogue.** To encourage open and dynamic exchanges, the interviews were conducted in a informal, adaptive style with open-ended questions. This allowed participants to naturally introduce topics of importance to them, providing insight into areas they found significant without being led by predefined prompts.
- **Question adaptation.** This mechanism was implemented by dynamically refining questions during and between interviews. Conversations were allowed to evolve organically, with participants' responses shaping the direction of the interview. Between sessions, questions were adjusted according to their effectiveness, with new prompts or questions added to explore emerging themes and less effective ones removed to maintain focus and flow.

3.2.2. Question-level interview design

The interview was divided into five components: an introduction, followed by producer and consumer focused sections, each containing a data conversation simulation and background questions. Each

participant was asked the introductory questions and depending on whether they identified themselves as a producer, consumer, or both, the corresponding sets of data conversation simulations and background questions were given. The introduction established the interview structure, clarified the relevant terminology, and collected information on the participant's research and data management experience. The introduction was directly followed by the data conversation component to avoid preparing the participant with any other questions. The data conversation component of the interview simulated a potential interaction with a data steward, focusing on the types of metadata participants created or referenced when producing or reusing datasets. These conversations aimed to capture additional metadata after the creation or use of the dataset, providing information that could help future users. For producers, this included metadata that could be useful for different audiences or use cases. For consumers, the focus was on identifying the challenges they faced that could inform future users. The background questions were then used to explore any details that did not fit in the previous components. A complete overview of the questions can be found in Appendix B, any changes to the original list of questions are also indicated here.

Introduction component

At the beginning of the interview, we explain the structure of the interview and introduce the relevant terminology to avoid miscommunication. This was followed by background questions to provide context, ensuring a clear understanding of each participant's research focus and experience level. For the complete interview introduction, refer to Appendix B.1.

As discussed in Section 2.2, metadata is not a well-defined term and especially for people from different domains it can have a different meaning. This thesis also considers information not typically seen as metadata, such as methodology or discussion. To make sure that every participant had a similar definition in mind, we introduced our definition of metadata at the beginning of the interview as follows.

Any data that describes or provides information/context to raw data. This includes details such as the creation date, the data owner, comments, data types, methodology, and more.

The terms data producer and data consumer were defined as these concepts are uncommon. In contrast, findability and reusability were not introduced explicitly, as they are well known and self-explanatory. However, participants were encouraged to seek clarification if needed.

After explaining the details of the interview, the participant was asked a couple of background questions about their research and data management experience; see Appendix B.1 for details. As discussed in section 2.2, the overall research experience can have a large impact on how a researcher creates and interacts with data. This is especially relevant for researchers who completed their degree several years ago, when data management was even less a part of the conversation. Knowing the research area and type of data provides the interviewer with context for the questions in the rest of the interview. Although many domains now deal with data extensively, it is not common for researchers to have had formal data management education. Thus, interviewees are asked about their last completed degree in their domain and whether data management was part of this education. Participants can also indicate they took more formal data management courses after completing their degree. The total data management experience is then roughly measured in years, including both formal education and practical experience.

Producer components

The producer interview questions focused on understanding how participants create metadata, including their considerations for findability, reusability, and future user and use cases. Key topics included addressing potential metadata gaps, evaluating the role of collaboration and feedback, and exploring sources of motivation to improve metadata quality, particularly through the lens of contextual metadata. Table 3.1 presents the key themes discussed in this part of the interview, while a complete list of questions and details can be found in Appendix B.2.

Data conversation simulation. To start the data conversation simulation part of the interview, the producer is first asked to focus on a specific dataset they helped create. Specifically, the participant was asked to focus on a dataset that was published with as wide an audience as possible. At times, datasets are published solely for internal use within an organisation, allowing data producers to make certain assumptions about future users. By focussing on widely shared datasets, we could better evaluate the

Table 3.1: Overview of the producer interview, categorised by section and theme. The data conversation simulation prompts the producer to focus on a specific dataset and reflect on their initial metadata considerations. They are then asked whether thoughts about potential future users or use cases influenced these considerations. Following this, the producer's general experience with metadata is explored, including whether they have encountered contextual metadata, the role of collaboration in metadata management, and their motivations. The full list of interview questions is provided in Appendix B.2.

Section	Theme	Explanation
Data conversation simulation	Dataset focus	Highlights the focus of the interview on widely shared datasets to assess considerations for unknown future users.
	Initial focus	Identifies key metadata priorities by exploring initial considerations for findability and reusability.
	Future user	Analyses how future users' knowledge and needs influence metadata decisions.
	Future use case	Explores how potential future applications of a dataset impact metadata choices and quality.
General (meta)data creation experience	Contextual metadata	Discusses the importance and perceived value of contextual metadata in data management.
	Collaboration	Explores the role of collaboration, including input during metadata creation and feedback afterward
	Motivation	Examines how future dataset uses can motivate better metadata creation.

types of considerations producers make in situations where future users and their needs are largely unknown. Producers were asked to imagine that they would have a conversation about this dataset on upload to a data repository, to simulate the data conversations being part of a larger data management system.

The data conversation began by asking the producer what considerations they remembered making regarding findability and reusability. This was posed as the first question to identify which information came to mind most readily for the participants.

Participants were then asked whether they had thought about the potential future user of the dataset, specifically regarding the user's familiarity with domain knowledge or data management practices. They were asked if these considerations had any impact on the metadata they chose to include. This question aimed to identify any metadata gap from the supply side, exploring whether producers took into account the needs of users on the demand side. Some producers may assume that future users have similar experience to their own, which could lead them to overlook certain metadata needs. Following this, producers were asked whether these considerations about future users influenced their decisions regarding metadata for increased findability and reusability.

Next, similar questions were posed, this time focussing on potential future use cases rather than future users. The intent was to explore whether thinking about various possible applications of the dataset might prompt producers to include additional information to support different research approaches. Participants were asked if they had considered potential future uses of their dataset and if not, to do so now. They were then encouraged to reflect on whether this awareness could have changed the information they chose to include. For instance, they were prompted to consider what data quality level might be necessary for different use cases, whether they had envisioned only extensions of their own research or entirely new approaches, and how these considerations could influence the metadata they provided. The participants were finally asked whether thinking about future use cases would affect the information they could include to improve findability and reusability.

General metadata creation experience. After the data conversation simulation, producers were asked about their general experiences with metadata management to provide context on their approach to creating metadata and to identify factors that could enhance their motivation for this task. With a clearer understanding of soft or contextual metadata, producers were asked if they had encountered similar types of metadata in their research. After discussing soft or contextual metadata in the simulation of data conversation, producers had a clearer understanding of these concepts. They were then asked whether they had encountered similar types of metadata in their research to assess its perceived value in data management practices.

The next set of questions focused on collaboration, examining two key aspects: whether producers

sought input from others during the metadata creation process and whether they received feedback afterward. The first question sought to determine whether producers recognised the value of multiple perspectives during metadata creation and if they would seek assistance when faced with challenges. The second addressed the lack of built-in feedback mechanisms in many data management systems, exploring whether receiving feedback made producers feel their efforts were valued. This feedback was hypothesised to encourage producers to invest more effort in metadata quality, driven by recognition or appreciation from peers. However, these topics often arose naturally during the data conversation simulation, and when asked separately, they rarely yielded additional valuable insights.

The final question also explored potential sources of motivation, specifically whether considering the future uses of their datasets could encourage producers to invest more time in creating high-quality metadata. The hypothesis was that producers would be more likely to provide detailed metadata if they felt connected to future users or perceived future applications of their data as meaningful.

Consumer components

The consumer interview questions focused on understanding how participants select, evaluate and reuse datasets, including their challenges with metadata, the role of collaboration, and their engagement with contextual metadata to identify areas for improvement in data reuse practices. Table 3.1 presents the key themes discussed in this part of the interview, while a complete list of questions and details can be found in Appendix B.2.

Table 3.2: Overview of the consumer interview, categorised by section and theme. The data conversation simulation prompts the consumer to focus on the challenges faced when reusing publicly available datasets without insider knowledge and to reflect on their metadata priorities during the initial evaluation. Consumers are also asked to explore gaps in metadata and additional resources needed for effective dataset reuse. Following this, the consumer's general metadata reuse experience is discussed, including the role of collaboration, demand gaps in locating required information, and their engagement with contextual metadata and methodologies to support future users. The full list of interview questions is provided in Appendix B.3.

Section	Theme	Explanation
Data conversation simulation	Dataset focus	Examines challenges consumers face when reusing publicly available datasets without insider knowledge.
	Initial evaluation	Identifies the metadata elements consumers prioritize when determining a dataset's suitability.
	Metadata reuse process	Explores gaps in metadata and the additional resources consumers need to reuse datasets effectively.
General (meta)data reuse experience	Collaboration	Assesses the availability and necessity of support or collaboration during dataset reuse.
	Demand gaps	Investigates how easily consumers locate required information and the common methods used to fill metadata gaps.
	Contextual metadata	Examines how consumers engage with contextual metadata and share their methodologies to support future users.

Data conversation simulation. To start the data conversation simulation, participants were again asked to focus on a specific dataset. This time, the focus was on an existing dataset that they had used for their own research. Participants were encouraged to select a dataset that was publicly available or intended for external audiences, as this required them to engage with metadata without the benefits of access to insider knowledge or informal communication with the producer. They were also encouraged to select a dataset that had caused them particular trouble when reusing. Because the data conversation was limited to one dataset per participant, this approach was intended to generate more insightful discussions. Consumers were asked to envision the start of this conversation occurring at the time of dataset download and later at the time of uploading their results.

The data conversation began by asking consumers to recount their initial criteria for selecting the dataset. This question aimed to capture the metadata elements that consumers prioritised when determining if a dataset is suitable for their needs.

The participants were then asked if they had experienced specific challenges with the metadata of the data set, including whether they felt compelled to seek additional information to address gaps. This question was intentionally broad and open-ended to see which issues the participants would raise first. This approach also provided insight into whether participants viewed metadata as an expansive concept

or if they tended to rely on a more narrow interpretation. After their initial responses, participants were offered prompts to clarify if they referenced external methodologies or sources, helping to determine if they engaged with metadata in ways they might not have initially recognised as metadata practices. The goal was to determine whether consumers identified metadata needs that are not currently recognised as essential for data reuse. This question also assessed whether the metadata provided was sufficient to support reuse or if consumers needed to seek additional resources to fill the gaps. These external resources could include clarifications on data types or details on the data production methodology. The responses highlighted metadata that was often needed but not provided, as well as the ways consumers address these gaps. If there is frequent overlap in these metadata needs, information collected by one user could benefit others, demonstrating how consumers can contribute to closing the metadata supply gap. This section concluded by asking whether their final method aligned with their initial plan, with the aim of gauging how well participants were able to predict their needs beforehand.

General (meta)data reuse experience. Similarly to the producer side, consumers were asked about their general experiences with metadata management to provide context on their approach to reusing metadata. However, the questions for consumers were more focused on a single dataset, highlighting the unique challenges and insights involved in metadata reuse.

Participants were first asked whether they had received any assistance with domain knowledge or data management while working with datasets. They were also asked whether understanding how others had used the data set, either for similar or different use cases, would have been helpful. These questions were aimed at uncovering whether collaboration or support was necessary and accessible.

To further explore the needs of dataset consumers, participants were asked if the information they required was easy to locate and how much extra time or effort was needed. They were also asked where they found the information to identify common methods for addressing gaps, such as metadata, methodologies, or external sources (e.g. Stack Overflow, GitHub).

To examine how consumers engage with contextual metadata, they were asked if they had published their methodology and if it was accessible to others using the same dataset. The underlying idea was that if participants found dataset reuse challenging, they might take steps to help future users. Finally, they were invited to share any additional difficulties they faced during dataset reuse and to suggest improvements that could make finding and reusing datasets easier in the future.

3.2.3. Process of conducting interviews

To ensure consistency, a standardised setup was used in all interviews. Research shows that the format of an interview can influence responses [36]. Each interview was conducted through Microsoft Teams to accommodate participants from various locations throughout the country. The setup included a high-quality webcam at eye level to enhance image quality and foster engagement. Although direct eye contact was not possible, which may have limited the sense of personal connection, non-verbal gestures still contributed to clarity. Despite occasional technical disruptions, such as video freezing, participants were familiar with the format of online meetings due to the COVID-19 pandemic.

Following each interview, Microsoft Teams generated an automatic transcript. The transcripts were then anonymised by removing identifiable details such as location names and company names. After anonymisation, they were sent to participants for review and approval, with adjustments made as necessary. To preserve participants' privacy, the transcripts are not provided.

Analysis of interview data

The interview data was analysed through a structured process to identify key insights into motivations, barriers, and data conversation mechanisms. The initial coding involved reviewing each transcript and noting insights for each question. Because adaptive questioning led to responses being out of sequence, insights were organised into a comprehensive table grouped by question. Then a thematic analysis was performed to identify recurring patterns and themes. Finally, the results were synthesised with findings from the literature to produce a detailed list of motivations, barriers, and sub-mechanisms.

3.2.4. Interview design limitations

The interview design was shaped by practical constraints and trade-offs, which influenced the consistency, scope, and specificity of the findings.

Evaluating subjective experiences is inherently challenging, as participants' interpretations and motivations are difficult to assess consistently. This limitation was compounded by a small participant

pool and the lack of a preliminary pilot test to refine the questions. Although a pilot could have improved reliability, time and resource constraints required an adaptive, iterative approach. This allowed real-time adjustments to maximise the depth of responses, but reduced overall standardisation.

To ensure broad applicability, the study avoided technology-specific questions, making the findings more relevant across domains. However, this limited the ability to identify specific soft metadata attributes critical for improving data reusability.

Despite these constraints, the interviews offered valuable insight into metadata practices and challenges, forming a strong foundation for further analysis.

3.3. Evaluating data conversation summaries through participant surveys

The context-bridging data conversation approach proposed in this thesis was designed to generate a large volume of complex metadata. The approach incorporates consumers as additional metadata contributors and includes detailed contextual metadata. Real-time dialogue introduces complexity with filler words, indirect responses, and clarifications. Adaptive questioning creates variability between interviews, further increasing the complexity.

To address these challenges, AI-generated summaries were used to extract relevant metadata from data conversation segments in the interview transcripts, making the information more concise and usable than the full transcript. This approach streamlined key details and helped identify relevant answers efficiently, avoiding the need to sift through entire transcripts.

This section describes the process of generating AI summaries and evaluating their quality and effectiveness using participant surveys. The primary objective was to explore whether AI summarisation shows potential for incorporating data conversations into a data management system, thus addressing Research Question 3.1.

3.3.1. Process of creating transcript summaries

This section outlines the rationale for choosing ChatGPT, the prompt engineering approach, the design considerations for the summaries, and the privacy measures taken to protect participant data. Additionally, key decisions regarding transcript processing are explained.

ChatGPT, which became publicly available in November 2022, quickly gained widespread usage. Since then, numerous other chatbots based on Large Language Models (LLM) have also gained popularity, each with its own unique characteristics and strengths [37]–[39]. For text summarization, research on the effectiveness of specific models remains limited, as highlighted by a recent preprint that compares several well-known models but does not include the most recent versions of ChatGPT [40]. Since this thesis focusses on broader trends rather than detailed evaluation of summarization quality, we selected ChatGPT for its accessibility and widespread adoption, prioritising practical exploration over model-specific comparisons. If future research identifies a chatbot better suited for the tasks described, integrating it into this workflow should be relatively straightforward. ChatGPT offers various LLM versions, each with different qualities. The transcript was summarised using ChatGPT-4o,¹ the latest version ChatGPT general purpose model available as of November 2024.

To ensure realistic conditions and reduce data processing time, typos were not removed from the transcripts. Previous research indicates that chatbots can effectively handle typographical errors in interview transcripts [41]. Furthermore, only the data conversation segments of each interview were summarised, which focused on interactions with a specific dataset rather than general participant experiences. This decision preserved participant privacy by excluding personal reflections or a broader context that could reveal identifiable information.

Due to the limited research on prompt engineering for transcript summarization and the frequent updates to the LLMs that power chatbots, we adopted a minimalist approach to prompt design. This aligns with recent findings suggesting that advanced reasoning LLMs may render complex prompt engineering techniques less effective or even counterproductive [42]. As a result, we focused on simple a simple prompt engineering approach. Each interview transcript had unique characteristics, making it difficult to create a single optimised prompt that worked universally across all cases. The goal was not to generate a perfect summary, but to assess the overall viability of using AI-generated summaries with

¹<https://openai.com/index/hello-gpt-4o/>

minimal effort. Generating just one summary per transcript allowed us to keep the evaluation workload manageable for participants, while maintaining focus on the main objective. The following prompt was used to generate the summaries, and the specific design choices are explained below.

The attachment contains several transcripts from interviews with data [type of data user] who have [type of data use] datasets in their research. Please provide a high-quality summary for each transcript using bullet points and dividing each summary into sections that best fit the topics discussed in that transcript. Emphasise accuracy, clarity, and conciseness in each summary.

In this prompt, [type of data user] was replaced with either *producer* or *consumer* and [type of data use] was replaced with either *produced* or *used* depending on the transcripts being summarised. This provided the chatbot with context on the transcripts. The prompt also instructed the chatbot to focus on accuracy and clarity, aligning with the qualities evaluated in the participant survey.

To ensure that the summaries were concise, the chatbot was instructed to use bullet points, enhancing readability, and reducing verbosity. This approach aimed to create short, digestible summaries compared to detailed interview transcripts.

The chatbot was also instructed to divide each summary into sections based on the content of the transcript to improve readability. This approach also allowed for evaluating the chatbot's ability to generate useful section headers. By comparing the headers used in different summaries and identifying which ones participants found most useful, the goal was to simulate an ideal data conversation implementation where information from one conversation could easily be connected to summaries of other conversations as well.

To protect participant privacy and summary consistency, we generated each summary in its own temporary chat session [43]. This ensured that ChatGPT had no access to previous conversations or stored memory and that the content was not used for model training.

Summarisation limitations

The use of AI for summarization introduced several challenges originating from the evolving nature of the technology and its inherent limitations. As ChatGPT models are updated over time, prompt engineering outcomes can vary between versions, leading to inconsistencies. There is also a risk of unwanted biases in the summaries, which could affect their neutrality and accuracy. Furthermore, summarising long transcripts increases the likelihood of oversimplification. Among these challenges, the most significant downside of using AI was the risk of hallucination, where the model fabricates content that is not present in the original transcripts.

One factor contributing to these limitations was the simplicity of prompt engineering. To streamline the process and reduce survey completion times, minimal prompt adjustments were made and only one summary version per participant was generated. Although this approach ensured efficiency, it likely constrained the model's ability to produce higher-quality outputs. In addition, no comparisons were made between different chatbot models to explore whether alternative tools could yield improvements in summary accuracy, consistency, or neutrality.

This simplicity in prompt design also influenced the risk of hallucination during the summarization process. For instance, early attempts to structure summaries using predefined section headers based on interview questions led to fabricated content, as ChatGPT created information to fit the given structure. Allowing the model to determine its own sections improved reliability, but the risk of hallucination persisted throughout the process.

Hallucinations also became evident during batch processing, where an attempt was made to generate all summaries in a single session to maintain consistency and enable pattern recognition. However, the model could only process up to 14 summaries at a time before stopping, and any attempts to continue resulted in fabricated summaries. Furthermore, when certain participant IDs were missing, for example, if participants did not contribute as both producers and consumers, ChatGPT generated hallucinated summaries for these non-existent IDs. To mitigate these issues, the transcripts were divided into four smaller batches, with separate processing for producers and consumers. However, this batching strategy introduced its own limitations, as the lack of context memory across chats reduced consistency between summaries.

Ensuring data privacy while maintaining consistency further complicated the summarisation process. Although running the model locally or using privacy-preserving APIs could have enhanced confidentiality, these options were not feasible due to resource constraints. Another potential method, using

a seeded chat interface to improve output consistency, was explored, but ultimately deemed unsuitable, as it reduced the model's adaptability and creativity, qualities critical to this research. As a result, the use of ChatGPT's built-in temporary chat feature was selected as a practical compromise, balancing privacy, simplicity, and accessibility, though this came at the expense of reduced consistency across sessions.

In summary, while the chosen approach effectively balanced privacy, practicality, and quality, challenges such as hallucination, batch processing inconsistencies, and privacy trade-offs highlight areas for improvement. Future research could refine prompt strategies, explore more robust batch processing methods, or utilise advanced privacy-preserving technologies to address these limitations.

3.3.2. Summary evaluation design

The purpose of the survey was two-fold: primarily to assess whether AI-generated summaries could streamline conversational data by reducing irrelevant details and secondarily to assess whether any insights from the data conversation resonated with the participants. Table 3.3 shows all the survey questions; these questions were repeated for both the producer and consumer data conversations summaries.

The first survey question asked whether the data conversation left a lasting impression on the participants and, if so, what those impressions were. This question explored whether the conversational approach influenced their thinking about metadata creation and reuse.

All other questions focused on evaluating the AI-generated summaries. Due to privacy concerns, each participant received only the summary of their own interview transcript. This approach was appropriate because participants were uniquely positioned to assess summaries, as they not only knew the content of the original transcript but also the intended meaning behind their responses. If a participant identified as both a producer and a consumer, they answered the set of questions twice, once for each role.

Participants evaluated the summaries both in isolation and in comparison to the full transcript; see the next two sections for details. The survey included a mix of questions on the Likert scale [44], open-ended questions, and ranking questions (where participants rank various options from most preferred to least) to collect quantitative and qualitative feedback. This combination of question types provided a balance between the ease of completion of the participants and the level of detail collected. The survey was administered using Microsoft Forms, allowing participants to complete it anonymously by providing their participant ID.

Summary usefulness in isolation

This section consists of three Likert scale questions, followed by two open-ended questions. The first question asks whether the summary accurately reflects the participants' actual experiences. The next two questions evaluate whether the summary would be helpful for future reference, either for the participants themselves or for others. The latter may yield different results, as participants have more context than someone else would have. Participants were also asked to suggest any improvements to the summary. Finally, they were asked to identify specific sections of the summary (since it was divided into sections) that they found particularly useful. This question aimed to assess whether ChatGPT can effectively create well-structured sections without explicit direction on what sections to include.

Summary usefulness compared to transcript

This section begins with two questions on the Likert scale. The first asks whether participants believe that the summary improves the findability and reusability of information compared to the full transcripts, as this is the main goal of the study. The second asks participants whether they generally prefer the summary or the full transcript, along with their reasoning, to determine if summarization is actually desired or even necessary. The next question asks participants to rank four alternatives according to their preferences for varying levels of AI integration and autonomy in decision-making about what information to present. They are also given the option to suggest their own alternative.

The final three questions address participants' comfort with sharing the information. This includes whether they are comfortable sharing the transcript or the summary, considering that the transcript may feel more personal or raw due to its detailed nature. Finally, participants are asked if they would like to remove anything before sharing, to ensure privacy, and address any concerns about sensitive content.

Table 3.3: Complete list of survey questions repeated for both producer and consumer data conversation summaries. Questions are organized into the same categories defined in the research design. An additional column specifies the question type, including Open (open text box), Likert scale (five levels from Strongly Disagree to Strongly Agree), Binary (e.g., Transcript vs. Summary, Yes or No), or Ranking. Ranking questions required participants to order four options: Semi-structured interview with AI-generated summary, Self-completed structured form, Self-made summary, and Chatbot-driven Q&A with AI-generated summary.

Section	Type	Question
Impression	Open	Did any part of the interview leave a lasting impression or change the way you think about or plan to work with metadata in the future?
Summary usefulness	Likert	The summary accurately reflects my experience.
	Likert	The summary will be helpful for me to find and/or reuse this dataset, or a similar one, in the future.
	Likert	The summary will be helpful for someone else to find and/or reuse this dataset, or a similar one, in the future.
	Open	What would you change or add to the summary to increase its accuracy and/or usefulness for you or future users?
	Open	Which of the summary sections are most useful for findability and reusability? Please explain why.
Summary vs transcript	Likert	The summary helps improve the findability and reusability of key information compared to the full transcript.
	Binary	When imagining yourself as a future user of this information, which would you prefer to access when reusing this or a similar dataset?
	Open	Why?
	Rank	Conversations can be used as a method to capture information about the creation and use of datasets, especially focusing on metadata that enhances findability and reusability. What method would you prefer?
	Open	Or is there some other method you would prefer? (not required)
	Likert	I am comfortable with sharing this summary publicly (or within the same organization as the dataset was published).
	Binary	Is there any information you would want to remove before sharing?
Open	If so, what? (not required)	

Analysis of survey data

Likert scale, binary, and ranking questions were examined to reveal overall patterns and notable differences between the groups. Open-ended responses were thematically analysed to uncover recurring suggestions, challenges, and insights, with direct quotes sparingly included to ensure privacy. This integrated quantitative and qualitative approach provided a well-rounded understanding of the perspectives of the participants on AI-generated summaries and metadata practices.

3.3.3. Summary evaluation limitations

The evaluation of AI-generated summaries faced several limitations due to ethical, methodological, and practical constraints.

The ethical guidelines set by the HREC committee at TU Delft restricted the sharing of summaries between participants, limiting the opportunity for cross-comparison. Consequently, each summary was evaluated by a single participant. Although this ensured participant privacy, it reduced the breadth of evaluative perspectives. However, since there was minimal overlap between the research areas of the participants, cross-comparison may not have been particularly beneficial even if it had been possible.

