CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

A Software Engineering Perspective on Data Quality Processes in Environmental Research

Recommendations Based on Software Engineering Practices Applied for Improving of Open Data Practices and Communication in Environmental Research

Master's Thesis in Computer Science and Engineering

MARKUS MOEN
MAX NORÉN

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2024

Master's thesis 2024

A Software Engineering Perspective on Data Quality Processes in Environmental Research

Recommendations Based on Software Engineering Practices Applied for Improving of Open Data Practices and Communication in Environmental Research

MARKUS MOEN
MAX NORÉN

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2024

A Software Engineering Perspective on Data Quality Processes in Environmental Research
Recommendations Based on Software Engineering Practices Applied for Improving of Open Data Practices and Communication in Environmental Research
MARKUS MOEN
MAX NORÉN

Supervisor: Hans-Martin Heyn, Computer Science and Engineering
Examiner: Birgit Penzenstadler, Interaction Design and Software Engineering, Computer Science and Engineering

Master's Thesis 2024
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2024

A Software Engineering Perspective on Data Quality Processes in Environmental Research
Recommendations Based on Software Engineering Practices Applied for Improving of Open Data Practices and Communication in Environmental Research
Markus Moen
Max Norén
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

The many fields within environmental research have been on the path towards open science and, most importantly, open data. With the increase in available data, there are opportunities to apply data-driven and data-intensive methods, including recent developments such as machine learning. However, the success of applying machine learning depends significantly on the quality of the available training data. The purpose of this thesis was to investigate the field of environmental research in regards to current views, practices and communication of data quality and to identify software engineering principles and practices that can form possible recommendations to progress data quality in environmental research. This process identified six challenges and proposed eight recommendations. The result shows a great deal of effort towards open data, with the FAIR principles as the main arbiter to achieve it. Most identified challenges are based on data quality handling, communication, and difficulties in achieving open science. We found suitable software engineering practices for four of the six challenges, with two key perspectives being derived from open source software and requirements engineering practices. Our results demonstrate that there is a willingness among environmental researchers to investigate and adopt software engineering practices in environmental research. Importantly, there is a broad agreement that open science is an improvement over to previous methods, and the stated challenges and recommendations need to preserve those advancements. The recommendations should be regarded as a first design iteration of these recommendations, and they should be explored further in terms of their applicability to different fields within environmental research.

Keywords: Software engineering, requirements engineering, data quality, environmental research, open science, open data, data-intensive, big data, FAIR, thesis

Acknowledgements

We would like to thank Hans-Martin Heyn, our supervisor in software engineering, for his great commitment and assistance in guiding our work on this thesis. He has been a great encouragement to us and has been very supportive throughout the thesis. We would also like to thank Michelle Nerentorp, our supervisor at IVL, for helping us at the start of the thesis and serving as the anchor for our understanding of the field of environmental research. Finally, we would also like to thank all of our interviewees and survey respondents for their invaluable insights and great enthusiasm for our work.

Markus Moen & Max Norén, Gothenburg, 2024-05-27

Glossary

Artificial intelligence (AI) The ability for a computer to perform tasks normally associated with intelligence. Artificial intelligence is an umbrella term for many related methods such as machine learning, machine reasoning, semantic annotation, and others.

Big data Massive amounts of heteregeneous data, usually from multiple sources which can be used for data-intensive or data-driven methods.

Citizen science Practices and methods that involve non-professionals to accomplish general goals within scientific research, often related to collection and/or analysis of data.

Data consumers Data consumers in the context of environmental research are any stakeholder that make use of environmental data to achieve certain goals. These stakeholders are not limited to environmental researchers but also includes cooperation, policy makers, governmental institutions and as an extension the public.

Data hosts Data hosts in the context of environmental research are stakeholders that provide databases or repositories of environmental data. These include smaller ones which may host data from their own project or big aggregators of data with extensive review processes for data.

Data producers Data producers are stakeholder Data producers in the context of environmental research are stakeholder that create or collect data, and process it for later use. This can include researchers gathering data for a specific project or monitoring efforts to contribute to long time series data.

Data quality Data quality reflects different quantifiable metrics or aspects relating to the usability and usefulness of data.

Data-intensive Data-intensive applications Methods or practices which utilize or necessitate large quantities of data for function or optimal results.

Environmental research A multidisciplinary field which includes many sciences, such as: air pollution monitoring, atmospheric science, biology, climate research, ecology, geology, oceanography [1].

FAIR A set of principles to achieve open data practices which include: Findability, Accessability, Interoperability and Reusability [2].

Git An open source version control system to track changes of computer files in an often shared and distributed environment. This is prevalent in software development in order to enable collaborations and tracking changes as software evolves over time.

Goal Question Metric (GQM) Goal-Questions-Metrics is a model that associated metrics to the goals they are meant to accomplish through questions.

Machine learning (ML) Various data-driven algorithms that can identify and make use of patterns in large datasets.

Machine readability The ability of data to be used with minimal processing or need for human interpretation, where a program can be run independently and work across different sources and utilize a variety of concepts.

Metadata Information about the data which associates to provide additional understanding of the data such as defining different quality aspects, context of the data collection process and other aspects that can be used such as dataset structure.

Metrics Different kinds of measurements in order to measure performance and quality. These are used to quantify and often compare different artifacts such as data and its associated data quality.

Open science Open science principles aim to provide data openly, facilitating greater amounts of data being available to anyone. Some benefits include increasing transparency and ability to reuse data for different purposes to a higher degree which also allows for further validation.

Requirements engineering (RE) A subfield of software engineering which studies the development of requirements which includes the process of eliciting, documenting, analyzing, verifying, and validating requirements based on stakeholder needs [3].

Software engineering (SE) The study of applying engineering, which encompasses development, operation, and maintenance, to software [4].

Stakeholders Stakeholders are people, organizations or other entities that have a vested interest in work that is being done or product such as in a project.

Team science Practices that includes a diverse set of people from different fields, with different expertise and perspectives, collaborating towards a common scientific goal [5].

Contents

List of Figures

List of Tables

1

Introduction

Environmental research encompasses a wide variety of sciences including ecology, oceanography, climate change, and more [1].

In addition, most environmental research relies heavily on large amounts of often heterogeneous data, and for certain fields, such as climate change research as illustrated by Easterbrook, the data is required to have an almost unprecedented temporal and spatial width [6].

With improved technology and increased efforts in open science to share datasets, the amount of data available to researchers has increased exponentially [7]. This has created new opportunities for research, perhaps most clearly in the incredibly complex climate models that accurately predict climate changes [6]. These insights can inform policymakers about appropriate action in the fight against global warming [8].

Environmental data are often collected and curated for a specific project. However, given the often high cost of data collection and the need for large datasets, researchers need to reuse and combine environmental data with different data quality from other projects [9]. Especially as new data-intensive and data-driven software approaches such as artificial intelligence (AI) and machine learning (ML) become ever more viable, the need for good quality data is increasingly important, and new skills are required to effectively employ these approaches [10]. This has spurred the development of additional data qualities and guidelines, such as the FAIR principles, to enable open science where access to data is increased both among researchers but also processing by computers through machine readability.

In addition, there are often many problems with contemporary data collection that lead to various weaknesses in the datasets [11]. Datasets are often created to optimize attributes such as accuracy in predictive models, but they often fail to account for various biases encoded in the data. Vulnerabilities in datasets, such as biases, are also often not clearly presented in open datasets, making it difficult for outside researchers to properly use the data and trust the results they obtain [12].

In particular, methods have been developed within the field of software engineering (SE) to improve not only the use of data-intensive research, but also the workflows associated with data collection, use, and storage [10]. Those methods are for example requirements engineering (RE) for ML [13], big data systems [14], and other practices described in SWEBOK [15]. Some of these methods might be of use to be applied to environmental research to improve workflow and increase data quality in their data-intensive research.

As environmental research continues to use software to process large amounts of data, it is reasonable to explore what techniques and practices can be brought over from SE. When it comes to methods for data-intensive processes and applications there is a continuous effort to take advantage of those opportunities, with climate research being very successful in data-intensive modeling as noted by Easterbrook [6] On the other hand, not all fields have the same history with data-intensive research, for example, ecology in the last decade incorporated big data practices [16]. Thus there is an opportunity to apply an SE perspective on environmental research, with a key aspect of this thesis being data quality. In terms of dealing with data quality, the field of RE is of particular interest as it provides a structured way of communicating needs and responsibilities and ensuring that those requirements are met [15]. Especially as new data-driven techniques such as ML change how some research is approached with new requirements on data quality [13]

This paper investigates the current challenges in environmental research in regards to data quality. Part of it includes investigating expectations and communication of data quality requirements between data consumers and data producers, the two main stakeholders of concern in this study. From this, potential recommendations from SE will be explored that can improve data quality processes and techniques to support data-intensive environmental research such as the application of ML.

2

Theory

This chapter provides an overview of work on data quality in environmental research, particularly in the context of open data. Efforts in SE to handle big data and more data-intensive practices are also presented. Furthermore, the purpose of the study and the research questions to be answered are introduced.

2.1 Environmental Research, Big Data, and SE

Below, we introduce relevant concepts for environmental research, SE, and RE as it relates to data quality in data-intensive research. The definitions for these terms and concepts may differ between fields, but throughout this thesis we will use the definitions detailed below.

Environmental Research as Data-Intensive Research

Environmental research is a multidisciplinary field which includes many sciences and includes but is not limited to [1]:

- Air Pollution Monitoring
- Atmospheric Science
- Biology
- Climate Research
- Ecology
- Geology
- Oceanography

As mentioned by Edwards in *A Vast Machine*, the natural world is a highly complex system and related research often spans large temporal and spatial scales, meaning that many of these fields deal with large amounts of heterogeneous data [7]. Farley et al. discuss how technological developments coupled with increased access to large volumes of data, known as big data [14], have spurred a wider adoption of more data-intensive applications in environmental research to harness the potential of the data [16]. This includes efforts to collect data from institutions and companies, new methods such as citizen science through increased availability of low-cost sensors,

efforts to standardize and share data in open science, and more. While these developments have opened up new opportunities and increased the capabilities of environmental research, they have also brought with them new challenges. Progress has been made, but as Hampton et al. discuss, tackling such a large task still requires further efforts, such as developing new computational skills within environmental research [10]. In addition, Balbi et al. discuss how improvements in data sharing and definition can increase the ability to reuse data [17], and Cheruvelil and Soranno highlight the importance of collaboration across research fields in data-intensive research [18].

Challenges With Big Data and Data Quality

Data-intensive applications refer to software or methods that require large amounts of data [10], with methods such as complex climate models [6] or, more recently, ML, which relies on a data-driven statistical approach to find patterns in data [13]. Paullada et al. highlight an important aspect, which is that data-intensive methods, such as ML, require a shift in current data collection practices to increase performance and reduce bias which ML is susceptible to [11]. This adjustment to data collection methods is even more varied in the field of environmental research, as different fields have different needs on both data and data quality, something that Easterbrook exemplifies with forecasting data requiring timeliness in the weather data rather than consistency while climate data is the reverse [6]. Data quality is the measure for assessing the properties of the data, where there is a wide range of definitions. For instance, Kahn et al. defined categories for data quality as information quality dimensions, which include related information quality attributes [19]. Although Kahn et al. use the term information quality, the term data quality is more common among environmental researchers such as Mccord et al. and Nguyen et al. [5, 9]. For the purpose of this thesis, these terms are interchangeable, and therefore the term data quality will be used for the remainder of this thesis, but use the dimensions and attributes defined by Kahn et al., as they were deemed to have an appropriate level of granularity.

With the increased use of data-intensive applications combined with open science, accessibility is not enough to promote reuse. Machine readability, where data formats are easy for both humans and computers to use, has led to an increased focus on interoperability and standards such as the FAIR principles [2]. This has been important even in environmental research as Balbi et al. note that the increasing amount of data available is beyond the scope of humans to go through it all, and increasing interoperability and machine readability is key to making effective use of the data [17] Important for data and data quality is the inclusion of metadata that provides additional information about the data [10]. Furthermore, to quantify and measure these data quality attributes, metrics are often used to assess and create better requirements for what is expected of the data in terms of data quality. For consistency, this paper will primarily use and refer to the list of information quality dimensions defined by Kahn et al. [19], shown in table 2.1.

Table 2.1: List of 16 data quality dimensions as defined by Kahn et al. [19].

Dimensions	Definitions
Accessibility	The extent to which information is available, or easily and quickly retrievable.
Appropriate Amount of Information	The extent to which the volume of information is appropriate for the task at hand.
Believability	The extent to which information is regarded as trie and credible.
Completeness	The extent to which information is not missing and is of sufficient breadth and depth for the task at hand.
Concise Representation	The extent to which information is compactly represented.
Consistent Representation	The extent to which information is presented in the same format.
Ease of Manipulation	The extent to which information is easy to manipulate and apply to different tasks.
Free-of-Error	The extent to which information is correct and reliable.
Interpretability	The extent to which information is in appropriate languages, symbols, and units, and the definitions are clear.
Objectivity	The extent to which information is unbiased, unprejudiced, and impartial.
Relevancy	The extent to which information is applicable and helpful for the task at hand.
Reputation	The extent to which information is highly regarded in terms of its source or content.
Security	The extent to which access to information is restricted appropriately to maintain its security.
Timeliness	The extent to which information is sufficiently up-to-date for the task at hand.
Understandability	The extent to which information is easily comprehended.
Value-Added	The extent to which information is beneficial and provides advantages from its use.

SE and RE in Big Data

SE is a field that focuses on processes surrounding the lifecycle of software, from design, to development and maintenance, as detailed in SWEBOK [15]. While the use of data-intensive research is not new to environmental research, especially climate research [6], it is becoming more widely used, and as explained by Hampton et al., SE knowledge could be applied to facilitate such a transition [10]. Additionally, the field of SE includes the field of RE, which contains elicitation, analysis, specification, validation, and management of requirements throughout the software lifecycle [15]. In addition to being useful skills applicable to research, RE can promote further clarity and communication of requirements for research. RE focuses on tailoring requirements to relevant stakeholders, which are entities, people, or organizations with a vested interest in the system or project. For example, in environmental research, the main stakeholders are data producers who create and collect data, data consumers who use data for analysis, and data hosts who store data. RE can be particularly useful for understanding and adapting to emerging needs as research is pivoting towards more data-intensive research, which differs from traditional research as explained by Vogelsang and Borg [13].

2.2 Open Data Practices and Related Works

The following section will present related work for both environmental research and SE as it relates to both open data practices and data-intensive techniques. This includes how open data practices and data-intensive techniques fit into environmental research and the current understanding of these methods in environmental research. This will in turn set the context for how this thesis is positioned to contribute a suitable SE perspective.

Data in Environmental Research

The relationship of environmental research to data has changed greatly, from quite limited local or national affairs to global efforts both in the accumulation of knowledge and in struggles such as climate change that span the globe, as investigated by both Easterbrook and Edwards [6, 7]. Environmental research has always been a data-intensive field in many respects, but the way in which the available data are used and harnessed can vary greatly between research areas. Climate research, for example, has been at the forefront of data-intensive applications, as evidenced by the advanced climate models presented in Easterbrook's *Computing the Climate* [6], that assess and predict scenarios. These scenarios are in turn often used to guide policy, as exemplified by a report on climate change by Lee et al. [8]. Meanwhile, other fields such as ecology have recently been in the process of adopting big data practices, as mentioned by Farley et al. [16]. This process does not only come with the technical challenges requiring new skills explored by Hampton et al. [10], but also cultural challenges, as discussed by McCord et al. [5], such as increased transparency of how data quality is handled as the scale of research increases to a global scale rather than small interpersonal research teams. In addition, in certain areas, such as in-situ

marine data as highlighted by Nguyen et al., the amount of data available in open datasets is growing while production remains very expensive, which presses the importance of reuse to increase the value provided by the data [9].

Reuse of data can also be of particular importance for data-intensive methods such as ML, due to the significant data needs, but reusability as a quality aspect is especially important in such cases as careless reuse of data can be problematic from both a technical and ethical standpoint, as discussed by Paullada et al. [11]. For while technology and efforts toward open science have made data more available than ever, change is not isolated to technical but also social challenges. Changes to think globally and contribute to shared global knowledge infrastructure and challenges accompanying it have caught observations from, for example, Hampton et al. to urge for environmental research to adopt computational skills [10]. These challenges are not isolated to individual research, but are aspects, as Cheruvelil and Soranno point out, that take full advantage of open science, team science, and collaboration between research fields [18]. Additionally, new skills will be needed as data-driven methods, such as AI and ML, continue to advance which provides new possibilities and capabilities to environmental research, as explained by Zhu et al. [20].

Rise of Open Science and International Standards

While Open Science has come a long way in establishing large open databases that adhere to rigorous standards and procedures to ensure both quality and trust, there is still room for improvement, as explored by Huber et al. [21]. For example, while there are many high-quality databases, they often use different standards, making it difficult to combine and synthesize data between them. Similarly, searching for data between databases is another challenge as data continues to grow, making it difficult to find and retrieve data for a specific purpose. These issues have spurred the development and adoption of the FAIR guidelines (findable, accessible, interoperable, and reusable), which are a key component of open science to not only make quality data available, but to actually increase its use by making it easy to both find, access, and use the data [2]. The increasing abundance of data is enabling data-driven methods, such as ML, to thrive, but it is also increasing the need for data to be machine-readable and interoperable to ensure that these methods can easily work with the available data without extensive pre-processing, as discussed by Balbi et al. [17].

Challenges With Data Quality in AI and ML

This increased use of big data has not come without its challenges, none more obvious than the need for large amounts of high-quality data. Nguyen et al. have investigated which data quality dimensions are important for data producers to consider, and the types of pre-processing that can increase the value of the data, whether for ML or reuse [9]. Sambasivan et al. underlined how data-driven methods require special attention to not undervalue the importance of data quality, as such practices in AI can cascade in the system with negative ramifications [12]. For example, Paullada stressed that whether or not the data itself is of sufficient quality with documentation

and metadata to even support scientific ventures with data-intensive applications has to be confirmed, to avoid drawing misleading conclusions [11]. Amershi et al. also clearly illustrate that traditional SE methods have needed to develop new processes and techniques to incorporate data as a key pillar of software development [22]. Vogelsang and Borg explain how RE is one of the SE practices that has had to adapt to the unpredictability of data-driven methods [13], which was important as RE plays a key role in any successful and effective software project, as mentioned in SWEBOK [15]. For example, data quality requirements will be important functional requirements for the training data in ML, as the data determines the capabilities of ML models. This includes new data quality requirements for data collection and processing, both to fit the needs of ML and to increase the ability to combine datasets from different sources to form bigger and more capable datasets for ML. The main goal of RE is to reduce the ambiguity of requirements and use metrics, in this case, data quality metrics, to quantify and create tangible requirements that can be both validated and verified. Furthermore, these systems may run over long periods of time and therefore need to adapt over time as improvements occur or changes in what is being studied are made.

2.3 Purpose of the Study and Research Questions

In summary, the field of environmental research has made broad efforts to use big data and ML [16], which has required a re-examination of how to deal with data quality, with papers by Hampton et al. and McCord et al. discussing both the technical and cultural changes that are necessitated by such a shift [5, 10]. These papers mainly explore singular fields or different ways of using and incorporating ML in environmental research with different frameworks and standards, but unfortunately they are field specific and rarely universal. Additionally, explicit communication of requirements is unclear in these papers. As environmental research increasingly uses software tools [16], there is a potential to learn from RE to better meet and share expectations of new ML processes and techniques to harness their power and avoid their pitfalls [20]. While Easterbrook has argued from an SE perspective that environmental research can use very advanced or even superior software techniques compared to SE, we would like to look beyond the field of climate research and apply an SE perspective on environmental research as a whole [6].

Environmental research is vast and many researchers have only recently adopted data-intensive software, such as the efforts in ecology to adopt big data, and environmental researchers like Hampton et al. have called for additional computational skills in environmental research [10, 16]. Furthermore, as Easterbrook illustrates in his book [6], SE studies in environmental research can lead to an exchange of ideas in both directions and benefit both fields. To explore this, the following research questions will be used and answered in this thesis:

RQ1: What is the understanding of data requirements for data-intensive models and software in environmental research from the perspective of data producers and data consumers?
Data used in different contexts and by different practitioners have different demands for quality. This question seeks to explore what those different demands are, and to examine the underlying context that leads to these different demands to gain a greater understanding of what creates the needs for data quality.

RQ2: What are the challenges of current methods for verifying and validating data quality for data-intensive models and software in environmental research?
Quality requirements serve little purpose if they are not being met. Verifying that the required data quality is reached, and validating that it has the desired effect is necessary to make effective use of the planned requirements. This question seeks to explore how this process is currently being approached by practitioners and what they are doing to ensure data quality, as well as what parts of this process needs improvements.

RQ3: What recommendations can we conclude based on approaches known to software engineering, such as requirements engineering, to improve the process of handling and managing data quality in environmental research?
There is a need for improvement with regards to how data quality is handled in environmental research. Based on the findings of **RQ1** and **RQ2**, this question is aimed at exploring what processes and techniques can be borrowed from the SE field to help with the process of communicating data quality and the needs thereof.

3

Methods

This section outlines the methods used to answer the research questions. These methods include interviews and a thematic analysis to understand data quality in environmental research to answer **RQ1** and **RQ2**. This is followed by a focus group and literature review in SE to formulate recommendations and finally a survey to validate and iterate the recommendations and challenges identified, in order to answer **RQ3**. How these methods relate to each other and the research questions can also be seen in figure 3.1.

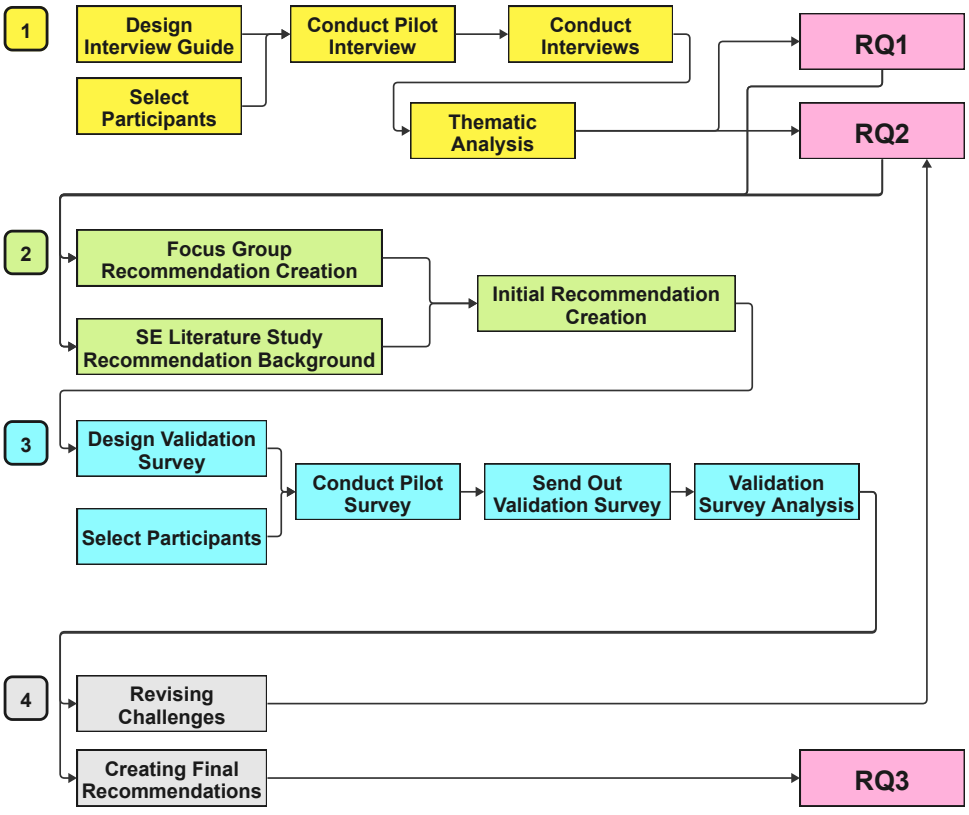


Figure 3.1: Overview of Methodology. The numbers correspond to different methods used. 1 = interview of data quality in environmental research, 2 = focus group and SE literature study to create initial recommendations, 3 = validation survey for challenges and recommendations, 4 = revision of challenges and final recommendations.

3.1 Interview Study With Environmental Researchers

A field study, as described by Stol and Fitzgerald, was conducted in the form of an interview study with the intention of creating a qualitative understanding of how data quality is handled in practice. This includes workflows, practices, standards, and difficulties that are currently encountered by environmental researchers [23]. This qualitative understanding provided preliminary answers to **RQ1** and **RQ2**.

Table 3.1: A table of all eleven interviewees with a unique ID for each interviewee and their associated role, research focus, and experience in environmental research.

