CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Data Quality Requirements Reporting of Open Datasets in Environmental Research

A sample study to identify the gaps in data quality reporting between open dataset stakeholders

Master's thesis in Computer science and engineering

Georgios Efthymiou

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

Master's thesis 2024

Data Quality Requirements Reporting of Open Datasets in Environmental Research

A sample study to identify the gaps in data quality reporting between open dataset stakeholders

Georgios Efthymiou

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

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

Data Quality Requirements Reporting of Open Datasets in Environmental Research
A sample study to identify the gaps in data quality reporting between open dataset stakeholders
Georgios Efthymiou

Supervisor: Hans-Martin Heyn, Computer Science and Engineering
Examiner: Jennifer Horkoff, 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

Data Quality Requirements Reporting of Open Datasets in Environmental Research
A sample study to identify the gaps in data quality reporting between open dataset stakeholders
Georgios Efthymiou
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Environmental research requires a large amount of data that are often provided by various private and public organizations. This ease of access introduces concerns about the quality of this data and how applicable they can be for the purposes of each individual task. This study aims to identify key data quality aspects (DQAs) and how they are viewed from the perspectives of three stakeholders: data portal maintainers, dataset providers, and dataset consumers.
The perspectives of these stakeholders were investigated using a sample study, which is also known as a survey study. Different populations were identified and contacted to partake in a questionnaire containing found DQAs from different sources and prompted to evaluate the importance from their role's perspective. The collected responses were analysed using Bayesian data analysis. It was identified that different roles have different perspectives when it comes to the data quality aspects that should be prioritized during reporting. This research emphasizes the importance of effective and transparent communication between stakeholders and enhance the quality of the data used in environmental research.

Keywords: Data Quality, Open Datasets, Environmental Research, Software Engineering, AI Engineering, Data Portals, Dataset Providers, Dataset Consumers.

Acknowledgements

I would like to thank my academic supervisor Hans-Martin Heyn for supporting me over the course of this thesis. His suggestions and inputs were very insightful and helped me improve my skill-set. In addition, I would like to appreciate my examiner Jennifer Horkoff, who provided invaluable feedback with the report. Furthermore, I would like to recognize my colleagues who have supported with the validation of different steps in this study. In particular, I would like to express my gratitude towards Felix Bölle, Markus Moen, Max Norén, Oleksandr Panasenko and Dmytro Siniukov, who supported with validating the different instruments created during the study. Also, I would like to acknowledge the invaluable input of the respondents to the questionnaire, as without them this thesis would not be possible. Finally, I would like to thank my family for supporting me and encouraging me during this endeavour.

Georgios Efthymiou, Gothenburg, June 2024

Contents

List of Figures **xi**

List of Tables **xv**

1 Introduction **1**
1.1 Background . . . 2
1.2 Statement of the Problem . . . 6
1.3 Research Questions . . . 7
1.4 Purpose of the study . . . 7
1.5 Outline of the study . . . 8

2 Related Work **9**

3 Methods **13**
3.1 Study Methodology Overview . . . 13
3.2 DQA Collection . . . 13
3.3 Sample Survey Process . . . 16
3.4 Bayesian Data Analysis Methodology . . . 20

4 Results **27**
4.1 DQA Collection . . . 27
4.2 Survey Results . . . 33

5 Conclusion **53**
5.1 Discussion . . . 53
5.2 Threats to Validity . . . 57
5.3 Suggestion For Future Work . . . 60
5.4 Conclusion . . . 61

Bibliography **63**

A Appendix 1 **I**

List of Figures

2.1 Recreation of example missing cases from McElreath 11

3.1 Example question in questionnaire for DQAs 19
3.2 Proposed DAG for each DQA investigated in the Ratings questions . 22
3.3 Prior predictive checks for rating questions regardless of DQA 23
3.4 Posterior predictive samples along with observed data for the Ratings model 23
3.5 Posterior density plots for the Ratings model 24
3.6 Proposed DAG for the Pick-5 most/least relevant DQAs question . . 24
3.7 Prior predictive checks for rating questions regardless of DQA 25
3.8 Posterior predictive samples along with observed data for the Pick-5 model 26
3.9 Posterior density plots for the Pick-5 model 26