The lack of technical detail in the transcripts also influenced the evaluation. Simpler content likely made it easier for ChatGPT to generate summaries, but prevented a meaningful comparison between AI-generated summaries and traditional metadata forms. Such a comparison could have tested whether participants provided less detailed responses in written metadata forms versus conversational transcripts.

Despite these constraints, the evaluation process provided a structured approach to assess the feasibility of using AI-generated summaries in a controlled context.

In summary, this methodology chapter describes a structured approach to investigating the metadata gap and strategies to bridge it through a combination of a literature review, interviews and surveys. The literature review established a theoretical foundation for understanding metadata practices and challenges. The interviews provided empirical insights into metadata creation and reuse while simulating context-bridging data conversations. Surveys evaluated the effectiveness of AI-generated summaries of these conversations, highlighting their potential to simplify metadata communication and improve usability. Together, these methods provide the methodological basis for exploring and refining metadata practices and advance data reusability.

4

Results

This chapter presents the findings of the interviews and surveys conducted for this study. Section 4.1 provides an overview of the participants, highlighting their domain expertise and experience with data management to offer context for their perspectives. Section 4.2 examines the interview results, focusing on key themes in the experiences of producers and consumers. It explores how these insights map the metadata gap and evaluate the data conversation mechanisms of involving consumers as co-creators and including contextual metadata. This section also highlights observations on the use of real-time dialogue and adaptive questioning. Finally, Section 4.3 reviews the survey findings, which evaluate the effectiveness of AI-generated summaries in improving the usability of data conversations.

4.1. Participant overview

This section provides an overview of the study's participants, including their roles, experience levels, and the challenges and themes that emerged during the study. The diversity of the group of participants highlights important trends in data management practices and underscores key limitations that affect the findings.

4.1.1. Participant characteristics and insights

A total of 18 individuals were interviewed for the study, of whom 15 identified as data producers and 16 as data consumers; many participants fulfilled both roles. One interview was excluded from the analysis, as P4 was found to lack relevant data management experience, due to a miscommunication during the interview planning process. The summarisation evaluation survey was completed by 12 of the participants; some did not respond to the survey on time.

Table 4.1 summarises the academic degrees, research areas, and experience of the participants in their primary research domains and data management. It also indicates their roles in the study, specifying whether they participated as producers, consumers, or both—and whether they took part in the survey.

The years of experience presented in Table 4.1 encompass both formal education and hands-on practice afterward. To simplify the interpretation of the results, participants were divided into two groups: junior and senior researchers (see Tables 4.2 and 4.3 for details on producers and consumers, respectively). This categorisation was based on a combination of the years of experience of the participants and the type of project discussed during the data conversation, providing a basic indication of their expertise and the resources available to them at that time. These project types included bachelor or master courses, bachelor or master theses, PhD research, industry research studies, and academic research studies. Participants who were still studying, in the early stages of a PhD or at the beginning of their careers, are categorised as junior, whereas those further along in their careers are categorised as senior.

This categorisation was necessary due to the challenges in defining and measuring experience during the interviews. Formal education and hands-on experience differ significantly; for instance, a year of working with data is not equivalent to a year of formal data management classes. In particular, no participant, except P4 (who studied computer science), reported more than a year of formal training

Table 4.1: This table provides details on the participants, including their research area, academic degree, current work status, and experience in their primary research domain (Dom.) and data management (Data mgmt.). The current status distinguishes participants as “Deg. IP” (degree ongoing), “Deg. done” (degree completed but not yet employed), “Work ind.” (working in industry), or “Work uni.” (working in academia). Both types of experience encompass are measured in years and include time during education as well as hands-on practice afterwards. Notably, none of the participants had more than a year of formal data management education, except for P4, who studied computer science. The table also indicates participants involved in interviews as producers (P) and/or consumers (C), with P4 excluded (NA) due to an unusable interview. Additionally, it identifies participants who completed the surveys, noting exclusions for late submissions.

ID	Research area	Degree	Current status	Data			
				Dom.	mgmt.	Int.	Surv.
P1	Technical medicine	PhD	Deg. IP	3	1.5	P	✓
P2	Statistics and data science	Master	Deg. IP	1	1	C	✓
P3	Bioinformatics	Master	Deg. IP	2	1	P&C	✓
P4	Computer science	Master	Deg. done	7	5	NA	
P5	Sanitation and technology	Master	Deg. done	2	0.5	P&C	
P6	Aerospace engineering	Master	Work ind.	7	5	P&C	✓
P7	Plant genetics	Master	Work ind.	5	5	P&C	
P8	Business analytics	Master	Work ind.	1.5	1.75	P&C	✓
P9	Catchment urban hydrology	Master	Work ind.	1	4	P&C	
P10	Biotechnology	PhD	Deg. IP	4	5	P&C	✓
P11	Bioinformatics	Master	Work ind.	7	7	C	✓
P12	Plant science	PhD	Work ind.	11	6	P&C	✓
P13	Plant science	PhD	Deg. IP	0.25	5	P&C	✓
P14	Civil engineering hydrology	Master	Deg. IP	0	0.5	P&C	✓
P15	Construction engineering	Master	Deg. done	5	0	P&C	✓
P16	Supply chains	Master	Work ind.	2	6	P&C	
P17	Paediatrics	PhD	Work uni.	5	3	P&C	✓
P18	Microbiology	PhD	Work uni.	25+	15	P&C	

in data management. However, P4’s results were excluded from the analysis because they lacked relevant data creation and reuse experiences. By distinguishing between junior and senior researchers, a more practical and interpretable framework was established to analyse the results in the context of the projects discussed during the study. For future work, it may be valuable to examine differences in experience levels more closely.

The years of experience presented in Table 4.1 encompass both formal education and hands-on practice afterward. To simplify the interpretation of the results, participants were divided into two groups: junior and senior researchers (see Tables 4.2 and 4.3 on producers and consumers, respectively). This categorisation was based on a combination of participants’ years of practice and the type of project discussed during the data conversation to give a basic indication of their expertise and the resources available to them at that time. These project types included bachelor or master courses, bachelor or master theses, PhD research, industry research studies, and academic research studies. Junior researchers typically included those in student or PhD roles or early in their careers, while senior researchers had more extensive experience gained over a longer period and in more advanced roles.

This categorisation was especially necessary due to the vagueness of the definition of experience during the interviews. Formal education differs significantly from hands-on experience; for example, a year of working with data is not equivalent to a year of formal data management classes. In fact, no participant reported more than a year of formal training in data management, except P4, who studied computer science but whose interview was excluded. By distinguishing between junior and senior researchers, a more practical and interpretable framework was created to analyse the results in the context of the projects discussed during the study.

A notable theme among the participants, regardless of the level of experience, was the importance of learning-by-doing in data reuse and management. This supports the hypothesis that many aspects of data creation and reuse can only be fully understood through hands-on experience. Consequently, valuable knowledge often remains undocumented, existing only in practice rather than in formal guide-

lines. This highlights a disconnect between the ostensive routines, which formal data management rules prescribe, and the performative routines, which researchers actually do (see Section 2.2 for more detail). Beginners, who have not yet accumulated substantial hands-on experience, were particularly impacted by this gap. For instance, P1 expressed uncertainty about the value of their data management practices, reflecting the steep learning curve faced by novices. In contrast, P12, an experienced researcher, emphasised that their data management skills were developed primarily through practical experience. This discrepancy underscores the challenges that beginners face and the importance of experiential learning to master effective data management.

4.1.2. Participant pool limitations

The composition of the participant pool introduced several limitations that influenced the scope and applicability of the study findings.

One clear constraint was the limited number of participants, which was further reduced for the survey phase due to non-responses. This smaller sample restricted the breadth of perspectives and feedback.

Another significant limitation was the absence of ideal producer-consumer pairs, where one participant provided data for another to use. Efforts to recruit such pairs, particularly within the CropXR context, were unsuccessful due to limited availability and logistical challenges. This omission hindered the potential for in-depth technical discussions about metadata exchange and reduced the comparability of findings. Including producer-consumer pairs would have offered more actionable insights into real-world data sharing practices and strengthened the relevance of the study to CropXR's objectives. However, the diversity of participants added value by revealing trends that appeared to transcend domain boundaries, making the findings more broadly applicable.

Interestingly, the varying levels of experience of the participants emerged as an important finding. Beginners often relied on advice from advisors, exposing knowledge gaps that limited their ability to address metadata challenges, while experienced researchers exhibited a deeper understanding gained through practical experience. This disparity highlighted the significant role of experience in effective data management, even as it introduced variability in the quality and depth of insights gathered.

In summary, while the participant pool had limitations in size and structure, the diversity and variation in experience levels provided valuable insights into metadata challenges across different domains and career stages.

4.2. Data producer and consumer interview results

This section presents the results of the interviews and offers information on how metadata is created, evaluated, and reused. On average, the interviews lasted 15 minutes per type of data stakeholder.

The analysis explores the supply side of the metadata gap by examining how producers approach metadata creation, including their motivational needs, the constraints they face, and the assumptions they make about future users. On the demand side, the analysis investigates how consumers locate and evaluate datasets, revealing the difficulties they encounter due to incomplete metadata or knowledge gaps. This perspective highlights the challenges of reusing data effectively and the strategies employed to overcome these barriers. For both producer and consumer interviews, a bullet list summarising notable results is presented at the beginning of each section. They should be interpreted with caution due to the limited number of participants. Furthermore, the adaptable nature of the interviews meant that some participants provided incomplete or inconclusive answers to certain questions. However, the lists highlight several key trends in the interview results.

Additionally, the role of real-time dialogue and adaptive questioning during data conversations is examined, demonstrating how these mechanisms enriched participant engagement and provided deeper insights. Although individual transcript excerpts are excluded to protect privacy, illustrative examples are included where relevant to highlight broader themes. The section concludes with a discussion of the result limitations.

4.2.1. Producers' metadata considerations and experiences

This section explores the results from the producer side of the interviews. It highlights their processes, motivations, and the barriers they face in creating high-quality metadata that focusses on motivational needs, time and space constraints, and assumptions about future consumers. In general, producers

indicated that they typically included basic metadata, such as production methodology and information needed for immediate project goals. However, more detailed or contextual metadata was often omitted due to various barriers. Table 4.2 summarises, for each participant, the type of dataset discussed during the data conversation, whether it was published publicly or internally, and whether the producer had a specific future user and/or use case in mind. This information is important because it provides insight into the constraints and goals associated with each project.

The following bullet points summarise notable results from the producer interviews.

- When the next consumer is known, producers tend to prioritise their specific needs, often disregarding broader reusability for other potential future users.
- Producers frequently assumed that future consumers would have the same or greater knowledge, leading to the exclusion of potentially useful metadata details.
- Contextual metadata was rarely, if ever, included in publications. When the next user was known, producers planned to explain details in person or kept personal notes for their own reuse, showing that this information was valuable but often omitted due to time and resource constraints.
- Producers rarely engaged in discussions about metadata quality during the project lifecycle and almost never received feedback on the dataset's reusability after publication.

Table 4.2: Summary of the dataset details discussed during producer interviews. Participants are categorised by experience level (XP): junior (those still in formal education or early in their careers) and senior (those with more extensive experience and advanced roles). The datasets they focused on during the data conversation are classified into the following types of work: bachelor or master courses, bachelor or master theses, PhD research, industry research studies, and academic research studies. The table also includes details on the publication method of the dataset (public or internal). The final column outlines the producer's considerations regarding the future user during the creation of the (meta)data. "Current colleague" refers to colleagues currently known to the producer. "Future colleague" indicates a user within the same organisation, though their identity is not yet clear. "Within own domain" signifies someone in the same research field. "Outside own domain" means that the producer also considered users beyond their direct field. These factors highlight differences in the experience of participants, the resources available to them, and their goals during the metadata creation process.

ID	XP	Project type	Dataset accessibility	User in mind
P1	Jr	PhD	Public	Current colleague
P3	Jr	Master thesis	Public	Within own domain
P5	Jr	Master thesis	Public	Outside own domain
P6	Jr	Master thesis	Internal	Future colleague
P7	Sr	Industry research	Public	Outside own domain
P8	Jr	Industry research	Internal	Current colleague
P9	Jr	Industry research	Internal	Future colleague
P10	Jr	PhD	Public	No
P12	Sr	Academic research	Public	No
P13	Jr	PhD	Internal	Current colleague
P14	Jr	Bachelor thesis	Public	Current colleague
P15	Jr	Master thesis	Public	Within own domain
P16	Jr	Industry research	Internal	Within own domain
P17	Sr	Academic research	Public	Future colleague
P18	Sr	Academic research	Public	No

Initial findability and usability considerations

Producers prioritised ensuring that data supported their own findings, which is understandable given their incentives. As a result, they generally put little effort into making datasets findable and reusable, often including only keywords or occasionally other metadata not essential to their own objectives.

Some producers relied on repository forms (P10) or added keywords to improve discoverability (P1, P6, P8, P10, and P16). Others included specific metadata, such as dates and models (P6), additional data explanations (P7) or pre-processed data (P9), to improve usability. Producers frequently defaulted to existing conventions or practices within their field, even when these were poorly defined. For instance, P1 and P17 noted that they followed what others had done before, despite limited formal

guidance on metadata standards. Interestingly, P18 indicated that they do want people to reuse their datasets but put minimal effort in adding metadata for findability and reusability “People can contact us, of course”.

Motivational needs of producers

The interviews highlighted several factors that motivate producers to create higher quality metadata.

The connection to the future user was identified as a key motivator. When participants knew for what purpose the data would be used and had some knowledge of the future user, often someone within their organisation—they tended to spend more time improving the reusability of the dataset (P7, P10, P13, P14, P18). In the case of P13 and P16, who were the future users of their own data, they kept private contextual metadata notes for their own future reference knowing that it would be necessary to efficiently reuse their own dataset. P14 suggested that knowing the future user or use case of their data could encourage better metadata practices, while P7 emphasised that this awareness is particularly important to maintain long-term motivation. In contrast, P9 indicated that a short link to the future user could decrease motivation. This was also evident in P17’s response, as they noted that while the future user was known, they planned to explain most details in person because documenting everything was considered too time consuming. In such cases, missing information could simply be explained afterward.

Collaboration, feedback, and recognition were expected to play an important role in motivating producers but were often lacking during or after metadata creation. For instance, P8 found direct communication with consumers motivating, but such interactions were rare. This lack of feedback led some participants, like P12, to adopt the mindset that “no feedback is good feedback”, leaving room for improvement in the promotion of collaborative environments. Producers like P10 and P14 expressed interest in receiving constructive feedback to improve their work and better support future users. Discussions about metadata within research groups were also minimal. P6, P10, P14, P15 and P16 indicated that such conversations were completely absent in their experience. In contrast, P13 appreciated the guidance of their supervisor, which motivated them to include more metadata; however, this level of support was rare. After project completion, most producers, such as P12 and P14, reported receiving little or no feedback, either positive or negative, on their metadata. As a result, there were few opportunities to improve or reinforce the quality of the metadata.

Research credibility was another significant driver. Transparency (P10) and the ability to interpret results (P13) through detailed metadata were important for many participants. For example, P13 had learnt during their bachelor that any detail could turn out to be important. Similarly, P15 described a sense of responsibility to contribute to the research community as a motivator for detailed metadata.

Personal negative reuse experiences further shaped motivation. P1 shared that their metadata practices were influenced by the desire to ensure that others did not face the same frustrations they had experienced. Similarly, P6 and P12 noted that previous challenges with poorly documented datasets motivated them to improve their own metadata for potential future users. P2 summarised it well: “You would love it if other people did it, but [you] don’t really want to do it yourself.” It is often easier to publish something quickly with the intention of fixing it later. This issue was also observed on the consumer side, where participants encountered datasets that were meant to be improved eventually but remained incomplete.

Finally, P16 shared an interesting perspective: they were hesitant to make their dataset too accessible because they were concerned that someone else might use it to pursue the research they intended to continue.

Time and space constraints

Producers consistently reported a lack of time and space to create detailed metadata as a significant barrier. Metadata creation was often considered only at the end of the project, when time and resources were most limited (e.g., P6, P8, P15, and P17). For example, P6 noted that they included some contextual metadata but could not devote enough time due to competing priorities.

The absence of formalised processes or enforcement mechanisms further deprioritised metadata creation. P14 and P16 noted that they had no prior experience explicitly accounting for contextual metadata. P6, P9 and P12 highlighted that the inclusion of metadata, especially contextual metadata, was often left to the individual’s discretion, leading to inconsistent practices. P13 reflected that, while they valued completeness, the lack of structure made it difficult to prioritise metadata tasks during busy

project periods. These results emphasise the need for structured support and formal standards to enable efficient and consistent metadata creation.

When data creation is a by-product of the research, the contextual information often does not fit neatly into the metadata. For example, in P1's project, they chose to publish an additional article dedicated to detailing the data production process, highlighting the substantial amount of extra information involved in data creation.

Limited by consumer assumptions

Producers' assumptions about future user knowledge and use cases shaped their metadata practices.

Participants had to make assumptions about the level of experience of the consumer. Most producers assumed that future users would have similar or greater domain knowledge (P9, P10, P13, P14, P16, and P18). In part, this is necessary as one can simply not include every detail every time one writes a paper. For someone like P5, who had policy professionals in mind as future users from a different domain, they spent extra time carefully considering what to include and how to explain things. However, P3 highlighted how familiarity with a dataset made it difficult to anticipate what others may not know, underscoring the challenge of addressing diverse user needs. P13 addressed this by trying to explain everything in the most basic terms. To balance this, P12, an experienced researcher, published two versions of their dataset: one raw and one processed. This approach provided advanced users with the flexibility to work with the data as they preferred while offering beginners a more accessible version.

Use-case considerations were also often limited. Producers rarely considered potential use cases beyond the immediate project goals unless explicitly prompted. For example, P13 and P14 focused primarily on their own future use cases or those of their immediate collaborators. In contrast, P5, P6, P7, P15 and P18, which had no specific future user in mind, included more metadata to accommodate a wider range of potential applications. Only P16 had their own use case in mind, but actively considered alternative options as well. They acknowledged that it is impossible to determine in advance exactly what metadata might be needed.

Some producers made additional efforts to improve usability when the dataset use case became clearer during the project. For example, P2 adjusted their practices to include more metadata when they realised that their dataset could be used in various contexts. However, producers like P10 admitted that they had not considered reusability during the project, but recognised its importance in hindsight.

Finally, discussion of metadata creation during the process was limited. Only one participant, P8, had frequent communication with the envisioned consumers, they indicated it greatly improved metadata quality for those specific users but did not encourage consideration of broader use cases. P7 and P12 did mention consulting colleagues, but for P12, these discussions were not particularly useful due to the specificity of their research. came to appreciate in hindsight.

The results from the producer interviews reveal a range of practices, motivations, and barriers to metadata creation. Although producers include basic metadata as standard practice, detailed and contextual metadata often takes a backseat due to time constraints, lack of feedback, and consumer assumptions. Motivational factors such as transparency, personal reuse experiences, and connections to future users highlight opportunities to encourage better metadata practices. However, addressing structural barriers and fostering awareness of diverse consumer needs will be essential to improve the overall quality of metadata in research.

4.2.2. Consumers' challenges and solution approaches

This section explores the results from the consumer side of the interviews. It looks at how consumers find and evaluate data, the challenges they face, often compounded by gaps in skill or knowledge, and the strategies they use to address them. Similarly to the producer side, Table 4.3 summarises each data conversation, detailing the project type, dataset publication method, and alignment of the consumer use case with the intended use case of the data producer. This information provides insight into the constraints and goals of each project.

The following bullet points summarise notable results from the consumer interviews

- Reuse challenges often arise from insufficient contextual metadata, and participants expressed a strong desire for additional details, such as methodologies used by other data consumers or

clarification of data conventions. This highlights a significant gap between the metadata provided and what is needed for efficient reuse.

- Direct contact with the data producer was frequently attempted, reflecting the value placed on personal communication to clarify missing or unclear information about the datasets. However, success varied depending on pre-existing relationships, time availability, or producer willingness.
- Research methodologies were frequently changed due to unforeseen problems with the dataset, demonstrating the significant impact that incomplete or inadequate metadata has on efficient reusability.
- Consumers often lacked the domain knowledge or data management expertise necessary to interpret poorly documented datasets quickly, exacerbating reuse challenges and requiring additional time and resources.
- Solutions to data reuse challenges were often not documented as they would not fit well into article methodologies, limiting opportunities for shared learning and potentially perpetuating similar issues for future users.

Table 4.3: Summary of the dataset details discussed during consumer interviews. Participants are categorised by experience level (XP): junior (those still in formal education or early in their careers) and senior (those with more extensive experience and advanced roles). The datasets they focused on during the data conversation are classified into the following types of work: bachelor or master courses, bachelor or master theses, PhD research, industry studies, and academic studies. The table also includes details on the Publication method of the dataset (public or internal) and whether the consumer considered their use case in line with what the original dataset creator would have had in mind. These factors highlight differences in the experience of participants, the resources available to them, and their goals during the (meta)data reuse process.

ID	XP	Project type	Dataset accessibility	Use case in line
P2	Jr	Master course	Public	In line
P3	Jr	Master thesis	Public	In line
P5	Jr	Master thesis	Public	In line
P6	Jr	Company research	Internal	In line
P7	Sr	Company research	Public	In line
P8	Jr	Company research	Internal	In line
P9	Jr	Company research	Public	In line
P10	Jr	PhD	Public	In line
P11	Sr	Company research	Public	In line
P12	Sr	Academic research	Public	In line
P13	Jr	Academic research	Public	Out of the box
P14	Jr	Master course	Public	Out of the box
P15	Jr	Master thesis	Public	In line
P16	Jr	Company research	Public	In line
P17	Sr	PhD	Public	In line
P18	Sr	Academic research	Public	Out of the box

Initial usability considerations

Consumers often began with a specific use case in mind, except in coursework settings, where exploration was more common because the focus was more on learning about a specific method (P2) than solving a real problem. Datasets were typically found through links given by a supervisor or through academic research search engines and repositories. P2 received links to various databases, P8 and P16 received datasets directly from their company, and P3, P5, and P9 relied on search engines. Other participants, such as P11, P12 and P15, located datasets through repository searches, while P15 used a repository after initial data access issues.

Consumers used both the data from the dataset itself and the corresponding metadata, published with the dataset or in the matching article, to evaluate whether a dataset was usable for their use case. Some participants relied on hard metadata published in data repositories (P5, P8), while others (P2, P3, P11, P15, and P17) used contextual metadata from articles that also used the dataset. Participants

also directly assessed datasets, focussing on aspects such as completeness, clarity, and specific requirements. For example, P12 prioritised datasets with replicates and well-documented processes, while P14 switched to a second dataset after discovering that the first lacked critical information. The most experienced researcher, P18, contacted the data producers directly, considering this the most efficient method based on their previous experience.

Constraints made worse by knowledge limitations

Consumers faced various challenges during reuse, often intensified by knowledge gaps. Lack of domain knowledge was a recurring issue. For example, P2 struggled with unclear terminology and variable construction, while P14 only realised that a dataset was unsuitable late in the process due to their own lack of domain knowledge.

Data management inexperience also presented significant challenges. P11 described difficulties in navigating a poorly formatted 100-page table. In contrast, P12 highlighted how exceptionally well-documented metadata made a dataset easy to use, even for an unconventional use case, underscoring the value of detailed metadata. P8 noted that a single conversation with an expert greatly improved their ability to reuse a dataset, although finding such an expert was time-intensive.

Some problems were attributed to the insufficient effort of producers. For example, P3 found a poorly maintained website with outdated images, while P9 and P16 spent extra time interpreting unexplained data conventions.

Solutions stay undocumented

Consumers used various strategies to resolve issues, but few documented their solutions for future use. Internet searches were a common method for resolving problems, explicitly mentioned by P2 and P6 and probably used by most participants. Direct contact with experts was effective when successful (P8 and P12). Internal discussions also offered support in some cases (P11, P12, and P13), with P12 benefiting from colleagues experienced in data management.

Attempts to contact producers for clarification had mixed success. P7, P9, P16, P17, and P18 found communication with producers helpful, while P3 and P11 were unable to reach anyone for support. However, even successful attempts had notable caveats. P16 found it easier to make contact because they already knew the producers. P17 needed raw data that could have been included initially, but was omitted because it was not directly beneficial to the original producer, forcing P17 to initiate a time-consuming personal exchange. For P18, they had to pay a fee for the assistance of the producer. Despite this, P18 emphasised that in their extensive experience, direct contact with producers is almost always the most effective approach.

Consumers frequently pivoted their research methodologies or goals when datasets failed to meet expectations (P3, P7, P9, P10, P11, P12, P13, and P14). P3 and P14 switched to entirely new datasets, while P12 adjusted their methodology to simplify the project. However, these adjustments were rarely documented in the consumers' published methodologies.

Most of the participants did not include their struggles or solutions in their own outputs, limiting opportunities for shared learning. P2, P5, P13, P14, and P16 did not document their problems, citing lack of time or considering the problems irrelevant to their research goals and thus unsuitable for inclusion in their articles. This trend was particularly evident in course or thesis projects, where participants viewed their research as less valuable or unlikely to be reused, further discouraging the effort to record challenges and solutions. For instance, P2 received a suggestion from their professor to fix an issue but chose not to implement it, as the course was already completed. Even participants who included detailed workflows, such as P10, rarely shared information about the challenges they faced along the way. P16 suggested that having a dataset comment section to discuss its qualities would have been helpful. Only P8 and P9 published at least parts of their struggles, P8 documented how they combined datasets but not failed paths, and P9 noted spending time creating workflows but also emphasised the impossibility of recording everything.