ID	Role	Research Focus	Experience*
P1	Data producer	Ocean monitoring	7 years
P2	Data consumer	Oceanography	20 years
P3	Data consumer	Climate modelling	1-2 years
P4	Data producer	Oceanography	7 years
P5	Data producer	Ocean monitoring	21-22 years
P6	Data producer	Air pollution monitoring	6-7 years
P7	Data producer	Air pollution monitoring	28 years
P8	Data producer	Oceanography	12-13 years
P9	Data consumer	Marine environmental monitoring	14 years
P10	Data producer	Environmental contaminants	15 years
P11	Data consumer	Deposition from air pollution	2 years
* Years of experience in environmental research			

The sampling procedure for the interviewees was conducted with a mix of purposeful sampling as outlined by Suri [24], convenience sampling following the approach of Sedgwick [25], and also snowball sampling as described by Naderifar et al. [26]. The final sample can be seen in table 3.1. The IDs of the interviewees were randomly allocated and do not reflect the order in which the interviews were conducted, to ensure the interviewees' anonymity. Known contacts were primarily recruited to be interviewees, but these contacts were also asked for further potential interviewees who were then recruited for the study. However, the purposeful part of the sampling was ensured by defining criteria for who was asked to participate in the study. The selected interviewees were all involved in, or worked closely with, environmental research, and they were all part of our two main stakeholders: data producers and data consumers. Though some interviewees had experience in both roles, all participants could classify themselves as at least one of the stakeholders.

The interviews were primarily conducted virtually through Microsoft Teams, though two of the interviews were conducted in person. To keep the interview format consistent, each interview had an interviewer and a note-taker present, though these roles were switched between each interview. All the interviews, including the interviews in person, were also recorded and transcribed through Microsoft Teams so that the interviews could be reviewed in detail after they had been conducted.

The interview study was designed with a *general interview guide approach* as suggested by Turner [27]. This entailed the preparation of an interview guide before the interviews were conducted. The interview guide, which can be found in appendix A, outlined a general structure of the interview and specific questions. However, the interviewers were allowed to deviate from the guide and ask follow-up questions as they deemed necessary during the interview. This method was used to create a consistent structure between all the interviews, while also allowing for flexibility. The interview guide also included a presentation that was used during the interview to clarify the structure and some of the questions. The first interview also acted as a pilot interview to verify the interview guide. As no major changes were made to the interview guide after the pilot interview, the data collected during the pilot interview was also used for the study.

The questions were divided into five categories:

1. **Professional Background**
   The main aim of this category was to gain an understanding of who the interviewee was and their personal history with environmental research. It consisted of questions regarding the interviewee's research focus, whether they would classify themselves as a data consumer or a data producer, and whether or not they had any experience working with ML or other data-intensive applications.

2. **Data Quality**
   This category brought the focus over to the interviewee's opinions and experiences in dealing with data quality and a few concepts that help to concretize data quality. This category made use of a list of information quality dimensions, referred to as data quality attributes in the interview guide as a more common term, specified by Kahn et al. in the paper *Information Quality Benchmarks: Product and Service Performance* to ask the interviewees which data quality attributes they thought were the most and least important [19]. Questions about what roles metrics and metadata take for ensuring data quality were also placed in this category.

3. **Communication**
   After questions relating to data quality, this category intended to investigate the communication of different data quality needs between data consumers and data producers. The majority of the questions in this category had two versions: one to be asked to data consumers and another for data producers. The interviewees did not always fit neatly into one of these roles, but in such cases the perspective that most closely matched their experience was used. For instance, an interviewee might have had experience as both a data consumer and as a data producer, but in such a case the perspective they had the most experience with, or the one they were mainly concerned with at the time of the interview, could be used.

4. **Data Standards and Directives**
   This category asked how the interviewees made use of data quality standards, and what roles such standards have in ensuring data quality. Like the *Communication* category, this was asked in different ways to data consumers and data producers, to ask for their specific perspective.

5. **Open Science and the Reuse of Data**
   The final category explores open datasets and databases, bringing our third stakeholder, data hosts more into focus. The questions were still angled for the perspectives of either data consumers or data producers, as appropriate for the interviewee, but they now focused more on how the interviewees interact with open data, and what benefits and challenges that brings compared to working with private datasets.

3.2 Thematic Analysis of the Interview Study

The data collected from the interviews were analyzed through the use of thematic analysis as described by Braun and Clarke [28]. Thematic analysis was chosen as the method of analysis due to its flexibility. The possibility to use it to find patterns of experiences in qualitative data is what made it suitable for this study. Additionally, thematic analysis is subjective, which is beneficial for this analysis as we want to view the field of environmental research, that has been captured in the interviews, through the perspective of software engineers. The freedom and flexibility that comes with thematic analysis enabled us to do this, while still maintaining a rigorous analysis through a structured iterative process. The collected data that was analyzed was in the form of recordings of the interviews and transcriptions that had been automatically generated through Microsoft Teams. The thematic analysis process was divided into six phases:

1. Familiarization
2. Coding
3. Generating initial themes
4. Developing and reviewing themes
5. Refining, defining, and naming themes
6. Writing analysis

However, the analysis was not a linear process, but rather an iterative one. Some phases were done multiple times, and earlier phases could always be revisited as necessary.

Familiarization
The familiarization phase is about gaining an understanding of the data, both through deep immersion and by examining the data critically [28]. An initial familiarization had already been done during the interviews, as we collected the data ourselves. Further familiarization was done through listening through the recordings and reading

through the transcripts. This familiarization also coincided with manual correction of the automatic transcripts, as they often had inaccuracies.

Coding

Once the data had been transcribed, the coding process began. The researchers first created codes individually and then traded them with each other to review, verify, and refine the codes. This was repeated so that each datapoint was reviewed at least twice, and all codes were then reviewed with in depth discussions. The codes were created manually, without the use of any particular qualitative data analysis software. As the main purpose of the interview study was to gain an understanding of the experiences and perspectives of environmental researchers, the coding process focused more on being inductive rather than deductive, as the aim was to gain an understanding of the interviewees' responses [28]. Initially, the codes were mainly semantic, but more latent codes were created during the later iterations of the coding process.

Creating Themes

Once the coding process was deemed sufficiently completed, the theming process was started. Much like the coding process, the theming was also done iteratively, which is emphasised by Braun and Clarke as theming spans over three of their six phases [28]. Initially, candidate themes were created in a wide scope to catch several avenues of looking at the data. These themes were created by grouping the codes by subject, and therefore also creating a better overview on what the interviews as a whole mentioned about certain subjects. This helped to highlight both what the interviewees in general agreed on, and where they had differing or conflicting opinions. Some codes were also discarded at this stage, as they were not relevant to the research questions. Key insights were then extracted from the clusters to create the candidate themes.

A second grouping was then done with the candidate themes to further refine the themes. This step provided a more holistic and clear overview of the candidate themes and how they are connected. Like in the previous grouping, some candidate themes were also discarded at this stage if they lacked depth, or were not relevant or useful in answering the research questions. These new groupings were then further refined to generate the final themes.

Writing and Formalizing Analysis

The final part of the analysis was to formalize and write up the findings. This part of the process took the looser and more implicit thoughts, notes, and verbally communicated ideas that lay behind the codes and themes and refined them into the more explicit and formal results presented in section 4.1. This part had a large overlap with the later stages of the theming process: much of the work to define and refine the themes was done by detailing them in writing and thus putting them in explicit terms.

3.3 Recommendation Design and Validation

The creation of the recommendations involved two parts to answer **RQ3**: a literature study in SE and a focus group with SE experts. The first step was to conduct a short literature study to investigate if the themes, which represent identified challenges, were also present in SE, and if so what measures had been taken to deal with them.

The second step of creating the recommendations was to hold a small focus group with SE experts, described as a judgment study by Stol and Fitzgerald [23]. During this focus group, the experts, two PhD students and a professor of SE, were presented with the findings from the interviews, and approaches were discussed for solving or amending the issues environmental research practitioners encounter. A set of questions was created before the focus group to act as a guideline for potential discussion points to aid the flow of the discussion. This focus group was based on the design steps presented by Hevner and Chatterjee [29], though it was adapted to be on a smaller scale, both due to limited access to SE experts and as this focus group was not meant to create the recommendations by itself, but rather simply generate ideas for them.

After the focus group, the literature study in SE was used to support the potential recommendations identified during the focus group and to potentially find additional recommendations. The literature study was also expanded upon to ensure that it covered all the topics discussed during the focus group. The literature served the purpose of both helping to refine the recommendations and to ground them in SE practices.

Recommendation Validation Survey

Once the initial recommendations were completed, a survey was created to validate the recommendations with environmental research practitioners. This form can be found in appendix D. As the survey was a validation of the interview study rather than broadly generalizing the findings, it was a continuation of the field study [23]. The purpose of this survey was to verify how feasible environmental research practitioners thought the recommendations were. The main concern was whether they thought the recommendations would be helpful and beneficial to their practices.

This survey was sent to both the interviewees and other contacts generated during the snowball sampling of the interviews, as well as additional snowball sampling done during the survey itself. Before the survey was sent to real respondents, a Master student dealing with similar themes of data quality in environmental research was used as a pilot respondent. The validation gained from the pilot respondent, along with feedback received from the supervisor of the thesis, helped to refine the survey into its final version. There were eleven respondents to the survey, with an overrepresentation of the following three categories: people with multiple roles, oceanographers, people with many years of experience, and previous interviewees. The respondents are presented in table 3.2. The survey was sent to a total of 30 practitioners, giving it a respondent rate of roughly 36.7%.

Table 3.2: All eleven survey respondents with a unique ID for each respondent and related information on roles, fields, years of experience, and whether they participated in the previous interview.

ID	Role	Field	Experience*	Interview participant
R1	Data producer Data host	Oceanography	6-8 years	Yes
R2	Data consumer	Air pollution monitoring	3-5 years	Yes
R3	Data producer	Air pollution monitoring Atmospheric science	9+ years	Yes
R4	Data producer	Oceanography Climate research	9+ years	Yes
R5	Data consumer	Oceanography	1-2 years	Yes
R6	Data producer Data consumer Data host	Oceanography Climate research	6-8 years	No
R7	Other**	Oceanography Climate research Ecology Other	9+ years	Yes
R8	Data producer Data consumer	Oceanography Meteorology Climate research	9+ years	Yes
R9	Data producer Data consumer	Air pollution monitoring	9+ years	Yes
R10	Data producer Data consumer Data host	Oceanography	9+ years	No
R11	Data producer Data consumer Data host	Oceanography	9+ years	No

* Years of experience in environmental research
** Management role

The survey was created and shared via Microsoft Form, with no login required to complete the survey to minimize the barrier to respond. The survey was open for one week with no time limit to complete the survey. The structure consisted of three parts: personal background, challenges and recommendations, and concluding remarks. The personal background included questions about roles, current environmental field, years of experience, and whether they participated as a respondent in the previous interview study. The challenges and recommendations where each presented individually to evaluate whether the participants agreed or disagreed with the statements. The concluding remarks were mainly for snowball sampling and whether they would like to receive the results of the thesis via e-mail.

Each recommendation had descriptions of the observed challenge in environmental research, background information on a similar aspect in SE, the recommendation itself to solve the asserted challenge, and an example of the recommendation in use. The challenges corresponded to the themes from the thematic analysis and were evaluated separately to verify the extent to which participants agreed that the challenge was a problem in environmental research.

This survey focused primarily on closed questions rather than open questions, as its main purpose was to evaluate the already created recommendations rather than to identify new ones [30], though it had the potential to do that as well. However, the questions did allow for an open response, for example, if a respondent wanted to comment on why they thought a particular recommendation would or would not work well, but such responses were not mandatory.

Recommendation Analysis

To analyze the survey we used descriptive statistics in order to extract the key insights from the results. According to Wienclaw, descriptive statistics can be used to summarize the dispersion of the data in a distribution which is useful for this survey [31]. Furthermore, due to the low number of responses, conventional inferences using frequentist approaches from the data are not feasible and thus we mainly describe the data, which is the purpose of descriptive statistics. Likert-scale questions were analyzed specifically with different groups in mind in order to compare responses and examine potential biases in the sample.

This included an analysis of the distribution of the survey respondents in terms of their roles, fields, and levels of experience, as well as if they were part of the interview study or not. Comparisons were made by dividing the respondents into groups, such as data consumers and data producers. Likewise, bias in the results could be investigated where a certain group was overrepresented in the data as it could be compared to the remaining sample or other groups. However, not all groups could be investigated in this way, as they were only represented by one or two respondents.

The open questions were mainly used to allow respondents to evaluate and add insights to our recommendations and themes. Of special interest were comments with negative sentiment or aspects not outlined in the recommendations, even though neutral or positive sentiment could also be highlighted if they provided additional insights beyond agreement with the recommendation.

Recommendation Refinement

With the validation from the survey on the recommendations, a final refinement iteration of the recommendations was conducted. This aimed to make final adjustments to the recommendations to improve their feasibility, understandability, and usefulness for environmental research. The survey provided insight into which parts of the recommendations were well received, and should therefore be kept or further emphasized, and which parts received more negative feedback and therefore needed to be changed. The open questions also provided tangible ideas for improvements that were incorporated into the recommendations. In short, the last step was to use the results of the survey analysis to improve the recommendations and make them more relevant to environmental research.

4

Results

This chapter presents the results including the themes derived from the interview study and recommendations based on the identified challenges. This includes the initial recommendations based on the focus group and SE literature, as well as the final recommendations based on the survey results, which are also detailed here.

4.1 Interview Themes

The analysis of the interviews resulted in six themes, labeled T1 to T6, where each theme can be further divided into three subthemes, illustrated in figure 4.1. These themes answer the research questions **RQ1** and **RQ2**. Here we will describe each theme and its subthemes.

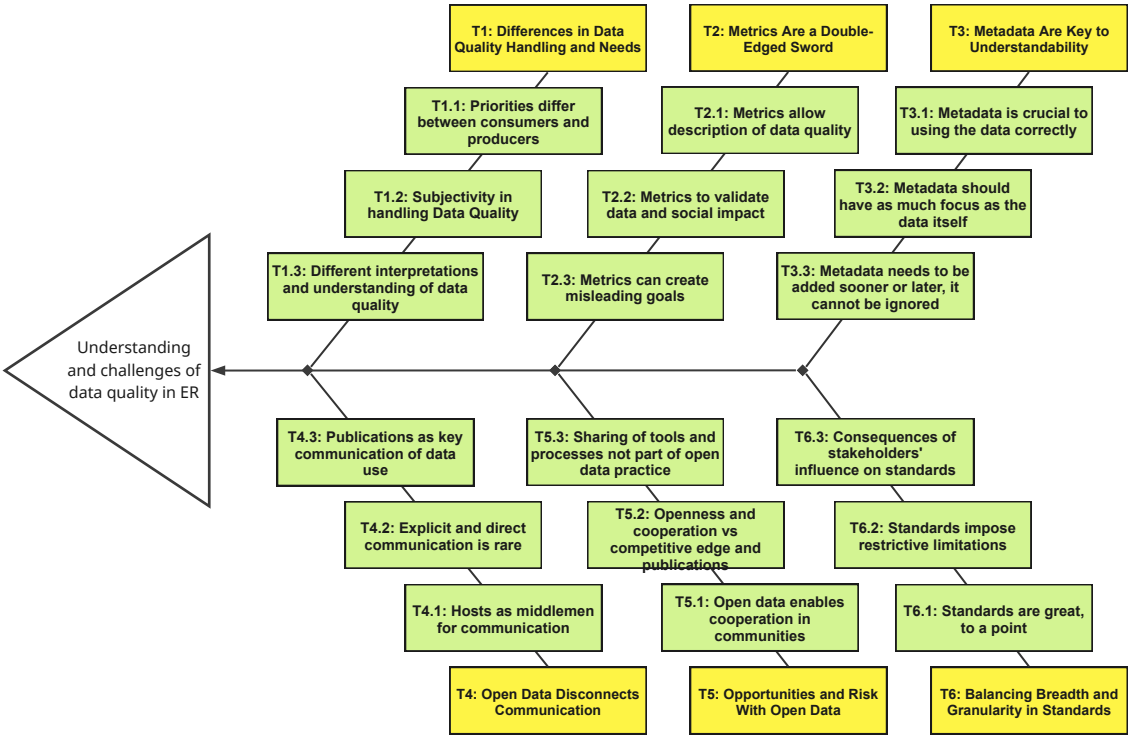


Figure 4.1: This diagram illustrates the themes and subthemes from the thematic analysis. The themes are placed at the outer edges with associated subthemes on the connected line.

T1 Differences in Data Quality Handling and Needs

This theme concerns how researchers with different individual needs and background may have different interpretations and understanding regarding the handling of data quality. Figure 4.2 shows which data quality attributes the interviewees thought were the most and least important. This figure highlights both the wide spread of opinions, but also how some attributes that are seen as the most important by some are seen as the least important by others, illustrating how perspectives can differ greatly between practitioners. It should be noted that data quality attributes being rated among the least important here does not mean that they are unimportant, several interviewees mentioned that all the listed attributes were important, but rather that those attributes are less prioritized.

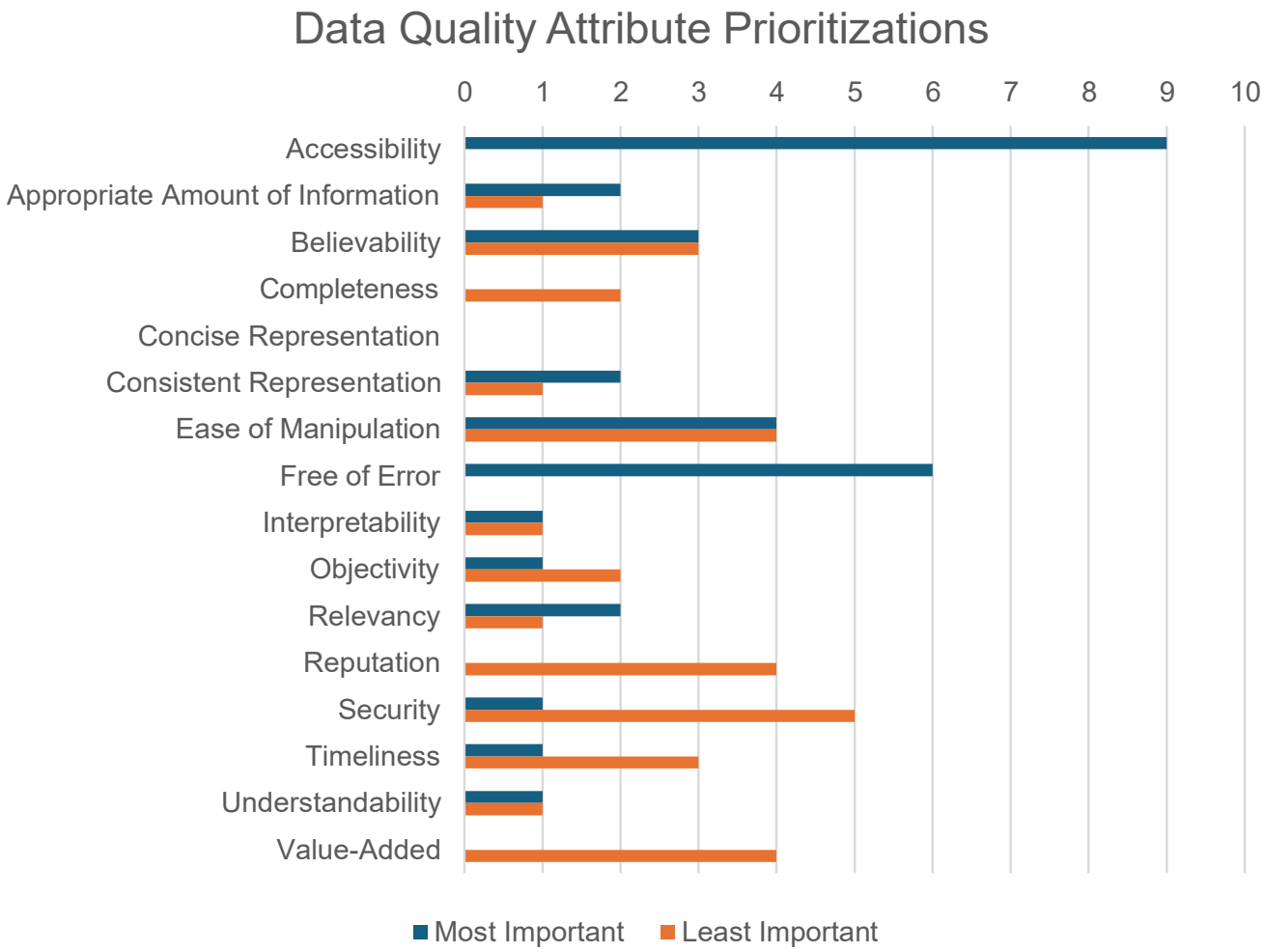


Figure 4.2: A graph of how data quality attributes were prioritized according to the interviews, where the interviewees were asked which attributes they thought were the three most important ones, and which they thought were the three least important ones.

T1.1 Priorities Differ Between Consumers and Producers

> "So a researcher, pure researcher, have other goals and other interests than the monitoring community and the data the respective community is uploading will be different." - P1

The different data quality attributes that are prioritized in environmental research data change from the point in time when data is collected to the point in time when the data are used. This leads to different prioritizations between data consumers and data producers as both their needs and their tasks differ, which can be seen in figure 4.3. While the interviewees rarely expressed any data quality attribute as unimportant, the relevancy of any attribute varied depending on their role. For example, that ease of manipulation may be less important to some data producers because it does not add additional information about the data. On the other hand, it provides a useful utility that may be critical for some data consumers to be able to use the data more easily. Likewise, how data quality attributes are prioritized can also vary between different data consumers depending on their field and experience, and similarly between different data producers.

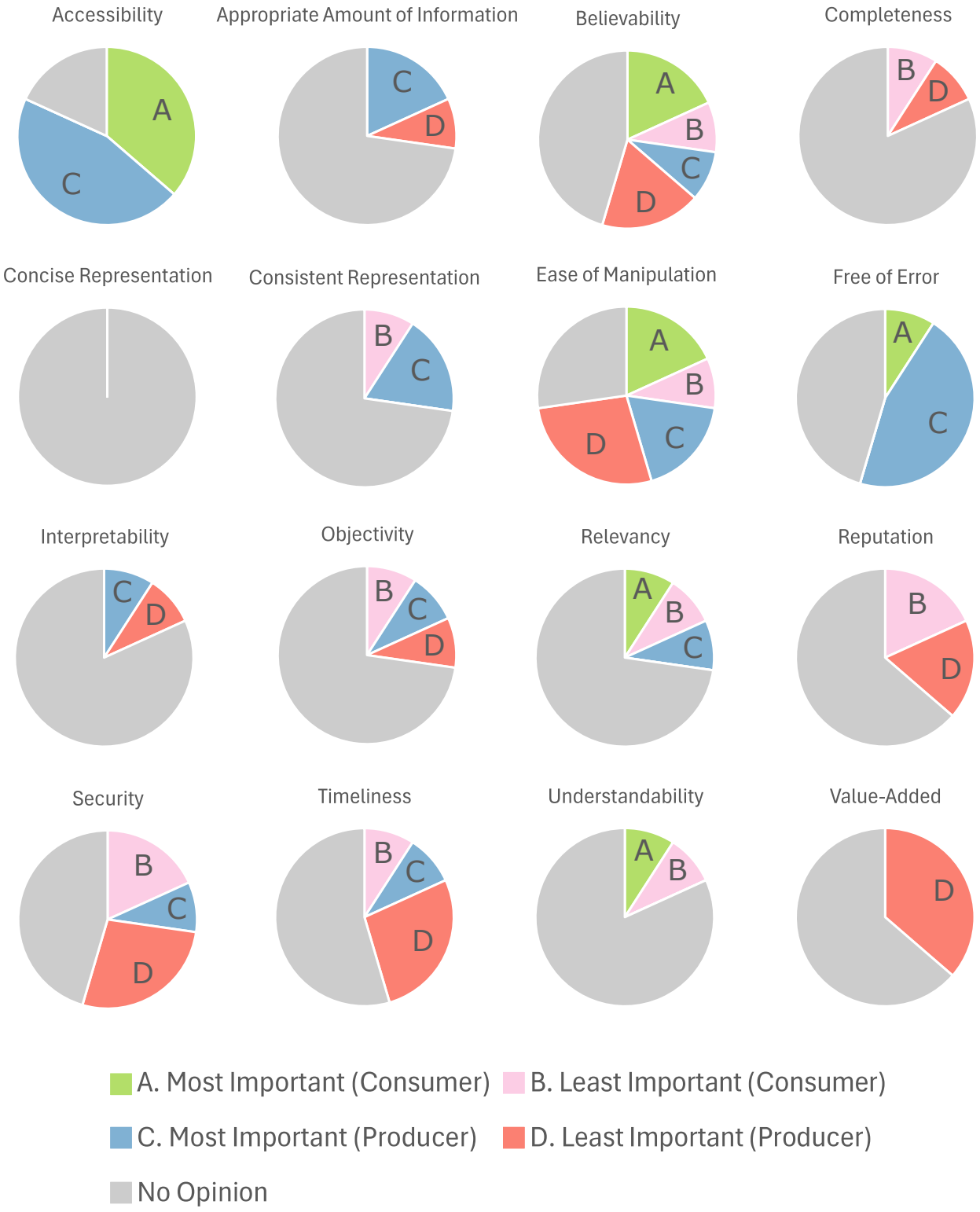


Figure 4.3: A set of pie charts showing the opinions of data consumers and data producers for what data quality attributes they thought were the most or least important for their work.