4.1 Publication year of literature selected for analysis. 27
4.2 Count of Updated DQAs in the investigated literature 29
4.3 Distribution of roles in the results datasets 35
4.4 Distribution of the data management experience in the dataset. . . . 35
4.5 Distribution of the data science experience in the dataset. 35
4.6 Distribution of the environmental research experience in the dataset . 35
4.7 Distribution of the data of the closed-ended rating questions. 41
4.8 “Most Relevant” Pick-5 responses separated by role. 42
4.9 “Least Relevant” Pick-5 responses separated by role. 42
4.10 Posterior predictive check of the Accessibility model. 43
4.11 Density plots of betas for the Accessibility model. 44
4.12 Forest plot representation of the Accessibility trace. 45
4.13 Forest plot representation of the Accuracy trace. 45
4.14 Forest plot representation of the Completeness trace. 46
4.15 Forest plot representation of the Consistency trace. 46
4.16 Forest plot representation of the Credibility trace. 46
4.17 Forest plot representation of the Findability trace. 47
4.18 Forest plot representation of the Interoperability trace. 47
4.19 Forest plot representation of the Interpretability trace. 47
4.20 Forest plot representation of the Relevance trace. 48
4.21 Forest plot representation of the Reusability trace. 48
4.22 Forest plot representation of the Security trace. 49
4.23 Forest plot representation of the Timeliness trace. 49

4.24 Forest plot representation of the Understandability trace. 50
4.25 Forest plot representation of the Usability trace. 50
4.26 Forest plot representation of the Value Added trace. 50
4.27 Forest plot representation of the “Most Relevant” trace. 52
4.28 Forest plot representation of the “Least Relevant” trace. 52

A.1 Snowball trail I
A.2 DQAs along with definitions from literature II
A.3 Output of Thematic Analysis after the review from the expert. III
A.4 Host evaluation descriptive results IV
A.5 Updated DQAs after clean up. Ordered by name V
A.6 Distribution of the data for Data Management - No Experience . . . XIV
A.7 Distribution of the data for Data Management - Less than 1 year . . XIV
A.8 Distribution of the data for Data Management - 1-2 years XV
A.9 Distribution of the data for Data Management - 3-5 years XV
A.10 Distribution of the data for Data Management - 5+ years XV
A.11 Distribution of the data for Data Science - No Experience XVI
A.12 Distribution of the data for Data Science - Less than 1 year XVI
A.13 Distribution of the data for Data Science - 1-2 years XVI
A.14 Distribution of the data for Data Science - 3-5 years XVII
A.15 Distribution of the data for Data Science - 5+ years XVII
A.16 Distribution of the data for Environmental Research - Less than 1 yearXVIII
A.17 Distribution of the data for Environmental Research - 1-2 years . . . XVIII
A.18 Distribution of the data for Environmental Research - 3-5 years . . . XIX
A.19 Distribution of the data for Environmental Research - 5+ years . . . XIX
A.20 Posterior predictive check of the Accuracy model. XX
A.21 Posterior predictive check of the Completeness model. XX
A.22 Posterior predictive check of the Consistency model. XXI
A.23 Posterior predictive check of the Credibility model. XXI
A.24 Posterior predictive check of the Findability model. XXI
A.25 Posterior predictive check of the Interoperability model. XXII
A.26 Posterior predictive check of the Interpretability model. XXII
A.27 Posterior predictive check of the Relevance model. XXII
A.28 Posterior predictive check of the Reusability model. XXIII
A.29 Posterior predictive check of the Security model. XXIII
A.30 Posterior predictive check of the Timeliness model. XXIII
A.31 Posterior predictive check of the Understandability model. XXIV
A.32 Posterior predictive check of the Usability model. XXIV
A.33 Posterior predictive check of the Value-Added model. XXIV
A.34 Density plots of betas for the Accuracy model. XXV
A.35 Density plots of betas for the Completeness model. XXVI
A.36 Density plots of betas for the Consistency model. XXVI
A.37 Density plots of betas for the Credibility model. XXVII
A.38 Density plots of betas for the Findability model. XXVII
A.39 Density plots of betas for the Interoperability model. XXVIII
A.40 Density plots of betas for the Interpretability model. XXVIII