Consumers face significant challenges in reusing datasets, often due to knowledge gaps, incomplete metadata, or inadequate producer effort. Although consumers employ various strategies to address these barriers, such as search on the Internet, expert consultations, or methodological pivots, these solutions are rarely documented, leading to repeated inefficiencies and missed opportunities for shared

learning. Addressing these issues requires improving metadata quality, fostering communication between producers and consumers, and encouraging the documentation of reuse processes to improve dataset usability and support the broader research community.

4.2.3. Real-time dialogue and question adaptation in data conversations

This section explores how the mechanisms of real-time dialogue and the adaptation of the questions were used during the interviews. These mechanisms mainly involved using the detailed real-time information provided during conversations and refining or removing questions between interviews to improve clarity and effectiveness.

One key advantage of real-time dialogue was the ability to clarify unclear questions or address participant hesitations during the interview. Non-verbal cues, such as hesitancy or confusion, prompted immediate clarification. For example, P1 initially struggled to interpret the question of how many datasets they produced until it was rephrased. Similarly, P8 found the motivation-related question in the end unclear, but after clarification they gave a meaningful response. Real-time dialogue also facilitated spontaneous discussions, providing richer insights. For example, during the conversation with P3, the interviewer empathised with the frustrations they encountered, which may have fostered a sense of connection and encouraged more open sharing.

Real-time dialogue mechanisms were used to prompt participants to provide more detailed and thoughtful responses during the interviews. For example, during P10's interview, additional follow-up questions encouraged them to expand on the criteria they used for evaluating metadata quality. Similarly, clarifications on the meaning of "help" in metadata production (P11) and suggestions for judging metadata practices (P14) helped participants better understand the intent of the questions. These real-time adjustments allowed participants to engage more fully with the interview process, improving the quality of the data collected without altering the overall structure of the interview for future participants.

Beyond real-time adjustments, the iterative nature of the interviews enabled a continuous refinement of the phrasing and structure of the questions for future sessions. Questions that consistently caused confusion were rephrased or clarified for subsequent interviews. For example, after P6 struggled to understand why the interview focused on a single dataset, an explanation was added to future interviews to emphasise its role in simulating a larger data management system. Similarly, examples were incorporated to provide additional context for certain questions, such as describing findability through a hypothetical scenario of a user trying to locate a dataset (P2).

These refinements were particularly important for the broad and abstract consumer-focused question about their entire data reuse process. Initially, prompts and examples were prepared in advance, but withheld until participants had attempted to answer on their own. However, it quickly became clear that this approach was ineffective. For example, P3 and P9 struggled to grasp the intent of these questions without concrete examples. To address this, examples were consistently included in all future interviews, ensuring that participants could engage more effectively.

Early interviews revealed that some questions unintentionally made participants feel criticised, particularly junior data producers with less experience in metadata management. For example, P1 showed visible discomfort when asked if they had considered specific metadata elements, seemingly interpreting the questions as criticism of their work. This reaction, inferred from hesitant responses and defensive explanations, indicated a flaw in the way the questions were formulated. The purpose of the interview was not to suggest that participants should have done things differently, but rather to understand their current practices and assess whether changes were actually needed. When the issue came up again during the interview with P3, a preamble was added to clarify that not considering certain metadata practices did not reflect poorly on their work. This adjustment not only made participants feel more at ease, but also improved the quality of their responses, as they better understood the goal of the questions.

In some cases, questions were removed entirely when they did not generate meaningful insights. For example, asking consumers whether they had considered alternative use cases for a dataset yielded no significant responses. Similarly, participants often gave polite but uninformative responses when asked about their thoughts on the interview itself, making this question ineffective and ultimately unnecessary.

The mechanisms of real-time dialogue and question adaptation ensured that unclear questions were clarified, participant discomfort was mitigated, and subsequent interviews benefited from lessons learnt

in previous sessions. These adaptations not only improved the quality of the collected data, but also fostered a more comfortable and engaging environment for participants, ultimately enhancing the depth and reliability of the study's findings.

4.2.4. Discussion of interview results

Several factors limited the quality, consistency, and comparability of the interview results, although they provided valuable information on metadata practices from different points of view.

The small sample size of the study was a primary limitation, restricting the ability to generalise the results. A larger and more diverse group of participants could have provided greater insight and increased the reliability of the conclusions.

The use of English as the interview language posed additional challenges, as it was a non-native language for most participants. This may have hindered the expression of nuanced responses. However, English was chosen intentionally to ensure accurate transcriptions and align with the language commonly used in publications and data communications. It also facilitated the use of tools like ChatGPT for generating summaries, which perform best in English.

The semi-structured interview format and conversational tone allowed flexibility in tailoring questions based on participant feedback but introduced variability in the depth and relevance of responses. Although this approach was used purposefully, it also led to inconsistencies between sessions, affecting comparability.

Participants were guided to consider the role of future users in metadata creation, and many expressed that this was motivating. However, this may be an example of how the phrasing of questions could influence responses, with participants potentially overstating their motivations to avoid appearing indifferent to future users' needs. Although efforts were made to ask questions naturally and neutrally, fully refining each question to minimise bias would have required additional effort and planning.

The existing connections of the interviewer with the participants, whether directly, through mutual acquaintances or via CropXR, probably influenced their willingness to participate and their level of openness during interviews.

The data management experience of the participants varied significantly and was often acquired sporadically through occasional coursework or practical work experience. This inconsistency made it difficult to assess the depth of each participant's data management skills. Furthermore, the introductory questions only asked about the participants' experience in general, which may have created a misleading impression of their experience as data producers or consumers separately.

Another limitation was the lack of consumer-producer pairs in the study, which limited the ability to assess direct links between the supply and demand of metadata. Without these pairs, the connection between metadata creation and reuse remained indirect and harder to evaluate. Additionally, only five participants from CropXR participated. This affected the ability to focus on CropXR-specific practices or expectations but made the results more generally applicable.

In relation to this, the diversity of professional domains among participants broadened the applicability of results across fields, but reduced the specificity of insights into domain-specific metadata practices. A more focused sample from a single domain could have allowed a deeper exploration of specific metadata requirements for certain types of dataset.

The interviews revealed valuable information on the interplay between metadata supply and demand, but also highlighted significant mismatches between producers and consumers. Producers often aspire to create high-quality metadata, but face challenges such as excessive workloads and too many options, leading to incomplete documentation. Consumers, in turn, are left to resolve issues themselves, typically succeeding but failing to document their solutions, causing valuable knowledge to be lost. Addressing these gaps will require better alignment of producer-consumer communication, more extensive contextual metadata standards, and incentives to document and share challenges and solutions.

4.3. Survey-based evaluation results of the data conversation summary

This section presents the results of the survey evaluating the effectiveness and usability of AI-generated data conversation summaries. The survey explored whether the summaries left a lasting impression,

evaluated their standalone accuracy and usefulness, and compared their effectiveness to alternative formats.

The survey included 12 participants, consisting of 9 producers and 10 consumers, as some participants fulfilled both roles. Among these, three were senior researchers (P11, P12, and P17). Unless otherwise noted, the results for producers and consumers are combined, as no significant differences were observed in most responses. Figure 4.4 shows the results for the questions on the Likert scale. The survey results for closed questions are provided for each participant in Appendix C, and the full summaries of data conversations are available in Appendix D. Open-ended responses are excluded from the appendix to safeguard the privacy of the participants.

Table 4.4: Results for the Likert scale survey questions, aggregated from both producer and consumer responses. The first column lists the survey statements, while the percentages of participants who (strongly) agreed and (strongly) disagreed are displayed in the left and right, respectively. Although 12 participants completed the survey, the table reflects 19 opinions in total, as each participant contributed as both a producer and a consumer. For individual responses, refer to Appendix C.

Survey Likert question	Result
The summary accurately reflects my experience.	90%  5%
This summary will be helpful for myself to find and/or reuse this dataset, or a similar one, in the future.	47%  21%
This summary will be helpful for someone else to find and/or reuse this dataset, or a similar one, in the future.	47%  16%
The summary helps improve the findability and reusability of key information compared to the full transcript.	74%  10%
I am comfortable with sharing this summary publicly	79%  5%
I am comfortable with sharing this transcript publicly	64%  22%

■ Strongly agree
 ■ Agree
 ■ Neutral
 ■ Disagree
 ■ Strongly disagree

4.3.1. Participant lasting impression of interviews

This section explores the results of the first survey question on whether the interviews left a lasting impression on participants and/or influenced their views on (meta)data management. Participants often struggled initially to understand how the interview topic could provide valuable information. However, post-interview feedback suggested that participants came to recognise the value of discussing their thoughts and practices surrounding metadata management, which is why this question was explicitly included in the survey.

Six out of nine data producers indicated that the producer part of the interview gave them insights into improving the presentation of data or the quality of metadata. The other three did not report any change in their approach, which included all senior researchers in this group (P12 and P17). P12, for example, noted that they were already following best practices in metadata management.

Fewer consumers reported a lasting impression, with only 4 out of 10 indicating a change in perspective. Most of these consumers acknowledged that the interview made them realise the challenges of data reuse, especially without high-quality metadata. One consumer (P15) mentioned that they would adapt their approach based on the interview. Again, none of the senior researchers (P11, P12, P17) in the consumer group reported that the conversation left a lasting impression.

The lower impact on consumers may be attributed to the nature of their conversations, which focused more on individual experiences. In contrast, the producer interviews included more thought-provoking questions designed to encourage reflection on metadata practices. It is also evident that experienced producers did not derive new insights, which aligns with our finding that experience plays a significant role in shaping metadata practices.

4.3.2. Summary usefulness in isolation

This section evaluates participants' perceptions of the summary's accuracy and usefulness as a standalone resource. There were no significant differences between producers and consumers in their responses to this section.

The participants generally agreed that the summary accurately reflected their interview experience. However, its usefulness was rated lower, with no major differences between its usefulness for personal

versus external purposes. When asked why the summary was not as useful, participants frequently cited a lack of detail about the dataset itself. Several respondents mentioned that they needed more context for the summary to be actionable or meaningful (Producers: P1, P3, P6, P8, P14, Consumers: P3, P6, P11, P12, P14). For example, P8 suggested including direct quotes from the transcript, with links to the full transcript for easy verification and additional context.

When asked about the usefulness of specific summary sections, no clear trends emerged. Participants often focused on the content of the information rather than the section headers themselves. The perceived usefulness of the summary was highly dependent on the specific interview and its content. For producers, sections related to data processing and future applications were most frequently highlighted as useful. Consumers, on the other hand, found the sections on challenges to be the most relevant.

4.3.3. Summary usefulness compared to alternatives

This subsection explores the preferences of participants for the summary compared to the complete transcript and other methods. The participants generally agreed that the summary improved the findability and reusability of key information compared to the full transcript.

When asked whether they preferred the summary or the full transcript, 13 out of 19 participants (7 producers and 6 consumers) chose the summary; producers were more likely to strongly agree with this statement, while consumers typically just agreed. Producers preferred the summary for its conciseness and ease of understanding. One participant noted that this was particularly valuable to them as a dyslexic individual. Even among producers who preferred the summary, some (e.g., P10 and P14) acknowledged that the full transcript might still be needed to fully understand all details. Among consumers, the preference for the full transcript stemmed from its ability to provide the detailed context necessary to fully understand and interpret the summarised information. For example, P3 stated “It [the full transcript] gives more details that are useful for the consumer to know, such as where the data were found and which aspects were difficult, etc. The summary is too broad.” P12 also cited concerns about the accuracy of the summary. Overall, the results indicate that while the summary is generally sufficient, access to more context is necessary for some participants, especially for data consumers.

Table 4.5: Results for the preferred metadata elicitation ranking question of the survey: “Conversations can be used as a method to capture information about the creation and use of datasets, especially focusing on metadata that enhances findability and reusability. What method would you prefer for gathering this type of information (e.g., challenges or insights during dataset creation or use)?” The first column lists the ranked options, while the percentages in the second column indicate how many participants placed each option in their top or bottom two preferences. Although 12 participants completed the survey, the table reflects 19 opinions in total, as each participant contributed as both a producer and a consumer.

Metadata elicitation method	Result
Semi-structured interview with AI-generated summary	78%  22%
Self-completed structured form	37%  63%
Self-made summary	42%  58%
Chatbot driven Q&A with AI-generated summary	42%  58%

■ 1st place
 ■ 2nd place
 ■ 3rd place
 ■ 4th place

Figure 4.5 presents the results of the question in which the participants ranked four metadata elicitation methods. The results reveal a clear preference for semi-structured interviews with AI-generated summaries, which participants consistently ranked as their top choice. The chatbot-only interview method was consistently ranked last by both producers and consumers. For the other two options, producers preferred structured forms slightly over self-made summaries, while consumers showed a nearly equal preference for the two.

The differences in rankings might reflect the nature of the interviews. Producers whose interviews were more structured may find it easier to translate their discussions into a structured form. Consumers, who had more free-flowing interviews, might see more value in creating their own summaries. The summary also allows for greater autonomy, which may have influenced these preferences as well. When asked for alternative methods, only P6 provided a response. Their suggestion involved having an AI bot provide prompts or examples during the interview to stimulate relevant thoughts, after which they could use that to write their own summary.

As for comfort with sharing, the summary was preferred over the full transcript, although the difference was minimal. The summaries provided were already pseudo-anonymized, and participants did not raise any concerns about privacy. Although it was expected that they might express concerns about the rawness of the full transcript, such as its exposure of their unedited thought processes, this did not occur. Two participants (P3 and P13) mentioned that they wanted to remove filler words or sentences from the transcript if they were shared. For example, P3 suggested “Single question + single answer, remove all the filler words/sentences.”

4.3.4. Discussion of survey results

The survey results provide useful information on the perceptions of the participants about metadata practices and tools, but limitations in sample size, participant diversity, and clarity of method descriptions highlight areas for refinement in both the study and the tools evaluated.

Not all participants completed the survey, which limits the sample size even further compared to the interviews, restricting the ability to draw definitive conclusions. Although no significant differences were observed between producers and consumers in most of the responses, a larger sample could have helped balance subjective variability. With more participants, it would have been possible to gain deeper insight into the differences between producer and consumer summaries and to assess the impact of experience more robustly. Additionally, analysing responses based on participants’ academic versus corporate backgrounds could have provided more understanding of how professional environments influence perspectives on metadata practices.

It is unclear whether the participants fully read the summaries or accurately remembered the details of the transcripts, especially since it is unlikely that anyone reviewed their entire transcript for a direct comparison. This is particularly significant because each summary was evaluated by only one person, due to privacy considerations, and due to the diversity of participants, participants would definitely have needed additional context to judge a summary of conversation about a dataset from a domain they are not familiar with at all.

The ranking question yielded some interesting results, but in hindsight, the options provided were too varied and vague to allow for a well-balanced comparison. For example, a combination of a structured conversation with a self-created summary could have been included as an additional option, offering both the guidance of a structured discussion and the autonomy to craft their own summary, allowing participants to decide on the level of detail they deemed necessary. Furthermore, much of the detail for each option was left open to the interpretation of the participants. Providing clearer explanations of what each option entailed would have been more effective, since only the method presented in the thesis was clearly described. This probably made it the most obvious choice to rank first. Although participants likely had previous experience with metadata forms to provide some context, the chatbot option was particularly problematic under these conditions. For the chatbot option, participants had to imagine how the conversation might unfold, including aspects such as potential hallucinations, response speed, and the level of guidance provided. This is especially significant if participants had negative perceptions of chatbots as unreliable, prone to inaccuracies, or overly verbose.

Overall, the survey results demonstrate that AI-generated summaries are a valuable tool to improve the usability and accessibility of metadata. Participants generally preferred summaries over full transcripts due to their conciseness and ease of understanding. However, both producers and consumers emphasised the need for additional context or links to the full transcript to enhance usability. Although there are limitations in participation and interpretation, the results offer practical recommendations for designing metadata workflows. Striking a balance between summary conciseness and contextual depth appears to be essential for meeting diverse user needs.

5

Findings

This chapter presents the findings of this study, synthesising the results of the literature review, interviews, and surveys to address the research questions. First, the focus is on mapping the metadata gap, which involves understanding the discrepancies between what data producers supply as metadata and what data consumers need for effective reuse. Section 5.1 explores this gap in detail, starting with the supply side, where the motivations and barriers influencing producers' metadata practices are analysed. Next, the demand side is examined, identifying the barriers that hinder dataset reuse. This dual perspective emphasises that metadata reuse is not dependent solely on one party but on effective communication between producers and consumers, with metadata serving as the common language.

Section 5.2 evaluates the potential of the four context-bridging data conversation mechanisms, tested through interviews, to address aspects of the metadata gap. The mechanisms are: integrating contextual metadata, involving consumers as co-creators, leveraging real-time dialogue, and adapting interview questions. By breaking down these mechanisms into their component parts and linking them to specific metadata challenges, this section demonstrates how they can address gaps in metadata quality and applicability, while also acknowledging their limitations. This chapter also examines the use of summarisation to manage the large volumes of data generated through interview transcripts, evaluating its role in capturing and communicating key insights. Together, these findings provide a comprehensive view of the metadata gap and propose actionable insights to improve data reusability.

5.1. Mapping the metadata gap

This section delves into the specifics of the metadata gap by examining its two key dimensions: supply and demand. On the supply side, it focusses on understanding what motivates data producers to create metadata and the barriers that hinder their ability to meet consumer needs. These insights are structured using self-determination theory (SDT), which categorises internal motivations into three key areas: relatedness, competence, and autonomy. In addition to motivations, this section also identifies barriers producers face when creating metadata, divided into two categories: constraints, systemic barriers such as limited resources or data complexity, and limitations such as gaps in knowledge or expertise.

On the demand side, this section examines the challenges consumers face when reusing datasets, categorising barriers into the same two groups as on the supply side. Constraints refer to the resources or information consumers rely on to understand the metadata or data enabling efficient reuse. Consumer limitations involve gaps in knowledge or expertise that further hinder their ability to reuse datasets effectively.

Table 5.1 provides a detailed overview of the metadata gap, highlighting the specific motivations, constraints, and limitations facing both producers and consumers. Although the list is not exhaustive and some categories may overlap, it represents the most comprehensive classification based on the available data. Each factor is supported by evidence from the literature review, interview results, or, in most cases, both. By systematically analysing both sides of the metadata gap, this section establishes a foundation for exploring targeted mechanisms, discussed later in the chapter, to address these challenges and bridge the metadata gap.

Table 5.1: Overview of the metadata gap, categorising the motivations, constraints, and limitations faced by data producers (supply) and consumers (demand). The Gap side column distinguishes between supply-side (producers) and demand-side (consumers). Motivational needs for producers are categorised under relatedness, competence, and autonomy, as discussed in Section 5.1.1, with two factors—S/M2 and S/M3—appearing in multiple categories. Barriers are divided into constraints and limitations to differentiate between systemic problems, which are external and structural, and limitations, which reflect internal or knowledge-based challenges. The supply side barriers are elaborated on in Section 5.1.2, and the demand side barriers in Section 5.1.3. This systematic breakdown lays the foundation for exploring the targeted mechanisms later in the chapter to address these challenges and bridge the metadata gap.

Gap side	Category	Sub-category	ID	Factor
Supply	Needs	Relatedness	S/M1	Enhancing communication and collaboration
			S/M2	Participating in grass-roots initiatives
			S/M3	Receiving peer recognition
			S/M4	Experiencing personal metadata frustrations
		Competence	S/M5	Building research credibility
			S/M6	Enhancing data quality
			S/M7	Developing data management skills
			S/M3	<i>Receiving peer recognition</i>
			S/M2	<i>Participating in grass-roots initiatives</i>
	Barriers	Constraints	S/C1	Time constraints
			S/C2	Space constraints
			S/C3	Data complexity
S/C4			Changing (meta)data standards and practices	
Limitations		S/L1	Lack of data management expertise	
	S/L2	Subjectivity in (meta)data creation		
	S/L3	Lack of insight into consumer needs		
	S/L4	Lack of ongoing commitment		
Demand	Barriers	Constraints	D/C1	Incomplete metadata
			D/C2	Scattered metadata
			D/C3	Divergent use cases
			D/C4	Un-reported data attributes
			D/C5	Limited access to producers
			D/C6	Inconsistent (meta)data standards
	Limitations	D/L1	Lack of domain knowledge	
		D/L2	Lack of data management skills	

5.1.1. Metadata supply: Motivations for creating high-quality metadata

Understanding metadata supply begins by examining the motivations that drive data producers to create metadata for their datasets. Although Section 2.1 addressed the broader motivations for making data reusable, this section focusses on practical incentives and barriers specifically related to creating high-quality metadata. High-quality metadata not only enhances the reusability of datasets, but also directly benefits producers by improving the quality and usability of their own work, as will be further explored in the competence section.

Motivations for creating high-quality metadata are analysed using the three fundamental human needs for internal motivation, as outlined by SDT [35]. By exploring these motivations and their practical implications, this section highlights opportunities to encourage the creation of high-quality metadata, ultimately supporting both producers and the broader research community.

- **Relatedness.** This need pertains to the producer's connection with the research community, including future consumers of their data, and highlights the factors that influence this motivational need.
- **Competence.** The producer's confidence in their research and data management skills highlights the intrinsic value of producing reliable and transparent work.
- **Autonomy.** Although less significant in the current context due to the inherently autonomous nature of research, external factors that restrict autonomy can hinder internal motivation.

Relatedness in the research community

Where competence and autonomy are closely related to intrinsic motivations, relatedness impacts mainly extrinsic motivations. Although intrinsic motivation often arises from personal interest or curiosity, extrinsically motivated behaviours are motivating in relation to others. For data producers, these others are the data consumers who reuse their datasets. Producers need a sense of connection to these consumers to feel motivated to address their needs [35]. Even if the needs of consumers do not align directly with those of producers, fostering a sense of relatedness can positively influence producers' internal motivation to create high-quality metadata.

This sense of relatedness is closely related to accountability, which can strengthen or weaken the connection between producers and consumers. As [3] notes, scientists are primarily focused on using data, "not on describing them for the benefit of invisible, unknown future users, to whom they are not accountable and from whom they receive little if any benefit"[3, p. 673]. This quote highlights how a lack of relatedness to data consumers can negatively impact metadata quality while also underscoring the opportunity to improve this connection.

Enhancing communication and collaboration. By emphasising accountability mechanisms that encourage horizontal relationships, such as collaboration with peers or direct engagement with identifiable stakeholders, producers can feel more connected to the broader research community [45]. Treating research data management more like a social network, where individuals actively share and exchange knowledge, further strengthens this sense of relatedness [46]. The interviewees noted that direct discussions with collaborators or intended data consumers encouraged them to produce higher-quality metadata. These interactions helped producers see the value of their contributions, making metadata creation feel less isolated and more purposeful. By promoting open communication and collective problem-solving, collaboration fosters a sense of shared purpose, aligning producer efforts with the needs of the research community.

Participating in grass-roots initiatives. Bottom-up collaboration involves engaging all stakeholders in the data management process, ensuring that diverse perspectives are considered and valued [4]. This includes involving stakeholders in decision-making, from the design of metadata templates to the identification of potential improvements. Such grass-roots initiatives rely on widespread participation and aim to include stakeholders at every stage, fostering collective efforts to improve metadata quality.

Bottom-up collaboration fosters ownership and autonomy by actively engaging stakeholders in the data management process [47]. By promoting recognition and data openness, it creates an environment where individual contributions are valued, encouraging stakeholders to become active stewards of their data. This is particularly effective for managing complex datasets where standardisation is difficult [47].

Community-driven projects, such as the R programming language, highlight the potential of bottom-up initiatives to address complexities, improve interoperability, and foster collaboration through shared tools [48]. These efforts not only improve metadata quality, but also support the generation of new research questions and provide a framework for sustainable data stewardship [47].

Building trust among stakeholders is key to strengthening bottom-up collaboration. Shared activities, such as co-authoring papers or engaging in joint projects, reinforce mutual understanding and shared goals [4], [48]. Equipping stakeholders with the necessary tools and knowledge ensures their active participation and meaningful contributions to the metadata management process [48].

Receiving peer recognition. Recognisance from peers reinforces relatedness by helping producers understand the value of their efforts within the research community. For example, several interview participants indicated that contributing to the broader research community motivated them to ensure that their metadata were accessible and useful. Still insufficient recognition of metadata efforts is a frequently raised issue [49], [47] notes that even small recognitions of effort can make a difference in a person's motivation, such as a thank you email [50]. Opportunities for mutual recognition, such as the participation of consumers in the creation of metadata, can further bridge the gap between producers and consumers while fostering stronger connections [48]. The factor of grass-roots initiatives can contribute here by reducing the gap between producers and consumers [48], bringing the two closer to achieve recognition. Bottom-up collaboration helps bridge the gap between various stakeholders, such as data producers and consumers, by fostering mutual recognition and shared responsibility [48].

Experiencing personal metadata frustrations. Beyond recognition of others' experiences, personal experience can also provide valuable insight to inform and enhance data management practices. Interview participants shared how frustrations with incomplete or unclear metadata inspired them to en-

sure that their own data was well-documented. Some also noted that anticipating their own future reuse needs prompted them to maintain well-organised metadata records during the data-creation process.