T1.2 Subjectivity in Handling Data Quality

> "So that could be, that we found a problem and we were not following their protocols because we felt we had a higher responsibility for the data." - P6

Along with different priorities, data quality is often handled in a subjective manner. Stakeholders often have at least some room to use their own judgment in what is and what is not good practice for handling data quality. This leads to situations in which even colleagues at the same workplace might take different stances on how certain aspects of data quality should be handled. This is beneficial as it allows for freedom of individual judgment, but it can also potentially be a risk to the consistency of the data quality.

T1.3 Different Interpretations and Understanding of Data Quality

> "How do they interpret what we write? For example, what would if we see some increased values but we know that it has been a forest fire close by and we write that. Will the data users really get that information and how will they use it?" - P6

Through the analysis, we also identified discrepancies between how different practitioners interpret different aspects of data quality. The data quality attributes from Kahn et al. that were used during the interviews were on a more granular scale than many of our interviewees were using in their own work [19], and instead they typically made use of principles such as the FAIR guiding principles or various data quality standards [2]. What practitioners are used to working with influences their perspective, meaning that when faced with something they are not familiar or accustomed to, practitioners might interpret the same thing in different ways.

T2 Metrics are a Double-Edged Sword

Our analysis of the interviews show that metrics for data quality in environmental research have a clear and useful purpose as a measure to quantify data quality. However, they can also create misleading goals by becoming a goal by themselves and thus obscuring the actual purpose of the data collection.

T2.1 Metrics Allow a Description of Data Quality

> "So for example, data can be flagged as bad or suspicious, and this is extremely important for our data users to see the quality flags." - P9

Metrics serve an important role of quantifying data quality, marking metrics as being clearly important because they create a common language for how data quality can be communicated. This common language is important as it both allows data consumers to communicate what level of data quality they need, and allows data producers to communicate the quality of the data they produce.

T2.2 Metrics to Validate Data and Social Impact

> "OK, well for me [metrics are] important because I need to be able to talk about what my scientific and societal impact is [...]" - P5

Metrics as a common language is also important to validate the data and the impact it has on a societal level. Without clear metrics, it is hard to tell what level of data quality to strive for, and it is also hard to tell if the data that are produced meets the expectations that were set for them. Metrics also serve as a measure of the work the data producers are doing, and as a way for them to present their productivity. This measure of social impact through metrics is an important evaluation of the usefulness and ethics of data collection, and plays a key role in the continued funding of projects and monitoring programs.

T2.3 Metrics Can Create Misleading Goals

> "[...] if you're trying to metricized some things, you can end up chasing a number that's kind of detached from reality." - P4

Even with the clear benefits and necessity of metrics, the analysis also indicated that metrics can become problematic if they are used without care. Metrics can give very clear goalposts of what is to be accomplished, but that can also make it easy to neglect to verify what the benefits of those goalposts are, and if they are useful. In such cases, the metrics themselves might become the goal, and the knowledge or benefit that was to be reached can become obscured.

T3 Metadata are Key to Understandability

The analysis highlights that metadata is an incredibly important complement to data and it is clear that metadata are prioritized and effort is oriented to create good metadata. Some challenges have been highlighted, but to what extent they are being resolved or remain a problem is not clear from the data collected from the interviews.

T3.1 Metadata is Crucial to Using the Data Correctly

> "[...] you also have metadata protocols that you can download, and I think that you can actually get a pretty good picture of what you actually look at, but you have to take the time to do that, you can't just use the data [...]" - P8

There is widespread agreement among the interviewees that without metadata it is difficult to make any use of data. That it includes crucial details needed in order to analyze and process the data. One such aspect is describing context which can be especially important for data errors and other irregularities. Furthermore, metadata is the primary way to validate data, especially for data consumers, and without appropriate metadata both trusting and using the data becomes a far more costly task, or it may even become impossible.

T3.2 Metadata Should Have as Much Focus as the Data Itself

> "I would say I spent probably as much time on the metadata as on the data itself when it comes to our processing." - P4

The importance of metadata also means a lot of effort is put towards creating good metadata. The general sentiment from the interviewees is that this can be quite resource intensive, with one interviewee even mentioning that they spend as much time on the metadata as they do on the data itself. The creation and documentation of metadata is thus both time-consuming and costly, but these are costs that are necessary for ensuring the usability of the data.

T3.3 Metadata Needs to Be Added Sooner or Later, It Cannot Be Ignored

> "And I've just heard a project recently, someone went back into these old archives and found a lot of new data that was just sitting there, but then often does data not have the right metadata. There will be hard to judge what has been done and so that comes with extra work and that increase uncertainties." - P2

The importance of metadata and associated cost also entails major consequences if metadata creation is postponed. Interviewees who identified as data consumers noted the difficulty in dealing with missing or lacking metadata, especially with old or private data it can be difficult to find or contact the source of the data. Likewise, most data producers in our interviews regarded themselves as the most capable in understanding and handling their data as they are a first-hand source of the data's creation. That also means that the ability to make use of and correctly verify the data decreases with time and when producers are not involved.

T4 Open Data Disconnects Communication

While communication between data consumers and data producers varies, the sentiment from the interviewees is that communication in open science is not particularly frequent. Therefore, expectations on data quality and responsibilities can vary depending on the shared involvement and cooperation among data producers and consumers.

T4.1 Hosts As Middlemen for Communication

> "So although I do talk to some direct data users, a lot of my time and effort is more talking to data aggregators and sort of organizers." - P4

As open science is increasingly embraced, data hosts have played an increasingly important role as aggregators of data. This has resulted in less communication between data producers and data consumers, as most communication is directed towards data hosts instead. Data hosts provide some communication opportunities by increasing the findability of datasets and, therefore, also the

associated data producers' contact information. However, these communication opportunities are often not taken advantage of.

T4.2 Explicit and Direct Communication Is Rare

> "So if we think about researchers who are gonna use the data, we're not often in directly in contact with those." - P10

> "So when I am a data user I, depends, but most of the time I prefer not to [communicate with producers]. So if I can, do avoid." - P2

Direct communication between data consumers and data producers often relies on social connections and the scope of a project. While avenues for communication, such as e-mail, are almost always present, they are rarely used. From our interviewees, seven out of eleven share the view that the expectations for responsibilities between data consumers and producers are not always clear, and rarely expressed explicitly. Similarly, whether these expectations are met or not is equally obscure.

T4.3 Publications as Key Communication of Data Use

> "[...] I would say we're still using the very traditional scientific channels of communication, which is you publish a paper, you give a talk and that's how people know what you're doing, yeah." - P4

Producers rarely receive direct feedback that their data is being used and instead gain this knowledge when papers citing their data are published. For data producers, this means the usefulness of their data is unknown until the point of publication of papers. This, along with open science often involving unknown data consumers, means that the time to receive feedback from data usage can be quite slow. Many interviewees that are producers were asking for more communication and feedback regarding data while many of the data consumers we interviewed preferred the opposite: to have adequate data quality that contains all necessary information, rather than having to rely on communication.

T5 Opportunities and Risks With Open Data

Overall the adoption and growth of open science is highly prioritized within the field of environmental research; a push that is mainly encouraged through standards like FAIR. Despite these efforts, there are still many challenges with open data regardless of the benefits of open science.

T5.1 Open Data Enables Cooperation in Communities

> "I think duplication of effort is a huge issue within our field where a lot of people just are privately working on the same methods and not communicating with each other where it would be much more effective to collaborate openly on it." - P4

All interviewees were hopeful in terms of increasing the amount of open data in environmental research. It is clear that the interviewees understand and regard the FAIR principles highly to enable and build towards open science. The interviewees also emphasized that, in their experience, open data means that the validation and use of data is increased, which is particularly important for data that is expensive to produce. Furthermore, to decrease the risks of data being produced and only used once, the interviewees express that environmental research as a field wants to spur a culture where open data is always the norm.

T5.2 Openness and Cooperation vs. Competitive Edge and Publications

> "There's a lot of pushback by some countries against open data, but that's mainly because if they're smaller research programs or from countries with developing research programs, they don't like the idea that there are more established nations and researchers in those nations who have more money and larger research teams come in and use their data and publish it before they might get the chance to." - P5

While open data does enable many types of cooperation, there are some challenges that can dissuade or delay researchers in making their data openly available. Publications are an important part of building a reputation as a researcher. If a researcher publishes their data early in a project, there is a fear that other researchers can publish analyses before the original researcher. The associated risk is that the original researcher would not gain the same credit for their data as they would by publishing their paper before releasing their data. This, in addition to publishing data being both laborious and time-consuming, means open data can often be postponed or neglected. Additionally, not all researchers have the resources to spend time on publishing data, and they might need to prioritize other tasks that are more financially beneficial to themselves.

T5.3 Sharing of Tools and Processes Not Part of Open Data Practice

> "I think it'll be great if people would upload their code and just accept other people coming through it, finding errors, putting in pull requests, and that being a good thing, not an embarrassing thing." - P4

One aspect that is rarely shared or included in metadata is the data processing and tools used for data handling. Some interviewees mentioned that environmental research does not have a culture of sharing software tools even though it could be of great benefit to the community. They also mentioned that there

is shame associated with sharing homemade tools. Environmental research is a field that relies on publications where scrutiny can lead to a paper being withdrawn, thus damaging a researcher's reputation. This means there is a perceived risk to reputation when sharing software tools that might contain bugs or other problems. Additionally, motivating work that does not relate to publishing papers can often be hard as it does not always lead to obvious benefits.

T6 Balancing Breadth and Granularity in Standards

Data standards have a very important role in environmental research; they are key to enabling various forms of cooperation for various practitioners, ranging from large companies to lone researchers. They determine what quality data needs to have, what metadata has to be associated with it, and what format the data is presented in, all of which are important aspects for combining or comparing data. However, as different practitioners have different needs and preferences, creating standards that fit many practitioners can be difficult. On one hand, standards need to be flexible, so that they can be used by many practitioners and cater to their needs. On the other hand, they need to be firm, to enforce that their expectations and formats are being met, that the standard is actually followed. Striking a balance between these aspects is challenging, with a standard losing what makes it useful if it is too flexible, and driving practitioners away if it is too strict.

T6.1 Standards Are Great, to a Point

> "So they have like a broad definition of people from everywhere should come and be able to take down the data, but that also means that it's less specific and it's harder to, for example, just pull up out our data and it's hard to have the metadata in there because you're collecting data from all of Europe, for example, and you added it into one database." - P10

As environmental research encompasses a wide range of fields, there is an even wider range of standards available. This can be problematic, as it can fragment the openly available data, scattering it over many different databases. Finding and combining data from many databases can be costly, and as such it is easier to work with the data when it is consolidated into fewer, larger databases. However, making broad standards that fit a wide variety of data can also be problematic, as that stops them from fitting the more specific details of the data. If databases become broad, they also become more shallow, and vice-versa, so finding an appropriate balance is difficult.

T6.2 Standards Impose Restrictive Limitations

> "When reporting, make metadata myself and it is a standardized way of reporting them that often doesn't fit to the data I have. So yeah, it's a bit of a hen and egg situation now." - P7

Standards can also be difficult to work with because while they both define what is needed to fulfill them, they also limit what a data producer can include when publishing their data. For instance, a data producer might collect metadata that they think is important but that has no room in the standard, and therefore cannot be included.

These limitations can lead data producers to either use other standards, to deviate from the standards, or to host a separate version of their data on their own database. While this allows them to share and include all data and metadata, it also feeds into the problem of fragmenting the publicly available data over many databases, rather than having it consolidated and easily accessible.

T6.3 Consequences of Stakeholders' Influence on Standards

> "So I think you can't always decide for yourself what is useful or not, because it's also it comes from a central place where hopefully smart people have thought about different aspects. [...] we don't have the choice for the reporting that we do of our data to really question the data standards." - P10

While data hosts, consumers, and producers all have influence for how data standards develop and change, data hosts have a larger role in deciding what standards are actually used and how they are enforced. As aggregators of data, data hosts are the arbiters of which data standards are used in practice, and therefore also what data can be accessed. This is especially true when research relies on big and reputable data hosts for their large datasets or long time series datasets where there is rarely an equivalent alternative. However, this can create friction when the standards that are used do not match the needs of data consumers and data producers.

4.2 Initial Challenges and Recommendations

This section presents the initial challenge definitions and their associated recommendations, as they were represented in the validation survey. The initial challenge definitions are condensed versions of the themes presented in section 4.1, while the initial recommendations are a processed result of the focus group. These initial recommendations represent a preliminary answer to **RQ3**, and the final version can be found in section 4.5. The main points and recommendations from the focus group were extrapolated and compared against SE literature from the literature study. This generated nine initial recommendations categorized in the different themes of our analysis with no recommendations for the challenges *Metadata are Key to Understandability* and *Balancing Breadth and Granularity in Standards* as

initial recommendations for these themes were found in neither the focus group nor the literature study. Some of these recommendations are mainly based on the points highlighted in the focus group and common SE practices, like using GitHub, while some have stronger associations with SE literature. Here the challenges and their associated recommendations will be presented. Each recommendation also has three additional components: a background about the practice or concept, a reference to how the recommendation relates to SE literature, and an example of how the recommendation might be used in environmental research. The *SE Literature* components of the recommendations were not present in the validation survey, but they are presented here to show how the recommendations connect to SE practices.

A Metrics are a Double-Edged Sword
Metrics play an important role in quantifying and communicating data quality. However, metrics can also create arbitrary targets that may be abstract, unfounded, or misleading, and therefore not represent useful goals.

Recommendation A-1
Environmental research should use the goal question metric (GQM) model to avoid the pitfalls of metrics that create arbitrary targets, and to clarify the context and purpose of metrics.

Background A-1 The GQM model is a method that connects metrics to their intended purpose by contextualizing what they are meant to accomplish.

SE Literature A-1 Basili et al. outline the GQM method as a more purposeful way to define metrics with clear traces between metrics and associated goals [32]. The method involves defining goals that are refined into smaller questions that can be answered and quantified by means of metrics.

Example A-1 Having a clearly stated goal allows a researcher to more clearly determine the usefulness of a metric, allowing them to better use the metric to fit their goal, rather than just finding a "good enough" value.

B Open Data Disconnects Communication
In open data which has been widely adopted in environmental research, data hosts play an important role as data aggregators, which has resulted in more communication being directed to data hosts at the expense of less communication between data producers and data consumers.

Recommendation B-1
Data hosts should act as community platforms that encourage communication, collaboration, and feedback sharing between data producers and data consumers.

Background B-1 GitHub is a platform where developers can store, share, and collaborate on code for software projects [33]. It allows open source developers to easily collaborate, track changes, manage projects, and contribute to open source software.

SE Literature B-1 Xiao et al. describes open source software as a *sociotechnical* system where even without formal processes the participants in open source software still interact to compute the necessary requirements and results for the project to succeed as a collaborative cognitive system [34].

Example B-1 Data hosts such as Copernicus could provide communication channels similar to those on github, where each project has an open discussion forum for questions and comments.

Recommendation B-2
Data producers should maintain a dataset after it is published. They can respond to feedback and adapt the dataset, for example by adding new processing or adding new metadata.

Background B-2 Requirements in software projects often change over time as new features are added or a deeper understanding of the problem is gained, requiring an iterative development approach, as reflected in GitHub projects that evolve based on community feedback.

SE Literature B-2 Ramirez-Mora et al. emphasize the importance of communication in solving problems in open source software, and the ability to both find problems and work on them continuously with appropriate communication to achieve both a good community and rapid resolution of problems [35].

Example B-2 A data consumer might report a potential improvement to the dataset or metadata, and a data producer could implement their feedback into the dataset.

C Differences in Data Quality Handling and Needs
Data producers and data consumers have different priorities, needs, and expectations for data quality, and their understanding of each other is hampered by infrequent communication. Data consumers often want to use data in other contexts than those it was originally produced for, while data producers are unaware of the additional contexts in which their data is being used.

Recommendation C-1
Data producers should make use of elicitation techniques to identify data consumers' needs and expectations for data quality. This should preferably be done through communication platforms provided by data hosts.

Background C-1 In requirements engineering elicitation is a set of techniques used to identify stakeholder needs and expectations.

SE Literature C-1 SWEBOK outlines several techniques involved in requirements engineering elicitation [15]. Linåker et al. highlight the importance of understanding stakeholders to increase utility and usefulness of the product being produced [36].

Example C-1 Data producers could interview potential data consumers at the beginning of a project, to identify their needs and expectations, so that the dataset can be adapted to be suitable for use.

Recommendation C-2
Data hosts should make use of branching, where raw data is the main dataset, and versions of the main dataset are available through branches that provide data processed in different ways for different users.

Background C-2 Branching is a method used in version control systems, such as GitHub, to create separate versions of the main program code (branches) to work on new features or fixes without affecting the main code [37].

SE Literature C-2 Branching can be done for a variety of reasons, as exemplified by Loeliger [37]: to save different versions or 'releases,' to encapsulate different phases of development, or to isolate changes to a particular area or problem.

Example C-2 Figure 4.4 provides an example of branching.

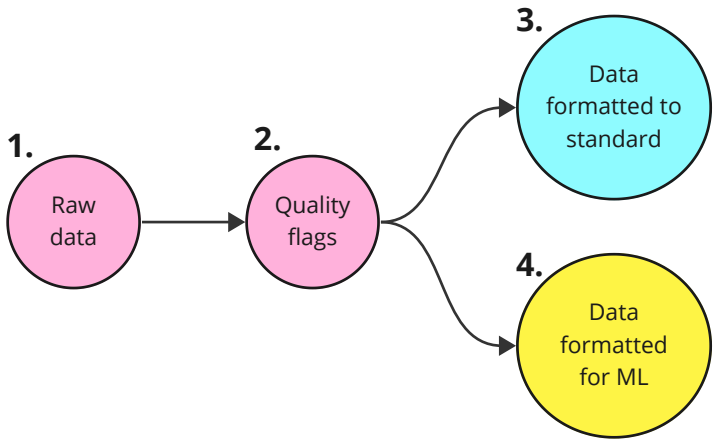


Figure 4.4: There is a main branch (1 and 2) with the raw data (1). Quality flags are then added to mark the quality of the data points (2). The data is then split into two branches: one to conform to a data standard (3) and one to be used for machine learning (4).

D Opportunities and Risks With Open Data
Environmental research is increasingly adopting open data practices following the FAIR principles. However, there remains a significant amount of data, tools, and process information that are not yet widely shared even though they could be useful in open data.

Recommendation D-1
The environmental research community should better encourage the sharing of data processing and analysis tools as open source code accompanying datasets, including homemade or custom tools. This would increase access to tools, improve the quality of tools through collaboration, and improve reproducibility, in line with the FAIR guiding principles.

> **Background D-1** In open source software there are many different programs and projects dedicated to solving small tasks and common problems with programs that automate menial tasks.
>
> **SE Literature D-1** Lamprecht et al. goes over how the FAIR guiding principles have mainly been applied to data, but how they should also be applied to research software [38]. Also sharing research software in accordance with the FAIR principles could help to increase the transparency of how data has been created and processed, while also increasing reusability and decreasing the need for duplicated work.
>
> **Example D-1** A data consumer with little coding experience needs to reformat a large dataset, which would take a long time to do manually. Instead, they find an openly shared tool to automate the process.

Recommendation D-2
Data producers should provide clear descriptions in their datasets about usage, including the purpose of collection, contributors, intended use, and used open data licenses, to ensure clarity and transparency for data consumers.

> **Background D-2** GitHub projects typically include a README file that explains the purpose of the program, reducing the risk of misuse [39]. Essential details that are also reported which include usage, contributors, and open source licenses.
>
> **SE Literature D-2** README files are essential to provide quick and accessible information as highlighted by Wang et al. including how the quality of README files can affect use or "popularity" of a project [40].
>
> **Example D-2** When a data consumer accesses a dataset, they can explicitly read how the data was collected and in what ways they can use it.

Recommendation D-3
Data producers should release raw data as soon as possible after data collection as a preliminary release, and continually update with new processed versions. This reduces the risk of duplicate data collection, both for projects with a similar context to the original use case and for projects where the data is applied in novel contexts.

> **Background D-3** Software projects often involve releasing early versions, such as alpha or beta versions, similar to a first draft. This allows for better collaboration, finding issues, and refinement before wider public use.

SE Literature D-3 The FAIR guiding principles promotes making all data openly available [2]. Keeping data private for longer than strictly necessary runs contrary to these principles, and prevents to data from being utilized where it could bring benefits.

Example D-3 Someone in a research project discovers that a data producer is currently working on a dataset that they could use, and contacts them to collaborate on adapting the dataset for their own project.

Recommendation D-4
Open data should make use of version control to make the history of changes to the data openly available. This would help to increase transparency and traceability of how the data has been changed and processed over time.

Background D-4 In open source software, maintaining a version history of the program makes it easy to track changes and to run previous versions of the program, known as version control [37].

SE Literature D-4 In the book *Version Control with Git* [37], Loeliger goes over the use and benefit of version control, specifically the version control tool git.

Example D-4 A data producer is reusing a dataset for a new project, and reviews the version history to investigate how the dataset was processed.

E Metadata are Key to Understandability
Metadata is crucial to the usefulness of data. A great deal of effort goes into creating correct and sufficient metadata, and without it, the value of the data is significantly reduced, or it may even become unusable.

F Balancing Breadth and Granularity in Standards
Standards are very important but there is a trade-off between broad general standards and the ability to use the collected data. If data standards are too general, data may be averaged or finer details may be lost, and if they are too granular, they are difficult to apply to a wide range of areas.

4.3 Validation Survey Results

The survey results provide an overview of the degree to which respondents agreed with the presented challenges as well as the degree to how applicable and beneficial they thought the recommendations would be to environmental research. Additionally, many respondents also provided comments on why they answered as they did, helping to highlight both what would not work in environmental research and why, and what might be particularly useful in the recommendations. The responses to the quantitative questions in the survey can be seen in figure 4.5. These insights resulted in the refinement of the challenge definitions presented later in this section, and the final recommendations presented in section 4.5.

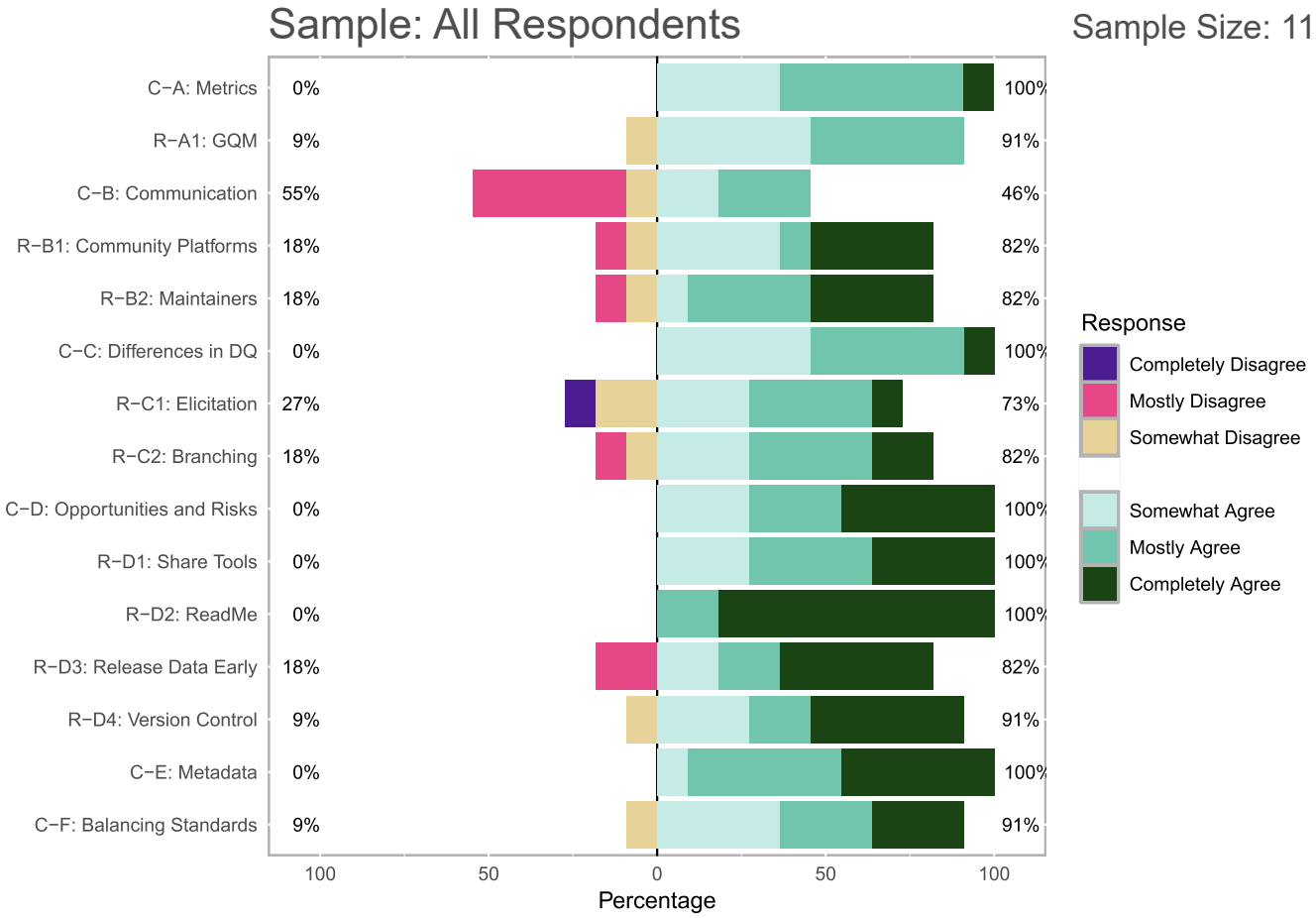


Figure 4.5: A Likert-scale plot showing how the respondents rated the challenges and recommendations described in the survey. Items prefaced with 'C' are challenges, while items prefaced with 'R' are recommendations. The sample size for this plot was eleven, representing all respondents of the survey.