A.41 Density plots of betas for the Relevance model. XXIX
A.42 Density plots of betas for the Reusability model. XXIX
A.43 Density plots of betas for the Security model. XXX
A.44 Density plots of betas for the Timeliness model. XXX
A.45 Density plots of betas for the Understandability model. XXXI
A.46 Density plots of betas for the Usability model. XXXI
A.47 Density plots of betas for the Value Added model. XXXII
A.48 Posterior predictive check of the Pick-5 "Most Relevant" model. . . . XXXVIII
A.49 Posterior predictive check of the Pick-5 "Least Relevant" model. . . . XXXVIII
A.50 Density plots of betas for the Pick-5 “Most Relevant” model. XXXIX
A.51 Density plots of betas for the Pick-5 “Least Relevant” model. XL

List of Tables

4.1 Sources and filters applied during literature review 27
4.2 Example DQAs grouped using Merriam-Webster 28
4.3 Data Portal Sources distribution 30
4.4 Data portals collection and initial evaluation 31
4.5 DQAs included in the Likert Scale of the survey 33
4.6 DQAs included in the Pick-5 section of the survey 33
4.7 Number of contacts for the different stakeholder groups 34
4.8 Updates from the proposed roles of the respondents 34
4.9 Respondents background for rating question comments 36
4.10 Thematic Analysis of Respondent comments on Accessibility 36
4.11 Thematic Analysis of Respondent comments on Accuracy 37
4.12 Thematic Analysis of Respondent comments on Completeness 37
4.13 Thematic Analysis of Respondent comments on Consistency 37
4.14 Thematic Analysis of Respondent comments on Credibility 38
4.15 Thematic Analysis of Respondent comments on Interoperability . . . 38
4.16 Thematic Analysis of Respondent comments on Relevance 38
4.17 Thematic Analysis of Respondent comments on Reusability 39
4.18 Thematic Analysis of Respondent comments on Security 39
4.19 Thematic Analysis of Respondent comments on Timeliness 39
4.20 Thematic Analysis of Respondent comments on Understandability . . 40
4.21 Thematic Analysis of Respondent comments on Usability 40
4.22 Thematic Analysis of Respondent comments on Value Added 40
4.23 Respondents' background for missing DQAs question 40
4.24 Thematic Analysis of Missed DQAs based on Respondents 41
4.25 Trace representation of Accessibility model prediction 44

A.1 Trace representation of Accuracy Model prediction XXXII
A.2 Trace representation of Completeness model prediction XXXIII
A.3 Trace representation of Consistency model prediction XXXIII
A.4 Trace representation of Credibility model prediction XXXIII
A.5 Trace representation of Findability model prediction XXXIV
A.6 Trace representation of Interoperability model prediction XXXIV
A.7 Trace representation of Interpretability model prediction XXXIV
A.8 Trace representation of Relevance model prediction XXXV
A.9 Trace representation of Reusability model prediction XXXV
A.10 Trace representation of Security model prediction XXXV

A.11 Trace representation of timeliness model prediction XXXVI
A.12 Trace representation of Understandability model prediction XXXVI
A.13 Trace representation of Usability model prediction XXXVI
A.14 Trace representation of Value Added model prediction XXXVII
A.15 Trace representation of pick-5 "Most Relevant" model prediction . . . XL
A.16 Trace representation of pick-5 "Least Relevant" model prediction . . . XLI

1

Introduction

Climate change is mostly undisputed nowadays, however there are concerns spanning almost half a century about the reliability of climate models and their outputs [3]. At the same time, Machine Learning (ML) is becoming more relevant in environmental research [28], as well as in other aspects of modern society [51]. This can introduce further issues considering that, as presented in multiple sources collected by Sambasivan et al. [53], statistical models and especially ML ones are fundamentally iterative and exploratory requiring a large amount of data that often require annotation and are separated into datasets [23]. The quality of the training data determines the quality of the future predictions of an ML model. This is commonly known in software engineering as "garbage-in, garbage-out" [39].

The quality of the data therefore plays a fundamental role in ML to ensure the appropriate behaviour of the models. Many datasets with environmental data are freely available on the internet. This study attempted to investigate the reporting of the quality aspects of such data and identifying which should be prioritized between the interactions of different stakeholder groups.