The participants highlighted that addressing their own frustrations not only benefited future users but also made their workflows more efficient. This proactive approach fosters empathy and a deeper commitment to improving metadata practices, especially when producers consider the challenges they themselves have faced or could face in the future.

Competence in research skills

SDT posits that feelings of competence, which refer to the sense of being capable and effective in one's tasks, can have a significant impact on a person's intrinsic motivation. Social-contextual factors, such as constructive feedback or rewards, can enhance these feelings, while demeaning evaluations can diminish them [35]. In the context of metadata supply, producers who feel confident in their research and data management skills are more likely to take pride in their work and strive for high-quality output, especially when their efforts are acknowledged and rewarded.

Building research credibility. One way metadata creation fosters competence is by increasing confidence in research skills. When producers document their metadata comprehensively, they improve the transparency and understandability, and thus the credibility of their research [51]. This was evident in interviews, where several producers expressed that maintaining detailed metadata gave them confidence in the robustness of their datasets and their ability to communicate findings effectively. High-quality metadata not only reflects a commitment to transparency, but also increases the likelihood that the research will be trusted and reused by others. This, in turn, reinforces the producer's sense of professional achievement, especially if they receive explicit positive feedback.

Enhancing data quality. Metadata creation also acts as a form of quality control, enabling producers to identify inconsistencies or gaps within their datasets during documentation [4]. Moreover, by providing contextual metadata they can also reevaluate whether the rationale behind their decisions was sound. By improving the dataset itself, metadata creation strengthens the overall research output, further contributing to the producer's sense of competence.

Developing data management skills. Confidence in data management skills is another important aspect of competence. For many producers, metadata creation can seem daunting, particularly if they lack training or experience [48]. This aligns with the interview findings, where less experienced participants often expressed uncertainty about the value of their insights, while more experienced producers demonstrated confidence in their data management practices. Tools and resources can help bridge this gap, streamlining the metadata creation process, and reducing frustration. Community-driven cyber-infrastructure can not only support the development of competencies, but also foster relatedness by encouraging collaboration and shared participation [48].

Receiving peer recognition. Peer recognition, discussed previously in the relatedness section, also plays a role in feelings of competence [52]. The recognition of peers reinforces the confidence of producers in their abilities and validates their efforts to create high-quality metadata. For instance, some producers noted that recognition from their supervisors provided reassurance that their work was valuable. However, it also became apparent that few producers received peer recognition for their work, which shows potential for improvement.

Autonomy in the metadata creation process

The third fundamental human need discussed in SDT is that of autonomy, which refers to the sense of being in control of one's actions and decisions. Even when individuals have a similar sense of competence, those with greater autonomy often exhibit higher levels of interest, excitement, and confidence [35]. This, in turn, results in improved performance, persistence, and creativity [53], [54].

Participating in grass-roots initiatives. Scientific research, by its nature, provides a significant degree of autonomy. Although researchers follow established guidelines and rely on peer evaluation to conduct and present research, much of the decision-making process, such as experiment design, interpretation of results, and selection of methodologies, remains the responsibility of the individual. This freedom fosters a sense of autonomy and ownership for the data producer [4]. These factors show the relation between autonomy and relatedness, as producers retain control over their work while benefiting from collective input and support, demonstrating how closely these motivational factors are intertwined.

However, external influences can sometimes inhibit autonomy, even as they aim to improve research quality. Securing funding, for instance, often requires compliance with mandatory requirements

set by funding bodies [31]. These may include adhering to specific methodologies, addressing particular research questions, or meeting detailed reporting standards, limiting the independent decision-making of researchers. Similarly, submitting datasets to repositories involves complying with metadata requirements and standards designed to ensure interoperability and usability for future users. Although such frameworks incentivise the production of high-quality work, overreliance on external standards risks diminishing the sense of autonomy critical to maintaining internal motivation.

In conclusion, relatedness is essential to motivate producers by fostering connections with the research community and future data consumers. Competence drives motivation by reinforcing confidence in research and data management skills through clear guidance and recognition. Autonomy supports motivation by allowing producers the flexibility to tailor metadata practices to their specific needs while maintaining ownership of their work. Addressing the individual factors discussed in each category can enhance producers' internal motivation to create high-quality metadata, thus improving the supply side of the metadata gap.

5.1.2. Metadata supply: Barriers to creating high-quality metadata

This section examines the barriers individual producers face in creating metadata that facilitate efficient data reuse. Time and space constraints play an important role, with greater expectations for high-quality metadata requiring more resources. Gaps in knowledge further compound these challenges, making it harder to use limited resources effectively. The interplay of these barriers underscores the difficulty producers face in meeting metadata expectations with constrained resources and incomplete guidance.

Metadata supply constraints to metadata creation

The creation of metadata is influenced not only by the skills and knowledge of the producer but also by the inherent characteristics of the data management process and the data itself. This section explores these barriers.

Time constraints. Creating comprehensive metadata is often perceived as a time-consuming task [48]. For many researchers, metadata documentation feels like an additional burden layered on top of their core research responsibilities [3]. This perception can lead to metadata being deprioritized, especially when researchers face tight schedules or competing demands for their attention.

Standards can streamline the process of providing metadata, but they also impact autonomy. Moreover, the more one tries to adhere to open data standards, the more work it is to provide metadata [2]. There are simply a lot of bureaucratic hurdles to overcome when providing metadata, which adds to the burden placed on the data producer [55]. During the interviews, producers noted that the time and effort required to meet these demands often left little room to consider additional metadata needs, such as those that might address diverse consumer needs. Some producers also indicated in hindsight that integrating metadata creation earlier in the project would have streamlined the process and improved quality.

Space constraints. When publishing an article, datasets are typically uploaded to data repositories. This creates two primary locations for storing metadata: the article itself and the data repository. However, each option presents distinct challenges that affect the quality and availability of metadata.

In articles, the inclusion of metadata depends on the role of the dataset in the research. Interviews revealed that if data production is secondary, articles often lack space to detail the rationale or methodology. Even when data production is central, methodologies are simplified to highlight key findings, leaving out alternative paths or rationale behind decisions made, which are often crucial to data consumers (Section 5.1.3).

Repositories, while offering space for metadata, often rely on standardised templates focused on structured fields. This approach promotes consistency, but can increase workload, especially when including more types of metadata [2]. Producers noted that although repositories sometimes allow contextual metadata, the lack of structure often leads to reliance on informal practices, such as copying what others have done. This contributes to variability and inconsistent quality.

Without designated spaces for contextual metadata in either articles or repositories, this information is often excluded, even though during the interviews producers themselves saw value in this information. Less experienced producers frequently depend on templates or conventions within the field, which may lack depth or are never formally defined. In contrast, more experienced researchers adapt

metadata to anticipated reuse needs, but even they deprioritise metadata creation when resources are limited or when platforms fail to accommodate detailed entries.

Balancing standardisation and flexibility is essential. Although standards provide a designated space for various types of metadata, enhancing interoperability and supporting novice researchers, a more flexible approach grants producers more autonomy. This flexibility can enable them to tailor the metadata to the specific needs of each dataset and ensuring that the metadata remains adaptable and relevant.

Data complexity. The diversity of data formats and structures is another barrier to effective metadata creation. Heterogeneous datasets, especially those that combine different types of information (e.g. numerical, textual, or geospatial data), make it difficult to apply consistent standards. The more complex the dataset, the greater the effort required to describe its characteristics accurately, leading to an increased workload for producers. This complexity can result in incomplete or inconsistent metadata, making it harder for future users to understand and integrate the data into their own workflows. Especially on scale, data complexity can become problematic [14], [47].

Changing (meta)data standards and practices. Metadata creation is an iterative process. As [3] argues, metadata as a product is good, but better results can be achieved by seeing metadata-as-process. Over time, needs might change, and communication styles should be adjusted according to the people involved. Confidentiality, as mentioned during interviews, can also influence the level of detail and accessibility of metadata. This includes concerns about data leakage risks, which can lead producers to limit or omit certain metadata details [55]. One producer expressed concern that making a dataset more accessible increases the likelihood that someone else might use it to pursue research that they were planning to conduct themselves. Moreover, standards and conventions can change. Even though the data itself could still be used, reuse is often degraded because the metadata has lagged behind. As noted in the discussion on space constraints, junior researchers highlighted during interviews that conventions within a domain are often unwritten, and understanding them can require significant time and effort.

Data producer limitations impacting metadata creation

The creation of high-quality metadata requires specialised knowledge and skills, which are often lacking in research settings. This section explores key areas where knowledge and skill deficiencies arise, their impact on metadata quality, and how they evolve over time.

Lack of data management expertise. Data management skills, as highlighted during interviews, are rarely taught comprehensively in university programmes outside of computer science. Instead, these skills are often acquired informally through experience or guidance from colleagues. Although researchers should not need to become full data management experts to produce high-quality standards-compliant metadata, tools should instead be designed to support them effectively [6]. However, having a basic understanding of data management principles would also be beneficial. For smaller projects, in particular, securing the expertise needed to establish and maintain proper data management practices can be challenging due to limited resources and personnel [56]–[58].

This leaves researchers to tackle metadata-related challenges to the best of their ability, often without formal data management expertise. Efforts to reduce technical barriers can improve data uploads; however, these efforts can inadvertently reduce metadata quality, compromising discoverability and reusability [47]. Finding a balance between accessible systems with low technical barriers and maintaining the high-quality outcomes associated with skilled data management remains a challenge.

Subjectivity in (meta)data creation. As [6] notes, researchers tend to organise their work in highly personal ways, tailoring metadata to their own needs or to the anticipated audience. They also vary the level and type of tagging depending on whether the metadata is intended for personal use or for others, further emphasising the subjective nature of metadata creation. The interviews confirmed that metadata creation is often shaped by the individual perspectives of the researchers, particularly when considering specific future users or use cases. This subjectivity is most evident in softer forms of metadata, such as contextual explanations, methodology descriptions, or interpretive notes, where domain experience heavily influences what researchers prioritise.

For more experienced researchers, this approach often works well. For example, one participant explicitly accounted for both novice and advanced future users when creating metadata, demonstrating the benefits of experience. However, less experienced researchers, relying on limited practices or examples, were more likely to produce inconsistent or incomplete metadata. Such variability in

approaches and priorities contributes to a lack of standardisation, making metadata harder to reuse across contexts.

This narrow, user-specific focus often limits the consideration of broader or unconventional uses of the data, restricting what metadata is included. The challenge lies in balancing the need for personalisation with the need for standardisation to ensure metadata can support diverse reuses effectively.

Lack of insight into consumer needs. Subjectivity is a challenge for both metadata producers and consumers, and the lack of interaction between the two exacerbates the problem. The interviews revealed that producers often lack a clear understanding of consumer needs due to minimal communication, except in cases where both are part of the same organisation. Producers often base their metadata practices on assumptions about future users, often expecting them to have similar or greater expertise. This can lead to insufficient detail for less experienced users or those outside the original research context, as will be discussed further in the next section. In terms of out-of-the-box use cases, consumers indicated primarily during interviews that their research aligned with what the original data producer had envisioned, making this less of a concern; however, this will be discussed in more detail in Section 5.1.3.

Lack of ongoing commitment. Metadata practices are not static; they evolve as the knowledge and skills of producers and consumers change over time [22]. New standards, tools, and research methodologies continually shape the metadata needed. However, producers often lack incentives to regularly update the metadata in their published datasets. This issue was illustrated during the interviews, where a consumer seeking clarification from the producer discovered that the creation of the dataset had occurred so long ago that the producer had forgotten many of the details. Such situations highlight the importance of ongoing commitments to ensure that metadata remain relevant and effective for future users.

In summary, metadata creation faces numerous barriers, including time and space constraints, gaps in producer knowledge, and limited understanding of consumer needs. The evolving nature of standards, tools, and practices further complicates these challenges, requiring ongoing commitments to maintain metadata quality. Addressing these barriers can simplify the process of creating and maintaining the metadata supply needed to support efficient and effective data reuse.

5.1.3. Metadata demand: Barriers to efficient reuse of datasets

High-quality metadata enables researchers to identify gaps and opportunities within datasets, inspiring new research questions while clarifying what a dataset can and cannot address to better align with specific use cases. Additionally, metadata plays a critical role in improving data management literacy by serving as both a reference and a learning tool for consumers who navigate complex datasets [4].

However, despite its importance, current metadata systems often do not meet consumer needs. These challenges not only introduce inefficiencies, but also result in untapped potential, as consumers are often forced to rely on ad hoc strategies to adapt datasets to their specific requirements.

Metadata demand constraints to metadata reuse

This section explores the challenges consumers face in reusing datasets, focussing on incomplete metadata, scattered metadata, divergent use cases, limited access to producers, unreported data attributes, and inconsistent metadata standards.

Incomplete metadata. From the range of metadata supply barriers, it is clear that supplying all necessary metadata is extremely challenging if not impossible. To overcome these gaps, consumers resort to various strategies, such as consulting similar studies, reaching out to colleagues, or contacting producers directly. However, contacting producers is often unsuccessful, especially when datasets are older or when communication barriers like language or domain-specific expertise arise. For instance, one consumer highlighted the challenges of using a poorly formatted dataset with missing contextual details, which required significant additional effort to interpret and integrate. The absence of adequate metadata leads to inefficiencies and delays, particularly when consumers must rely on trial and error or external expertise to adapt datasets to their needs. Even when successful, these efforts are rarely documented, perpetuating the same challenges for future users.

Scattered metadata. To gather all the necessary metadata, consumers often need to consolidate information from various sources. Metadata is frequently scattered across multiple locations, making access and usability more complicated for consumers. Structured metadata is typically available in

repositories, but these vary significantly in their organisational standards. Domain-specific repositories often provide clearer guidance, making metadata easier to access and reuse, whereas datasets hosted on personal websites or general-purpose repositories often suffer from inconsistent formats. For example, some consumers noted difficulties in navigating disorganised repositories or outdated websites.

Softer contextual metadata, such as the rationale for data collection or intended use, is often missing from repositories and is instead embedded in accompanying articles or supplementary materials. Links between the dataset and these materials are not always explicit, forcing consumers to search for relevant articles or repository documentation. This fragmentation delays understanding and reuse.

Divergent use cases. Consumer needs of course depend on the use case. It is easier if the use case is in line with ideas the producer had, they might even include information that is useful even. Sometimes someone might come up with ideas for a dataset completely from a new angle, and this, of course, would make things more complicated.

When the use of a consumer aligns with the original purpose of the producer for the dataset, metadata is more likely to include relevant and helpful details. Fortunately, this was the case for most of the consumers interviewed; however, this alignment may reflect a selection bias, as consumers likely chose datasets that they believed could suit their needs. Given that these datasets were selected due to reuse challenges, there still appears to be some misalignment between what producers deemed necessary to include and what consumers ultimately needed. When consumers approached datasets with truly unconventional use cases, the challenges of reuse became even more pronounced. Despite these risks, such out-of-the-box use cases can uncover new areas of interest, ultimately enhancing the value and applicability of datasets.

Un-reported data attributes. When consumers approach datasets with use cases that differ significantly from the producer's original intent, the likelihood increases that a critical data attribute will be missing. Producers often decide which data attributes to include based on their perceived importance. These decisions are shaped by the intended use of the dataset by the producer, which may not align with the needs of future consumers. As a result, consumers indicated during interviews that they would sometimes encounter problems when specific data attributes they expected were missing. Unreported attributes not only hinder the immediate reuse of datasets, but also limit the potential for innovative applications, as consumers struggle to adapt datasets without a clear understanding of their structure or limitations.

Limited access to producers. Direct contact with data producers can be an effective way to resolve issues by leveraging their expertise and addressing missing details [48]. However, some consumers reported attempting to contact the producers of the dataset for clarification or additional information with mixed success. Some producers provided helpful information, clarifying conventions, or explaining missing details, while others did not respond or were unavailable due to changes in roles or affiliations. For example, consumers reported difficulties reaching producers when the datasets were outdated or when the producer had left the organisation. These limitations underscore the importance of robust metadata that minimises reliance on direct producer input for dataset reuse.

Inconsistent (meta)data standards. Variability in metadata standards between repositories and domains introduces additional challenges. Some repositories enforce strict guidelines for metadata creation, making datasets more standardised and easier to reuse, while others have looser or undefined standards, resulting in inconsistent metadata. Consumers indicated that they had to rely on their own expertise to interpret metadata or cross-reference multiple sources to fill gaps, which is time-consuming and error-prone.

Data consumer limitations impacting data reuse

Understanding consumer needs is crucial to ensure that metadata is useful and accessible, yet producers often underestimate these needs. This variation stems from differences in use cases, domain expertise, and levels of experience in data management, which can create significant barriers to data reuse.

Lack of domain knowledge. A dataset created for one research domain may also have applications in another, but interdomain and interdisciplinary uses make metadata needs even more difficult to anticipate. Consumers who work outside the producer's domain often struggle with unfamiliar conventions or implicit assumptions that producers took for granted. Even if these conventions are widespread within a domain, subtle differences in definitions between organisations can create challenges [16].

Even within the same domain, differences in experience levels can create challenges. Novice researchers may lack familiarity with established conventions or informal practices, making it harder to interpret metadata and fully leverage the dataset. For example, a consumer highlighted that a lack of detailed metadata forced them to pivot their methodology, while another reported that inadequate documentation delayed their ability to identify key variables. These examples underscore how disparities in domain knowledge exacerbate metadata reuse challenges.

Lack of data management skills. In addition to domain-specific challenges, varying levels of experience in data handling significantly influence what consumers need from metadata. Junior researchers, in particular, often lack formal training in data management, requiring them to develop these skills during their projects. For example, a participant struggled with a dataset that lacked a clear format or organisation, highlighting how stronger data management skills could have alleviated the problem. Another participant noted that their limited coding knowledge prolonged the data processing stage, adding unnecessary delays.

Research by [5] highlights that a better understanding of data management tools not only streamlines processes but also allows researchers to explore new research questions that could otherwise remain inaccessible. Bridging gaps in data management skills and providing clearer guidance on tool usage could greatly improve dataset reusability, particularly for less experienced consumers navigating complex or poorly documented systems.

Consumers encounter various constraints and limitations when reusing datasets, which significantly impede the potential of metadata to facilitate efficient and innovative research. Incomplete and unreported (meta)data restrict consumers' ability to adapt datasets for novel use cases, limiting opportunities for interdisciplinary research and innovative applications. When use cases diverge from the producer's original intention, these issues become even more pronounced, and contacting the producer is often not a practical solution. Inconsistent standards exacerbate the challenges, particularly for less experienced consumers who are constrained by gaps in domain knowledge and data management expertise. Addressing these barriers is essential to bridge the gap between metadata supply and demand, ultimately making dataset reuse more efficient and impactful.

5.2. Bridging the metadata gap

In this section, we address four key aspects of the proposed implementation of the data conversation, explored during the interviews, and discuss by which mechanism each element addresses the metadata supply and demand gaps outlined previously. The four mechanisms are: involving consumers as metadata co-creators, integrating contextual metadata, leveraging real-time dialogue, and incorporating question adaptation.

The interviews revealed that while data conversations occur frequently in practice, they are typically informal and seldom recorded. For instance, when future users are within the same organisation, producers often sit down to explain dataset details or make themselves available to answer questions. However, these valuable exchanges are rarely documented, which limits their usefulness for broader reuse and hinders metadata improvement.

For each mechanism, we explain how the proposed data conversation format facilitates this aspect and outline the sub-mechanisms for each (see Table 5.2). The subdivision of data conversation mechanisms into smaller sub-mechanisms serves two primary purposes. First, it provides a structured and detailed framework to illustrate how context-bridging data conversations can address various components of the metadata gap identified in this thesis. By breaking down each mechanism into specific sub-mechanisms, the relationship between the mechanisms and the metadata challenges becomes clearer and more actionable. Second, this structured approach offers guidance for future research, enabling researchers to focus on particular sub-mechanisms depending on which aspects of the metadata gap they aim to address. While overlap between sub-mechanisms is unavoidable, this framework provides some structure for future research, offering a practical guide to exploring the role of data conversations in addressing metadata challenges. We then summarise the impacts of each mechanism on these gaps in Table 5.3. It is important to note that the tables highlight the most critical links between mechanisms and metadata gaps, but may not be exhaustive, leaving room for future exploration and refinement. Finally, we examine the limitations specific to each element.

Table 5.2: Context-bridging data conversation mechanisms broken down into parts. Breakdown of context-bridging data conversation mechanisms into their sub-mechanisms, highlighting specific strategies to address metadata challenges. Each sub-mechanism addresses parts of the metadata gap as detailed in Table 5.3. Section 5.2 provides additional insights into the application of these mechanisms.

Mechanism	ID	Sub-mechanism
Consumers as co-creators	C1	Integrate more perspectives
	C2	Record and share consumer experiences
	C3	Learn continuously through consumer participation
	C4	Create feedback loops between producers and consumers
	C5	Encourage social connection through collaboration
Contextual metadata	M1	Capture implicit knowledge
	M2	Document obstacles and paths not pursued
	M3	Provide a designated space for contextual metadata
Real-time dialogue	D1	Speed up metadata creation through dialogue
	D2	Capture nuanced details using (non)-verbal cue
	D3	Resolve ambiguities with real-time feedback
	D4	Build trust and engagement through interaction
Question adaptation	A1	Adapt questions to changing (meta)data standards
	A2	Tailor questions to reflect evolving user needs
	A3	Improve response quality by refining question clarity

Table 5.3: This table links motivations (M), limitations (L), and constraints (C) from both the metadata supply (S) and demand (D) sides of the metadata gap to the sub-mechanisms of consumers as co-creators (C), contextual metadata (M), real-time dialogue (D), and question adaptation (A). Refer to Table 5.2 for descriptions of the sub-mechanisms corresponding to the IDs used here, and see Table 5.1 for more details about the metadata gap factors. Section 5.2 provides further explanation of how each sub-mechanism addresses these factors and contributes to bridging the metadata gap.

ID	Metadata gap factors	C	M	D	A
S/M1	Enhancing communication and collaboration	C5	M1, M3	D4	A3
S/M2	Participating in grass-roots initiatives	C5		D4	A2
S/M3	Receiving peer recognition	C4			
S/M4	Experiencing personal metadata frustrations	C4	M2		
S/M5	Building research credibility		M1		
S/M6	Enhancing data quality	C4	M1		
S/M7	Developing data management skills and tools	C4	M1		A1
S/C1	Time constraints	C1		D1	A3
S/C2	Space constraints		M3		
S/C3	Data complexity	C1			
S/C4	Changing (meta)data standards and practices		M3		A1
S/L1	Lack of data management expertise		M1	D3	A1
S/L2	Subjectivity in (meta)data creation		M1	D3	
S/L3	Lack of insight into consumer needs	C2	M1	D3	A2
S/L4	Lack of ongoing commitment	C3-4			
D/C1	Incomplete metadata	C2	M1	D2	
D/C2	Scattered metadata	C2	M3		
D/C3	Divergent use cases	C2	M1		A2
D/C4	Un-reported data attributes		M1	D2	A2
D/C5	Divergent use cases		M1-2		A2
D/C6	Inconsistent (meta)data standards		M1, M3		
D/L1	Lack of domain knowledge	C2	M1-2	D3	A2
D/L2	Lack of data management skills	C2	M1-2	D3	A2

5.2.1. Consumers as metadata co-creators to bridge the gap

Involving consumers in the metadata creation process is a key mechanism to bridge the gap between metadata supply and demand. By integrating consumer needs and perspectives from the outset, this approach promotes shared responsibility [4] and promotes collaboration between producers and consumers.

Consumers vary widely in their use cases, domain expertise, and data management experience. Although this diversity poses challenges for tailoring metadata, it also provides an opportunity to capture a wealth of potentially valuable insights each time a dataset is reused. By sharing these varied perspectives, the contextual understanding of the data becomes richer and more comprehensive, facilitating the sharing and reuse of knowledge [59]. This aligns with the findings of [60], which highlight the critical role of capturing user experiences to enhance the utility of data.

Sub-mechanisms through which consumers as co-creators can help bridge metadata gaps:

- **C1. Integrate more perspectives:** Capture more and different perspectives by increasing the number of people involved in creating and maintaining metadata.

→ S/C1, S/C3

Incorporating more people can help address the time and data complexity constraint by better dividing the metadata supply burden.

- **C2. Record and share consumer experiences:** Collect insights and feedback on metadata demand gaps from data consumers.

→ S/L3, D/C1-3, D/L1-2

When consumers encounter scattered or incomplete metadata needed for their use case, recording and communicating their corrections or improvements to both future producers and consumers can help inform future data creators and users.

- **C3. Learn continuously through consumer participation:** Leverage repeated data reuse by collecting metadata with each consumer reuse to iteratively improve its quality.

→ S/L4

As consumers continue to reuse a dataset, their evolving needs can be captured through recorded experiences, helping to ensure that metadata remains relevant and up to date .

- **C4. Create feedback loops between producers and consumers:** Establish practices for collaborative metadata refinement between data producers and consumers to create a shared understanding of (meta)data requirements.

→ S/M3-4, S/M6-7, S/L4

Recording consumer experiences allows positive feedback to serve as recognition for producers, showing the impact of their metadata efforts. Moreover, consumers' negative experiences can potentially elicit a reaction similar to what a producer might feel if they encountered such issues themselves. This shared understanding fosters empathy and a greater emphasis on addressing these challenges. Additionally, feedback mechanisms can enhance data quality and compensate for any deficiencies in the producer's data management expertise. Finally, feedback can keep the data producer engaged for longer.