Many of the respondents selected multiple roles. Seven identified as data consumers, eight as data producers, four as data hosts and one as neither of those roles. The responses from the data producers and the data consumers can be seen separately in figure 4.6, though as many of the respondents selected multiple roles there was a large overlap between these two groups. The sample was judged as too small to group these roles exclusively, as there were only three data producers who were not also data consumers, and only two data consumers who were not data producers. Similarly, a large portion of the respondents were in the field of oceanography, had nine or more years of experience, and had participated in the interview study, making the sample too small for groupings along those axes as well.

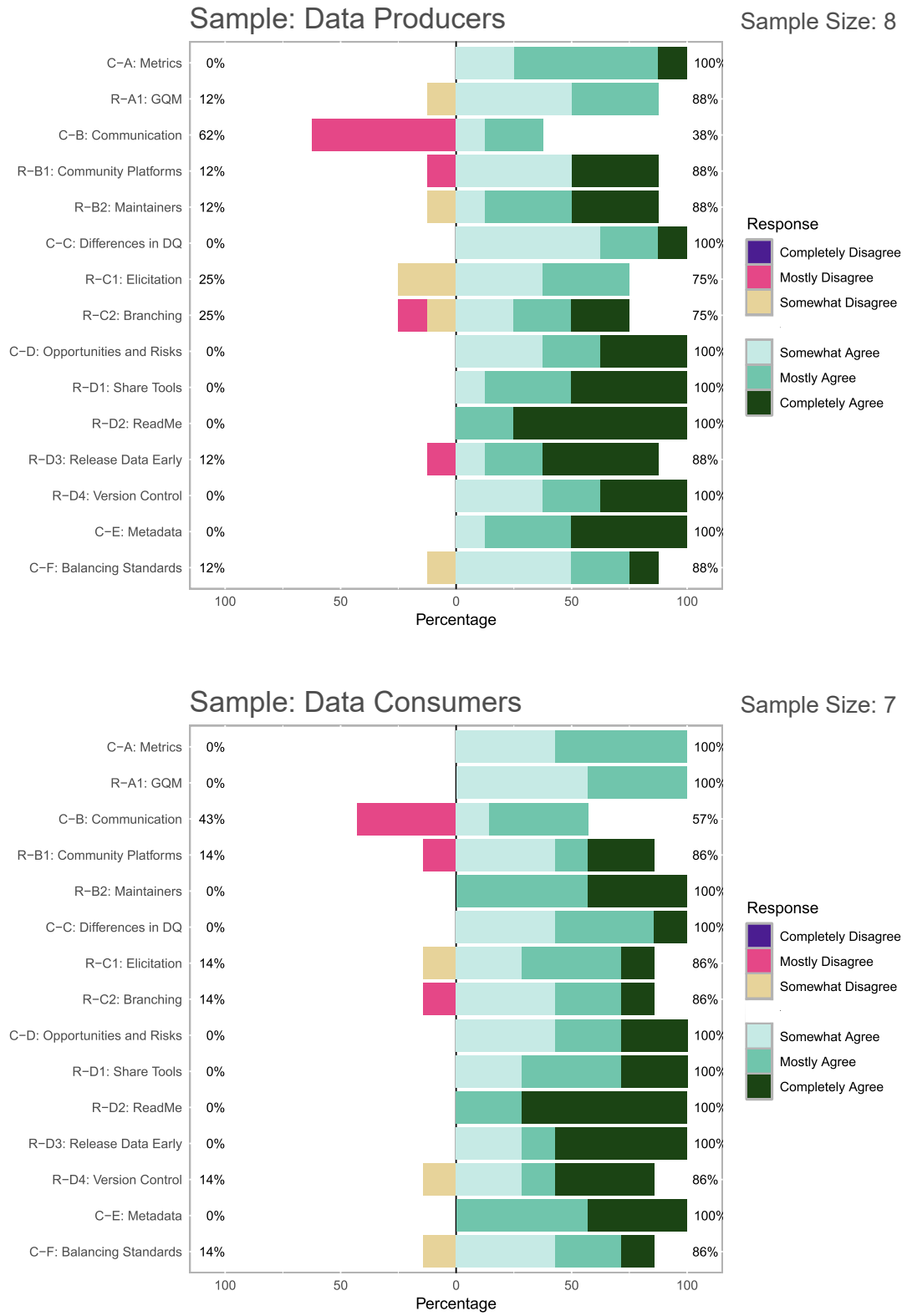


Figure 4.6: Two Likert-scale plots showing how the respondents who categorized themselves as data producers and data consumers rated the challenges and recommendations separately. Items prefaced with 'C' are challenges, while items prefaced with 'R' are recommendations. The sample size for this plot was seven data consumers, this includes people identifying with multiple roles.

Following is how the survey evaluated each challenge and its associated recommendations:

A **Metrics are a Double-Edged Sword**
Challenge A and its recommendation were generally well received. Data producers agreed more strongly with the challenge than data consumers, though the difference is small. Recommendation A-1 was similarly well received, with only one respondent disagreeing based on a concern that the GQM model could still lead to metrics becoming useless goals.

B **Open Data Disconnects Communication**
Challenge B was the most disagreed point of the survey, with data producers disagreeing more than data consumers. The comments in the survey indicate that this disagreement comes from respondents who have not experienced this challenge and respondents who have but think that the benefits of open data outweigh the challenge. For instance, one respondent said:

> "The benefit is the open data. If the data is not open you don't know what data is produced. Then I rather have this challenge." - R4

The recommendations, in contrast, were generally well received, with Recommendation B-2 being particularly well received among data consumers.

C **Differences in Data Quality Handling and Needs**
This challenge was generally well received, with data consumers agreeing more strongly than data producers. There were some concerns about the increased workload that Recommendation C-1 would bring for data producers, while there was an alternative solution presented for Recommendation C-2, making that recommendation obsolete.

D **Opportunities and Risks With Open Data**
Challenge D and its recommendations were one of the most strongly agreed parts of the survey, with comments mainly highlighting difficulties that need to be accounted for when following the recommendations. The exception to this is Recommendation D3, where comments raised concerns for data being used incorrectly or in ways that take credit away from the original data producer. Recommendation D2 was the most strongly agreed point of the survey, with all but two respondents reporting that they 'completely agree.'

E **Metadata are Key to Understandability**
This was the most strongly agreed challenge, stressing the importance of metadata in environmental research. R1 voiced the concern that how metadata is used is not always known, highlighting the need for more communication between stakeholders.

F **Balancing Breadth and Granularity in Standards**
All respondents agreed with this challenge, except for one who mentioned that they did not understand it. One respondent brought up the idea of standards with varying granularity as a potential solution to this challenge.

4.4 Revised Challenge Definitions

The validation from the survey resulted in the finalization of the challenge definitions which serve as the answer to **RQ2**, as presented here:

A Metrics are a Double-Edged Sword
Metrics play an important role in quantifying and communicating data quality. However, metrics can also create arbitrary targets that may be abstract, unfounded, or misleading, and therefore not represent useful goals. This is sometimes referred to as Goodhart's Law [41].

Motivation
This challenge was changed after the survey as all respondents agreed with the description. However, a reference to Goodhart's law was added to the statement to clarify it further, as the law also describes the challenge.

B Open Data Changes Communication
Open data creates many opportunities for collaboration and communication. However, open data introduces different conditions compared to traditional research, and adapting to these conditions could make open data even more beneficial. In open data, communication between data consumers and data producers is often rare or infrequent, as most communication is instead directed towards data hosts.

Motivation
This challenge definition was changed after the survey. The original description of the challenge was unclear and gave the impression that open data was a problem. The challenge definition was, therefore, changed to better reflect open science as a new and improved environment that is good for communication in general, and focus the challenge on a specific issue where communication could be further improved. The name of the challenge was similarly changed to reflect that there can be communication challenges within open data, not that open data is a communication challenge.

C Differences in Data Quality Handling and Needs
Data producers and data consumers have different priorities, needs, and expectations for data quality, and their understanding of each other is hampered by infrequent communication. Data consumers often want to use data in other contexts than those it was originally produced for, while data producers are unaware of the additional contexts in which their data is being used.

Motivation
This challenge remained unchanged after the survey. All respondents agreed with the challenge description, and related comments from the survey highlighted the need for proper handling and interaction between data consumers and data producers.

D Opportunities and Risks With Open Data
Environmental research is increasingly adopting open data practices following the FAIR principles. However, there remains a significant amount of data, tools, and process information that are not yet widely shared even though they could be useful in open data.

> **Motivation**
> This challenge remained unchanged after the survey. All respondents agreed with the challenge description. R7 commented on how funding, which is "notoriously bad" for data management, has a big impact on what can and cannot be done, which will be discussed further in section 5.1.

E Metadata are Key to Understandability
Metadata is crucial to the usefulness of data. A great deal of effort goes into creating correct and sufficient metadata, and without it, the value of the data is significantly reduced, or it may even become unusable.

> **Motivation**
> This challenge remained unchanged after the survey. All respondents agreed with the challenge, as did the comments related to it.

F Balancing Breadth and Granularity in Standards
Standards are very important but there is a trade-off between broad general standards and the ability to use the collected data. If data standards are too general, data may be averaged or finer details may be lost, and if they are too granular, they are difficult to apply to a wide range of areas.

> **Motivation**
> This challenge remained unchanged after the survey. Most respondents agreed with the challenge, with R1 noting the universality of the problem and R11 bringing up possible ways to address it, which are discussed further in section 5.2.

4.5 Final Recommendations

This section presents the final recommendations for the challenges described in the previous section, along with motivation for why they were either changed or kept as they were based on the insights for the validation survey. These final recommendations are the answer to **RQ3**. Despite being largely well received, **Recommendation C-2: Branching Datasets** was removed as the survey highlighted an existing solution already present in environmental research, which will be discussed in section 5.2.

> **Recommendation A-1: GQM model for Metrics**
> Environmental research should use the GQM model to avoid the pitfalls of metrics that create arbitrary targets and to clarify the context and purpose of metrics.

Motivation
This recommendation was kept unchanged from the initial recommendations. There were no significant comments against the recommendation, only one concern from R1 that GQMs would just work as any other metric, which will be discussed in section 5.2.

Recommendation B-1: Data Hosts as Community Platforms
Data hosts should act as community platforms that encourage communication, collaboration, and feedback sharing between data producers and data consumers. This could enable more frequent and open discussions about datasets and increase transparency compared to potential private conversations via e-mails. Large data hosts have a unique opportunity to link this communication to datasets.

Motivation
This recommendation was changed from the initial recommendations. While most respondents agreed with the recommendation, a comment from R11 highlighted that arguments for how the recommendation can add value were missing. The recommendation was changed to emphasize the opportunity for frequent open discussion with data hosts as the most appropriate facilitator.

Recommendation B-2: Dataset Feedback Through Data Maintainers
Data producers should maintain open datasets after it is published. This could enable better response to feedback, whether data quality feedback such as reporting data issues, or adding new value by responding to improvements to the dataset such as alternative processing and adding new metadata.

Motivation
Most respondents agreed with this recommendation, but it was changed to further emphasize that it should be applied to open datasets as commented on by R7, and that it should connect to a feedback cycle as emphasized by R11.

Recommendation C-1: Elicitation of Stakeholder Needs
When producing data, elicitation techniques should be used by data producers to identify the needs and expectations for data quality of data consumers, including external stakeholders. This can increase both the usefulness and overall quality of the data, but also the ability to reuse the data by providing producers with information on how their data/metadata are being used.

Motivation
This recommendation was rephrased to further highlight how elicitation techniques could increase both the usefulness of the data and its reusability. R7 commented on how this recommendation might be difficult or impossible to follow due to the required time and costs, but in section 5.2 we will argue that the benefits of identifying stakeholder needs outweigh the costs.

Recommendation D-1: Share Tools Openly
The environmental research community should better encourage the sharing of data processing and analysis tools as open source code accompanying datasets, including homemade or custom tools. This would increase access to tools, improve the quality of tools through collaboration, and improve reproducibility, in line with the FAIR guiding principles.

Motivation
As this recommendation received a very high level of agreement, it was kept unchanged. However, section 5.2 will discuss a few prerequisites that were mentioned by R1, R2, and R7, which would help this recommendation be applicable to a wider range of researchers.

Recommendation D-2: READMEs for Data
Data producers should provide clear descriptions in their datasets about usage, including the purpose of collection, contributors, intended use, and used open data licenses, to ensure clarity and transparency for data consumers.

Motivation
This recommendation was kept unchanged as it received near-complete agreement from the survey respondents.

Recommendation D-3: Release Raw Data Early
Data producers should release raw data as soon as possible after data collection as a preliminary release, and continually update with new processed versions. This reduces the risk of duplicate data collection, both for projects with a similar context to the original use case and for projects where the data is applied in novel contexts.

Motivation
This recommendation was kept unchanged due to the high level of agreement from the survey respondents. Section 5.2 will discuss some of the difficulties in following this recommendation, and what larger changes to both policies and culture might be needed to enable more data producers to follow this recommendation.

Recommendation D-4: Version Control for Open Data
Open data should make use of version control to make the history of changes to the data openly available. This would help to increase transparency and traceability of how the data has been changed and processed over time.

Motivation
As this recommendation received a high level of agreement from the respondents and a very low level of disagreement, it was kept unchanged from the initial version. R11 also mentioned that some data hosts already make use of version control, but that this is rarely open for external users. Section 5.2 will also go over an alternative solution mentioned by R1, that some data hosts already make use of.

5

Discussion

This chapter will discuss the results and their implications in more detail. This will include further discussions on the validation of both the challenges and recommendations, but also on how they can be interpreted and developed further. Additionally, aspects that SE can learn from environmental research will be discussed, as well as threats to validity. Finally, the research questions will be answered.

5.1 Challenges Around Data Quality in Environmental Research

Based on the analysis of the interview and survey results, it is clear the environmental research community has made great efforts to adopt open science, with much success. While there are still challenges and cultural changes that need to be addressed, there is widespread agreement that open science is a net positive, as highlighted in the survey results for *Challenge B Open Data Changes Communication*, and that everyone in the field should follow the FAIR principles, as reflected in the interview results. These aspects reflect the type of challenges found in the results, which are best represented by *Challenge D Opportunities and Risks With Open Data*, and emphasize the need for all recommendations to be adjusted to fit within open science and open data practices.

In general, all of the challenges are related to each other. The main goal or challenge in focus is *Challenge D*, as it concerns how best to take advantage of the many opportunities of open data and negate its consequences. This has spurred the increased availability of data and big datasets that, as explained by Balbi et al. [17], are beyond the scope of any individual researcher to understand without both metrics and metadata that describe or include key aspects about the data. These are represented by *Challenge A Metrics are a Double-Edged Sword* and *Challenge E Metadata are Key to Understandability*, respectively. In order to improve both the quality and the consistency of metrics and metadata, it is important that data quality handling is consistent, and that standards are widely used and meet the different needs of the field. Edwards describes standards as a lubricant, reducing friction in data collection and communication [7]. These requirements are represented by *Challenge C Differences in Data Quality Handling and Needs* and *Challenge F Balancing Breadth and Granularity in Standards*, respectively.

While many of these challenges can be dealt with individually, from an SE perspective, one key aspect is missing: communication. In order to provide useful metrics, metadata, and standards, there must first be a clear understanding of the needs of different stakeholders, a practice that is central to requirements engineering as explained in SWEBOK [15]. The best way to create such an understanding is, of course, by communicating with the stakeholders. Such communication does, of course, exist in environmental research, though the form it takes differs compared to SE due to the fields' unique needs and norms.

Agile development, which is a common and popular workflow in SE, makes heavy use of work that involves many iterations that requires frequent communication to avoid misunderstandings and faulty requirements [42]. This naturally leads to different priorities than environmental research that relies on long time series data to enable comparisons, for example climate research that Edwards explains needs 100 years or more of data [7]. This was clearly illustrated in the comments from the survey for *Challenge B* and its recommendations: *Recommendation B-1* and *Recommendation B-2*. *Challenge B* illustrates that, unlike SE, environmental research needs more deliberate actions for change, as careless changes can break the consistency needed for long time series. This means that for a change to occur, the change must add enough value over current practices to justify updating those practices. How big of a problem *Challenge B* is, and to what extent the recommendations can solve it, is not clearly known in this thesis, as statistical inference could not be performed. However, from an SE perspective, it is clear that frequent open discussions including all stakeholders regarding data would be a clear improvement over infrequent and private e-mails.

On the other hand, our interviews indicate that not only is communication infrequent, but that data quality expectations are not always met, responsibilities on data quality are rarely communicated, and it is still an uncommon practice to share tools. This means that despite the disagreement with *Challenge B* in our validation survey, we still think there is value to continue investigating the matter and how it can be applied in the most appropriate way for environmental research. It is important to understand that the recommendations are not end goals, but merely a starting point from an SE perspective to be included in the progression of the field of environmental research.

Lastly, in addition to the challenges outlined in section 4.4, both the interviews and the survey analysis emphasized the effect of funding as a factor in research. This includes the reliance on funding in order to both collect data and also to release it openly, which requires extra resources. Any additional data, quality improvements, or analysis of stakeholders require more resources, which are not always available. The funding aspect was not dealt with in the challenges as it was not considered to be related to a research problem, as it is a political one instead. While this thesis cannot affect such decisions directly, we can highlight the potential value of certain efforts as worthwhile or encourage further investigation to more efficiently adopt potentially helpful practices borrowed from SE.

5.2 Recommendation Considerations and Implementations

The discussion of the recommendations in this section responds to some of the issues and questions highlighted in the survey results. This reasoning was the basis for our final recommendations. The section also discusses how some of the recommendations can be interpreted and potentially implemented in environmental research. In the survey, R11 provided a potential recommendation for *Challenge F Balancing Breadth and Granularity in Standards*. The solution discussed standards that could cover varying levels of granularity, but this was not included in our final recommendations as it was not based on SE practices and was not expressed or explored by any other respondent. While many of these recommendations do not focus directly on data consumers, this does not negate the need for data consumer involvement as they have a responsibility to express their data needs and use datasets responsibly.

Recommendation A-1: GQM Model for Metrics
In the survey, one respondent asked:

> "Isn't a GQM metric just another metric that ceases to be a useful measure once it has become a target?" - R1

Of course, that could be the case; a researcher could focus on only the metric part of the GQM model, in which case it could cease to be useful in line with Goodhart's law [41]. However, the strength of the GQM model is that it is goal-oriented. In addition to using metrics, the GQM model also provides the questions those metrics are intended to answer and the goal those answers are meant to lead to. This added context allows practitioners to evaluate whether what they are doing is relevant; do the metrics help to answer the questions, and do the answers help the goal? If they do not, then the questions and metrics should be changed to better align with the goal. The GQM model does not provide automatic protection against Goodhart's law, as practitioners must make this evaluation to ensure the usefulness of their metrics, but the model does help to provide them with the context necessary to do so and to guide the work toward the goal it is intended to achieve.

Recommendation B-1: Data Hosts as Community Platforms
Data hosts are central to open data as they are the ones who aggregate the data. This gives them a unique opportunity to act as community platforms, as they are already who data consumers and data producers go to for their data. Currently, data hosts often provide contact details, such as e-mail addresses, to enable communication. However, we argue that a platform that enables open discussions would be an improvement from this, as it would help to increase transparency and encourage the participation of many different practitioners. These open discussions could allow more community engagement with asynchronous communication and feedback, in contrast to private e-mail conversations which are typically the form this communication takes currently, that do not engage or allow for the participation of anyone other than the individuals holding the conversation.

Recommendation B-2: Dataset Feedback through Data Maintainers
In SE, there are often unexpected issues that are only discovered after a product is finished and released to the final customers. Therefore maintaining software, at least for a period of time after the project is completed, is important. Given the sentiment from the interviews that many data producers are unaware of how much their datasets are used and if they are used correctly, we think that dataset maintainers could encourage such communication. While e-mails can be great for reporting errors, they do not guarantee answers for other kinds of feedback. Some examples of other feedback might be general questions or suggestions for making the datasets more usable, such as tweaks to processing or formatting to increase compatibility with certain applications. This can be particularly useful as certain aspects are best handled by the data producers as they often have intrinsic knowledge of the data from when it was produced.

Recommendation C-1: Elicitation of Stakeholder Needs
This recommendation can be difficult to follow, as it suggests a significant task that can be both costly and time-consuming, in addition to the existing data collection process.

> "[...] it seems completely unrealistic for data producers to undertake this level of work before collecting their data. Most research projects have a short life-span and researchers are busy enough just to do the research [...]" - R7

Nevertheless, we argue that the benefits of stakeholder elicitation, such as enabling reuse, outweigh these difficulties. A different respondent explained the issue clearly:

> "Problem is that as a data producer you often don't know who and for what they will use your data in the future" - R9

Elicitation techniques could help data producers to solve this problem and gain an understanding of who would use their data and how. This would help improve the quality of the data for its primary use case and its potential for reuse, which is an important aspect of the FAIR guiding principles. While not all data producers may have the time and resources to spend on elicitation, we would encourage those that do to elicit the needs of those who would use their data. We can compare this to private data, where datasets are often only used once and then forgotten in a hard drive, whereas open science makes that data available to more people. Similarly, elicitation techniques can increase the availability of data by making datasets useful to more researchers and applicable to more applications. Given that interoperability and reusability are key pillars of FAIR, it is important to consider external stakeholders and different applications in order to thoroughly follow the FAIR principles and ensure the openness of data.

Recommendation C-2: Branching Datasets
This recommendation was removed, not because it is not useful to have different versions of the data available, but because there is already an existing solution used by some data hosts in environmental research, as presented by one of the respondents:

> "Better to use dynamic data serving platforms e.g. ERDDAP, Xpublish, Thredds, that dynamically fulfill user data requests based on one, well structured, data product. Host one version, be ready to serve many formats" - R1

Therefore, instead of applying SE concepts, we would encourage data hosts to use this existing method of adapting data to the different needs of different researchers.

Recommendation D-1: Share Tools Openly
Sharing software tools, in this context to analyse and process environmental data, is an efficient way to reduce unnecessary work, but one concern that was commented on by the survey respondents is how useful tools would be found and evaluated. This is a particular concern for practitioners with little to no coding experience, who therefore have no way to evaluate the tools on their own. Systems for rating and reviewing tools would be necessary to make it easy for researchers to find useful and relevant tools. Even so, those with no previous coding experience will often be unable to make use of many of the tools that would be shared in this way.

A cultural shift to include more software development in environmental research educations, as discussed by Hampton et al. [10], would be necessary to enable a wider range of researchers to follow this recommendation, but even without such a shift, this recommendation can still help those who do have the knowledge to use shared software tools. Such a cultural shift, along with the FAIR principle accessibility, could further encourage those who develop software tools to include guides to make the tools easier to use for researchers with little or no software experience. Additionally, sharing software tools can also create opportunities for community-based feedback that can lead to improvements in the tools. This kind of improvement is something that open source software development clearly exemplifies, and it can lead to tools that are more useful both to the people who created them and to other researchers who want to use them.

Recommendation D-2: READMEs for Data
In the survey, several respondents made a connection between a README file and metadata. However, we would differentiate between the two. Metadata is information about each datapoint, such as where, when, and how it was collected, whereas we would categorize a README as information about the dataset: who collected it, how it can be used, and so on. A README file is also a format that is easy to standardize, which can be beneficial in ensuring that the required information is covered in an appropriate way. However, this separation is not necessarily an important one. One respondent commented:

> "A separate file might not be the best way to do it, I'd rather have this information together with the data in the same file." - R11

The specific format is not the key part of this recommendation; what is important is that information about the dataset and how it should be used is easily available to those who wish to use it. In open source software, the README files act as an entry point that is accessible, easy to read, and provide first-glance information. We think this could be applicable to environmental research, regardless of the exact implementation and format.

Recommendation D-3: Release Raw Data Early
Releasing raw data early is beneficial for open science as it allows external stakeholders to provide feedback to the data producers before the data collection process is finalized, and in that way helps to make the data more useful. In addition, it can also help to inform other data producers about the work currently being conducted, which can influence the work they will do themselves. For example, a data producer might avoid collecting a certain type of data from a specific location if a different data producer has already collected that data, and in that way avoid duplicating work, but this is not a decision they can make if they are not aware of the other data producer's work. If the data is released early, efforts can instead be focused on more useful work or collaboration. Raw data released early also has to be clearly marked as unfinished, as it may not have been verified or validated yet and is therefore not suitable for use. Even so, the opportunities for feedback and openness make it useful to release the data early.