Earlier studies indicate that data quality for environmental research can be subject to many issues, as identified for example by Nguyen et al. [46]. Unfortunately, these are not always communicated correctly [56] and different frameworks have been proposed to understand the risk associated with quality [2] or putting it in the spotlight of the requirements' collection [6]. In other fields, such as software engineering, data, or software code, play a critical role in sharing quality information. As a well established field in research, software engineering has established approaches in order to enable practitioners to perform such actions. Therefore, the aim of this study is to investigate how different stakeholders can understand and prioritize their needs.

To achieve this, data portals were investigated using document analysis, with emphasis on ones relating to environmental research. Also, a literature review with systematic elements was performed to identify aspects of data quality that currently exist. After accumulating a representative set of different data quality aspects (DQAs), a survey questionnaire was created and shared to open data stakeholders in order to gather their views on these aspects. The results were analysed using Bayesian data analysis in order to identify the priorities of the investigated stakeholders. The conclusion of this study was identifying which of these aspects should be prioritized in the stakeholder interactions and how software engineering knowledge and techniques can be used in order to best enable them.

1.1 Background

This section provides concepts and terms mentioned throughout this thesis.

Environmental Research

Environmental research made its first appearance in the late 19th century in Sweden, and since then it has gone through many transformations. From small independent efforts on specific aspects of the environment, to struggling attempts of unifying formats and standards, to a global interconnected system of hundreds of scientists who work on defined standards to help create the most accurate models as possible [13, 15]. Through these endeavours, climate scientists have learned a lot on the potential of creating a reliable representation of the climate and how it can affect the ecosystems in the future [15].

Climate Models

A climate model, or any mathematical model for that matter, is a set of equations that aims to understand the relationships described in the theory surrounding said model and using appropriate data to correctly investigate values that have not been observed before [13]. For many years, there has been a discourse around the reliability of the climate models' output [3]. With data being such a critical component of such solutions [53], it is important to consider using software engineering practices in order to increase the reliability of their predictions [17, 39].

Data Procurement

Data are usually viewed as just being part of a platform but in reality they are living in an "ecosystem" with different practices maturing alongside them which can introduce difficulties in their procurement[1]. Such activities include practices surrounding data collection, data representation, data quality assessment and data sharing [46].
It is difficult to assess the quality of the data that are collected. A good understanding of the environment is needed [23] to evaluate them, and the sensors must be capable of capturing information sufficiently [13]. Sensors unfortunately can fail due to brittleness inherently present in the physical world [53] leading to missing data and values [38, 41]. This raises the need for requirements and design decisions that should be taken before data collection starts to create feedback loops and raise awareness to the potential issues regarding data quality [8]. If these steps are not taken, different issues can potentially be compounded on top of each other, leading to a tangled situation where the root cause is difficult to identify. These events are known as "data cascades" [53].
This general uncertainty of the quality and quantity is introducing a trade-off when annotation costs are part of the procurement effort [23]. Providing consistent an-

[1]Procurement refers to the process of obtaining supplies of something - link: Oxford Definition of Procurement.

notations is also a challenge, with a need to define specifications and guidelines. Ambiguous definitions introduce human errors due to judgment calls from the annotators. These sources of bias can prove a major challenge for data scientists [39].

Data Quality

Before trying to improve data quality, it is important to understand its definition. It can vary in the literature, but one of the most common ones presented is **Fitness For Use** [30, 60, 65]. In some cases, data quality is viewed from a straightforward perspective by placing emphasis on assuring "completeness" which is an important DQA but not the only one that should be considered. Completeness is achieved by defining strategies to handle missing data points or values. This can be done with different approaches, including imputations, meaning replacements based on different data sources [3, 39] or removing missing cases all together [39]. Extra care must be taken in such cases, as they can introduce biases if the relationships between the data is not fully understood [38].
Data quality, even from this straightforward prospective, is more complicated than what it initially seems. As a critical component of climate models, it is important to acknowledge this. A software engineering view can potentially help identify the present gaps and suggest improvements with some potential adjustments [5, 26].