- **C5. Encourage social connection through collaboration:** Strengthen relationships between producers and consumers by fostering closer connections to build a more engaged data community.

→ S/M1-2

By fostering stronger communication and collaboration between producers and consumers, a greater sense of connection and shared purpose can be created helping motivate the data producer. And by involving stakeholders more pro-actively, namely, data consumers, metadata management becomes more of a grass-roots effort.

Consumers as metadata co-creators limitations

While involving consumers as co-creators has significant benefits, there are also limitations. Firstly, this thesis did not explicitly test whether the experience of one consumer could directly benefit another.

Since the data conversations were not designed to delve deeply into technical details, it remains uncertain whether insights from one data conversation would be transferable or beneficial to subsequent users. The interview findings do seem to suggest that at least for junior data consumers there will be overlap as they all have less domain and data management experience. It is likely that this will be project- and dataset-dependent.

Secondly, as suggested by a producer, involving consumers may unintentionally discourage producers from providing detailed metadata, as they might be tempted to rely on consumers to fill in the gaps. Although producers may have gaps in their knowledge, they can still provide valuable information. Improving the data conversation process as proposed in this thesis aims to reduce metadata supply barriers sufficiently to keep producers motivated to provide high-quality metadata despite this potential limitation.

Thirdly, adding consumers as co-creators adds tasks to the metadata supply side. Although the proposed methods aim to minimise this burden and highlight the benefits, the success of the system depends on broad participation.

5.2.2. Contextual metadata to bridge the gap

Contextual metadata offers detailed background information, rationale, and domain-specific explanations essential to understand and use datasets effectively. By capturing the often unspoken knowledge that producers have about their datasets, contextual metadata ensures that nuances are preserved, enhancing both the usability and credibility of the data. This element not only emphasises the importance of including contextual metadata, but also ensures that there is a dedicated space to record these details. Without this, valuable information can be lost, overlooked, or scattered across various sources, reducing the long-term utility of the dataset.

Sub-mechanisms through which an increased focus on contextual metadata can help bridge metadata gaps:

- **M1. Capture implicit knowledge:** Expand the metadata supply by including contextual metadata details that are often assumed, such as domain-specific conventions or definitions, and the rationale behind key decisions.

→ S/M1, S/M5-7, S/L1-3, D/C1, D/C3-6, D/L1-2

This sub-mechanism improves the metadata gap by increasing producer awareness of their strengths and areas for growth through explicit documentation of implicit information. It enhances transparency, enabling consumers to better understand the rationale behind decisions and exposing any gaps in knowledge. Furthermore, it helps consumers assess whether the metadata meets their needs, reducing their reliance on personal expertise by providing access to the collective knowledge of producers and previous dataset users.

- **M2. Document obstacles and paths not pursued:** Expand the metadata supply by including contextual metadata details that are often assumed, such as challenges encountered during data collection and processing, along with the rationale behind decisions to forgo specific approaches.

→ S/M4, D/C5, D/L1-2

By documenting the challenges encountered during the data creation process, producers can better recognise what information could be valuable to future users. Additionally, consumers can benefit from the producers' experiences, gaining insights not only into what was successful but also into what was not, helping them avoid repeating approaches that producers have already determined to be ineffective.

- **M3. Provide a designated space for contextual metadata:** Establish a dedicated area within the data management system to ensure that contextual information is systematically included, easily accessible, searchable, and modifiable for both producers and consumers.

→ S/M1, S/C2, S/C4, D/C2, D/C6

Providing a dedicated space for this type of metadata, which is often scattered across various locations, offers both producers and consumers a centralized resource to share and access information.

Contextual metadata limitations

Although contextual metadata offers many advantages, there are limitations to consider as well. The first limitation is similar to one mentioned for the mechanism of consumers as co-creators. Although the thesis did not test whether one consumer's experience could directly benefit another, the same uncertainty applies to the contextual metadata mechanism, whether the details and rationale provided by one user could meaningfully help a wide range of consumers. However, the results of the interview suggest that consumers already seek and use this type of metadata independently, indicating that aggregating it into a single, accessible location could be a practical and beneficial approach. Uncertainty remains regarding the level of detail and participation needed to ensure that contextual metadata is useful, as well as whether the time required to capture this information will prove to be a significant barrier.

The second limitation also mirrors the challenges identified with the consumers-as-co-creators mechanism. Focussing on contextual metadata inevitably increases the volume of data collected, raising concerns about potentially obscuring critical information. However, the interviewees highlighted that it is often impossible to predict which details will prove useful in future scenarios. This reinforces the idea that collecting a wider range of metadata may still be valuable despite the risks of overloading users [4].

5.2.3. Real-time dialogue to bridge the gap

Instead of using a static written form to collect metadata, we tested whether using dialogue could help bridge some of the identified metadata gaps. This section does not focus on the specific content of the dialogue; this is primarily discussed in the section on contextual metadata (5.2.2). The method of adapting questions between interviews is covered in detail in Section 5.2.4. The hypothesis was that dialogue allows for greater detail through non-verbal cues, less polished responses, and opportunities for immediate feedback, all of which enhance the quality and richness of the metadata collected (see 3).

Sub-mechanisms through which real-time dialogue can help bridge metadata gaps:

- **D1. Speed up (meta)data creation through dialogue:** Facilitate faster metadata creation by enabling immediate verbal exchanges, leveraging the real-time nature of dialogue to streamline the process and reduce delays.

→ S/C1

This efficiency helps alleviate time constraints for producers by allowing them to provide metadata in a more streamlined and less time-consuming manner, and it would do the same if consumer also become metadata suppliers.

- **D2. Capture nuanced details using (non)-verbal cue:** Combine the unpolished, detailed nature of verbal responses with additional context provided by body language, tone, and other non-verbal cues to capture richer and more nuanced metadata, made possible through real-time dialogue.

→ D/C1, D/C4

The interviews demonstrated that the richness of the dialogue, including additional context and opportunities for follow-up questions, facilitated new insights into what metadata was important to include.

- **D3. Resolve ambiguities with real-time feedback:** Allow participants to clarify ambiguities and refine their responses instantly during conversations, using the interactive nature of real-time dialogue for immediate resolution.

→ S/L1-3, D/L1-2

This sub-mechanism can help clarify and rectify any knowledge gaps during the metadata creation process.

- **D4. Build trust and engagement through interaction:** Strengthen collaboration by fostering a sense of connection and mutual understanding, an outcome directly supported by the interactive and personal nature of real-time dialogue

→ S/M1-2

By making the metadata creation process a real-time collaborative process we can enhance the sense of community of metadata producers, increasing their sense of relatedness.

Real-time dialogue limitations

Despite its advantages, real-time dialogue as a method for metadata extraction also has limitations. First, all interviews were conducted in English, which was not the first language of most of the participants. Although much of scientific work is conducted in English, as is in the CropXR environment, this limitation may have reduced the potential advantages of dialogue by restricting participants' ability to express themselves fully. Allowing participants to respond in their native language and translating the responses afterwards could help address this limitation in future studies. This approach could improve the richness and precision of responses, particularly for participants less comfortable with spoken English.

Second, in addition to language differences, cultural variations can also influence the effectiveness of communication. Differences in communication styles can lead to misunderstandings. However, since the content of these conversations does not involve particularly sensitive or contentious topics, the impact of such barriers is likely to be minimal.

Third, real-time dialogue requires both the interviewer and the participant to be available at the same time, which can be difficult to arrange. Even in this study, where the interviewer was widely available, scheduling interviews remained a challenge. Although conducting data conversations online reduces logistical hurdles, scheduling still poses a barrier compared to traditional metadata forms. Additionally, the need to coordinate meetings introduces scalability challenges, as this approach requires the time and effort of two people instead of one. However, this method offers the advantage of reducing the producer's cognitive load, as the interviewer guides the conversation and identifies what is needed, streamlining the process.

If the primary focus is on guiding questions and speed of dialogue, an alternative could be to use a chatbot, especially with recent advances in realistic voice bots. Although this option might address some scheduling and scalability issues, it would make the process less personable. This limitation was reflected in the survey results, where chatbot-based metadata elicitation methods were ranked the lowest by participants. Maintaining a human element in the process is crucial to foster engagement and build trust, which are key factors that foster feelings of relatedness.

5.2.4. Adaptive data conversation questions to bridge the gap

As described in the interview methodology (see Section 3.2), interview questions were adapted both between and during the conversations. The goal was to maximise the potential of the conversational format by leaning toward the quality of adaptability while still extracting the important information. The questions were adjusted according to the responses of the participants, skipped if already addressed earlier in the conversation, or rephrased to improve clarity. In a broader sense, adaptability ensures that metadata processes, standards, and tools evolve alongside changing needs, technologies, and contexts.

Sub-mechanisms through which adaptive data conversation questioning can help bridge metadata gaps:

- **A1. Adapt questions to changing metadata standards:** Ensure that questions remain relevant and up-to-date by aligning them with evolving industry and research norms, leveraging the flexibility offered by adaptive questioning.

→ S/M7, S/C4, S/L1

Incorporating metadata standards into data conversations could alleviate some of the burden on producers by providing clear guidance on which (meta)data standards to follow and how to implement them effectively.

- **A2. Tailor questions to reflect evolving user needs:** Customise questions to reflect changing data consumer requirements.

→ S/M2, S/L3, D/C3-5, D/L1-2

By adapting interview questions based on consumer needs, the metadata supply can be better aligned with the metadata demand.

- **A3. Improve response quality by improving question clarity:** Refine data conversation questions to eliminate ambiguity, ensuring clearer responses and better insights through the iterative process of adaptive questioning.

→ S/M1, S/C1

Enhances the data conversation experience by making the questions more dynamic and responsive to the input of the participants, while also reducing the overall process.

Adaptive data conversation questions limitations

Despite its strengths, adaptability in the question adaptation mechanism presents two main challenges. The first limitation lies in the tension between flexibility and standardisation. Adapting questions introduced variability into the data conversations, which conflicts with the need for standardised metadata management practices. This flexibility increased the risk of skipping or overlooking critical details during conversations, potentially reducing metadata completeness and leading to additional workload due to follow-up clarifications. The variability between data conversations also resulted in inconsistencies in the metadata collected, making it harder to compare and integrate information across datasets. Although adaptability offers clear advantages, these findings underscore the persistent challenge of balancing flexibility with the need for standardisation in data management.

The second limitation concerns the reliance on interviewer skill and experience. The effectiveness of the adaptive approach depends heavily on the interviewer's ability to balance natural conversational flow with the need to extract critical information. Less experienced interviewers may struggle to identify key gaps in metadata or adapt their questioning dynamically, which can compromise the quality and completeness of the metadata collected. This dependence on skill highlights the need for extensive interviewer training and potentially standardised prompts or guidelines to mitigate the risks associated with interviewer variability.

5.2.5. Transcript summarisation to bridge the gap

The survey results provide valuable insights into the effectiveness of AI-generated summaries and their role in improving the usability of context-bridging data conversations. Participants generally found summaries to be more concise and accessible, demonstrating that this method can indeed reduce the time needed to communicate (contextual) metadata effectively.

However, access to the full transcript for additional details was also considered beneficial. Since complete transcripts can introduce significant identification risks, we recommend an alternative approach: combining a short summary, similar to the one used in this study, with sanitised excerpts from the transcript. These excerpts could be presented through collapsable menus, featuring direct quotes from the transcript that have been refined to remove stop words and unnecessary clarifications, focussing solely on the question-and-answer content. This flexibility ensures that the summary meets the needs of both those seeking a quick overview and those requiring a more in-depth exploration.

Providing participants with the option to proofread and refine their generated summaries before finalisation could further improve the quality and ensure a sense of ownership over the content. This additional step would offer participants greater autonomy while allowing those who have the time to make meaningful contributions to the metadata process.

The survey results also suggest that the summary section headers might be more effective in scenarios involving technical data or conversations centred on similar datasets. In this study, the diversity of the datasets and the varied nature of the conversations may have limited the perceived utility of these headers.

Lastly, the survey revealed a consistent preference for semi-structured interviews with AI-generated summaries over chatbot-only conversations, which were ranked lowest. Future work should explore why chatbot-based methods were less favoured and investigate which combinations of conversational styles and summarization techniques work best. This exploration should also account for the varying needs of participants based on their level of experience, as the junior researchers in the survey appeared to benefit more from structured guidance than their senior counterparts.

These findings underscore the importance of balancing conciseness, contextual depth, and participant autonomy in designing effective data conversation systems. Incorporating these recommendations into metadata workflows could significantly improve usability and user satisfaction while addressing key concerns raised in the survey.

In conclusion, these findings highlight the importance of addressing metadata supply and demand gaps through structured and context-sensitive data conversation mechanisms. By focussing on involving consumers as co-creators, integrating contextual metadata, leveraging real-time dialogue, and incorporating adaptive questioning, this approach provides actionable strategies for improving metadata quality and reusability. Each mechanism and its sub-mechanisms offer unique contributions to addressing key challenges, such as enhancing collaboration, capturing implicit knowledge, and adapting to evolving needs. The proposed framework emphasises the dynamic and collaborative nature of metadata creation. It illustrates how data conversations can transform informal exchanges into structured processes that bridge the metadata gap. However, the limitations of each mechanism highlight the need for further testing and refinement in real-world applications.

The survey findings offer valuable insights for refining the integration of context-bridging data conversations into data management systems. Although summaries were identified as a more accessible alternative to full transcripts, the results underscore the need to balance conciseness, contextual depth, and participant autonomy to effectively meet the diverse needs of users.

It is essential that each project assesses the specific aspects of the metadata gap that most require attention and identifies the components of the data conversation framework that best address those gaps within their unique context. Different projects may find it easier to implement certain mechanisms based on their resources, goals, and data types, and they must tailor their approaches accordingly to achieve the greatest impact. Through these mechanisms, this thesis lays the foundation for future research and practical implementation. It offers CropXR and similar initiatives a robust framework for improving metadata practices in diverse and interdisciplinary contexts. By addressing both the technical and social dimensions of metadata, this approach ensures that metadata creation becomes not only more efficient, but also more meaningful and impactful for all stakeholders involved.

6

Discussion & Future Work

This chapter synthesises the findings of this thesis by examining the benefits, limitations, and opportunities of context-bridging data conversations. It highlights their value in capturing undocumented practices into metadata workflows while identifying areas that require further development. Throughout the thesis, various limitations have been noted; however, this chapter focusses on three overarching challenges: transitioning from proof-of-concept to functional prototypes, managing the increased volume and complexity of metadata, and addressing the social dynamics of collaborative approaches. These challenges provide the foundation for future research directions.

Section 6.1 discusses the interpretation of the findings, focussing on the gap in metadata documentation and the importance of capturing undocumented consumer workarounds. Section 6.2 examines the key limitations of context-bridging data conversations, including scalability challenges, metadata volume management, and social dynamics, and explores how these issues might be addressed in future work.

6.1. Interpretation of findings: bridging the metadata gap

This study highlights a critical gap in metadata research: the lack of documentation for the performative routines that consumers rely on during data reuse. This under-documentation stems from the normalisation of leaving (meta)data creation and reuse workarounds undocumented. These workarounds, while discussed informally in producer-consumer interactions or noted for personal use, are rarely incorporated into metadata forms or research articles, leaving subsequent data consumers to repeatedly solve the same issues.

On the supply side, most barriers identified in this thesis are well-supported in the literature, with the notable exception of "Experiencing personal metadata frustrations." This exception is significant because it demonstrates how producers' challenges in metadata creation could play a vital role in bridging the metadata gap if they were incorporated into metadata management systems. On the demand side, this lack of documentation is even more pronounced. Of the six constraints identified in this thesis, only "Limited access to producers" is supported by existing research, likely because it directly intersects with supply-side challenges. The remaining five constraints, which focus on ad hoc routines used by consumers to address metadata gaps, reflect the informal and under-explored nature of workaround strategies.

The interview results reinforced this pattern, showing that the participants often did not recognise the value of documenting these workarounds until prompted by the data conversation. Many participants initially underestimated the importance of sharing such insights, only to realise during discussions how beneficial this information would have been to themselves or others. This highlights the need for tools and mechanisms, such as context-bridging data conversations, to capture these undocumented routines and integrate them into metadata workflows efficiently. Traditional static approaches to metadata creation do not accommodate these changes, leaving data consumers unsupported when reusing datasets for their unique requirements.

Involving data consumers as co-creators fosters collaboration, bridging gaps in understanding between stakeholders, expanding metadata utility, and incorporating diverse perspectives. Increasing the

focus on contextual metadata emphasises the importance of capturing the performative steps people take to create and reuse data, documenting these actions to provide richer, more actionable metadata. Integrating real-time dialogue enables detailed, rich communication, helping to reduce misunderstandings. And by adapting questions throughout the process, metadata can be tailored to the evolving and diverse needs of consumers, while also reducing the burden on producers. Through these mechanisms, context-bridging data conversations enable metadata systems to make data reuse more inclusive and efficient for both data producers and consumers, enabling greater collaboration and innovation in research.

6.2. The future of context-bridging data conversations

The applicability of the context-bridging data conversation in (meta)data management systems also has limitations addressed in this section with proposed mitigation strategies and future research directions.

The first limitation addresses the transition from proof-of-concept to functional prototype. Although the research offers valuable information, it remains theoretical in nature and lacks the scalability testing and practical implementation required for real-world applications. Advancing to a fully operational prototype will require further testing.

The second limitation concerns managing the increased volume of complex metadata generated through data conversations. Although these methods produce rich and detailed metadata, they also introduce challenges in prioritising and synthesising this information. Addressing this metadata haystack will require advanced tools for summarization and effective strategies for streamlining data.

The third limitation focusses on the complexities of social interactions in collaborative and community-driven approaches. Effective communication, inclusion, and conflict resolution are critical for success, particularly in diverse teams. Structured frameworks and ongoing support are essential to mitigate these challenges and foster productive collaboration.

6.2.1. Moving beyond the proof-of-concept to a prototype

The data conversation simulations conducted in this study provide valuable insight as a proof-of-concept, demonstrating the potential of each of the sub-mechanisms. However, these simulations do not align fully with the envisioned data conversation model, integrated into a data management system, and tested during multiple reuses of data sets. Instead, the study focused on isolated interactions within a controlled interview setting, leaving challenges to be addressed in future work.

One limitation lies in the absence of scalability testing. Although the proof-of-concept was effective in the controlled interview setting, it has not been applied as part of a larger data management system. Real-world systems involve complexities such as varied user expertise, different organisational workflows, and the need to manage significantly larger datasets, none of which were explored in this study.

Furthermore, due to the wide diversity of participants and limited resources, the study lacked a strong focus on technical details. As a result, many proposed benefits of the mechanisms, such as cross-conversation synthesis and adaptive metadata creation, remain largely untested. Whether these mechanisms would perform as envisioned in a large-scale data management project is still uncertain.

To move beyond this proof-of-concept, future work should emphasise iterative prototyping and real-world validation. A small-scale pilot involving producer-consumer pairs and using diverse datasets with high reuse potential could provide a focused opportunity to explore technical details and assess whether shared experiences across datasets yield practical benefits. Involving novice researchers as interviewers in the pilot would allow them to learn from the data conversations while contributing to metadata refinement. Providing these interviewers with pre-prepared questions, adapted from information collected in previous conversations, would ensure consistency and relevance while also serving as a learning tool. To reduce redundancy and improve the specificity of data conversations, information from articles associated with datasets should be used, particularly those that detail data creation or reuse processes. Key sections, such as introduction, methodology, and discussion, can offer context to align data conversation questions with existing knowledge and address gaps effectively. The survey results also indicate that AI-generated summaries have value, but highlight the need to balance concise overviews and detailed explanations. Future work could explore techniques for finding this balance, possibly with the help of feedback mechanisms in the system. Structured tools for collecting consumer feedback, such as comment sections in metadata repositories or usability surveys, could further enhance metadata practices by providing producers with actionable insights.

6.2.2. Managing the larger metadata haystack

The adoption of data conversations enriches metadata by capturing detailed and diverse information, but also amplifies the complexity of managing and navigating these contributions. This phenomenon of creating a larger metadata haystack poses several challenges to effective metadata usage.

The additional volume of metadata generated during conversations increases the “signal-to-noise ratio,” making it more difficult to extract the most relevant insights [29], [52]. As more metadata is introduced, the signal—the meaningful, actionable information—becomes harder to isolate from the noise—the extraneous or less useful data. Moreover, the signal itself is highly dependent on the perspective of the user; different users may find different aspects of the metadata relevant to their specific needs, further complicating the identification of key insights. This challenge is compounded by the processing limitations of current tools. For instance, while automated summarisation techniques tested in this study offer a promising solution, they struggle to achieve a balance between brevity and detail. Another approach would be to connect related discussions across datasets to reduce redundancy and enhance the overall usability of metadata.

To address the growing metadata haystack, future work should focus on developing tools that prioritise metadata, maintaining both metadata clarity and richness. A key strategy involves adopting customisable approaches to the presentation of metadata. Tiered metadata systems that provide basic metadata for quick reference along with advanced metadata for detailed exploration can help users with diverse levels of expertise navigate metadata effectively without feeling overwhelmed.

Cross-dataset integration offers a compelling approach to minimise redundancy by automatically linking related (meta)data. For instance, conventions used across multiple datasets could be documented once and referenced via metadata links, rather than duplicated in each dataset’s contextual metadata. This not only streamlines metadata management, but also improves navigation, possibly revealing before unnoticed connections, or discrepancies, between datasets.

Finally, regular evaluation and user feedback are essential to refine these tools and approaches, ensuring that they remain adaptable to evolving needs. Data conversation questions should incorporate this feedback to avoid redundancy and ensure relevance.

6.2.3. Navigating the complexities of social interactions

From the beginning, this thesis has argued that making data more reusable is not just a technical challenge but also a social one, rooted in the communication of information. Involving more participants, data consumers, and fostering a dynamic, real-time process through face-to-face dialogue can help bridge the gap. However, these mechanisms also introduce challenges related to communication, inclusion, and conflict resolution that require careful attention.

The conversational approach depends on effective communication, but barriers such as cultural differences, domain-specific terminology, and varying levels of data literacy can hinder understanding. These challenges are exacerbated in diverse teams, where translating individual perspectives into shared knowledge often leads to misunderstandings. Furthermore, fostering inclusivity within these conversations requires deliberate efforts to prevent dominance by certain individuals or groups, ensuring that all participants contribute meaningfully [47], [60].

Conflict resolution is another critical challenge. Disagreements over metadata standards, priorities, or interpretations can slow progress and create friction. These conflicts require skilled mediation and alignment on shared goals, which can be resource-intensive and difficult to achieve. As collaboration scales, these difficulties multiply, adding layers of complexity.

Despite these challenges, the collaborative potential of data conversations remains significant. They offer an opportunity to foster innovation, share knowledge across disciplines, and build a sense of community among stakeholders. To address these complexities, the context-bridging data conversation approach proposed in this thesis must be tailored to meet the specific needs of the community that implements it. Training and support for participants can improve communication skills, while the integration of technology to document informal exchanges ensures that valuable insights are not lost. Building trust through regular feedback loops and recognition of contributions further strengthens collaboration, paving the way for more inclusive and effective metadata ecosystems.

This chapter has highlighted how the mechanisms of context-bridging data conversations, consumer co-creation, contextual metadata, real-time dialogue, and adaptive questioning address critical gaps in metadata practices by providing a method to efficiently record undocumented routines of both pro-

ducers and consumers. At the same time, the chapter acknowledges significant challenges, including scalability, metadata complexity, and the social dynamics of collaboration, emphasising the need for iterative refinement and real-world testing.

The proposed strategies for future work, such as tiered metadata systems, cross-dataset integrations, and tools for capturing feedback, provide a roadmap for addressing these challenges while enhancing metadata systems' efficiency and inclusivity. By focusing on balancing clarity with depth and fostering communication within diverse teams, these strategies aim to create metadata workflows that are adaptable to evolving research needs.

Ultimately, this chapter underscores the transformative potential of context-bridging data conversations to make metadata practices more efficient, equitable, and sustainable. Building on the insights and recommendations presented here, future efforts can further develop scalable and innovative solutions, paving the way for richer, more collaborative data ecosystems. The practical application of the findings of this thesis to real-world scenarios is further explored in Chapter 7.

7

Case study: Mapping and bridging metadata gaps in CropXR

This chapter applies the findings of this thesis to a real-world research project with extensive data management needs: CropXR.¹ CropXR is a Dutch institute dedicated to advancing resilient crop research through innovative technologies, and this chapter demonstrates how theoretical insights can be translated into actionable recommendations to improve metadata practices.

The chapter is structured into three sections: Section 7.1 provides an overview of the organisational structure and the CropXR data management plan. Section 7.2 explores how CropXR's initiatives address the metadata gaps identified earlier in this thesis (Section 5.1), emphasising their practical impact. Finally, Section 7.3 offers actionable recommendations tailored to CropXR, combining the analysis from previous sections with the findings of this thesis. By connecting research insights to practical applications, this chapter serves as both a case study and a guide for addressing metadata challenges in interdisciplinary and collaborative projects like CropXR.

7.1. The data management landscape of CropXR

CropXR is a Dutch institute that integrates plant biology, computational modelling, and artificial intelligence to develop resilient, sustainable, and climate-adaptive crops. Its interdisciplinary approach unites universities, research institutions, and industry partners to tackle global challenges such as food security and climate change. Operating under a 10-year roadmap, CropXR generates vast amounts of complex data that must remain accessible and relevant to diverse stakeholders over time. This section provides an overview of the organisational structure of CropXR and its current and planned data management strategies. By examining these fundamental elements, this section lays the foundation for understanding how the findings of this thesis can address the challenges of CropXR metadata and support its long-term research goals.