One hurdle to this recommendation, identified in both the interviews and the survey, was that releasing data early could be unfair to the original data producer, as research groups that are larger or have more resources may be able to analyze and use the data before the original data producer can, and might in that way claim credit for the data. The interviews highlighted that publishing papers with associated data plays a key role in building a researcher's reputation. Therefore, data producers might wait to publish their data to prevent others from using it before them, but this goes against open data principles such as the FAIR principles. Therefore, this requirement might require a cultural shift in environmental research, supported by policy changes, to make it feasible for a wider range of data producers to follow this recommendation.

Recommendation D-4: Version Control for Open Data
In the survey, R11 mentioned that some data hosts already use version control, but only for internal use. Having version control at all is a good sign, though making the different versions publicly available would both enable more use cases and increase transparency. However, as datasets are often very large, storing multiple versions of them can require a lot of resources. To alleviate this problem, R1 suggested the possibility of having only one version of the data and making use of version controlled scripts to transform the data into different versions. This method would require less storage space and therefore be more resource efficient, but it would also lose the traceability provided by a

version history of the data. These scripts could be run on the data host's end, and convert the data for researchers to download. However, if the scripts are instead designed to be downloaded by researchers, this may create an additional barrier to using the data, and clear instructions on how to use the scripts may be needed. These are trade-offs that data hosts can consider if they wish to follow this recommendation.

5.3 Future Work

An important aspect to understand for this thesis is that the recommendations should be considered a first iteration in a design process. In SE, it would be considered a prototype where we have performed an analysis of the problem, created an initial design, received some feedback on that design, and altered the design according to that feedback. That this still simply one iteration and the initial steps of a second one. To make these still quite general recommendations more applicable to environmental research, additional design iterations would be needed to refine how to implement these recommendations appropriately in the different fields of environmental research. Most importantly, some of the recommendations could benefit from being used in practice to further evaluate what value they provide in relation to additional resources that they might require. Likewise, given our sample participants, future work should focus on increasing generalizability across more fields in environmental research and different levels of experience. This includes how to adapt SE principles to an environmental research context, as not all practices can or should simply be copied, but implemented in a way that makes sense for researchers in the field. Another point is to investigate environmental researchers with more experience in data-intensive research, including ML, to better understand what is already being implemented but may need further widespread use. While our results provide an overview of data and data quality understanding and practices, which is relevant as many participant's data are open data and thus can be used for data-intensive applications, it does not investigate the direct measures and methods that have been applied and changed in environmental research to accommodate these data-intensive and data-driven methods beyond statistical analysis.

5.4 What Can SE Learn From Environmental Research?

For this thesis, it is also interesting to explore what SE can learn from environmental research in the spirit of exchanging knowledge between the fields, even if this is outside the scope of the thesis. This section will touch on a few of the ideas from environmental research that were found during the course of this thesis, that might be useful for SE.

One aspect is the persistent use of Digital Object Identifiers (DOI) and especially their use along with data publication papers. While this is not a foreign concept to SE, there is a lot of data and documentation that is stored without a DOI that

could help in retrieving online digital objects. This could apply to documentation in software projects, which currently often are linked to GitHub accounts that, if deleted, would remove existing documentation. Another aspect from environmental research that SE could do is to release data publication papers to a greater extent, which improves early communication and documentation of data in SE. Since many research projects rely on mining data repositories, providing additional DOI references for associated data publication papers could provide useful context about the data.

Overall, the FAIR principles could also be applied to SE to help the field consider how data and software are shared and used over time. The prevalence and adoption of agile methodology illustrates how SE favors quick and iterative development in order to adapt to new technologies, APIs, and standards. On the other hand, this means that the field of SE has little experience with projects and datasets meant to last over long periods of time. In contrast, environmental research often uses long time series data that might date back decades. Environmental research, therefore, has much to teach SE about longitudinal studies.

5.5 Threats to Validity

This section discusses threats to validity as identified by Runeson and Höst, which include internal validity, construct validity, external validity, and reliability [43].

Internal Validity

To address bias in the interview study, an interview guide was used to improve consistency along with a pilot interview to evaluate the interview questions and format. The interview guide was validated with in-depth discussions and a review from our supervisor. The pilot helped to assess the viability of the questions in an interview setting that was conducted in the presence of both supervisors. In addition, the interviews were always kept to at most one hour in order to maintain the attention of the participants. However, since most of the interviews were conducted through Microsoft Teams, environmental factors could not be controlled. By scheduling the interview far enough in advance, the interviewees had the possibility to plan an appropriate location. However, when conducting in-person interviews, disruptions could still occur. This did happen at one of our in-person interviews, where we had a fire alarm interrupt the interview, which can affect the validity even though the interview was still completed.

To keep the interviews consistent, there were always two people present for each interview, in addition to the interviewee: one to conduct the interview and one to take notes. All interviews were conducted in English, which all participants were fluent in, and transcripts were automatically generated and manually edited after the interview to match the transcripts to the videos through auditory inspection. A systematic and iterative method of analysis was used to reduce bias in the coding [28]. In addition, the codes were reviewed at least twice separately by two different people and then discussed in depth. The results were also elaborated and confirmed by the subsequent survey study to better understand and validate the interview agreements and disagreements.

For the focus group, it was important that the moderators facilitated discussions. This was accomplished through preparatory open questions that could be asked whenever appropriate, in addition to explicitly asking for different opinions. Similarly, if discussions stalled the moderators could provide additional context learned from the interview study to help spur discussions.

Construct Validity

The pilot interview provided valuable experience and knowledge for conducting the remaining interviews. This included clear procedures and preparation, as well as dealing with confusion and misunderstandings. Prior to all interviews, participants were provided with the same preparation slides outlining the structure of the interview and introducing the terms stakeholders, data-intensive applications, and the list of data quality attributes. The terms stakeholders and data quality attributes were repeated during the interview for clarity, as these were central concepts for the interview. In addition, prior to the interview, the participants were clearly informed of both the purpose of our thesis and the objectives of the interviews. During the interviews, participants were free to ask questions and seek clarification, which was also provided when the interviewer observed confusion or lack of understanding. These explanations were limited to ensure that all answers were given by the participants themselves and not influenced by the interviewer.

To help the focus group participants understand the identified challenges, information about the thesis and the themes was provided to the participants when they were invited. In addition, a brief presentation was given at the beginning of the focus group to provide more context and detail regarding the purpose of the thesis and the various themes.

The survey did not have the same possibility to easily provide clarification, though respondents did have the option to ask for clarification via e-mail. Examples were used in conjunction with the described recommendations to make the recommendations clearer and easier to interpret for the respondents. The survey was validated by an SE expert and a Master student before it was sent out, which helped to ensure that the examples supported consistent interpretation and understanding. In addition, the various stakeholder terms used in this survey were introduced and explained in the questions where they were asked.

External Validity

The interview study used convenience, purposeful, and snowball sampling, as random sampling was not feasible because the study required certain field-specific experience. Personal networks and contacts were used for this purpose. We argue that the sample is still representative as the participants had different levels of experience, institutions, and research backgrounds. Similarly, the results are not intended to be generalizable beyond environmental research, thus reducing the risk of threats to external validity. The main threat to validity is that not all fields of environmental research were interviewed, with many participants specifically in the field of oceanography, and the inherent risk of interviewing only available respondents. To compensate for the threat to validity posed by the interview sample, a follow-up survey was conducted that, besides using convenience sampling, also had purposeful sampling with the explicit goal of including fields beyond those represented in the interviews.

The focus group only included three SE experts, excluding the moderators. These included two PhD students, one with longer and one with shorter experience, and one SE expert with prior knowledge and a vested interest in the thesis. Though this expert might have some bias due to their prior knowledge of the thesis, this risk was mitigated by including the PhD students who had no prior knowledge of the thesis before being invited to the focus group. While this sample limits the generalizability of the focus group results themselves, this was not considered a major issue because these focus group results were the recommendations that were externally validated by the survey. In addition, the recommendations were also supported by a literature study, which helped to ground them in existing SE practices.

The survey itself was useful in validating both the analysis of the interviews and the recommendations. However, the survey only had eleven respondents, eight of whom had participated in the interview study, thus limiting the generalizability outside of that sample despite a response rate of 36.7%. Additionally, many of the respondents were from the field of oceanography, making the recommendations potentially less applicable outside of that field. On the other hand, the distribution between data producers and data consumers was more balanced in the survey than in the interviews, allowing for a more representative response in the validation. These distributions were considered in the analysis where overrepresented groups were compared to underrepresented groups, though the lack of respondents in underrepresented groups makes it impossible to fully account for bias in the sample. Additionally, given that no login was required to respond to the survey, there is a risk that a respondent could answer multiple times. All responses were visually inspected and confirmed to be distinct from one another, which reduces the risk of multiple answers, although it cannot be completely eliminated.

Reliability

Replication of the interviews and surveys is limited because responses will vary among different participants with different experiences and perspectives. These aspects may change even within the same sample over time and even as environmental research evolves. Even so, an interview guide was used to ensure that the interviews

were conducted in a consistent fashion, and the interviewees were sent preparatory materials before the interviews. The interview guide and a presentation that was used during the interviews can be found in appendix A and appendix B, respectively. The slides sent as preparation to the interviewees can be seen in appendix C, and descriptions of the methodology detailing data collection and analysis are found in section 3.1 and section 3.2. Furthermore, the chosen method of analysis is subjective as the collected data was qualitative, which makes replication difficult, even though the method is structured in phases.

The thematic analysis of the interviews relied heavily on the subjective perspectives and experiences of the researchers, making it difficult to reproduce exactly the same results. However, thematic analysis is a well-established method and its iterative process helps to remove undue bias.

Similarly, the focus group and its results relied heavily on the perspectives and experiences of the participants and the moderators. Questions and discussion points were prepared in advance of the focus group, but these were used mainly as guidelines to help keep the discussion going rather than as a strict structure. However, the results of the focus group were grounded in SE literature, which helped to make them more reliable and less subjective.

Informed Consent

All participants were informed of the data collection and handling procedures for the interview study prior to the interview. This included information about the Microsoft Teams recordings, which would be deleted upon completion of the study and used only for the purposes of the study. They were also promised that the data would be anonymous. This information was repeated during the interview, and participants were asked for explicit verbal consent during the interview for both data collection and recording. After the interview, participants were also provided with a reminder of how we would handle the collected data and what they had consented to.

The focus group participants were given the option of being acknowledged in the thesis or be anonymous. The data collected in the survey was anonymous.

Usage of Generative AI in This Thesis

The AI-based tool *DeepL Write* was used to improve language and grammar in this paper. All text was written by the authors, and DeepL was used to correct grammar and, in some cases, word choice. The output of this tool was considered as a recommendation for improving the text, and the authors' own judgment was used to determine which of these recommendations were and were not followed. DeepL was similarly used to improve the language of the questions for the interview guide and survey. *ChatGPT 3.5* from *OpenAI* was used to help compress the formulations of the challenges and the recommendations for the survey. However, all of the rephrasings generated in this way also removed important context from the definitions, so the context was then added back manually.

5.6 Answers to the Research Questions

This section summarizes the results in terms of how they answer our research questions. The answers for **RQ1** and **RQ2** were collected simultaneously and cover data quality in environmental research. Although the research questions are highly connected to each other, with a slight intersection, they are answered separately for clarity.

RQ1: What is the understanding of data requirements for data-intensive models or software in environmental research from the perspective of data producers and data consumers?
While both the environmental literature and practitioners have clear ambitions to work with ML or other data-intensive methods in the future, it is not yet a widespread practice, and few in our interview sample had actual experience working with such methods. This limits the ability of this thesis to directly relate data-intensive models or software to data quality requirements, although there are still some important insights to be found from an SE perspective. In particular, we found that while everyone agrees on the importance of data quality and metadata as being essential to the field, there are still differences in how data quality is viewed and handled. This, combined with the infrequent direct communication between data producers and data consumers, can lead to ambiguity in requirements from an SE perspective. Consequently, while open science has vastly increased the availability of data, there are still opportunities to improve the usefulness of the available data by better understanding the needs of different stakeholders and, by extension, the data quality requirements in the field.

RQ2: What are the challenges of current methods for verifying and validating data quality for data-intensive models and software in environmental research?
Six challenges were identified in this thesis. They are listed as follows, with the prefix *C* standing for *Challenge*:

C-A: Metrics are a Double-Edged Sword

C-B: Open Data Changes Communication

C-C: Differences in Data Quality Handling and Needs

C-D: Opportunities and Risks With Open Data

C-E: Metadata are Key to Understandability

C-F: Balancing Breadth and Granularity in Standards

These challenges are defined in section 4.4.

RQ3: What recommendations can we conclude based on approaches known to software engineering, such as requirements engineering, to improve the process of handling and managing data quality in environmental research? To address the challenges in **RQ2**, eight recommendations were designed. They are listed as follows, with the prefix *R* standing for *Recommendation*:

R-A1: GQM model for Metrics

R-B1: Data Hosts as Community Platforms

R-B2: Dataset Feedback through Data Maintainers

R-C1: Elicitation of Stakeholder Needs

R-D1: Share Tools Openly

R-D2: READMEs for Data

R-D3: Release Raw Data Early

R-D4: Version Control for Open Data

These recommendations are defined in section 4.5.

5.7 Conclusion

As technology advances and data-intensive technologies become more common and sophisticated, environmental research has had to evolve. Environmental research has made great progress in adopting open science practices, and more recently has worked to include more data-intensive methods into its workflows. However, the understanding of data-intensive methods, such as ML, remains dependent on the individual researcher and field, as many data-intensive methods are still not widely used.

In this thesis, we investigated what challenges environmental researchers experience when dealing with data quality. These challenges were investigated through interviews with environmental research practitioners, which resulted in six challenges being identified that mainly concerned communication, open data, and data quality handling. To approach these challenges, we designed a set of eight recommendations based on SE practices. When these recommendations were evaluated, they were generally viewed positively by environmental researchers, indicating that there is a willingness among environmental researchers and also an opportunity to apply SE practices to environmental research. However, our recommendations are only the result of an initial design phase, and more work is needed both to validate them with a wider range of practitioners and to find the best way to implement the recommendations in environmental research.

Bibliography

[1] *Aims and scope - Environmental Research | ScienceDirect.com by Elsevier*. [Online]. Available: https://www.sciencedirect.com/journal/environmental-research/about/aims-and-scope (visited on 03/01/2024).

[2] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," en, *Scientific Data*, vol. 3, no. 1, p. 160 018, Mar. 2016, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. [Online]. Available: https://www.nature.com/articles/sdata201618 (visited on 01/31/2024).

[3] A. Vegendla, A. N. Duc, and S. Gao, "A Systematic Mapping Study on Requirements Engineering in Software Ecosystems," en,

[4] "Ieee standard glossary of software engineering terminology," *IEEE Std 610.12-1990*, pp. 1–84, 1990. DOI: 10.1109/IEEESTD.1990.101064.

[5] S. E. McCord *et al.*, "Provoking a Cultural Shift in Data Quality," en, *BioScience*, vol. 71, no. 6, pp. 647–657, Jun. 2021, ISSN: 0006-3568, 1525-3244. DOI: 10.1093/biosci/biab020. [Online]. Available: https://academic.oup.com/bioscience/article/71/6/647/6188799 (visited on 01/23/2024).

[6] S. M. Easterbrook, *Computing the Climate: How We Know What We Know About Climate Change*, en, 1st ed. Cambridge University Press, Aug. 2023, ISBN: 978-1-316-45976-8 978-1-107-13348-8 978-1-107-58992-6. DOI: 10.1017/9781316459768. [Online]. Available: https://www.cambridge.org/core/product/identifier/9781316459768/type/book (visited on 01/25/2024).

[7] P. N. Edwards, *A vast machine: computer models, climate data, and the politics of global warming*, en. Cambridge, Mass: MIT Press, 2010, OCLC: ocn430736496, ISBN: 978-0-262-01392-5.

[8] K. Calvin *et al.*, "IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.," en, Intergovernmental Panel on Climate Change (IPCC), Tech. Rep., Jul. 2023, Edition: First. DOI: 10.59327/IPCC/AR6-9789291691647. [Online]. Available: https://www.ipcc.ch/report/ar6/syr/ (visited on 02/16/2024).

[9] N.-T. Nguyen *et al.*, "Synthesized Data Quality Requirements and Roadmap for Improving Reusability of In-Situ Marine Data," en, in *2023 IEEE 31st International Requirements Engineering Conference (RE)*, Hannover, Germany: IEEE, Sep. 2023, pp. 65–76, ISBN: 9798350326895. DOI: 10.1109/RE57278.2023.00016. [Online]. Available: https://ieeexplore.ieee.org/document/10260772/ (visited on 01/23/2024).

[10] S. E. Hampton *et al.*, "Skills and Knowledge for Data-Intensive Environmental Research," en, *BioScience*, vol. 67, no. 6, pp. 546–557, Jun. 2017, ISSN: 0006-3568, 1525-3244. DOI: 10.1093/biosci/bix025. [Online]. Available: https://academic.oup.com/bioscience/article/67/6/546/3784601 (visited on 01/23/2024).

[11] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," en, *Patterns*, vol. 2, no. 11, p. 100 336, Nov. 2021, ISSN: 26663899. DOI: 10.1016/j.patter.2021.100336. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666389921001847 (visited on 01/23/2024).

[12] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI," en, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–15, ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445518. [Online]. Available: https://dl.acm.org/doi/10.1145/3411764.3445518 (visited on 02/07/2024).

[13] A. Vogelsang and M. Borg, *Requirements Engineering for Machine Learning: Perspectives from Data Scientists*, en, arXiv:1908.04674 [cs], Aug. 2019. [Online]. Available: http://arxiv.org/abs/1908.04674 (visited on 01/23/2024).

[14] A. Davoudian and M. Liu, "Big Data Systems: A Software Engineering Perspective," en, *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–39, Sep. 2021, ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3408314. [Online]. Available: https://dl.acm.org/doi/10.1145/3408314 (visited on 04/08/2024).

[15] P. Bourque and R. E. Fairley, Eds., *SWEBOK: Guide to the Software Engineering Body of Knowledge*. IEEE Computer Society, 2014.

[16] S. S. Farley, A. Dawson, S. J. Goring, and J. W. Williams, "Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions," en, *BioScience*, vol. 68, no. 8, pp. 563–576, Aug. 2018, ISSN: 0006-3568, 1525-3244. DOI: 10.1093/biosci/biy068. [Online]. Available: https://academic.oup.com/bioscience/article/68/8/563/5049569 (visited on 02/20/2024).

[17] S. Balbi *et al.*, "The global environmental agenda urgently needs a semantic web of knowledge," en, *Environmental Evidence*, vol. 11, no. 1, p. 5, Dec. 2022, ISSN: 2047-2382. DOI: 10.1186/s13750-022-00258-y. [Online]. Available: https://environmentalevidencejournal.biomedcentral.com/articles/10.1186/s13750-022-00258-y (visited on 01/30/2024).

[18] K. S. Cheruvelil and P. A. Soranno, "Data-Intensive Ecological Research Is Catalyzed by Open Science and Team Science," en, *BioScience*, vol. 68, no. 10, pp. 813–822, Oct. 2018, ISSN: 0006-3568, 1525-3244. DOI: 10.1093/biosci/biy097. [Online]. Available: https://academic.oup.com/bioscience/article/68/10/813/5088531 (visited on 01/23/2024).

[19] B. K. Kahn, D. M. Strong, and R. Y. Wang, "Information quality benchmarks," en, vol. 45, no. 4, pp. 184–192, Apr. 2002, ISSN: 0001-0782. DOI: 10.1145/505248.506007.

[20] J.-J. Zhu, M. Yang, and Z. J. Ren, "Machine Learning in Environmental Research: Common Pitfalls and Best Practices," en, *Environmental Science & Technology*, vol. 57, no. 46, pp. 17 671–17 689, Nov. 2023, ISSN: 0013-936X, 1520-5851. DOI: 10.1021/acs.est.3c00026. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.est.3c00026 (visited on 02/07/2024).

[21] R. Huber *et al.*, "Integrating data and analysis technologies within leading environmental research infrastructures: Challenges and approaches," en, *Ecological Informatics*, vol. 61, p. 101 245, Mar. 2021, ISSN: 15749541. DOI: 10.1016/j.ecoinf.2021.101245. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1574954121000364 (visited on 01/23/2024).

[22] S. Amershi *et al.*, "Software Engineering for Machine Learning: A Case Study," en, in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, Montreal, QC, Canada: IEEE, May 2019, pp. 291–300, ISBN: 978-1-72811-760-7. DOI: 10.1109/ICSE-SEIP.2019.00042. [Online]. Available: https://ieeexplore.ieee.org/document/8804457/ (visited on 03/05/2024).

[23] K.-J. Stol and B. Fitzgerald, "The ABC of Software Engineering Research," en, *ACM Transactions on Software Engineering and Methodology*, vol. 27, no. 3, pp. 1–51, Jul. 2018, ISSN: 1049-331X, 1557-7392. DOI: 10.1145/3241743. [Online]. Available: https://dl.acm.org/doi/10.1145/3241743 (visited on 01/23/2024).

[24] H. Suri, "Purposeful Sampling in Qualitative Research Synthesis," *Qualitative Research Journal*, vol. 11, no. 2, pp. 63–75, Jan. 2011, Publisher: Emerald Group Publishing Limited, ISSN: 1443-9883. DOI: 10.3316/QRJ1102063. [Online]. Available: https://doi.org/10.3316/QRJ1102063 (visited on 03/01/2024).

[25] P. Sedgwick, "Convenience sampling," en, *BMJ*, vol. 347, no. oct25 2, f6304–f6304, Oct. 2013, ISSN: 1756-1833. DOI: 10.1136/bmj.f6304. [Online]. Available: https://www.bmj.com/lookup/doi/10.1136/bmj.f6304 (visited on 03/01/2024).

[26] M. Naderifar, H. Goli, and F. Ghaljaie, "Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research," en, *Strides in Development of Medical Education*, vol. 14, no. 3, Sep. 2017, ISSN: 1735-4242. DOI: 10.5812/sdme.67670. [Online]. Available: http://sdmejournal.com/en/articles/67670.html (visited on 04/25/2024).

[27] D. Turner, "Qualitative Interview Design: A Practical Guide for Novice Investigators," en, *The Qualitative Report*, Nov. 2014, ISSN: 2160-3715, 1052-0147. DOI: 10.46743/2160-3715/2010.1178. [Online]. Available: https://nsuworks.nova.edu/tqr/vol15/iss3/19/ (visited on 01/23/2024).

[28] V. Braun and V. Clarke, *Thematic analysis: A Practical Guide*. SAGE, Oct. 2021.

[29] A. Hevner and S. Chatterjee, *Design Research in Information Systems: Theory and Practice* (Integrated Series in Information Systems), en. Boston, MA: Springer US, 2010, vol. 22, ISBN: 978-1-4419-5652-1 978-1-4419-5653-8. DOI: 10.1007/978-1-4419-5653-8. [Online]. Available: https://link.springer.com/10.1007/978-1-4419-5653-8 (visited on 03/25/2024).

[30] J. Horkoff, "DIT831 Research Methods in Software Engineering," en,

[31] R. A. Wienclaw, "Descriptive Statistics.," *Salem Press Encyclopedia*, 2021. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=89185422&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[32] V. R. Basili, G. Caldiera, and H. D. Rombach, "THE GOAL QUESTION METRIC APPROACH," en, 1994.

[33] *Build software better, together*, en, 2024. [Online]. Available: https://github.com (visited on 04/25/2024).

[34] X. Xiao, A. Lindberg, S. Hansen, and K. Lyytinen, ""Computing" Requirements for Open Source Software: A Distributed Cognitive Approach," en, *Journal of the Association for Information Systems*, pp. 1217–1252, 2018, ISSN: 15369323. DOI: 10.17705/1jais.00525. [Online]. Available: https://aisel.aisnet.org/jais/vol19/iss12/2 (visited on 04/08/2024).

[35] S. L. Ramírez-Mora, H. Oktaba, H. Gómez-Adorno, and G. Sierra, "Exploring the communication functions of comments during bug fixing in Open Source Software projects," en, *Information and Software Technology*, vol. 136, p. 106 584, Aug. 2021, ISSN: 09505849. DOI: 10.1016/j.infsof.2021.106584. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950584921000665 (visited on 04/08/2024).

[36] J. Linåker, B. Regnell, and D. Damian, "A method for analyzing stakeholders' influence on an open source software ecosystem's requirements engineering process," en, *Requirements Engineering*, vol. 25, no. 1, pp. 115–130, Mar. 2020, ISSN: 0947-3602, 1432-010X. DOI: 10.1007/s00766-019-00310-3. [Online]. Available: http://link.springer.com/10.1007/s00766-019-00310-3 (visited on 04/08/2024).

[37] J. Loeliger, *Version control with Git: powerful techniques for centralized and distributed project management*, en, 1. ed. Beijing Köln: O'Reilly, 2009, ISBN: 978-0-596-52012-0.