7.1.1. Project overview

The institute's organisational structure brings together domain expertise and facilitates communication and transfer of knowledge across CropXR. It is divided into five key components: PlantXR, focused on crop resilience research; DataXR, responsible for developing robust data infrastructure; AgroXR, exploring agricultural applications; EduXR, focussing on education and workforce development; and TransferXR, which ensures efficient knowledge sharing with industry partners. Overseeing these initiatives is the CropXR central office, which manages coordination and ensures alignment with the project's overarching vision.

The project is designed to span a 10-year timeline, divided into two distinct phases. During the first five years, efforts have focused on research and data production, with researchers generating

¹Information about the CropXR project presented in this thesis is based on the author's attendance at project meet-ups and review of materials shared on the project's private SharePoint, some of this information is also available on the project website (<https://cropxr.org>). Although this thesis is not formally affiliated with the project, the access provided allowed the inclusion of this case study to support the research.

datasets and conducting scientific studies. Simultaneously, the necessary data infrastructure will be established to support data storage, accessibility, and sharing for the remainder of the project. In the last five years, the focus will shift toward practical applications. Plant breeding companies will test and implement the research findings to validate their real-world applicability. These companies will bring their own metadata requirements, which could differ from those of academic researchers.

A major challenge for CropXR is that many of the researchers involved are used to working on a much smaller scale. Historically, for many, their data management needs have been simple and could be handled using spreadsheet editors. However, CropXR's plans will require them to transition to working with vastly larger datasets, where proper database systems and management practices are essential. This leap in scale represents more than just a technical change: it demands new skills, workflows, and mindsets, marking a significant departure from their previous experience. Highlighting the need for targeted support and capacity building to ensure success in this new large-scale environment.

Given the complexity of its stakeholder landscape, this thesis groups CropXR participants into three archetypes: senior academic researchers, junior academic researchers, and industry professionals. For each archetype, we focus on the people involved in generating and using data within the PlantXR initiative, rather than the people primarily responsible for setting up the data management infrastructure. We assume that these participants are engaged in research and have a background in plant sciences. These archetypes simplify the analysis of the project data management needs. They are broad generalisations rather than exhaustive definitions.

- **Senior academic researchers:** Senior researchers are experienced academics who prioritise the advancement of scientific knowledge through research. Their primary incentives include publishing their findings in peer-reviewed journals, contributing to the scientific community, and improving the reputation of their institution. These stakeholders are typically familiar with data publishing and reuse within (domain-specific) data repositories. They value open dissemination of research results.
- **Junior academic researchers:** Junior researchers include early career PhD candidates, university students, and students from universities of applied sciences, all with a focus on plant sciences. Under the mentorship of senior researchers, they focus on skill development and addressing research questions that align with academic and industry objectives. Although they have domain knowledge, their experience with formal data management is often limited, as this is not typically emphasised during their studies. This group exhibits significant variation in expertise, reflecting differences in educational backgrounds and research roles.
- **Industry professionals:** Industry professionals represent plant breeding companies involved in applying CropXR's research to develop resilient crop varieties. Their priority lies in translating scientific findings into commercially viable and sustainable products. These stakeholders are more accustomed to internal data management systems and may be less familiar with academic repositories. Some industry professionals have previous academic experience that can bridge these gaps. Their focus on practical outcomes requires effective collaboration and streamlined access to actionable data.

CropXR faces significant data management challenges, primarily arising from the intersection of various stakeholders and extensive interdisciplinary data requirements. The project must manage large volumes of complex data generated over an extended period that span multiple scientific domains while ensuring that these findings are later translated into practical applications. This effort must accommodate stakeholders with varying needs and priorities, while fostering collaboration and innovation across academic and industry boundaries. By addressing these challenges, CropXR aims to establish a sustainable and collaborative framework to advance crop resilience.

7.1.2. Current data management plan

Effective data management is at the heart of CropXR's efforts to advance crop resilience. The project has already established several foundational tools and processes to handle the volume and complexity of data generated by its research activities.

Currently, during this development phase, CropXR utilises the SURF Research Drive for data storage. The drive functions as a data lake, providing a space to store data before further processing and serving as a testing ground for various data management approaches. A pilot phase is in progress

to refine workflows for data storage, metadata annotation, and validation, establishing a foundation for the project's long-term data management strategy.

To standardise data practices across the project, CropXR has established a code of conduct that outlines ethical and technical principles for data management. These guidelines emphasise consistency and usability in data storage, sharing, and annotation, ensuring that the data remains valuable for current and future users. A dedicated team, the Standards and Metadata (SAME) group, is developing metadata frameworks for two key data types, phenotyping and sequencing data, which currently lack widely accepted standards. Their focus is on "hard metadata," including structured technical details such as sample identifiers and measurement units, while laying the foundation for integrating contextual information in the future.

Collaboration remains a cornerstone of CropXR's operations. Regular biannual conferences and smaller component-specific meet-ups foster dialogue and ensure alignment across teams. Slack is used for communication across the entire project and within smaller sub-groups, while SharePoint serves as a repository for project documents, including conference presentations and procedural guidelines.

The project is guided by a 10-year roadmap organised into work packages that outline key milestones and deliverables. CropXR's central office ensures coordination and oversight, aligning efforts between stakeholders, and maintaining a unified vision for the project.

CropXR also includes two critical initiatives to ensure long-term success. First, EduXR collaborates with universities to engage the next generation of researchers, creating a closer connection to future scientists in plant science. By working with these institutions, EduXR gains insight into the needs and priorities of students, who represent a key group of future contributors to the project. Meanwhile, DataXR focusses on developing the project data infrastructure and serves as a knowledge hub for all data management related questions. The DataXR team actively engages future system users in its development, ensuring usability and alignment with stakeholder needs.

In summary, CropXR has established a solid foundation for data management through tools such as the SURF drive, the metadata standards of the SAME group, and collaborative initiatives such as EduXR and DataXR. The project also benefits from a centralised organisation with a clear vision, regular meet-ups, and a mostly centralised communication platform in the form of Slack, all of which foster personal connections and enhance collaboration. The next section will explore CropXR's vision for the future of its data management systems.

7.1.3. The CropXR vision

The vision of CropXR is focused on creating a robust and sustainable data infrastructure to advance crop resilience, with the Resilient Hub at its core. This comprehensive repository is designed to house datasets, experimental protocols, and metadata, making it one of the world's largest resources on crop resilience. The hub aims to ensure the accessibility, usability, and long-term relevance of data for current and future stakeholders.

By integrating datasets from various sources and harmonising them through consistent metadata practices developed by the SAME group, the Resilient Hub aligns contributions from CropXR components. This integration facilitates interdisciplinary research and practical applications, bridging the gap between academic research and industry needs while fostering innovation in resilient crop development. Furthermore, the hub is designed to remain a valuable resource beyond the project ten-year timeline, ensuring that its impact continues to support future research efforts.

Complementing the Resilient Hub is the Meta Buddy initiative, an AI-driven tool inspired by the concept of context-bridging data conversation explored in this thesis. Meta Buddy is envisioned as a chatbot that guides researchers through the metadata collection process. Initially, it will focus on collecting "hard metadata" (e.g., sample identifiers, measurement units) as determined by the SAME group through a text-based interface. Future iterations may incorporate voice-to-text capabilities for enhanced accessibility.

Meta Buddy will draw on existing datasets stored in the SURF drive and documentation hosted on SharePoint to ensure alignment with both project-specific and broader metadata needs. Meta Buddy is still in the pre-development stage, with a proof of concept expected soon.²

²CropXR conference, October 2024: <https://cropxr.org/grand-success-the-first-cropxr-conference/>

In conclusion, CropXR has a vision of creating a robust and sustainable data infrastructure, which they are already working toward balancing current practical tools, such as the SURF drive and metadata guidelines, with ambitious future innovations such as the Resilient Hub and Meta Buddy. By advancing these tools, CropXR positions itself as a leader in agricultural research and data management innovation. However, significant challenges remain. The diversity of stakeholders introduces varied requirements and expectations, making it challenging to design tools and workflows that meet all needs. Furthermore, the sheer volume and complexity of the data, covering multiple domains and use cases, further complicates the development of a cohesive and efficient management system. The next section explores how CropXR's current approach and future visions address the metadata gap.

7.2. Mapping the CropXR metadata gap

CropXR employs a range of initiatives to tackle the metadata gap, focussing on practical strategies and mechanisms to address specific challenges. In the following, I detail how each initiative contributes to resolving these gaps, focussing on the processes and methods involved. See Appendix E for a table indicating which parts of the metadata gap are addressed by which CropXR initiatives.

7.2.1. Connection through the centrally managed network

CropXR strengthens interpersonal and professional networks through recurring meet-ups, the use of Slack as a collaboration tool, and centralised project management. Meet-ups provide structured, yet informal environments for researchers to share challenges, brainstorm solutions, and exchange best practices, fostering communication and collaboration (S/M1). Although meeting notes are often published on SharePoint, some interactions remain unrecorded and are not formally captured.

Slack serves as an asynchronous platform for discussion and collaboration, fostering continuous engagement and grass-roots participation (S/M2). Most Slack activity occurs in closed channels, but CropXR indicates that it is well used. Similarly, while SharePoint hosts agendas and outcomes from in-person meetups, the extent to which these documents are accessed or used beyond the directly involved teams appears to be limited.

Centralised management ensures that individual contributions align with broader project objectives. This structure provides visibility into metadata workflows, encouraging peer recognition (S/M3) and exposing common metadata needs and frustrations by including data users in infrastructure development (S/M4). By facilitating alignment and maintaining organised connections, centralised management also supports ongoing commitment by keeping participants engaged and on track with the project objectives (S/L4). These insights lead to actionable solutions, such as identifying divergent use cases (D/C3) and facilitating contact between producers and consumers (S/L3, D/C4). By consistently fostering these interactions and effectively managing networking connections, CropXR cultivates a collaborative ecosystem that enhances data quality (S/M6) and strengthens research credibility (S/M5).

7.2.2. Knowledge fostering and dissemination through EduXR and DataXR

EduXR and DataXR focus on fostering and disseminating domain and data management knowledge essential for both current and future CropXR researchers. EduXR emphasises the preparation of the next generation of crop researchers by collaborating with universities to ensure future participants gain the skills required for CropXR projects (D/L1). By engaging educational institutions, EduXR helps design systems that cater to the needs of future researchers (S/L3).

DataXR serves as a hub of expertise for data management, providing support to researchers at all levels (S/M7, S/L1, and D/L2). It addresses common constraints in data management, such as limited familiarity with best practices, and ensures alignment between project needs and researcher workflows. Critical initiatives, including the SAME group, Resilient Hub, and Meta Buddy, fall under DataXR's purview and will be detailed in subsequent sections due to their importance.

7.2.3. Metadata standards by the SAME group

The SAME group creates metadata standards tailored to CropXR's key data types, such as phenotyping and sequencing data, which currently lack widely accepted norms. Since the standards are developed within the organisation, adapting them to future needs will be more straightforward (S/C4).

SAME standards reduce subjectivity in metadata creation (S/L2) and integrate data consumer requirements, removing individual producers of the burden of addressing these needs independently

(S/L3). Additionally, these standards help address challenges related to data complexity (S/C3), further easing the workload for producers. By establishing and promoting these standards, the SAME group directly addresses incomplete metadata (D/C2), unreported data attributes (D/C5), and inconsistencies in metadata standards (D/C6), creating a unified framework for interoperability and completeness.

7.2.4. A central long-term data hub through the Resilient Hub

The Resilient Hub addresses the challenges of metadata storage and accessibility by serving as a central repository for all CropXR data, including all types of metadata. It mitigates space constraints (S/C2) by allowing CropXR to control what data are retained and prioritised. The hub addresses the challenge of ongoing commitment (S/L4) and consolidates scattered metadata (D/C1) by providing a centralized, easily accessible system that the entire organization can collectively manage, ensuring metadata remains relevant, usable, and sustainable beyond the project's ten-year timeline.

7.2.5. Enhanced metadata elicitation and communication via Meta Buddy

Meta Buddy simplifies metadata creation through an interactive interface that guides researchers step-by-step. If the tool can learn from existing documentation and datasets, it could automate many of the advantages the SAME, EduXR, and DataXR groups bring. It would have the ability to explain and adapt metadata templates while also serving as a teaching tool to disseminate domain and data management knowledge. By automating these processes, Meta Buddy has the potential to save significant time (S/C1).

In summary, CropXR is effectively addressing many aspects of the metadata gap through targeted initiatives. Efforts such as meet-ups, Slack, and centralised management improve communication, grass-roots participation, and peer recognition. The SAME group advances metadata standardisation, reducing subjectivity, and aligning with consumer needs. EduXR and DataXR strengthen data management skills and bridge knowledge gaps, ensuring that future and current researchers are equipped to handle complex data. The Resilient Hub provides a scalable and long-term solution for metadata storage and accessibility, addressing issues of commitment and space constraints. Finally, Meta Buddy streamlines metadata creation and fosters better communication between data producers and consumers.

The next section will explore how the Meta Buddy concept can be refined and enhanced, drawing on the findings of this thesis. These recommendations aim to address the remaining gaps and maximise the impact of CropXR on metadata management and crop resilience research.

7.3. Recommendations for Meta Buddy development

Effective metadata management is critical to advance resilient crop research at CropXR. This thesis has demonstrated that metadata challenges are not only technical, but also social, requiring innovative solutions that foster collaboration, improve accessibility, and encourage continuous improvement. The Meta Buddy prototype presents a perfect opportunity to address these challenges by integrating insights gained from research findings into its design. These recommendations are grounded in metadata gap analysis (Section 5.1), context-bridging data conversation mechanisms tested in this study (Section 5.2), and the practical challenges of real-world implementation (Chapter 6).

Although these recommendations are based on the findings of this thesis, they have not yet been tested in a real-world environment at scale. Each recommendation outlines potential features for CropXR to test in their Meta Buddy and Resilient hub prototypes, aiming to establish a metadata management system that improves data reusability for users of all experience levels, unlocking the full potential of the Resilient Hub. For readers interested in exploring the underlying insights and challenges in greater depth, the detailed findings presented in this thesis provide additional context and support for these recommendations.

7.3.1. Recognise motivational needs of metadata producers

Motivating metadata producers is critical to maintaining high-quality metadata creation, as discussed in Section 5.1.1. Significant barriers, such as resource constraints and the uncertainty or complexity of determining what metadata is needed for reusability, often discourage producers from prioritising metadata tasks. These challenges can undermine motivation, particularly when producers feel their

efforts go unrecognised or are disconnected from the broader research community. However, fostering a sense of relatedness and helping producers understand the value of their work and its impact on others can greatly enhance their motivation. Additionally, increasing producers' feelings of competence in research skills and ensuring that they retain a sense of autonomy further support their engagement.

Meta Buddy should address both sides of this issue: reducing barriers by providing clear incentives and support, while improving intrinsic motivation through mechanisms that emphasise recognition, competence, and autonomy. Features that highlight the impact of producers' efforts and acknowledge their contributions can strengthen their sense of purpose and connection to the research community.

- Include metrics within Meta Buddy to show how producer contributions improve dataset reuse (e.g., citation counts or user feedback).
- Offer public acknowledgement, such as a “Contributors” feature or recognition emails, to celebrate producers' efforts.
- Create opportunities for peer-to-peer feedback to build a stronger sense of community.

7.3.2. Involve data consumers as metadata co-creators

Metadata creation has traditionally been the responsibility of data producers, often leading to gaps between the metadata provided and the needs of consumers. Section 5.2.1 emphasises that involving consumers as co-creators can enrich metadata by incorporating diverse perspectives and fostering collaboration. Consumers provide unique insights, particularly when they interact with datasets in ways that producers may not have anticipated. For CropXR, involving consumers in metadata creation could lead to more robust and adaptive metadata that supports both routine and novel use cases. This approach also encourages shared ownership of metadata, fostering a sense of community between producers and consumers.

However, involving consumers as co-creators also introduces social complexities that must be carefully managed. Section 6.2.3 highlights challenges such as inclusivity, communication barriers, and conflict resolution, which can arise during collaborative metadata creation. These dynamics are particularly important in interdisciplinary environments like CropXR, where diverse stakeholders bring varying levels of expertise and expectations. Meta Buddy must support this collaboration by providing tools and guidelines that ensure the metadata co-creation process is equitable, inclusive, and productive.

- Enable consumers to document their experiences directly within Meta Buddy using guided prompts or forms.
- Introduce peer-to-peer communication features to support collaboration, especially between beginner data consumers.
- Provide feedback loops so that consumers' contributions visibly improve metadata quality, creating a shared sense of purpose.
- Provide guidelines for conflict resolution during metadata co-creation processes.
- Foster inclusivity by designing features that encourage contributions from all users, regardless of the level of experience.

7.3.3. Capture and centralise (contextual) metadata

Contextual metadata—such as the rationale behind decisions, challenges encountered, and conventions followed, play a crucial role in making datasets reusable; see Section 5.2.2 for more details. Without centralised metadata, users face significant additional effort to reuse datasets, as metadata often becomes scattered across articles, repositories, informal interactions, or may not be collected at all. For CropXR, where data must remain relevant over time and serve various types of user, establishing a central location for contextual metadata is essential for long-term, efficient data reusability.

The Resilient Hub should act as the centralised repository for all metadata, with Meta Buddy serving as the primary tool for capturing and organising this information. This single source-of-truth approach would make metadata easier to access, interpret, and adapt to diverse use cases. Metadata from communication platforms (e.g., Slack), document repositories (e.g., SharePoint), and article publications should be automatically reviewed, with relevant information extracted and consolidated into the Resilient Hub. Particular attention should be paid to documenting data conventions and challenges, as

these were identified as critical points for junior researchers. Using an adaptable metadata elicitation tool like Meta Buddy will further enable CropXR to identify gaps, determine user needs, and ensure metadata completeness.

The OpenMetadata open source project offers features that CropXR could use as a reference, such as automated metadata extraction, customisable schemas for different data types, and tools for data producers and consumers to communicate in relation to a dataset. Especially Collate, an AI-driven data management solution extending upon the open source OpenMetadata project, bears many similar features as discussed here and by CropXR itself for the Resilient Hub. These functionalities can guide the development of the Resilient Hub and ensure that it meets the diverse needs of its users.

- The Resilient Hub should consolidate, preferably automatically, all the metadata currently scattered in various locations, such as shared drives, articles, or meeting notes.
- Provide clear guidelines for linking metadata across systems when full consolidation is not feasible.
- Enable Meta-Buddy to document and communicate the following:
 - Data conventions and domain-specific knowledge.
 - Challenges and paths not taken during data collection and analysis.

7.3.4. Plan for metadata iteration over time

Metadata is not static; it evolves over time as new consumer needs, research standards, and data complexities emerge. Meta Buddy should support ongoing updates, allowing producers and consumers to collaborate on improving metadata beyond the initial creation phase. This ensures that metadata remains relevant and adapts to changing requirements.

- Implement features that track and suggest updates to metadata based on consumer feedback or changes in standards.
- Enable collaborative tools within Meta Buddy for producers and consumers to refine and expand metadata iteratively.
- Provide notifications or reminders for producers to revisit and update metadata when appropriate.

7.3.5. Accommodate diverse use cases and users

Metadata systems must meet the needs of a diverse range of use cases and users, both of which present distinct but interconnected challenges. Diverse use cases arise when metadata must support novel or interdisciplinary research applications, often extending beyond the producer's original intent. In contrast, diverse users include individuals with varying levels of expertise, from seasoned researchers to beginners, who require different types of guidance and support to navigate datasets effectively. Section 5.1.3 explores these barriers.

Meta Buddy can address these challenges by tailoring metadata workflows and prompts to account for the diversity of both use cases and users. For diverse use cases, it is crucial to collect more detailed information about the dataset's intended and unintended uses, along with the rationale behind these assessments. Producers should be encouraged to reflect on what the dataset is and is not useful for, while consumer feedback about their experiences and attempts to use the dataset can help future users as well. Additionally, recording contextual data, such as failed paths and not taken roads, provides invaluable information for novel and interdisciplinary applications.

For diverse users, Meta Buddy must provide tools and explanations tailored to their experience levels. Beginners benefit from simplified guidance and explanations, while advanced researchers may require detailed prompts and access to raw data alongside more processed versions. Including a centralised space in the Resilient Hub for documenting domain-specific conventions, terminology, and methodologies will help bridge knowledge gaps and ensure metadata is accessible and adaptable to a wide range of research activities.

- Include Meta Buddy prompts for producers to record the intended and unintended uses of the data set, as well as contextual data such as failed paths and not taken roads.
- Create a centralised space within the Resilient Hub to document domain conventions, terminology, and methodologies, ensuring accessibility for all users.

- Meta Buddy should be designed to answer questions from research novices, ensuring that the metadata extracted achieves a level of detail similar to that provided by experienced producers.
- Incorporate features that allow users to navigate complex datasets with varying levels of granularity, ensuring accessibility across experience levels.
- Use feedback from users to continuously refine metadata prompts and workflows, incorporating adaptive questioning mechanisms that dynamically adjust metadata collection processes based on user responses and experience levels.

7.3.6. Use real-time dialogue

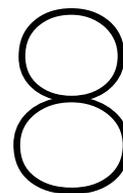
Dialogue is a powerful mechanism for enriching metadata creation, offering a dynamic and collaborative alternative to traditional written methods. Section 5.2.3 highlights that real-time dialogue facilitates immediate clarification, captures nuanced details, and fosters deeper engagement between producers and consumers. For CropXR, where collaboration is key to addressing complex challenges in crop research, integrating dialogue-based metadata tools can significantly enhance the Meta Buddy prototype.

Currently, CropXR plans to focus on chat-based solutions as a starting point for real-time dialogue. Although chat offers scalability and ease of integration, we expect that real-human interaction will often be more effective for capturing nuanced insights and implicit knowledge. Such interactions could also serve as a teaching opportunity for junior researchers, allowing them to develop skills in metadata management and collaboration while contributing to the metadata creation process.

Real-time dialogue can address gaps in metadata creation by allowing producers to verbally explain their decisions, challenges, and rationale in greater depth. This process is particularly valuable for capturing implicit knowledge and context that may not be fully conveyed through written documentation. Additionally, dialogue encourages trust and collaboration among team members, creating a stronger sense of shared responsibility for metadata quality. To ensure effectiveness, the system should incorporate both AI-driven tools and human oversight to balance automation with quality control.

- Pilot dialogue-based metadata creation by assigning junior researchers to act as data stewards, facilitating conversations with producers.
- Integrate AI-driven dialogue tools to streamline metadata collection, ensuring that they include a human-in-the-loop approach for validation and refinement.

The recommendations presented in this section provide a roadmap for addressing CropXR's metadata challenges through the development of the Meta Buddy and Resilient Hub prototypes. By recognising and mitigating barriers to metadata creation, fostering collaboration between producers and consumers, centralising and enhancing metadata accessibility, and embracing iterative processes, these tools can ensure that metadata meet the diverse needs of CropXR stakeholders. Furthermore, features such as real-time dialogue and adaptive workflows create opportunities for dynamic interactions and knowledge sharing, particularly for junior researchers and interdisciplinary teams. Although CropXR's initial focus on chat-based solutions provides a scalable starting point, integrating human interaction could unlock further potential as both a metadata enrichment and teaching tool. These recommendations emphasise the importance of testing and refining these features in real-world settings, allowing CropXR to build a robust and inclusive metadata system that fully supports its mission of advancing resilient crop research.



Summary

Data reusability is fundamental to scientific progress, as each new discovery builds upon prior research. Enhancing data reusability streamlines this process, reducing redundancy and accelerating innovation. To ensure that research data can be reused efficiently, it must be accompanied by information that makes it accessible and comprehensible to researchers of diverse disciplines and levels of expertise. Metadata serves as the medium through which data producers communicate the context necessary for efficient reuse, transferring knowledge from producers to consumers. It acts as a bridge, enabling efficient data reuse. However, existing metadata practices often do not fully meet consumer needs, creating a gap between the metadata supplied and the metadata demanded. This gap results in inefficient reuse, requiring additional time and effort from the consumer, or, in some cases, rendering reuse entirely unfeasible. This thesis investigates this "metadata gap," analysing the challenges faced by producers and consumers, and proposing a novel solution: context-bridging data conversations.

The metadata gap becomes apparent when the metadata supplied by producers do not align with the needs of consumers. This study examined the internal motivations driving producers to provide metadata for reusability, as well as the barriers they face in doing so. Producers are incentivised to prioritise their own research goals, with limited motivation to invest in metadata that enhance long-term reusability, as such efforts offer little immediate benefit. Additionally, data producers are constrained by time, resource limitations, and a lack of understanding of consumer requirements. On the demand side, these challenges result in incomplete or inconsistent metadata, making it difficult and time consuming for consumers to interpret and reuse datasets, particularly in interdisciplinary or unfamiliar contexts.

This thesis addresses the metadata gap by introducing context-bridging data conversations. Designed to simulate a dialogue with a data steward, these conversations with data producers and consumers explore the metadata they create, reference or require during dataset production and reuse. They bridge the metadata gap by integrating consumer needs, capturing detailed contextual information, fostering real-time dialogue about metadata, and continuously refining the process. Four mechanisms were evaluated to achieve these goals and address the metadata gap, leading to the following feature recommendations for a data management system that wants to bridge the metadata gap.

- **Involve consumers as metadata co-creators.** Integrating consumers into (meta)data management provides greater insight into the demand for metadata.
- **Recognise and incorporate contextual metadata.** Capturing contextual information, such as decision rationales and domain conventions, improves the clarity and interpretability of the data for data consumers of all levels of knowledge.
- **Use real-time dialogue.** Facilitating live interactions in the metadata elicitation process helps resolve ambiguities and bridge knowledge gaps immediately and efficiently.
- **Continuously adapt data conversations questions.** Refining questions throughout the process ensure that the collected metadata remain relevant and the process stays efficient.

The CropXR institute provides a practical example of applying these mechanisms in a real-world research environment. Its interdisciplinary nature highlights the complexities of metadata creation and the importance of collaborative and adaptive practices. By analysing the institute's current initiatives

and mapping their impact on the metadata gap, it becomes evident where improvements are needed and which recommendations would most benefit the project. These findings demonstrate how context-bridging data conversations can generate actionable insights and improve metadata practices in similar research settings.

Beyond its practical implications, this thesis advances theoretical understanding by the dynamic, perspective-dependent needs of metadata for reusability. It emphasises the social dimensions of metadata management, framing it as a process of communication between researchers. The work underscores the discrepancy between the metadata recorded by producers, including associated published articles, and the information actually used by consumers during data reuse. By reframing metadata management as a communication challenge between humans, rather than merely a technical issue, this thesis extends existing research on metadata practices.

Future research should focus on developing context-bridging data conversations from proof-of-concept to scalable prototypes integrated into real-world metadata management systems. Testing these prototypes in small-scale environments with diverse producer-consumer pairs will provide valuable insights into their impact and technical feasibility. Emerging technologies, such as AI-driven metadata prioritisation, should also be explored to address challenges like managing complex metadata and balancing detail with usability. By refining these approaches and tailoring them to specific research contexts, future work can further enhance metadata practices, fostering more effective and inclusive data reuse.

This thesis positions metadata creation as a collaborative, context-aware, and dynamic process. By bridging the gap between metadata supply and demand, it provides a framework to enhance the efficiency of data reuse and broaden data accessibility. Context-bridging data conversations address existing challenges and pave the way for future advances, contributing to open and reproducible science that benefits researchers in all domains and levels of experience.

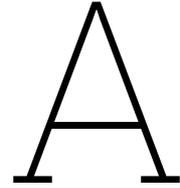
References

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, “The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, no. 1, pp. 1–9, 1 Mar. 15, 2016. DOI: 10.1038/sdata.2016.18.
- [2] T. J. Hostler, “The invisible workload of open research,” *Journal of Trial & Error*, 2023. DOI: 10.36850/mr5.
- [3] P. N. Edwards, M. S. Mayernik, A. L. Batcheller, G. C. Bowker, and C. L. Borgman, “Science friction: Data, metadata, and collaboration,” *Social Studies of Science*, vol. 41, no. 5, Aug. 15, 2011. DOI: 10.1177/0306312711413314.
- [4] F. Urbano, F. Cagnacci, and, “Data Management and Sharing for Collaborative Science: Lessons Learnt From the Euromammals Initiative,” *Frontiers in Ecology and Evolution*, vol. 9:727023, 2021. DOI: 10.3389/fevo.2021.727023.
- [5] M. Koho, T. Burrows, E. Hyvönen, *et al.*, “Harmonizing and publishing heterogeneous premodern manuscript metadata as Linked Open Data,” *Journal of the Association for Information Science and Technology*, vol. 73, no. 2, pp. 240–257, 2022. DOI: 10.1002/asi.24499.
- [6] S. Kanza, N. Gibbins, and J. G. Frey, “Too many tags spoil the metadata: Investigating the knowledge management of scientific research with semantic web technologies,” *Journal of Cheminformatics*, vol. 11, pp. 1–23, Mar. 21, 2019. DOI: 10.1186/s13321-019-0345-8.
- [7] G. Alemu, B. Stevens, and P. Ross, “Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries: A social constructivist approach,” *New Library World*, vol. 113, no. 1/2, pp. 38–54, 2012. DOI: 10.1108/03074801211199031.
- [8] J. Grewe, T. Wachtler, and J. Benda, “A bottom-up approach to data annotation in neurophysiology,” *Frontiers in neuroinformatics*, vol. 5:16, 2011. DOI: 10.3389/fninf.2011.00016.
- [9] Wiley. “Researchers on open access practices,” Wiley. (Aug. 2024), [Online]. Available: <https://www.wiley.com/en-us/network/publishing/research-publishing/open-access/researchers-on-open-access-practices-2024> (visited on 01/28/2025).
- [10] M. Boeckhout, G. Zielhuis, and A. Bredenoord, “The FAIR guiding principles for data stewardship: Fair enough?” *European Journal of Human Genetics*, vol. 26, no. 7, pp. 931–936, 2018. DOI: 10.1038/s41431-018-0160-0.
- [11] C. J. Markiewicz, K. J. Gorgolewski, F. Feingold, *et al.*, “The OpenNeuro resource for sharing of neuroscience data,” *Elife*, vol. 10, e71774, 2021. DOI: 10.7554/eLife.71774.
- [12] D. M. Leigh, A. G. Vandergast, M. E. Hunter, *et al.*, “Best practices for genetic and genomic data archiving,” *Nature Ecology & Evolution*, pp. 1–9, May 24, 2024. DOI: 10.1038/s41559-024-02423-7.
- [13] H. A. Piwowar, “Who Shares? Who Doesn’t? Factors Associated with Openly Archiving Raw Research Data,” *PLoS ONE*, vol. 6, no. 7, C. Neylon, Ed., e18657, Jul. 13, 2011. DOI: 10.1371/journal.pone.0018657.
- [14] R. A. Hackett, M. W. Belitz, E. E. Gilbert, and A. K. Monfils, “A data management workflow of biodiversity data from the field to data users,” *Applications in Plant Sciences*, vol. 7, no. 12, e11310, Dec. 2019. DOI: 10.1002/aps3.11310.
- [15] K. Broman, M. Cetinkaya-Rundel, A. Nussbaum, *et al.*, “Recommendations to funding agencies for supporting reproducible research,” in *American Statistical Association*, vol. 2, 2017, pp. 1–4.
- [16] J. Furner, “Definitions of “Metadata”: A Brief Survey of International Standards,” *Journal of the Association for Information Science and Technology*, vol. 71, no. 6, E33–E42, 2020. DOI: 10.1002/asi.24295.

- [17] J. Pomerantz, *Metadata*. Cambridge, MA: MIT Press, 2015.
- [18] A. J. Gilliland, "Setting the stage," in *Introduction to Metadata*, M. Baca, Ed., Third edition, Los Angeles: Getty Research Institute, 2016, pp. 1–19.
- [19] M. L. Zeng and J. Qin, *Metadata*, 2nd edition. Chicago: Neal-Schuman, 2016.
- [20] J. Riley, "Understanding metadata: What is metadata, and what is it for," *Washington DC, United States: National Information Standards Organization*, vol. 23, pp. 7–10, 2017.
- [21] I. M. Faniel and E. Yakel, "Significant Properties as Contextual Metadata," *Journal of Library Metadata*, vol. 11, no. 3–4, pp. 155–165, Jul. 1, 2011. DOI: 10.1080/19386389.2011.629959.
- [22] H. Ulrich, A.-K. Kock-Schoppenhauer, N. Deppenwiese, *et al.*, "Understanding the Nature of Metadata: Systematic Review," *Journal of Medical Internet Research*, vol. 24, no. 1, e25440, Jan. 11, 2022. DOI: 10.2196/25440.
- [23] E. G. Carayannis, E. Grigoroudis, M. Del Giudice, M. R. Della Peruta, and S. Sindakis, "An exploration of contemporary organizational artifacts and routines in a sustainable excellence context," *Journal of Knowledge Management*, vol. 21, no. 1, pp. 35–56, Jan. 1, 2017. DOI: 10.1108/JKM-10-2015-0366.
- [24] M. S. Feldman and B. T. Pentland, "Reconceptualizing Organizational Routines as a Source of Flexibility and Change," *Administrative Science Quarterly*, vol. 48, no. 1, pp. 94–118, Mar. 1, 2003. DOI: 10.2307/3556620.
- [25] A. Wibisono, D. Sammon, and C. Heavin, "Plausible pictures for data governance: A narrative network approach," Association for Information Systems, May 2022.
- [26] C. L. Borgman and P. T. Groth. "From Data Creator to Data Reuser: Distance Matters." arXiv: 2402.07926 [cs]. (Aug. 28, 2024), pre-published.
- [27] D. Plotkin, *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*, Second edition. London: Academic Press, 2021.
- [28] K. M. Hüner, B. Otto, and H. Österle, "Collaborative management of business metadata," *International Journal of Information Management*, vol. 31, no. 4, pp. 366–373, Aug. 1, 2011. DOI: 10.1016/j.ijinfomgt.2010.12.002.
- [29] S. Bateman, C. Brooks, and G. McCalla, "Collaborative tagging approaches for ontological metadata in adaptive e-learning systems," in *Proceedings of the Fourth International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL 2006)*, 2006, pp. 3–12.
- [30] H. Hata, N. Novielli, S. Baltés, R. G. Kula, and C. Treude, "GitHub Discussions: An exploratory study of early adoption," *Empirical Software Engineering*, vol. 27:3, pp. 1–32, Jan. 2022. DOI: 10.1007/s10664-021-10058-6.
- [31] C. Clare, M. Cruz, E. Papadopoulou, J. Savage, M. Teperek, and Y. Wang, *Engaging Researchers with Data Management: The Cookbook*. Open Book Publishers, Oct. 9, 2019. DOI: 10.11647/obp.0185.
- [32] F. Edwards, B. Cowie, and S. Trask, "Using colleague coaching to develop teacher data literacy," *Professional Development in Education*, pp. 1–14, 2022. DOI: 10.1080/19415257.2022.2081247.
- [33] L. Koesten, E. Kacprzak, J. Tennison, and E. Simperl, "Collaborative Practices with Structured Data: Do Tools Support What Users Need?" In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19, New York, NY, USA: Association for Computing Machinery, May 2, 2019, pp. 1–14. DOI: 10.1145/3290605.3300330.
- [34] K. B. Read, A. Surkis, C. Larson, *et al.*, "Starting the data conversation: Informing data services at an academic health sciences library," *Journal of the Medical Library Association : JMLA*, vol. 103, no. 3, pp. 131–135, Jul. 2015. DOI: 10.3163/1536-5050.103.3.005. PMID: 26213504.
- [35] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," *American Psychologist*, vol. 55, no. 1, pp. 68–78, 2000. DOI: 10.1037/0003-066X.55.1.68.
- [36] S. Kvale, *Interviews: Learning the craft of qualitative research interviewing*. Sage, 2009.

- [37] V. Mavrych, A. Yaqinuddin, and O. Bolgova, "Claude, ChatGPT, Copilot, and Gemini Performance versus Students in Different Topics of Neuroscience," *Advances in Physiology Education*, Jan. 17, 2025. DOI: 10.1152/advan.00093.2024.
- [38] K. Cho, Y. Park, J. Kim, B. Kim, and D. Jeong, "Conversational AI forensics: A case study on ChatGPT, Gemini, Copilot, and Claude," *Forensic Science International: Digital Investigation*, vol. 52:301855, 2025. DOI: 10.1016/j.fsidi.2024.301855.
- [39] A. Sobo, A. Mubarak, A. Baimagambetov, and N. Polatidis, "Evaluating LLMs for Code Generation in HRI: A Comparative Study of ChatGPT, Gemini, and Claude," *Applied Artificial Intelligence*, vol. 39:2439610, Dec. 31, 2025. DOI: 10.1080/08839514.2024.2439610.
- [40] F. Kirstein, J. P. Wahle, B. Gipp, and T. Ruas. "CADS: A Systematic Literature Review on the Challenges of Abstractive Dialogue Summarization." arXiv: 2406.07494 [cs]. (Jun. 12, 2024), pre-published.
- [41] Z. W. Taylor, "Using Chat GPT to clean interview transcriptions: A usability and feasibility analysis," *Available at SSRN 4437272*, 2023.
- [42] G. Wang, Z. Sun, Z. Gong, *et al.* "Do Advanced Language Models Eliminate the Need for Prompt Engineering in Software Engineering?" arXiv: 2411.02093 [cs]. (Nov. 4, 2024), pre-published.
- [43] OpenAI, *Temporary chat faq*, 2024. [Online]. Available: <https://help.openai.com/en/articles/8914046-temporary-chat-faq> (visited on 11/11/2024).
- [44] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, 1932.
- [45] S. Lee, "When Tensions Become Opportunities: Managing Accountability Demands in Collaborative Governance," *Journal of Public Administration Research and Theory*, vol. 32, no. 4, pp. 641–655, Oct. 1, 2022. DOI: 10.1093/jopart/muab051.
- [46] R. C. Amorim, J. A. Castro, J. Rocha Da Silva, and C. Ribeiro, "A comparison of research data management platforms: Architecture, flexible metadata and interoperability," *Universal Access in the Information Society*, vol. 16, no. 4, pp. 851–862, Nov. 2017. DOI: 10.1007/s10209-016-0475-y.
- [47] I. Aubin, F. Cardou, L. Boisvert-Marsh, E. Garnier, M. Strukelj, and A. D. Munson, "Managing data locally to answer questions globally: The role of collaborative science in ecology," *Journal of Vegetation Science*, vol. 31, no. 3, pp. 509–517, 2020. DOI: 10.1111/jvs.12864.
- [48] I. Fer, A. K. Gardella, A. N. Shiklomanov, *et al.*, "Beyond ecosystem modeling: A roadmap to community cyberinfrastructure for ecological data-model integration," *Global Change Biology*, vol. 27, no. 1, pp. 13–26, 2021. DOI: 10.1111/gcb.15409.
- [49] B. Bond-Lamberty, A. P. Smith, and V. Bailey, "Running an open experiment: Transparency and reproducibility in soil and ecosystem science," *Environmental Research Letters*, vol. 11:084004, no. 8, 2016. DOI: 10.1088/1748-9326/11/8/084004.
- [50] J. M. Serra-Diaz, B. J. Enquist, B. Maitner, C. Merow, and J.-C. Svenning, "Big data of tree species distributions: How big and how good?" *Forest Ecosystems*, vol. 4:30, no. 1, Dec. 2017. DOI: 10.1186/s40663-017-0120-0.
- [51] F. J. van Rijnsoever and L. K. Hessels, "Factors associated with disciplinary and interdisciplinary research collaboration," *Research Policy*, vol. 40, no. 3, pp. 463–472, Apr. 2011. DOI: 10.1016/j.respol.2010.11.001.
- [52] L. D. Hughes, G. Tsueng, J. DiGiovanna, *et al.*, "Addressing barriers in FAIR data practices for biomedical data," *Scientific Data*, vol. 10, no. 1, pp. 1–7, Feb. 23, 2023. DOI: 10.1038/s41597-023-01969-8.
- [53] E. Deci and R. Ryan, "A Motivational Approach to Self: Integration in Personality," *Nebraska Symposium on Motivation*, vol. 38, pp. 237–88, Feb. 1, 1990.
- [54] K. M. Sheldon, R. M. Ryan, L. J. Rawsthorne, and B. Ilardi, "Trait self and true self: Cross-role variation in the Big-Five personality traits and its relations with psychological authenticity and subjective well-being," *Journal of personality and social psychology*, vol. 73, no. 6, pp. 1380–1393, 1997. DOI: 10.1037/0022-3514.73.6.1380.

- [55] D. Mittal, R. Mease, T. Kuner, H. Flor, R. Kuner, and J. Andoh, "Data management strategy for a collaborative research center," *GigaScience*, vol. 12, pp. 1–25, Jan. 1, 2023. DOI: 10.1093/gigascience/giad049.
- [56] Y. Gil, C. H. David, I. Demir, *et al.*, "Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance," *Earth and Space Science*, vol. 3, no. 10, pp. 388–415, Oct. 2016. DOI: 10.1002/2015EA000136.
- [57] A. Culina, M. Baglioni, T. W. Crowther, M. E. Visser, S. Woutersen-Windhouver, and P. Manghi, "Navigating the unfolding open data landscape in ecology and evolution," *Nature ecology & evolution*, vol. 2, no. 3, pp. 420–426, 2018. DOI: 10.1038/s41559-017-0458-2.
- [58] R. B. Waide, J. W. Brunt, and M. S. Servilla, "Demystifying the landscape of ecological data repositories in the United States," *BioScience*, vol. 67, no. 12, pp. 1044–1051, 2017. DOI: 10.1093/biosci/bix117.
- [59] V. T. Nunes, F. M. Santoro, and M. R. Borges, "A context-based model for Knowledge Management embodied in work processes," *Information Sciences*, vol. 179, no. 15, pp. 2538–2554, 2009. DOI: 10.1016/j.ins.2009.01.033.
- [60] J. M. Pawlowski, M. Bick, R. Peinl, *et al.*, "Social Knowledge Environments," *Business & Information Systems Engineering*, vol. 6, no. 2, pp. 81–88, Apr. 2014. DOI: 10.1007/s12599-014-0318-4.



Informed consent forms

Informed consent form interview

You are being invited to participate in a Master Thesis titled “Collaboration driven data stewardship for better long-term data re-usability”.¹ This study is being conducted by Sara Op den Orth, a Computer Science Master student from TU Delft.

The participants in this study are students, graduates, PhD students, and researchers from Dutch scientific and practical universities, as well as employees of various companies with current or past experience in data management. Part of the participants are connected to the CropXR project.

The purpose of this research study is to collect data on how both consumers (users) and producers (creators) of data experience and think about the (re-)usability of datasets. The interview will take approximately 30-45 minutes to complete, depending on the participant. We will ask you to answer various questions.

First, we will talk about some background information regarding your experience with data management in your research. Second, if you have produced data, we will ask questions about your experience with creating and/or publishing datasets and your considerations regarding data re-use. Third, if you have used data, we will ask questions about your experience using existing data, and how much effort and/or time reusing a dataset took. The interview will contain mostly open questions and is designed to flow naturally—there are no wrong ways of answering the questions.

The data will be used to corroborate the challenges and opportunities in data management with the goal of better reusability. The pseudo-anonymised transcripts of the interview will be entered into Chat-GPT. The goal is to explore the potential value of a chatbot interview process to enhance data management collaboration and improve data reusability.

As with any online activity, the risk of a breach is always possible. We will minimize these risks by removing, wherever possible, any mentions of specific people, datasets, data repositories, or organizations. Before sending any data to ChatGPT, we will send you the transcript so you can verify whether the pseudo-anonymisation is satisfactory. At the end of the project, the pseudo-anonymised transcripts will also be published.

Your participation in this study is entirely voluntary, and **you can withdraw at any time**. You are free to omit any questions.

Corresponding researcher: Sara Op den Orth

Responsible researcher: Christoph Lofi

¹The working title of this thesis

Explicit Consent points

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICIPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated 16-07-2024, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	<input type="checkbox"/>	<input type="checkbox"/>
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	<input type="checkbox"/>	<input type="checkbox"/>
3. I understand that taking part in the study involves: <ul style="list-style-type: none"> - Answering questions about my experience with creating or using datasets. - Having the interviews recorded so they can be transcribed to text (after which the recording will be destroyed). The transcription will be pseudo-anonymised and I will be sent the transcript before it is processed. I will have a week to object in case the pseudo-anonymised was not satisfactory. - Having the pseudo-anonymised transcripts entered into Chat-GPT. 	<input type="checkbox"/>	<input type="checkbox"/>
4. I understand that I will not be compensated for my participation.	<input type="checkbox"/>	<input type="checkbox"/>
5. I understand that the study will end one week after I am send the transcript.	<input type="checkbox"/>	<input type="checkbox"/>
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
6. I understand that taking part in the study also involves collecting specific personally identifiable information (PII) the participant names, contact details, information in informed consent and any identifiable details in their answers and associated personally identifiable research data (PIRD) audio recordings of the interviews, participant demographics (age range, research area, expertise), pseudonymous transcripts with the potential risk of my identity being revealed.	<input type="checkbox"/>	<input type="checkbox"/>
7. I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach. The researcher will attempt to pseudo-anonymise the interview transcript by redacting references to specific people, datasets, data repositories or organisations.	<input type="checkbox"/>	<input type="checkbox"/>
8. I understand that personal information collected about me that can identify me, such as name and contact information needed for the consent form, will not be shared beyond the study team.	<input type="checkbox"/>	<input type="checkbox"/>
9. I understand that the (identifiable) personal data I provide will be destroyed after 10 years.	<input type="checkbox"/>	<input type="checkbox"/>
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
10. I understand that after the research study the de-identified information I provide will be used for a Computer Science Master Thesis. Direct quotes may be used to validate the challenges and opportunities in datamanagement reusability. The transcript will be entered into Chat-GPT, to try and automatically summarise and then possibly connect summaries that have similarities. With the goal of connecting dataset users who have similar experiences or challenges.	<input type="checkbox"/>	<input type="checkbox"/>
11. I agree that my responses, views or other input can be quoted anonymously in research outputs.	<input type="checkbox"/>	<input type="checkbox"/>
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
12. I give permission for the de-identified interview transcripts that I provide to be archived in 4TU.ResearchData repository so it can be used for future research and learning.	<input type="checkbox"/>	<input type="checkbox"/>
13. I give permission for the de-identified interview transcripts that I provide to be included in the Master Thesis which will be archived in the TU Delft education so it can be used for future research and learning.	<input type="checkbox"/>	<input type="checkbox"/>
14. I understand that access to this repository is open.	<input type="checkbox"/>	<input type="checkbox"/>

Signatures

Name of participant [printed] Signature Date

I, as researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

Researcher name [printed] Signature Date

Study contact details for further information: *[Name, phone number, email address]*

Informed consent form survey

You are being invited to participate in a Master Thesis titled “Collaboration driven data stewardship for better long-term data re-usability”.² This study is being done by Sara Op den Orth, a Computer Science Master student from TU Delft.

The participants in this study are students, **graduates**, PhD students and researchers from Dutch scientific and practical universities and **employees of various companies with current or past experience in data management**.

The purpose of this research study is to collect data on how both consumers (users) and producers (creators) of data experience and think about the (re-)usability of datasets, and will take you approximately *10 minutes* to complete.

During this survey you will be presented with the Chat-GPT generated outputs of the first interview to evaluate its potential usefulness. The interview will contain a mix of closed and open questions, there are no wrong ways of answering the questions. The data will be used for corroborating whether the Chat-GPT generated outputs would improve dataset reusability and discoverability.

As with any online activity the risk of a breach is always possible. We will minimise any risks by removing whenever possible any mentions of specific people, datasets, data repositories or organisations. **For this survey only the aggregated results and direct quotes will be used in the thesis. All other data collected will be destroyed after publication.**

Your participation in this study is entirely voluntary and **you can withdraw at any time**. You are free to omit any questions.

Corresponding researcher: Sara Op den Orth - S.M.OpdenOrth@student.tudelft.nl

Responsible researcher: Christoph Lofi - C.Lofi@tudelft.nl

²The working title of this thesis

Explicit Consent points

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICIPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated 12-09–2024, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	<input type="checkbox"/>	<input type="checkbox"/>
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	<input type="checkbox"/>	<input type="checkbox"/>
3. I understand that taking part in the study involves: - Answering questions about my opinion about the usefulness of the Chat-GPT generated results for increasing data reusability.	<input type="checkbox"/>	<input type="checkbox"/>
4. I understand that I will not be compensated for my participation.	<input type="checkbox"/>	<input type="checkbox"/>
5. I understand that the study will end after this survey .	<input type="checkbox"/>	<input type="checkbox"/>
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
6. I understand that taking part in the study also involves collecting specific personally identifiable information (PII) the participant names, contact details, information in informed consent and any identifiable details in their answers and associated personally identifiable research data (PIRD), survey answers, with the potential risk of my identity being revealed.	<input type="checkbox"/>	<input type="checkbox"/>
7. I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach. The researcher will attempt to pseudo-anonymise any direct quotes used from this interview by redacting references to specific people, datasets, data repositories or organisations. After relevant aggregated results and direct pseudo-anonymised quotes have been extracted the other collected survey results will be destroyed.	<input type="checkbox"/>	<input type="checkbox"/>
8. I understand that personal information collected about me that can identify me, such as name and contact information needed for the consent form, will not be shared beyond the study team.	<input type="checkbox"/>	<input type="checkbox"/>
9. I understand that the (identifiable) personal data I provide will be destroyed after 10 years.	<input type="checkbox"/>	<input type="checkbox"/>
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
10. I understand that after the research study the de-identified information I provide will be used for a Computer Science Master Thesis. Aggregated results and direct anonymous quotes may be used to in the thesis to evaluate the usefulness of the Chat-GPT generated results.	<input type="checkbox"/>	<input type="checkbox"/>
11. I agree that my responses, views or other input can be quoted anonymously in research outputs.	<input type="checkbox"/>	<input type="checkbox"/>
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		
13. I give permission for the de-identified direct quotes that I provide to be included in the Master Thesis which will be archived in the TU Delft education so it can be used for future research and learning.	<input type="checkbox"/>	<input type="checkbox"/>
14. I understand that access to this repository is open.	<input type="checkbox"/>	<input type="checkbox"/>

Signatures

Name of participant [printed]

Signature

Date

I, as researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

Researcher name [printed]

Signature

Date

Study contact details for further information: [*Name, phone number, email address*]

B

Interview protocol

B.1. Prologue statements and questions

The following list gives the full prologue information given to each participant at the beginning of the interview so they knew what to expect.