[38] A.-L. Lamprecht *et al.*, "Towards FAIR principles for research software," en, *Data Science*, vol. 3, no. 1, P. Groth, P. Groth, and M. Dumontier, Eds., pp. 37–59, Jun. 2020, ISSN: 24518492, 24518484. DOI: 10.3233/DS-190026. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/DS-190026 (visited on 04/03/2024).

[39] *About READMEs*, en. [Online]. Available: https://docs.github.com/en/repositories/managing-your-repositorys-settings-and-features/customizing-your-repository/about-readmes (visited on 04/25/2024).

[40] T. Wang, S. Wang, and T.-H. (Chen, "Study the correlation between the readme file of GitHub projects and their popularity," en, *Journal of Systems and Software*, vol. 205, p. 111 806, Nov. 2023, ISSN: 01641212. DOI: 10.1016/j.jss.2023.111806. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0164121223002017 (visited on 04/22/2024).

[41] *Goodhart's law*. [Online]. Available: https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095859655 (visited on 05/09/2024).

[42] W. Alsaqaf, M. Daneva, and R. Wieringa, "Quality requirements in large-scale distributed agile projects – a systematic literature review," in *Requirements*

Engineering: Foundation for Software Quality, P. Grünbacher and A. Perini, Eds., Cham: Springer International Publishing, 2017, pp. 219–234, ISBN: 978-3-319-54045-0.

[43] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," en, *Empirical Software Engineering*, vol. 14, no. 2, pp. 131–164, Apr. 2009, ISSN: 1382-3256, 1573-7616. DOI: 10.1007/s10664-008-9102-8. [Online]. Available: http://link.springer.com/10.1007/s10664-008-9102-8 (visited on 03/07/2024).

A

Interview Guide

This appendix contains the text and questions for the interview guide. The interview guide is divided into five categories: (1) *Personal Background*, (2) *Data Quality*, (3) *Communication*, (4) *Data Standards and Directives*, (5) *Open Science and Reuse*. The first part of each section is the introduction to its relevant category, which was presented to the interviewees. The questions in the categories are here presented in the numbered list. Additionally, category 3-5 had different questions for data consumers and data producers, many of which were similar but framed for different perspectives. These role-specific questions will here be presented in parallel.

A.1 Personal Background

Hello and welcome! Nice to have you here! We are Max and Markus and we are master students at Chalmers University of Technology. In this interview we will be exploring data quality in the field of environmental research, including some of the processes surrounding data quality, especially communication. We hope to use what we learn from these interviews to help create recommendations for data quality based on software engineering practices and knowledge. We're going to go through a total of five categories as mentioned on this slide (show the preparation and go through the categories),

Before we begin, we would like to ask your consent to record this interview. It will only be used for the purpose of this thesis and deleted once the thesis is complete. Furthermore all data we collect will be anonymized. Is this okay?

And, of course, we're going to start with some questions about your professional background, about 5 minutes, before we move on to data quality.

1. Briefly explain: What is your current research focus in environmental research?
2. How long have you been active in environmental research?
3. (Present Stakeholders, see figure A.1) – Throughout this interview, we'll be talking about stakeholders. Stakeholders are people, organizations or other entities that have an interest in something like a project.

4. Of course in our case for this interview its regarding data quality within environmental research. For this purpose we have considered these three stakeholders, data producers who create data, data consumers who use the data, and finally data hosts who keep and store the data.
5. Do you agree with this picture, or is there anything you would like to change or add?
6. Would you consider yourself primarily a data consumer, a data producer?
   (a) Discuss the stakeholders regarding their research and choose one, and ask if it is okay to continue the interview from that perspective.
7. Do you have previous experience in working with or developing machine learning or other data-intensive applications?

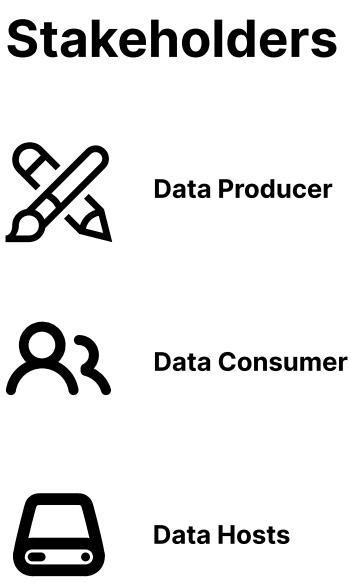


Figure A.1: The three groups of stakeholders of data quality in environmental research that were presented during the interviews.

A.2 Data Quality

Now let's get into the main parts of this interview, starting with data quality, which should take about 10-15 minutes. Here we have a list of data quality attributes(see figure A.1), and we'll give you some time to look through them and you may ask if you have any questions or if you disagree with some definition.

1. Before we start we want to ask you if you would use some other term for these or if you also know them as data quality attributes?
   (a) * Use their term instead of data quality attributes if they have one.
2. Which of these attributes do you think are the three most important ones?
3. And which ones would you say are the three least important ones?
   (a) If they don't motivate reasons for the order: Ask them to explain their line of thought.
4. Do you feel that there is any attribute* missing from the list of data quality attributes* that you would like to add?
5. What metrics do you use to measure the three most important attributes you've chosen?
6. What role do metrics play in data quality?
7. Do you experience any challenges using metrics?
8. What role does metadata play in data quality?
9. Do you experience any challenges for using metadata?

Table A.1: List of 16 data quality dimensions as defined by Kahn et al. [19].

Dimensions	Definitions
Accessibility	The extent to which information is available, or easily and quickly retrievable.
Appropriate Amount of Information	The extent to which the volume of information is appropriate for the task at hand.
Believability	The extent to which information is regarded as trie and credible.
Completeness	The extent to which information is not missing and is of sufficient breadth and depth for the task at hand.
Concise Representation	The extent to which information is compactly represented.
Consistent Representation	The extent to which information is presented in the same format.
Ease of Manipulation	The extent to which information is easy to manipulate and apply to different tasks.
Free-of-Error	The extent to which information is correct and reliable.
Interpretability	The extent to which information is in appropriate languages, symbols, and units, and the definitions are clear.
Objectivity	The extent to which information is unbiased, unprejudiced, and impartial.
Relevancy	The extent to which information is applicable and helpful for the task at hand.
Reputation	The extent to which information is highly regarded in terms of its source or content.
Security	The extent to which access to information is restricted appropriately to maintain its security.
Timeliness	The extent to which information is sufficiently up-to-date for the task at hand.
Understandability	The extent to which information is easily comprehended.
Value-Added	The extent to which information is beneficial and provides advantages from its use.

A.3 Communication

Now that we've talked a bit about data quality, we would like to ask some questions about the communication between data consumers and data producers that we mentioned at the beginning. Remember this (show figure A.1), where you mentioned that you are primarily a data consumer/producer, so we are going to ask your perspective as a data consumer/producer.

We expect this category to take around 10 minutes.

Data Consumer:

1. In your experience, are you able to communicate directly with the data producer regarding data quality issues?
   (a) At what frequency does this communication happen?
2. How do you communicate your data quality expectations to the data producer?
3. Do you experience problems with data producers not meeting your data quality expectations?
4. As a data consumer, what do you see as your responsibilities and what are the responsibilities of data producers with respect to ensure data quality?
5. In your opinion, how often are these responsibilities explicitly communicated?
6. In your experience as a data consumer, do you always have an understanding of the source of the data as well as how the data was collected and produced? Like how can you trust the data?
7. If working with statistical modelling or ML: How do you decide which variables of the data to include in the model as prediction variable and which variables to omit.

Data Producer:

1. In your experience, can you communicate directly with the data consumer on data quality issues?
   (a) At what frequency does this communication happen?
2. As a data producer, are you often aware of the expectations of data consumers?
   (a) Do you find them reasonable and actionable in terms of how you should produce the data?
3. How do you verify that the data you produce meets the data quality expectations of the data consumers? (verify as in feedback, different from initial expectations)
4. As a data producer, what do you see as your responsibilities and what are the responsibilities of data consumers with respect to ensure data quality?
   (a) In your opinion, who is responsible for ensuring data quality?
   (b) In your experience, who usually takes the responsibility?
5. In your opinion, how often are these responsibilities explicitly communicated?
6. As a data producer, do you typically know who will use the data you produce and what they will use it for? How do you establish trust in your data?

A.4 Data Standards and Directives

Now we move on to some questions about standards and directives. This category is a little shorter than the others and should only take about 5 minutes.

Data Consumer:

1. As a data consumer, what role does data quality standards have to you?
   (a) If a dataset follows a standard, does that influence the likelihood that you will use the dataset?
2. Which one(s) do you follow?
3. Do these standards work well for you, or is there something you find problematic?
4. How do you evaluate whether the standard is useful? What are in your opinion important aspects of a useful data quality standard?

Data Producer:

1. In your opinion, what is the role of standardization and directives to ensure data quality?

A.5 Open Science and Reuse

And now we'll move on to our final section on open science and reuse, which should only take about 10 minutes.

Data Consumer:

1. Have you ever reused data from another project for a different purpose?
   - If yes: What did you think worked well when reusing data?
   - If yes: Where there any challenges to reusing existing data?
   - If no: Do you know of any challenges when reusing existing data?
2. As a data consumer, how do you ensure that the data you use is suitable for a project?
3. Are there any particular challenges you've encountered when looking for open datasets?
4. Have you ever combined data from multiple sources in your research?
   (a) If so, were there any challenges?
5. Lastly, we would like to ask you if you've ever used citizen science as a data source?
   (a) Are there any challenges or considerations when using citizen data as a data consumer?

Data Producer:

1. Have you contributed data to an open dataset?
   - If yes: Were there any particular challenges you encountered when doing so?
   - If yes: Was there anything you liked about working with an open dataset compared to working with a private dataset,
   - If no: What would be needed for you to contribute data to an open dataset?
2. If the data you produce is publicly available, how do you know if it is being used and if it is useful for the data consumer?
   (a) Is there feedback, and if so, where is it coming from?
3. How does producing data for internal use compares to publishing data? Especially perhaps when it comes to using tools for cleaning data?
4. Lastly we would like to ask if you've ever produced data through citizen science?
   (a) Did you encounter any particular challenges compared to conventional methods?

B

Interview Guide Presentation Slides

This appendix contains the presentation slides used during the interview along with the interview guide.

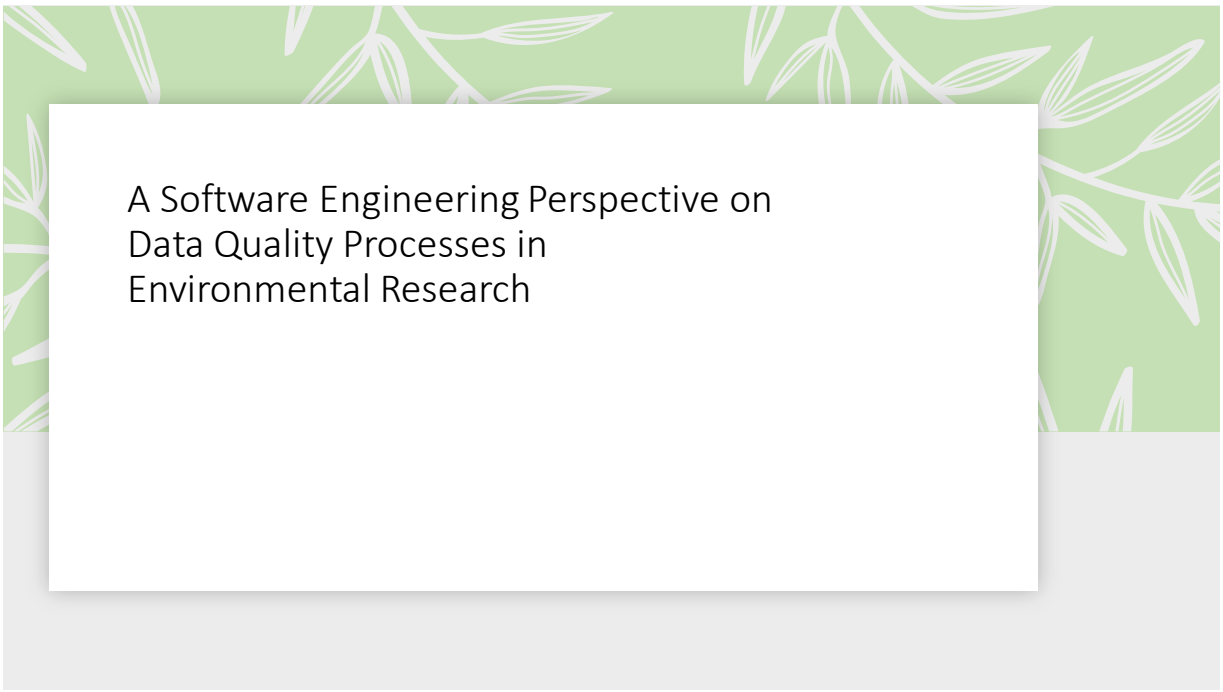


Figure B.1: Slide 1 of the presentation used during the interviews.

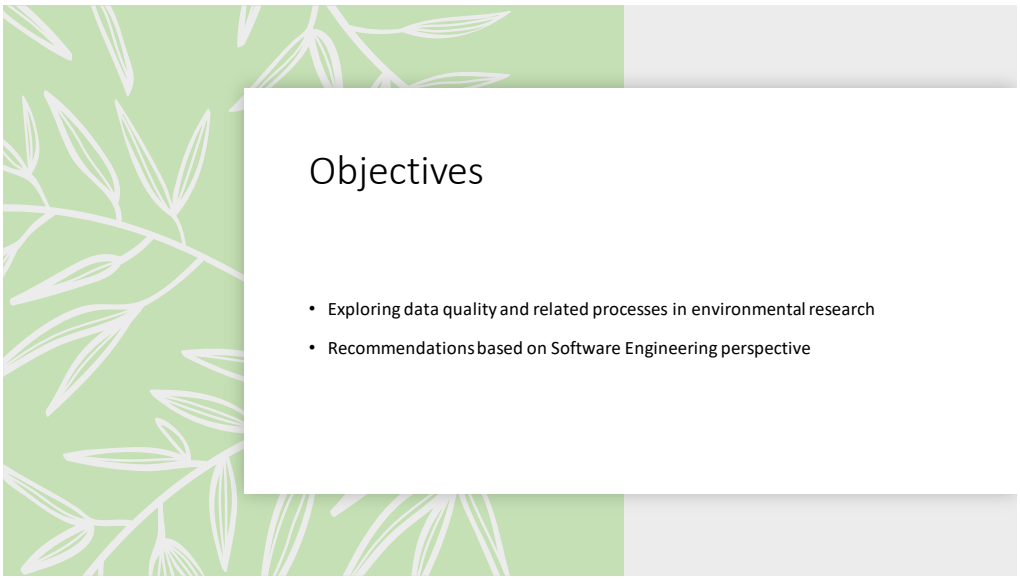


Figure B.2: Slide 2 of the presentation used during the interviews.

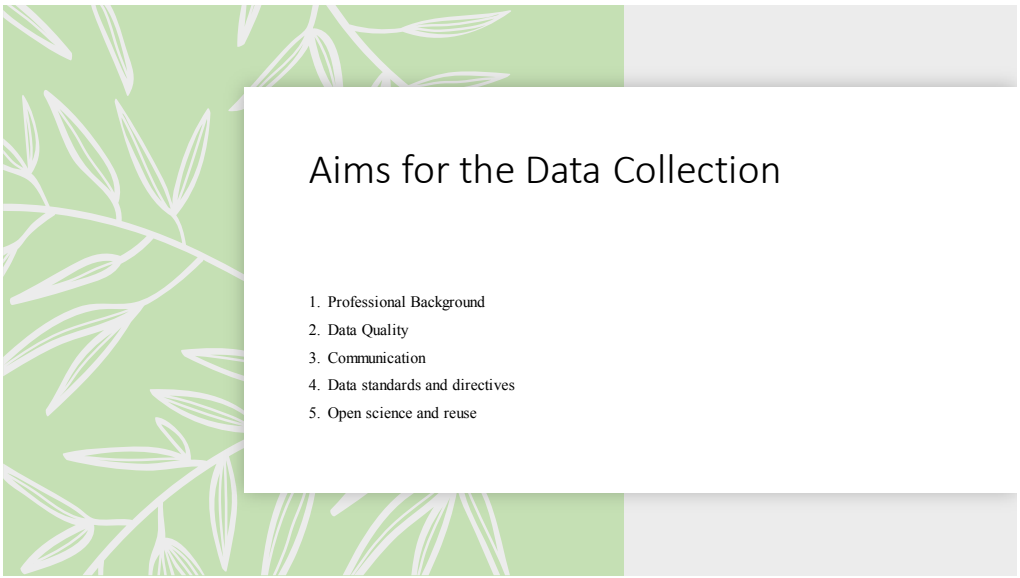


Figure B.3: Slide 3 of the presentation used during the interviews.

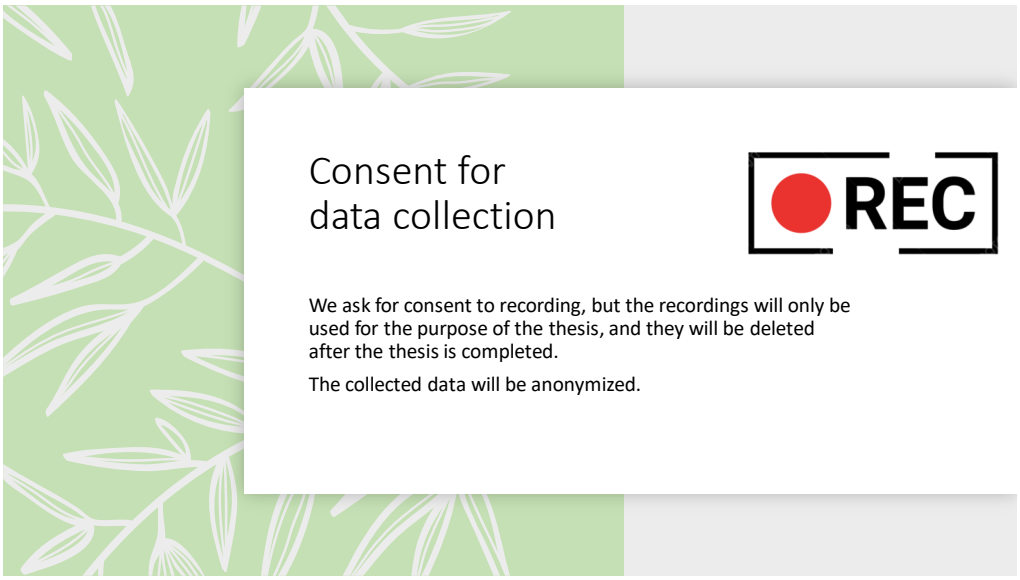


Figure B.4: Slide 4 of the presentation used during the interviews.

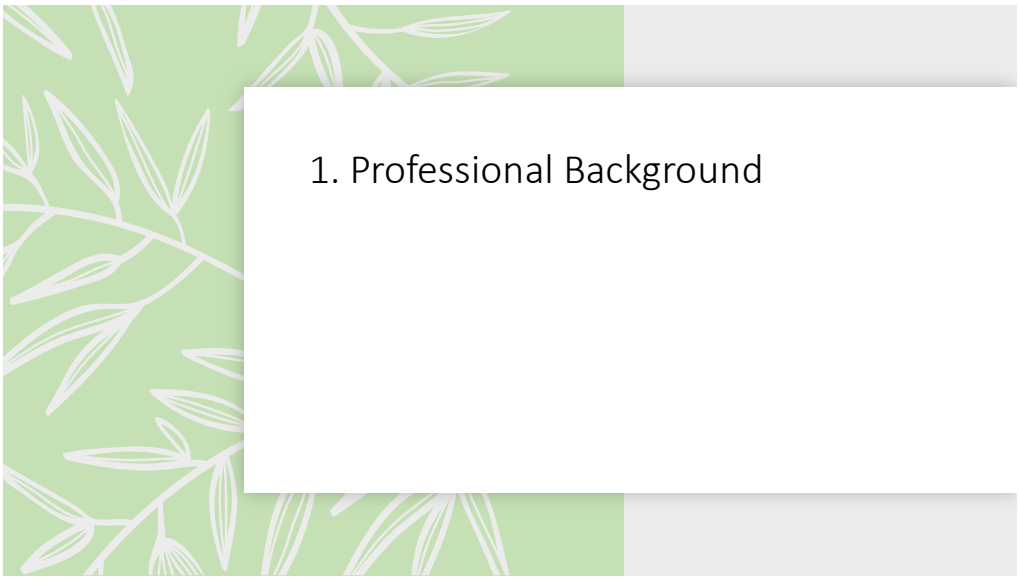


Figure B.5: Slide 5 of the presentation used during the interviews.

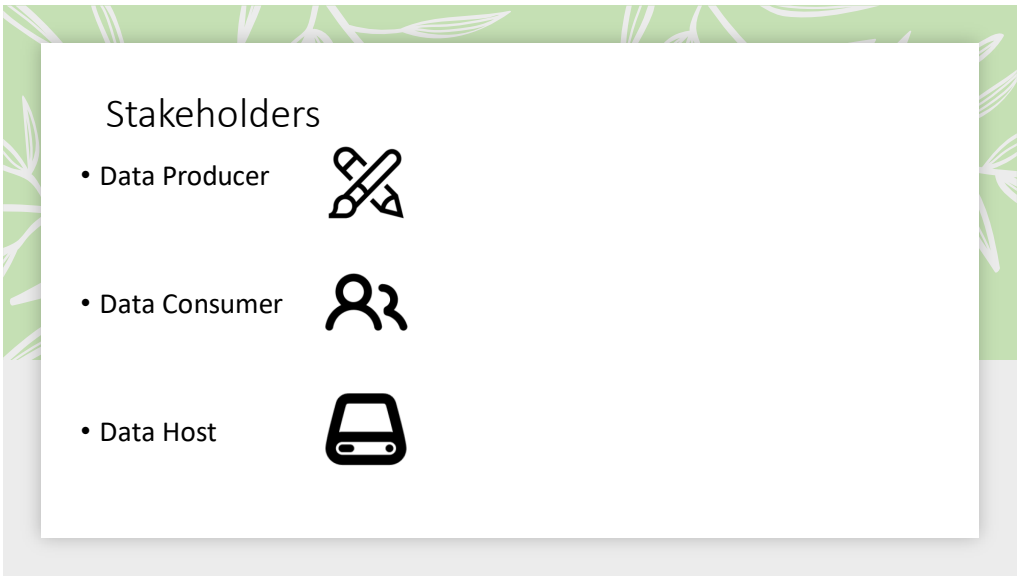


Figure B.6: Slide 6 of the presentation used during the interviews.

Figure B.7: Slide 7 of the presentation used during the interviews.

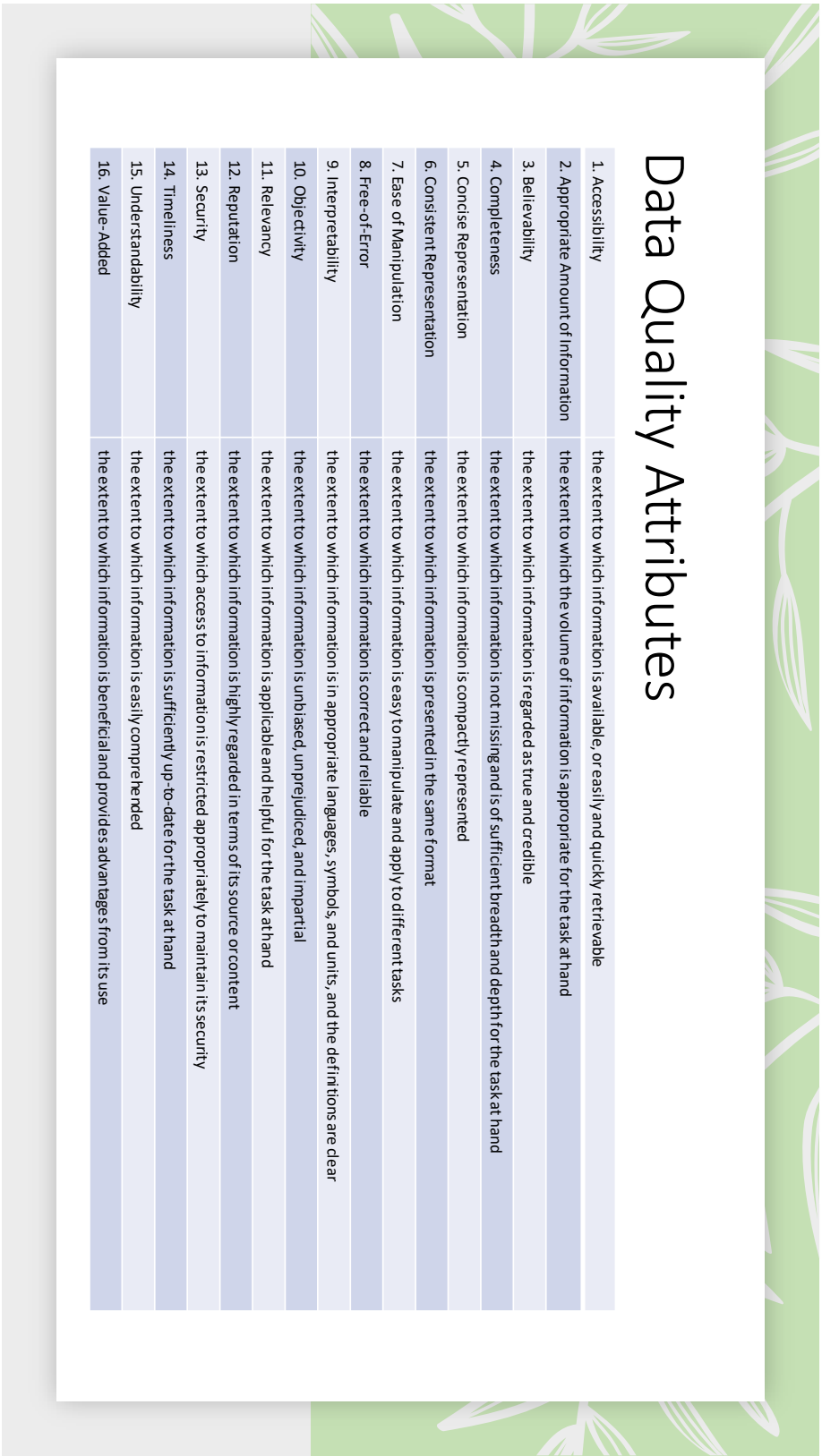


Figure B.8: Slide 8 of the presentation used during the interviews.

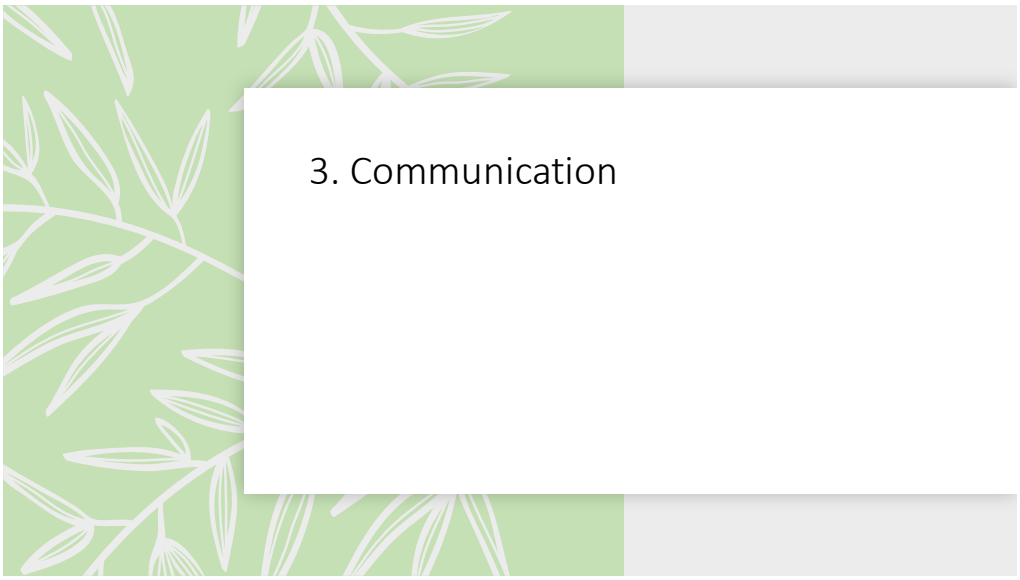


Figure B.9: Slide 9 of the presentation used during the interviews.

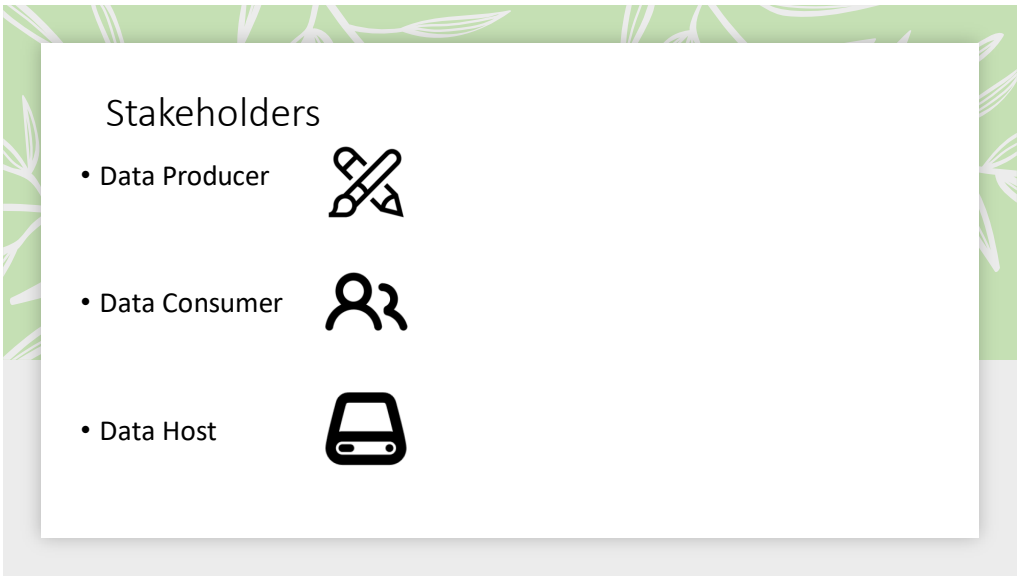


Figure B.10: Slide 10 of the presentation used during the interviews.

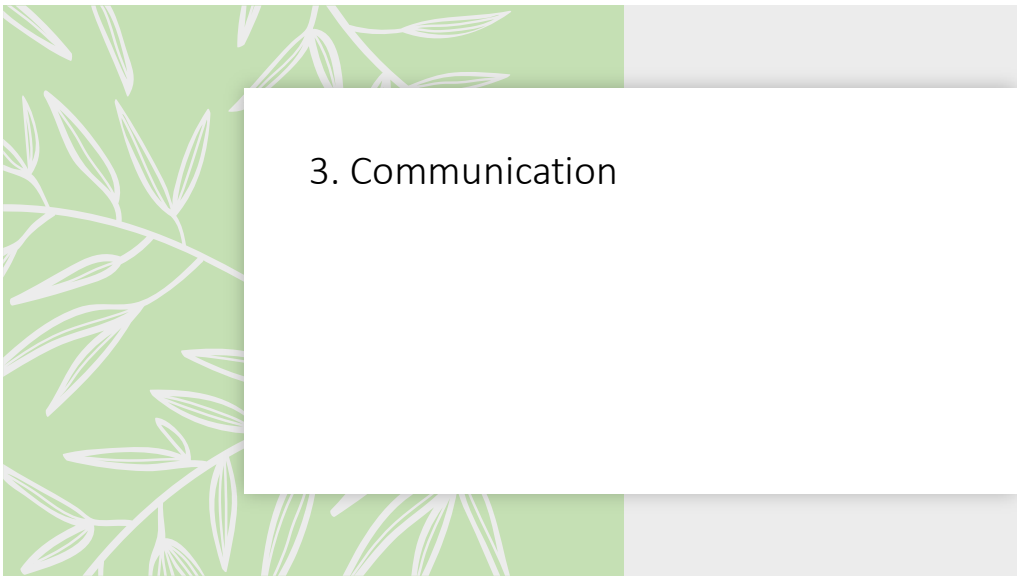


Figure B.11: Slide 11 of the presentation used during the interviews.

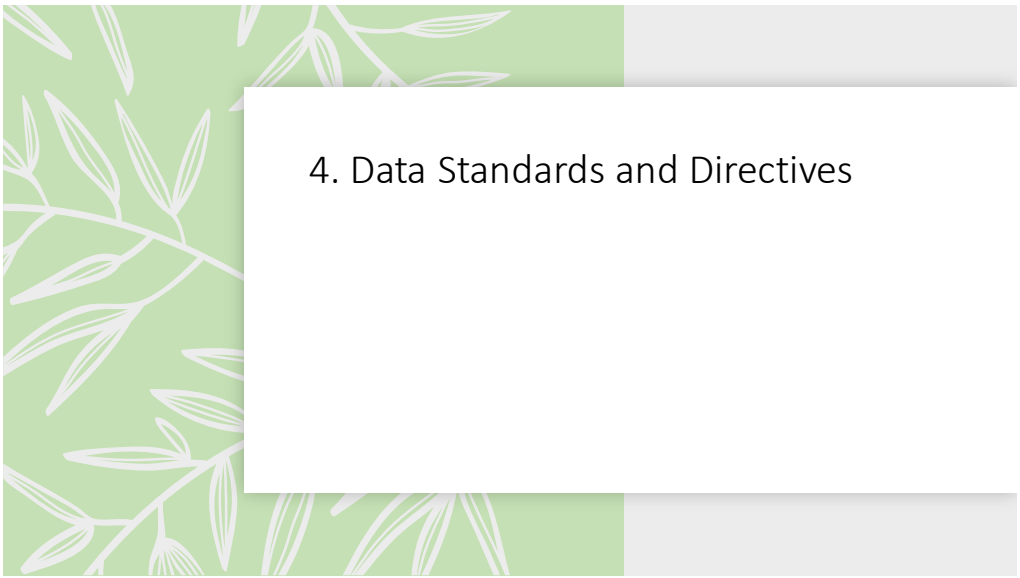


Figure B.12: Slide 12 of the presentation used during the interviews.

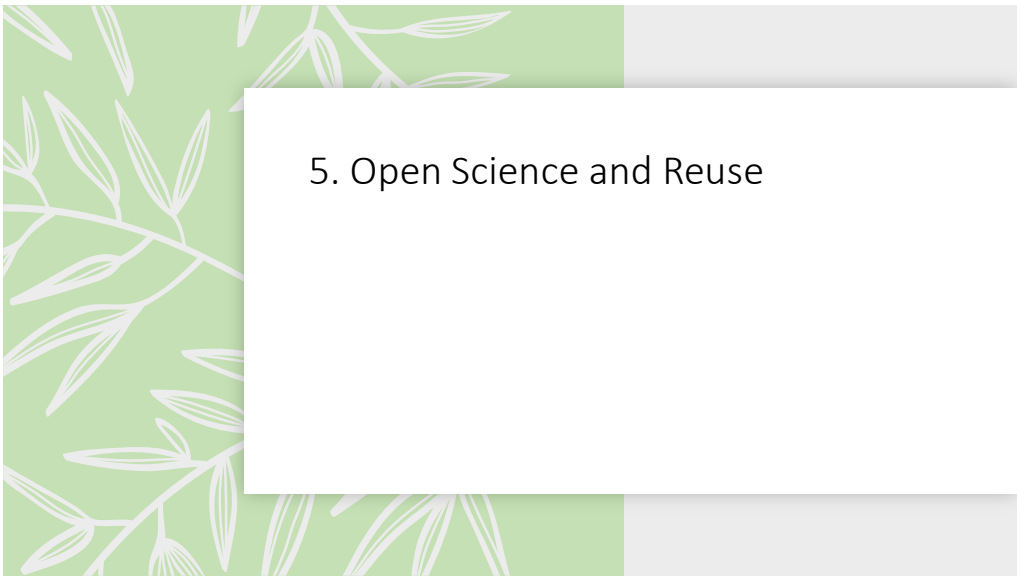


Figure B.13: Slide 13 of the presentation used during the interviews.

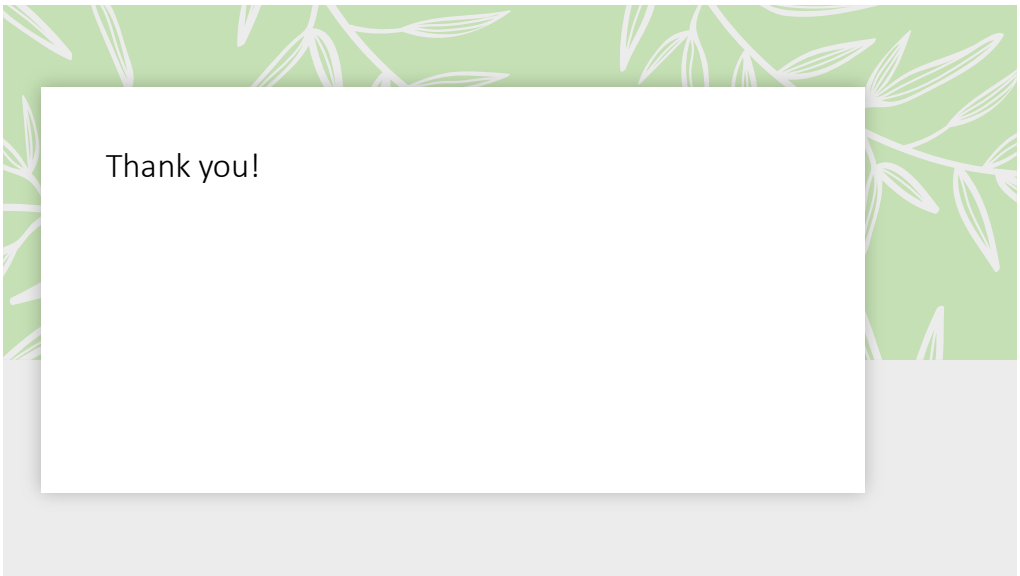


Figure B.14: Slide 14 of the presentation used during the interviews.

C

Interview Preparation Slides

The slides presented in this appendix were sent out to the interviewees as preparatory material to inform them of what the interviews would go over.

Figure C.1: Slide 1 of the presentation sent out to inform the interviewees of what the interview would go over.

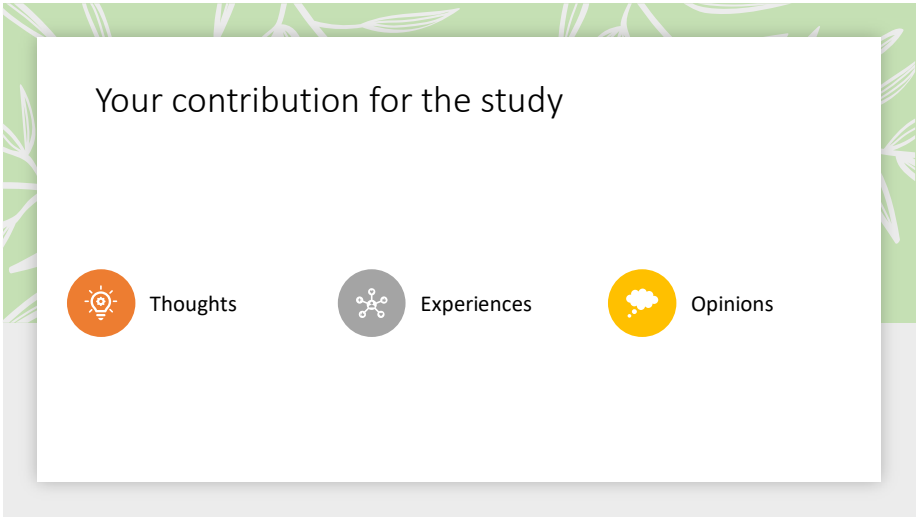


Figure C.2: Slide 2 of the presentation sent out to inform the interviewees of what the interview would go over.

Figure C.3: Slide 3 of the presentation sent out to inform the interviewees of what the interview would go over.

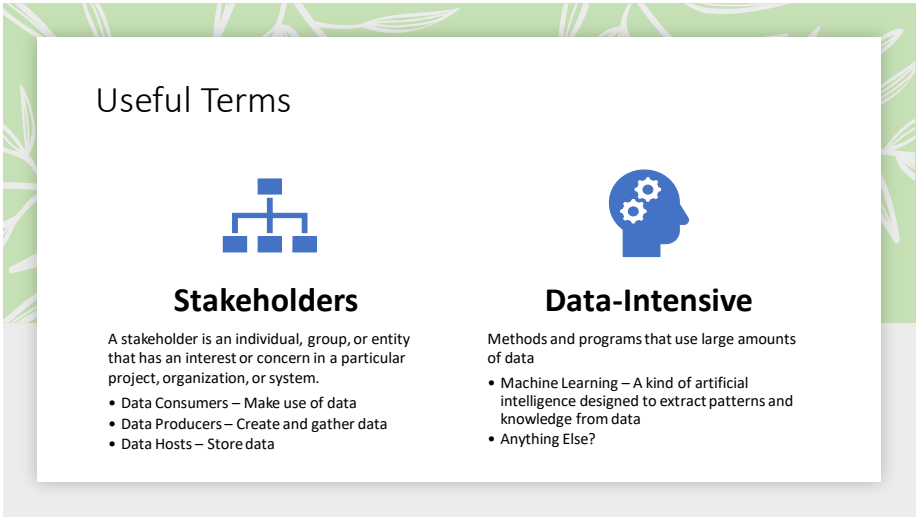


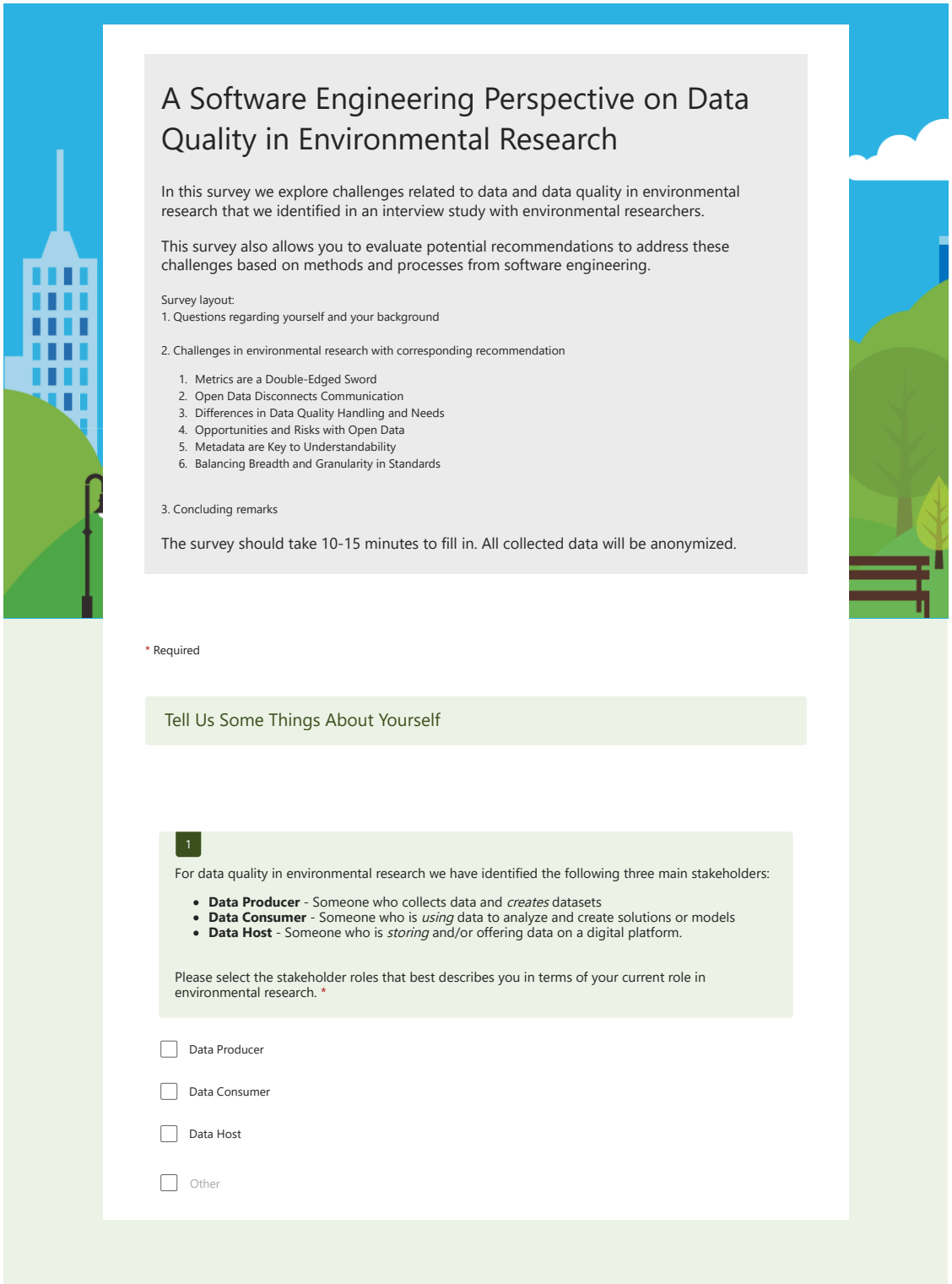
Figure C.4: Slide 4 of the presentation sent out to inform the interviewees of what the interview would go over.

Figure C.5: Slide 5 of the presentation sent out to inform the interviewees of what the interview would go over.

D

Survey Form

This appendix contains the validation survey form.



A Software Engineering Perspective on Data Quality in Environmental Research

In this survey we explore challenges related to data and data quality in environmental research that we identified in an interview study with environmental researchers.

This survey also allows you to evaluate potential recommendations to address these challenges based on methods and processes from software engineering.

Survey layout:
1. Questions regarding yourself and your background

2. Challenges in environmental research with corresponding recommendation

1. Metrics are a Double-Edged Sword
2. Open Data Disconnects Communication
3. Differences in Data Quality Handling and Needs
4. Opportunities and Risks with Open Data
5. Metadata are Key to Understandability
6. Balancing Breadth and Granularity in Standards

3. Concluding remarks

The survey should take 10-15 minutes to fill in. All collected data will be anonymized.

* Required

Tell Us Some Things About Yourself

1

For data quality in environmental research we have identified the following three main stakeholders:

- **Data Producer** - Someone who collects data and *creates* datasets
- **Data Consumer** - Someone who is *using* data to analyze and create solutions or models
- **Data Host** - Someone who is *storing* and/or offering data on a digital platform.

Please select the stakeholder roles that best describes you in terms of your current role in environmental research. *

- [ ] Data Producer
- [ ] Data Consumer
- [ ] Data Host
- [ ] Other

Figure D.1: Page 1 of the validation survey form.

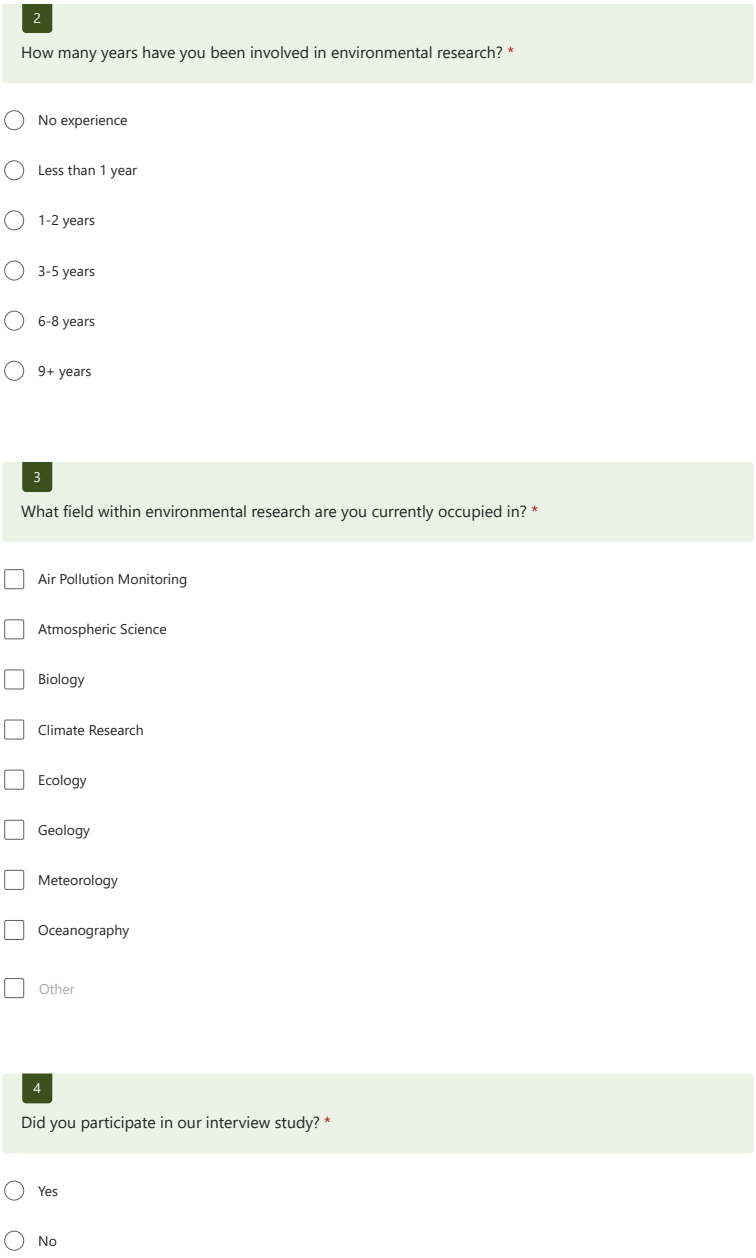
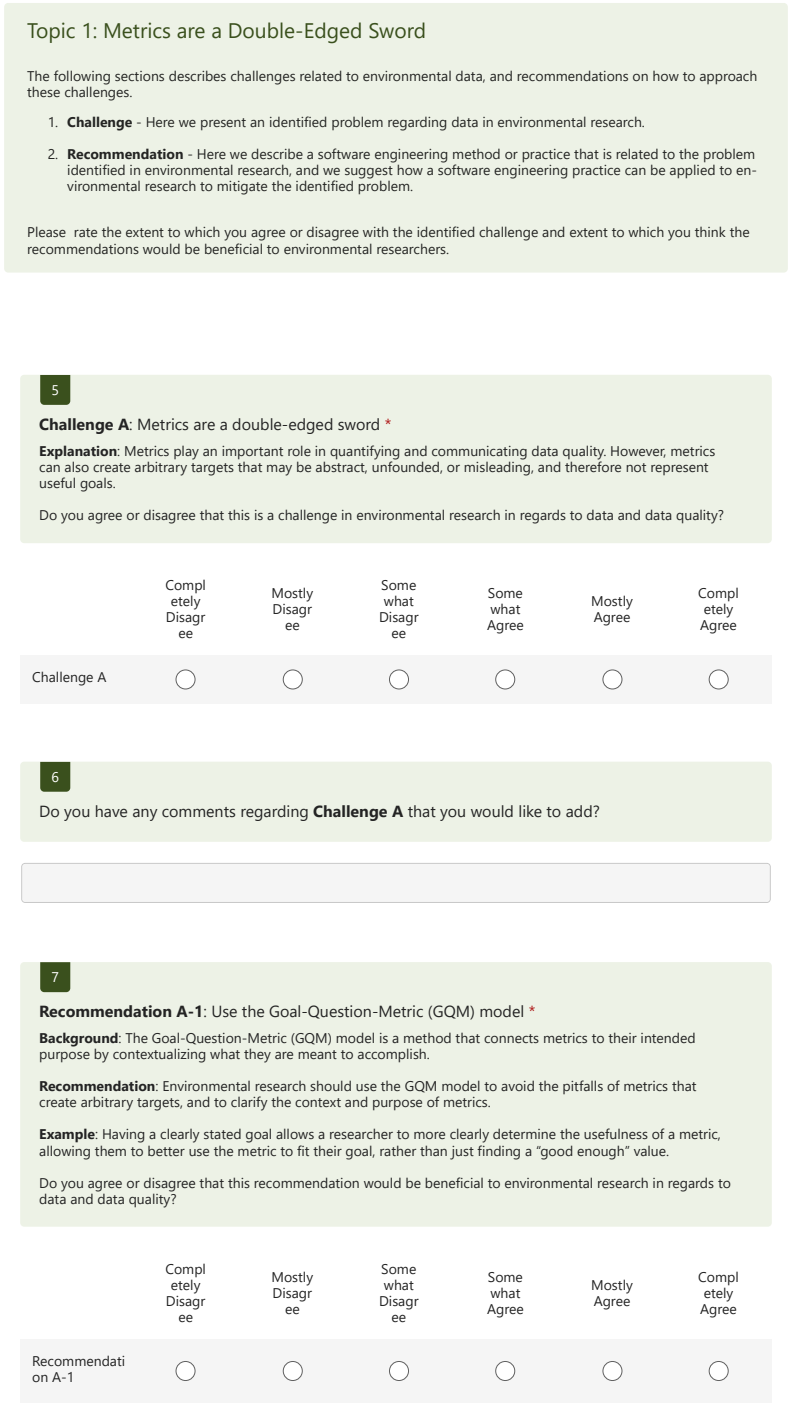


Figure D.2: Page 2 of the validation survey form.