- **Interview Details.** The interview will be in English. The interview will be mostly open questions, focussing on your experience with creating or using datasets in the context of findability and reusability, specifically regarding the metadata included in the dataset you worked on. The questions are sometimes in a strange order because parts will be processed in a further way using Chat-GPT.
- **Added clarification.** None of the questions is meant as a critique of how you have handled the datasets.
- **Definitions.**
 - Metadata includes any data that describes or provides information or context to raw data. This includes details such as the creation date, the data owner, comments, data types, methodology, and more.
 - Producers, those who create new data that is published or reused.
 - Consumers, Those who use existing datasets, one can be both of course.
- **Other.** Any questions?

Table B.1 shows the questions used to collect demographic and professional details to contextualise the perspectives of the participants. The question about the type of data was removed because it was not particularly relevant and could, in some cases, be identifying. Evaluation of data management experience was occasionally challenging, as participants often had sporadic education or experience in the field.

Table B.1: Introductory interview questions, with notes indicating which questions were removed and why.

Question	Notes
What is your primary research area?	
What type of data do you work with? (e.g., genomics, ocean density)	Removed, did not provide interesting detail.
What is your highest completed degree within your current research domain?	
How many years of experience do you have in your field?	
Do you have a formal degree in data management? If so, what is the highest level completed?	
How would you rate your data management skills?	

B.2. Data producer questions

Table B.2: Producer interview questions, with notes indicating which questions were removed or changed and why. Suggestions were added if it was clear the question on its own was too vague.

Question	Notes
Did you produce data from your own measurements or by combining existing sources of data?	Removed, no interesting insight.
How many datasets have you produced?	Removed, this was not a good indicator of experience as there was no time to ask about how extensive these other projects were.
How was the dataset published?	Suggestions: on a repository or internally published.
Did you have help when creating the metadata, either in domain or data management knowledge?	
How did you imagine a user might find the dataset?	
How easy do you think the dataset would be to reuse?	
Did you include how the dataset was produced?	Often removed, no interesting insight.
Did you consider who might reuse the dataset?	
Did you consider the future users domain experience?	Suggestion: for example a student.
Did you consider the future users data management experience?	
Did/Does considering the future user change what metadata you include to increase findability?	Removed or merged with next question: as the answers would overlap.
Did/Does considering the future user change what metadata you include to increase reusability?	Suggestion: if the participant indicated they knew the future user they were asked this question again imagining an unknown future user.
Did you consider what the dataset might be used for in the future, a future use case?	Suggestion: was there an additional data attribute you chose to include, for example.
Did/Does considering the future use case change what metadata you include to increase findability?	
Did/Does considering the future use case change what metadata you include to increase reusability?	
Was this conversation useful to you, in what way?	Removed: it was too vague of a question.
Have you encountered contextual metadata as we discussed here before?	
Have you considered adding contextual metadata as discussed here to your publication?	
Do you often discuss what metadata to include with others?	
Do you receive feedback on the quality and usefulness of your metadata?	
Do you get the sense creating high-quality metadata is appreciated by others?	Removed or merged with previous questions: as the answers would overlap.
Does considering the future user and use case motivate you to spend more time on metadata creation?	
Did this conversation give you any new insights?	Removed: no interesting responses and was moved to the survey instead.

B.3. Data consumer questions

Table B.3: Consumer interview questions, with notes indicating which questions were removed or changed and why. Suggestions were added if it was clear the question on its own was too vague.

Question	Notes
How many projects have you worked on that involved dataset processing?	Removed: this was not a good indicator of experience as there was no time to ask about how extensive these other projects were.
How was the dataset published?	Suggestions: on a repository or internally published.
Did you already have a use case in mind for this dataset? What criteria did you use to judge whether the dataset was useful?	
Where did you find the datasets you needed?	
Did you require additional information besides included metadata and the producer's methodology?	Suggestions: contacted the producer, had to use search engine to understand metadata or used other methodologies.
Was your final use of the dataset in line with the producer's original intent?	
Was this conversation useful to you, in what way?	Removed: it was too vague of a question.
Were you able to use the dataset as you had planned?	
Was the information you needed to reuse the dataset easy (and/or fast) to find?	Removed: this would come up during previous questions most of the time.
Would it have been useful to have information on other people's use of the same dataset?	
Did you publish your methodology?	
Did you publish the challenges you encountered in the process?	
Do you have any suggestions on what would have made the process easier and more efficient?	

C

Survey protocol

Due to the size of the table, it is printed on the next page.

Table C.1: Responses of survey participants to all closed questions. Each row represents an individual participant, and columns correspond to the specific questions. The table is divided into two sections: responses provided by participants in their roles as producers or as consumers. Responses are coded based on the survey's predefined scales or options.

ID	Producer							Consumer						
	Accuracy	Helpful self	Helpful other	Summary improv.	Summary vs transcript	Sharing summary	Sharing transcript	Accuracy	Helpful self	Helpful other	Summary improv.	Summary vs transcript	Sharing summary	Sharing transcript
P1	Strong agree	Agree	Agree	Strong agree	Full transcript	Strong agree	Agree	Agree	Strong agree	Agree	Strong agree	Strong agree	Agree	Agree
P3	Strong agree	Neutral	Strong agree	Strong agree	Summary	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong disagree	Strong disagree
P6	Agree	Agree	Neutral	Agree	Full transcript	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree
P8	Agree	Disagree	Agree	Strong agree	Summary	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong disagree	Strong disagree
P10	Strong agree	Strong agree	Disagree	Agree	Summary	Agree	Agree	Agree	Strong agree	Strong agree	Strong agree	Strong agree	Agree	Agree
P12	Agree	Neutral	Neutral	Neutral	Summary	Neutral	Neutral	Neutral	Neutral	Neutral	Agree	Agree	Agree	Agree
P14	Neutral	Agree	Agree	Strong agree	Summary	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree
P15	Agree	Agree	Agree	Strong agree	Summary	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree
P17	Agree	Disagree	Disagree	Disagree	Summary	Disagree	Disagree	Disagree	Disagree	Disagree	Neutral	Neutral	Agree	Neutral
Consumer														
P2	Agree	Neutral	Neutral	Agree	Summary	Agree	Agree	Agree	Agree	Agree	Neutral	Neutral	Disagree	Disagree
P3	Agree	Disagree	Strong agree	Disagree	Full transcript	Disagree	Disagree	Disagree	Disagree	Disagree	Disagree	Disagree	Neutral	Neutral
P6	Agree	Agree	Neutral	Agree	Full transcript	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree
P10	Agree	Strong agree	Neutral	Strong agree	Summary	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree
P11	Agree	Agree	Agree	Agree	Full transcript	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree
P12	Disagree	Neutral	Neutral	Neutral	Full transcript	Neutral	Neutral	Neutral	Neutral	Neutral	Agree	Agree	Agree	Agree
P13	Agree	Neutral	Agree	Strong agree	Summary	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Strong agree	Disagree	Disagree
P14	Agree	Neutral	Neutral	Neutral	Summary	Neutral	Neutral	Neutral	Neutral	Neutral	Strong agree	Strong agree	Strong agree	Strong agree
P15	Agree	Agree	Agree	Agree	Summary	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree
P17	Agree	Disagree	Disagree	Agree	Summary	Disagree	Disagree	Disagree	Disagree	Disagree	Neutral	Neutral	Agree	Neutral

D

Data conversation transcript summaries

D.1. Data producer transcript summaries

Summary producer P1

- **Dataset Production and Usage**

- Has produced datasets for two experiments, with one actively used for research while the other remains unprocessed.
- Primarily uses datasets for internal purposes due to medical data sensitivity, adhering to fair use policies with pseudo-anonymization.

- **Metadata and User Accessibility**

- Metadata planning is minimal, expecting future users to be internal or team members with domain familiarity.
- Anticipates that interest in the dataset may arise through publications, where access would be available on request.

- **Consideration for Future Users**

- Future documentation for wider usability is limited, as dataset use is envisioned for someone familiar with the research context.
- Reflects on the potential for broader dataset applications, though no clear strategy for making it accessible beyond immediate research circles.

Summary producer P3

- **Publication and Access**

- Dataset published alongside an article, with access primarily intended through the article itself.

- **Reusability and Audience**

- Anticipated users are bioinformaticians with assumed familiarity with relevant terms and domain-specific knowledge.
- Simplified data structure (CSV format) with minimal data management considerations, relying on standard tools for accessibility.

- **Broader Applicability**

- Limited considerations for reuse outside bioinformatics, but dataset kept general enough for slight modifications to enable cross-domain use.

Summary producer P5

- **Dataset Scope and Publication**
 - Dataset was published in a university repository and is primarily intended for internal project stakeholders and policy analysts.
- **Data Accessibility and Methodology**
 - Emphasized step-by-step explanations for ease of use, especially considering the unique data collection method unfamiliar to broader audiences.
 - Pseudonymization includes basic demographic indicators, carefully balanced to maintain anonymity.
- **Future Use Considerations**
 - Data included additional demographic elements (e.g., proximity to water sources) to potentially support secondary research beyond the immediate study scope.

Summary producer P6

- **Dataset Context and Accessibility**
 - Dataset stored internally within a company; limited public accessibility due to confidentiality and dataset size.
 - Metadata focuses on identifiers relevant to atmospheric and location-based data, aiding internal searchability.
- **Audience and Skill Level**
 - Intended primarily for other students and researchers within the same organization; assumes comparable skill levels and domain knowledge.
- **Future Application and Documentation**
 - Enhanced documentation included in scripts to improve future usability for internal users, with annotations to clarify processing steps and data management practices.

Summary producer P7

- **Data Storage and Retrieval**
 - Dataset organized in SQL for phenotypic data, accessible through internal apps primarily for breeding analysis and genetic marker development.
- **Metadata and User Familiarity**
 - Metadata tailored based on department needs; visualizations developed for breeders with less technical expertise, whereas technical metadata is structured for bioinformaticians.
- **Future Use Consideration**
 - Metadata includes data quality indicators relevant to future genetic research applications, maintaining flexibility for varied uses within the organization.

Summary producer P8

- **Collaborative Dataset Production**
 - Produced in consultation with teams who would use the data (application and dashboard teams) to ensure relevance and usability.
- **Data Accessibility and Confidentiality**
 - Direct discussions with users influenced the data's structure and accessibility, accounting for limitations due to confidentiality and testing phases.
- **Long-term Usability**
 - Dataset created with user feedback in mind, adapting to various needs within the organization and ready for future revisions based on evolving requirements.

Summary producer P9

- **Internal Use and Metadata**
 - Dataset incorporates essential metadata for internal cataloging, particularly for resampling and integrating open-source meteorological data.
- **User Assumptions**
 - Dataset is aimed at professionals within the same organization, presuming similar expertise in data handling and domain knowledge.
- **Use Case Limitations**
 - Minimal attention to diverse use cases, focusing mainly on the primary purpose of supporting internal modeling and predictive applications.

Transcript: P10

- **Dataset Publication and Metadata Consideration**
 - Dataset published within an article but limited metadata was included, mainly due to it not being common practice in their field.
 - Metadata primarily kept offline in a lab notebook for personal reference rather than comprehensive reuse.
- **Audience and Accessibility Consideration**
 - Data was shared with supervisors, the primary envisioned users, though no extensive efforts were made to ensure usability for external researchers.
 - Assumed future users would possess a similar educational level.
- **Reusability and Use Case Consideration**
 - Considered future use mainly to aid understanding in the context of the original research.
 - Limited efforts to make data easily interpretable or reusable for those outside the immediate research context.

Transcript: P12

- **Dataset Publication and Metadata**
 - Data published in domain-specific repositories, including NCBI's short read archive and a Max Planck Society Library repository.
 - Metadata and keywords were chosen with findability in mind, particularly in domain-specific contexts.
- **Audience Considerations**
 - Targeted both advanced users (capable of analyzing raw data) and less experienced users (benefiting from processed data).
 - Processed data was made available to enhance accessibility for users without deep bioinformatics expertise.
- **Reusability and Use Cases**
 - Emphasis on transparency and reproducibility, with adjustments made to data format for ease of use.
 - Thought given to facilitating various future analyses, though not extensively customized for unknown use cases.

Transcript: P13

- **Dataset Publication and Metadata Inclusion**
 - Dataset published internally in Elab Journal for colleagues, following an internal data management plan.
 - Metadata aimed at ensuring colleagues can locate and interpret data efficiently, with data organized to align with project work packages.
- **Audience and Accessibility Considerations**
 - Intended for colleagues and future students, with data structured to be user-friendly within the research team.
 - Assumed users would have similar expertise, and naming conventions were tailored for ease of access within the internal framework.
- **Reusability and Use Cases**
 - Minimal consideration of external or alternative use cases beyond internal reusability.
 - Data specifically designed for the internal growth chamber, limiting broad applicability.

Transcript: P14

- **Dataset Publication and Metadata**
 - Published alongside an article, primarily organized for personal retrieval and usability rather than broader accessibility.
 - Metadata and categorization adjusted minimally to ensure that others within the same project could understand it.
- **Audience Considerations**
 - Next users were expected to be students or internal team members with similar educational backgrounds.
 - Data was indirectly accessible via supervisors, limiting external use without direct facilitation.
- **Reusability and Future Use Cases**
 - Made efforts to categorize and label data for easier internal use, anticipating reuse within the same project.
 - Use cases envisioned were limited to follow-up student projects or theses, with no specific adjustments made for unknown external users.

Transcript: P15

- **Dataset Publication and Accessibility**
 - Planned to publish data in a university repository, with considerations for both academic and industry use.
 - Early considerations for data's visibility, influenced by engagement with industry contacts interested in the research.
- **Audience Considerations**
 - Envisioned users included both students and industry professionals, with diverse experience levels anticipated.
 - Assumed users would have foundational civil engineering knowledge, but data management expertise levels were varied.
- **Reusability and Broader Impacts**
 - Hopes to make data a starting point for broader industry and academic applications, though limited by time constraints as a student project.
 - Some anonymization and adjustments made to enhance accessibility, yet constrained by project resources.

Transcript: P17

- **Dataset Publication and Metadata**

- Dataset is available upon request, associated with an article but kept internal for privacy reasons due to clinical data restrictions.
- Metadata includes standard clinical outcomes, allowing comparability within the field without additional keyword optimization for accessibility.

- **Audience Considerations**

- Anticipated users primarily include researchers within the clinical domain, with similar domain expertise.
- Limited usability for those with less data management knowledge, as data was structured without novice accessibility in mind.

- **Reusability and Future Use Cases**

- Data prepared for internal continuity, particularly for a new PhD student in the research line.
- Limited efforts to enhance general accessibility, as time constraints prioritized the current research over extensive documentation.

Transcript: P18

- **Dataset Publication and Metadata**

- Data published alongside an article, without dedicated efforts to enhance findability beyond general keywords.
- Metadata was basic, aiming primarily to document data locations while keeping identifying information anonymized.

- **Audience Considerations**

- Target audience includes soil scientists with assumed knowledge for interpreting the data.
- Limited adjustments were made for users with minimal data management skills due to the dataset's simplicity.

- **Reusability and Use Case Consideration**

- While future use cases were anticipated, no additional information was included to support unknown applications.
- Research plans involve expanding the dataset in collaboration with other institutions, potentially necessitating future adjustments for broader accessibility.

D.2. Data consumers transcript summaries

Summary consumer P2

- **Project Overview**

- Worked on three projects involving data processing, focusing on one that was particularly challenging.
- Project centered on predicting forest fire occurrences using a dataset found online.

- **Dataset Selection**

- Chose the dataset based on ease of handling, as it had clean data and was contextually understandable (forest fire data).
- Decision influenced by a similar prediction-focused article using the same dataset.

- **Data Processing Challenges**

- Initial steps involved understanding each variable, with some needing clarification from external articles due to vague descriptions.
- Encountered issues with predictive model performance, attributed later to a missing critical component after feedback from a professor.

- **Project Outcome**

- Despite setbacks, the project was completed and presented.
- Considered alternative use cases but stayed focused on prediction-related analysis.

Summary consumer P3

- **Data Discovery and Selection**

- Located an older version of a specific dataset for use in validating results produced by a tool.
- Dataset choice was based on availability as it was the only dataset suited to the intended analysis.

- **Methodology and Adaptation**

- Relied heavily on the original methodology from the dataset's creators.
- Required extensive searching for additional explanations due to limited metadata and vague original documentation.

- **Challenges and Adjustments**

- Encountered outdated or missing components within the dataset's tool, leading to a custom tool adaptation.
- Significant time investment was necessary to ensure accuracy due to limited support from the original data producers.

- **End Result**

- Final method diverged from initial plans due to tool limitations.
- Updated the analysis method using a self-created tool alongside the original dataset.

Summary consumer P5

- **Dataset Identification and Usage**

- Located demographic data from a national statistical bureau, accessed in the local language to find specific regional datasets.
- Key factors in dataset selection included completeness and relevance to research topics (e.g., population density, income levels).

- **Process and Accessibility**

- Ensured data alignment with research needs by organizing data systematically and cross-referencing with other sources.
- Straightforward data download options enhanced ease of use, despite occasional page crashes.

- **Data Suitability and Alignment**

- Data aligned well with research goals as it served as contextual support for a social research study.

Summary consumer P6

- **Initial Dataset Challenges**

- Received an undocumented dataset from a former researcher who abruptly left, making the data and scripts challenging to interpret.
- Data comprised numerous Python scripts, which required line-by-line debugging due to compatibility issues.

- **Methodology and Support**

- Utilized former research papers for guidance but relied heavily on trial-and-error to rewrite code due to missing instructions.

- Encountered additional complexities working remotely, with unfamiliar server setups during the pandemic lockdown.
- **Project Resolution**
 - Successfully adapted the code for project use, despite substantial delays due to the undocumented scripts.

Summary consumer P7

- **Dataset Discovery and Selection Criteria**
 - Originally selected a dataset to answer a specific research question, but the data's lack of metadata complicated understanding.
 - Relied on personal contact with the data producer to gain necessary insights, though some details remained unclear due to elapsed time.
- **Outcome and Adjustment**
 - Ultimately discontinued use of the dataset due to documentation challenges.
 - Project objectives shifted as the initial dataset was unsuitable for intended use.

Summary consumer P8

- **Data Accessibility and Use Case**
 - Predetermined use case focused on predictive modeling with a dataset provided by a colleague.
 - Faced issues with unclear variable names and numerous missing values, necessitating external expertise and supplementary data sources.
- **Supplementary Research and Resolution**
 - Consulted with a domain expert and used an additional dataset to resolve data gaps.
 - Final methodology deviated significantly from initial plans due to the data's limitations.
- **Project Alignment**
 - Use case was in line with the data's intended application, although methodology required improvisation due to initial dataset quality.

Summary consumer P9

- **Data Gap and Dataset Search**
 - Dataset was found through online search to address an existing data gap in research.
 - Metadata was insufficient, leading to extensive searches to confirm data validity and source identification.
- **Methodology and Hurdles**
 - Invested considerable time in identifying data conventions and proper labeling.
 - Communication with data producers was occasionally required for clarification.
- **Conclusion**
 - Final method adapted to the complexity of verifying dataset conventions, though anticipated issues with data management contributed to initial planning.

Summary consumer P10

- **Dataset Selection and Initial Purpose**
 - Sourced a bioinformatics dataset from a well-established public repository.
 - Dataset chosen based on previous lab experiment outcomes.
- **Data Handling and Accessibility Issues**
 - Faced difficulties downloading the dataset in a suitable file format, making processing challenging.
 - Simplified data extraction using copy-paste as an alternative to prolonged data formatting.
- **Project Outcome**
 - Adjusted methods based on file access challenges.
 - Although not formally documented, these adaptations were noted for potential future use.

Summary consumer P11

- **Data Source and Accessibility**
 - Data was published on a website via a public article, making it publicly accessible but challenging to download fully.
 - Original attempts to obtain complete data from researchers were unsuccessful; researchers were too busy to assist.
- **Usage Challenges**
 - Difficulties in downloading data due to pagination and limited accessibility prompted consideration of web scraping.
 - Unable to automate data gathering, P11 had to search the website manually, limiting data utility.
- **Data Quality and Format**
 - The data quality and types were sufficient for their needs, but the unusual data format required extra effort.
 - A comprehensive download would improve efficiency, but the dataset was not available in standard genomic formats.
- **Alignment with Data Producer's Intent**
 - The use case aligned with the anticipated functions of the dataset, though more accessible formats would better support various research applications.

Summary consumer P12

- **Data Source**
 - Accessed data from a public repository (NCBI) with programmatic options, allowing straightforward data access.
 - Metadata and treatment-related details were readily available in the repository.
- **Data Application**
 - Conducted a meta-analysis on omics datasets, using data replicates for analysis consistency.
 - Maintained their analytical methods to prevent study biases, enabling the application of a standardized methodology across datasets.
- **Challenges and Observations**
 - Repository standards facilitated the application of familiar methods, simplifying data processing.
 - No significant barriers encountered, as the dataset format was well-known in their research field.

Summary consumer P13

- **Data Source**
 - Utilized a gene database as a primary resource for researching gene expressions.
 - Easy to locate specific genes, but connecting genes to specific research scenarios (e.g., flowering or drought resistance) was challenging.
- **Data Extraction Challenges**
 - The database lacked scenario-specific metadata, leading to reliance on external publications for context.
 - Required additional effort to extract and interpret data relevant to specific biological traits, limiting efficiency.
- **Application Outcome**
 - Used the database as a background resource for building a gene interaction network.
 - Data was for personal research understanding, so documentation of extraction steps was deemed unnecessary.

Summary consumer P14

- **Data Discovery and Use Case Formation**
 - Discovered datasets without a specific use case; initial exploration influenced the research direction.
 - Later determined that datasets encountered did not suit the finalized research purpose.
- **Criteria for Data Evaluation**
 - Evaluated datasets based on data content and metadata from related articles, but ultimately found them insufficient for specific needs.
- **Method Adjustments**
 - Shifted focus after further literature review, identifying alternative datasets that better aligned with the project requirements.

Summary consumer P15

- **Data Access and Substitution**
 - Intended to use a dataset referenced in multiple reports but could not access it directly.
 - Found alternative data by using synonyms and related keywords in Scopus, gathering a similar dataset through different reports.
- **Data Usage**
 - Did not replicate methodologies directly but used report insights to build a foundational understanding for their research.
 - Skimmed methodological sections, focusing instead on findings relevant to their research goals.
- **Adaptations**
 - The search process required flexibility in terminology, which proved effective in gathering comparable data insights.

Summary consumer P17

- **Data Source and Judgment**

- Found datasets through published articles on relevant research subjects due to limited options within the topic.
- Relied on the articles' alignment with project topics as a criterion for data selection.

- **Data Processing**

- Needed raw data for subgroup analyses, which was not included in publications; reached out to original researchers to obtain this.
- Managed to acquire necessary data, although response delays were common.

- **Method Consistency**

- Anticipated needing additional data but faced challenges due to delays in researcher responses.
- Final analysis required adaptation due to the logistical challenges in obtaining raw data in a timely manner.

Summary consumer P18

- **Data Access through Networking**

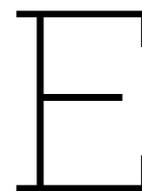
- Accessed field experiment data through direct contact with researchers, as the dataset wasn't publicly available.
- Maintained long-term relationships with the data producers, which facilitated the data acquisition process.

- **Data Application and Adjustments**

- Used data for calculations involving updated models, with methodology evolving based on discussions with the data provider.
- Faced limitations as only partial data was shared, impacting the extent of analysis.

- **Challenges and Collaboration**

- Encountered delays due to miscommunication and logistical challenges, emphasizing the value of strong interpersonal connections.
- Financial considerations impacted the availability and assistance in data usage, as researchers' time and guidance required funding.



CropXR metadata gap

Table E.1: Overview of CropXR metadata gap as presented in Section 7.2. Linking CropXR initiatives to metadata gap factors as identified in Section 5.1.

ID	Metadata gap factor	CropXR initiative
S-M1	Enhancing communication and collaboration	Managed network
S-M2	Participating in grass-roots initiatives	Managed network
S-M3	Receiving peer recognition	Managed network
S-M4	Experiencing personal metadata frustrations	Managed network
S-M5	Building research credibility	Managed network
S-M6	Enhancing data quality	Managed network
S-M7	Developing data management skills and tools	EduXR & DataXR
S-C1	Time constraints	Meta Buddy
S-C2	Space constraints	Resiliency Hub
S-C3	Data complexity	SAME group
S-C4	Shifting data standards and practices	SAME group
S-L1	Lack of data management expertise	EduXR & DataXR
S-L2	Subjectivity in metadata creation	SAME group
S-L3	Lack of insight in consumer needs	Managed network, EduXR & DataXR, SAME group
S-L4	Lack of ongoing commitment	Managed network, Resiliency Hub
D-C1	Scattered metadata	Resiliency Hub
D-C2	Incomplete metadata	SAME group
D-C3	Divergent use cases	Managed network
D-C4	Limited access to producers	Managed network
D-C5	Unreported data attributes	SAME group
D-C6	Inconsistent (meta)data standards	SAME group
D-L1	Lack of domain knowledge	EduXR & DataXR
D-L2	Lack of data management skills	EduXR & DataXR