Topic 1: Metrics are a Double-Edged Sword

The following sections describes challenges related to environmental data, and recommendations on how to approach these challenges.

1. **Challenge** - Here we present an identified problem regarding data in environmental research.
2. **Recommendation** - Here we describe a software engineering method or practice that is related to the problem identified in environmental research, and we suggest how a software engineering practice can be applied to environmental research to mitigate the identified problem.

Please rate the extent to which you agree or disagree with the identified challenge and extent to which you think the recommendations would be beneficial to environmental researchers.

5

Challenge A: Metrics are a double-edged sword *

Explanation: Metrics play an important role in quantifying and communicating data quality. However, metrics can also create arbitrary targets that may be abstract, unfounded, or misleading, and therefore not represent useful goals.

Do you agree or disagree that this is a challenge in environmental research in regards to data and data quality?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Challenge A	◯	◯	◯	◯	◯	◯

6

Do you have any comments regarding **Challenge A** that you would like to add?

7

Recommendation A-1: Use the Goal-Question-Metric (GQM) model *

Background: The Goal-Question-Metric (GQM) model is a method that connects metrics to their intended purpose by contextualizing what they are meant to accomplish.

Recommendation: Environmental research should use the GQM model to avoid the pitfalls of metrics that create arbitrary targets, and to clarify the context and purpose of metrics.

Example: Having a clearly stated goal allows a researcher to more clearly determine the usefulness of a metric, allowing them to better use the metric to fit their goal, rather than just finding a "good enough" value.

Do you agree or disagree that this recommendation would be beneficial to environmental research in regards to data and data quality?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Recommendation A-1	◯	◯	◯	◯	◯	◯

Figure D.3: Page 3 of the validation survey form.

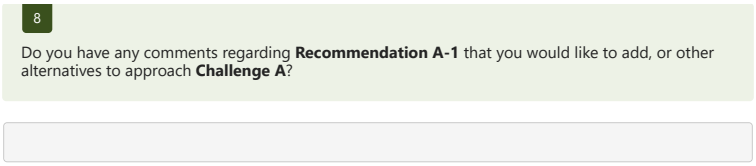


Figure D.4: Page 4 of the validation survey form.

Topic 2: Open Data Disconnects Communication

9

Challenge B: Open data disconnects communication *

Explanation: In open data which has been widely adopted in environmental research, data hosts play an important role as data aggregators, which has resulted in more communication being directed to data hosts at the expense of less communication between data producers and data consumers.

Do you agree or disagree that this is a challenge in environmental research in regards to data and data quality?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Challenge B	◯	◯	◯	◯	◯	◯

10

Do you have any comments regarding **Challenge B** that you would like to add?

11

Recommendation B-1: Data hosts should allow for community building around data *

Background: GitHub is a platform where developers can store, share, and collaborate on code for software projects. It allows open source developers to easily collaborate, track changes, manage projects, and contribute to open source software.

Recommendation: Data hosts should act as platforms for community building in order to encourage communication, collaboration, and feedback sharing between data producers and data consumers.

Example: Data hosts such as Copernicus could provide communication channels similar to those on github, where each project has an open discussion forum for questions and comments.

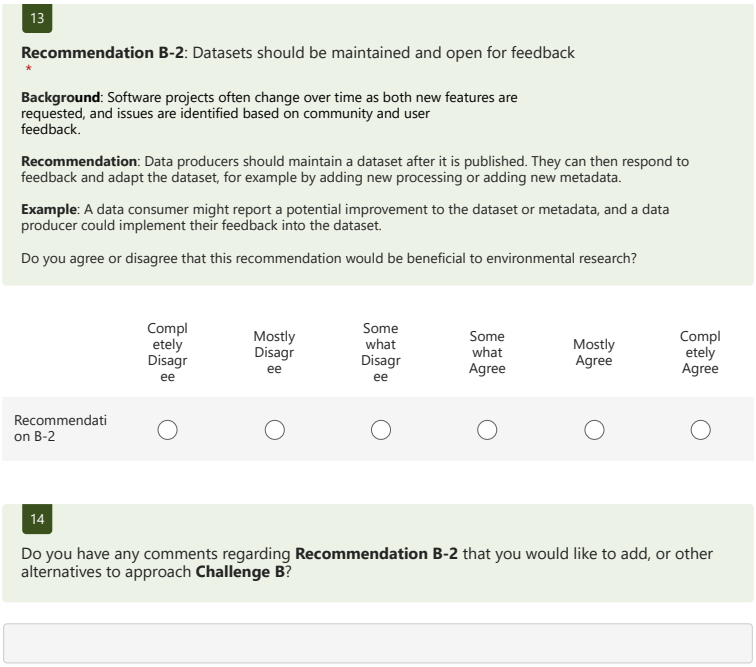
Do you agree or disagree that this recommendation would be beneficial to environmental research?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Recommendation B-1	◯	◯	◯	◯	◯	◯

12

Do you have any comments regarding **Recommendation B-1** that you would like to add?

Figure D.5: Page 5 of the validation survey form.



13

Recommendation B-2: Datasets should be maintained and open for feedback
*

Background: Software projects often change over time as both new features are requested, and issues are identified based on community and user feedback.

Recommendation: Data producers should maintain a dataset after it is published. They can then respond to feedback and adapt the dataset, for example by adding new processing or adding new metadata.

Example: A data consumer might report a potential improvement to the dataset or metadata, and a data producer could implement their feedback into the dataset.

Do you agree or disagree that this recommendation would be beneficial to environmental research?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Recommendation B-2	◯	◯	◯	◯	◯	◯

14

Do you have any comments regarding **Recommendation B-2** that you would like to add, or other alternatives to approach **Challenge B**?

Figure D.6: Page 6 of the validation survey form.

Topic 3: Differences in Data Quality Handling and Needs

15

Challenge C: Data quality needs and handling differ *

Explanation: Data producers and data consumers have different priorities, needs, and expectations for data quality, and their understanding of each other is hampered by infrequent communication. Data consumers often want to use data in other contexts than those it was originally produced for, while data producers are unaware of the additional contexts in which their data is being used.

Do you agree or disagree that this is a challenge in environmental research in regards to data and data quality?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Challenge C	◯	◯	◯	◯	◯	◯

16

Do you have any comments regarding **Challenge C** that you would like to add?

17

Recommendation C-1: Use elicitation techniques from requirements engineering *

Background: In requirements engineering elicitation is a set of techniques used to identify stakeholder needs and expectations.

Recommendation: Data producers should make use of elicitation techniques to identify data consumers' needs and expectations for data quality. This should preferably be done through communication platforms provided by data hosts.

Example: Data producers could interview potential data consumers at the beginning of a project, to identify their needs and expectations, so that the dataset can be adapted to be suitable for use.

Do you agree or disagree that this recommendation would be beneficial to environmental research?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Recommendation C-1	◯	◯	◯	◯	◯	◯

18

Do you have any comments regarding **Recommendation C-1** that you would like to add?

Figure D.7: Page 7 of the validation survey form.

19

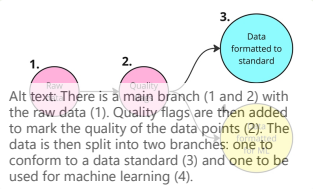
Recommendation C-2: Use branching to have multiple versions of the data available *

Background: Branching is a method used in version control systems, such as GitHub, to create separate versions of the main program code (branches) to work on new features or fixes without affecting the main code.

Recommendation: Data hosts should make use of branching, where raw data is the main dataset, and versions of the main dataset are available through branches that provide data processed in different ways for different users.

Example: As seen in the figure, there is a main branch (1 and 2) with the raw data (1). Quality flags are then added to mark the quality of the data points (2). The data is then split into two branches: one to conform to a data standard (3) and one to be used for specific use cases, for example machine learning (4).

Do you agree or disagree that this recommendation would be beneficial to environmental research?



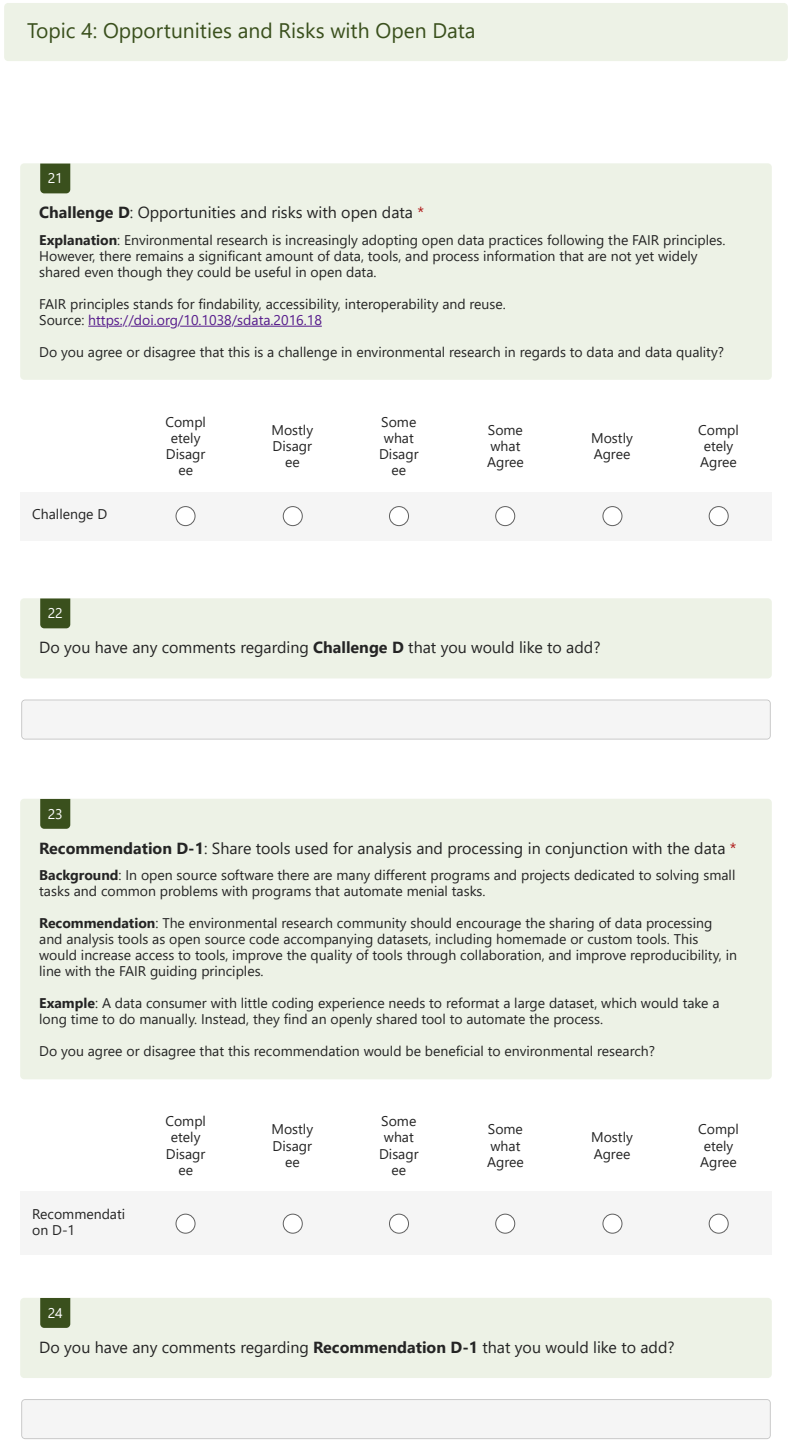
Alt text: There is a main branch (1 and 2) with the raw data (1). Quality flags are then added to mark the quality of the data points (2). The data is then split into two branches: one to conform to a data standard (3) and one to be used for machine learning (4).

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Recommendation C-2	◯	◯	◯	◯	◯	◯

20

Do you have any comments regarding **Recommendation C-2** that you would like to add, or other alternatives to approach **Challenge C**?

Figure D.8: Page 8 of the validation survey form.



Topic 4: Opportunities and Risks with Open Data

21

Challenge D: Opportunities and risks with open data *

Explanation: Environmental research is increasingly adopting open data practices following the FAIR principles. However, there remains a significant amount of data, tools, and process information that are not yet widely shared even though they could be useful in open data.

FAIR principles stands for findability, accessibility, interoperability and reuse.
Source: https://doi.org/10.1038/sdata.2016.18

Do you agree or disagree that this is a challenge in environmental research in regards to data and data quality?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Challenge D	◯	◯	◯	◯	◯	◯

22

Do you have any comments regarding **Challenge D** that you would like to add?

23

Recommendation D-1: Share tools used for analysis and processing in conjunction with the data *

Background: In open source software there are many different programs and projects dedicated to solving small tasks and common problems with programs that automate menial tasks.

Recommendation: The environmental research community should encourage the sharing of data processing and analysis tools as open source code accompanying datasets, including homemade or custom tools. This would increase access to tools, improve the quality of tools through collaboration, and improve reproducibility, in line with the FAIR guiding principles.

Example: A data consumer with little coding experience needs to reformat a large dataset, which would take a long time to do manually. Instead, they find an openly shared tool to automate the process.

Do you agree or disagree that this recommendation would be beneficial to environmental research?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Recommendation D-1	◯	◯	◯	◯	◯	◯

24

Do you have any comments regarding **Recommendation D-1** that you would like to add?

Figure D.9: Page 9 of the validation survey form.

25

Recommendation D-2: Include a ReadMe in datasets to communicate important information about the data *

Background: GitHub projects typically include a ReadMe file that explains the purpose of the program, reducing the risk of misuse. Essential details that are also reported which include usage, contributors, and open source licenses.

Recommendation: Data producers should provide clear descriptions in their datasets about usage, including the purpose of collection, contributors, intended use, and used open data licenses, to ensure clarity and transparency for data consumers.

Example: When a data consumer accesses a dataset, they can explicitly read how the data was collected and in what ways they can use it.

Do you agree or disagree that this recommendation would be beneficial to environmental research?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Recommendation D-2	◯	◯	◯	◯	◯	◯

26

Do you have any comments regarding **Recommendation D-2** that you would like to add?

27

Recommendation D-3: Release data early and update it as it develops *

Background: Software projects often involve releasing early versions, such as alpha or beta versions, similar to a first draft. This allows for better collaboration, finding issues, and refinement before wider public use.

Recommendation: Data producers should release raw data as soon as possible after data collection as a preliminary release, and continually update the dataset with new processed versions. This allows for an early use of new data by other researchers. It also reduces the risk of duplicate data collection, both for projects with a similar context to the original use case and for projects where the data is applied in novel contexts.

Example: Someone in a research project discovers that a data producer is currently working on a dataset that they could use, and contacts them to collaborate on adapting the dataset for their own project.

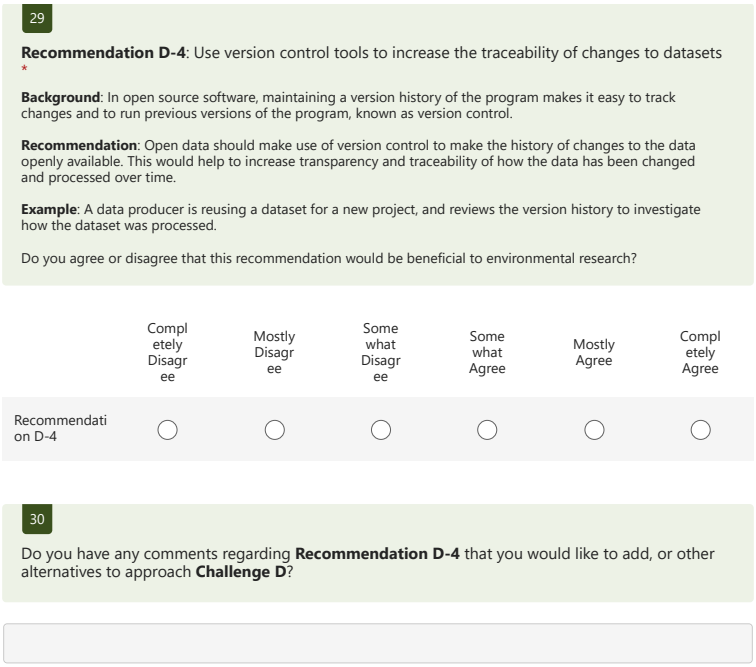
Do you agree or disagree that this recommendation would be beneficial to environmental research?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Recommendation D-3	◯	◯	◯	◯	◯	◯

28

Do you have any comments regarding **Recommendation D-3** that you would like to add?

Figure D.10: Page 10 of the validation survey form.



29

Recommendation D-4: Use version control tools to increase the traceability of changes to datasets *

Background: In open source software, maintaining a version history of the program makes it easy to track changes and to run previous versions of the program, known as version control.

Recommendation: Open data should make use of version control to make the history of changes to the data openly available. This would help to increase transparency and traceability of how the data has been changed and processed over time.

Example: A data producer is reusing a dataset for a new project, and reviews the version history to investigate how the dataset was processed.

Do you agree or disagree that this recommendation would be beneficial to environmental research?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Recommendation D-4	◯	◯	◯	◯	◯	◯

30

Do you have any comments regarding **Recommendation D-4** that you would like to add, or other alternatives to approach **Challenge D**?

Figure D.11: Page 11 of the validation survey form.

Topic 5: Metadata are Key to Understandability

31

Challenge E: Metadata are key to understandability *

Explanation: Metadata is crucial to the usefulness of data. A great deal of effort goes into creating correct and sufficient metadata, and without it, the value of the data is significantly reduced, or it may even become unusable.

Do you agree or disagree that this is a challenge in environmental research in regards to data and data quality?

	Completely Disagree	Mostly Disagree	Somewhat Disagree	Somewhat Agree	Mostly Agree	Completely Agree
Challenge E	◯	◯	◯	◯	◯	◯

32

Do you have any comments or suggestions for approaches to **Challenge E** that you would like to add?

Figure D.12: Page 12 of the validation survey form.

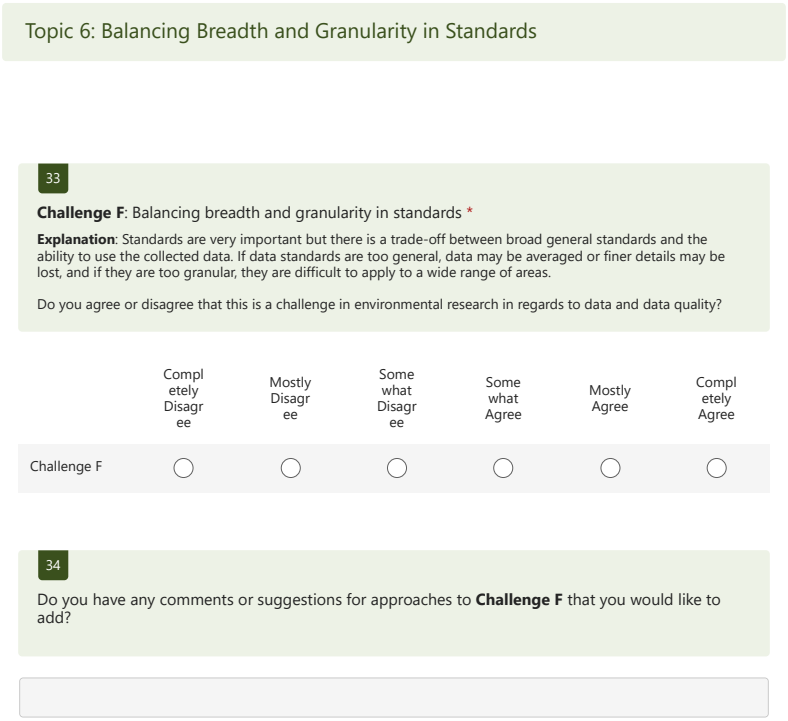


Figure D.13: Page 13 of the validation survey form.

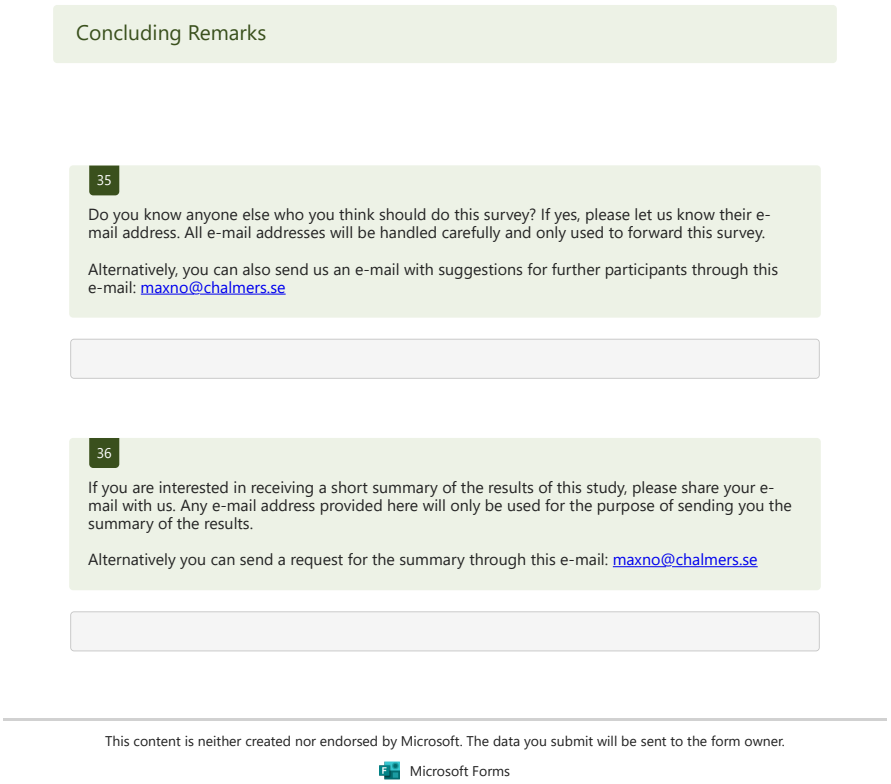


Figure D.14: Page 14 of the validation survey form.