Poor Data Quality

Poor data quality can negatively impact the climate models, leading to unreliable predictions [59]. There is also a financial incentive to handling quality. For example, it was discovered in an investigation in the US that poor quality reduces effectiveness and efficiency in various businesses, leading to billions of dollars in loses [2].
Poor quality can appear in many forms, like with the introduction of noise, bias, inaccuracies, heterogeneity in the data or imbalances, among other issues that lead to misleading and unreliable representations of reality [17, 68]. These issues are considered to be a significant challenge by practitioners [33]. Quite often they have to spend a lot of time on querying, cleaning and shaping data among other activities [60] which can be considered a mundane and deglamoured activity [33]. Such issues appear due to multiple reasons relating to faulty sensors, wrong or irrelevant practices, and bad communication between the stakeholders as they are not able to convey their needs [13, 33, 53].
Even when data are not subject to such issues initially, they should still be considered as physical objects, which are subjected to "friction" and "corruption" as they are moved to different storage warehouses or are not fit for their expected use in the long run [48]. Reliable data are an accomplishment and not a natural state, and it must be earned with the expertise and judgment of the stakeholders. Unfortunately, when these tasks are done well, the work is often invisible and not appreciated [60].

Data Requirements

Data Requirements are essential when developing a data intensive system, due to the high impact of data on their output. This raises the need for elicitation practices and

Non-Functional Requirements [5, 26]. Incomplete or incorrect requirements can lead to *garbage-in, garbage-out*, where the output of a model does not meet expectations because of unsuitable or insufficient input data for training [64]. In a worse case this can lead to *garbage-in, package-out* which happens when a model is deployed in the production environment as part of a product and behaving in an unexpected manner [39].
Data coming directly from the collection can often be considered "raw" and thus mistakenly considered to be ground truth [42, 47, 53]. Data are collected and shaped based on the existing practices for collection, curation and sense-making, giving them a socio-political dimension that is often ignored. Simply defining technical specifications around data quality requirements will not be enough to address the issues related to defining the necessary traceability of ML model output to training data quality.
There is a need for improved articulation and coordination to successfully convey these requirements to the stakeholders [60]. Establishing traceable communication while following a defined approach [42, 46] can help avoid ambiguities, judgement calls [39] and broken practices [53]. Version and revision practices need to be established to reduce complexity and the difficulties around maintainability of the existing data [43].
The requirements regarding the data collection, data formats and the ranges of data need to be defined clearly and concisely. This can mitigate the accidental selection of unsuitable data and help define mechanisms that address data and data quality requirements in a structured manner [46, 53, 64]. In addition, version and revision practices need to be established to reduce complexity and the difficulties around maintainability of the existing data [43].

Data Requirements in Software Engineering

Data requirements have an important role in software engineering and its knowledge areas [8]. They include procedures like elicitation, analysis, specification, and validation&verification [26, 64]. These are needed for completing relevant tasks like: (1) identifying quality data sources, (2) facilitating productive discussions between stakeholders, (3) explicitly specifying data quantity and quality needs, (4) defining strategies to handle anomalies in the collections and handling confidentiality among others [33, 40].
This requires knowledge that is usually found in *software requirements*. Ensuring data quality mirrors *software quality* along with *mathematical and engineering foundations* to analyse the collected data. Testing of data aligns with *software construction&testing*. While managing data frictions and corruption [48], can be considered part of Maintenance.
Also, data collection requires coordination, management and planning [25, 33, 39, 46] which are part of *software engineering management*. Increasing data volumes requires appropriate resource allocations aligning with the need for *software engineering economics*. Making the data accessible afterwards can align with the *software engineering process*. Defining data formats that can be seamlessly integrated in the produced software show the need for *software design*.
In summary, there is a multifaceted relationship between data requirements and

software engineering that needs to be recognized. Good practices can elevate the value of data and mitigate potential challenges in software development.

Software Engineering in Environmental Research

Climate scientists had many years to refine their methods and learn to implicitly make use of software engineering practices [13, 15]. They also have some distinct advantages compared to their commercial colleagues. They are not competitors and can freely share information, which increases transparency and promotes higher work standards [13]. This leads to a remarkable set of test design practices as discovered by Easterbrook [13], with new standards proposed often evolving their practices based on publicly available knowledge [3].
Environmental research can be elevated through innovation and explicit relation to software engineering [51]. This is especially important with the raise of ML [28]. Data need to represent reality as even small imbalances can be impactful, and it is important to be able to identify issues as early as possible [26, 39, 46].

AI Engineering

AI engineering [7] is an extension of software engineering that is introducing new processes and technologies relating to the development of AI systems. In this model emphasis is based on aspects of ML development including data quality. Data versioning and dependency management in architecture and development appears in the form of Data-Ops. Also, having procedures for creating a data trail and data generation procedures is promoted for traceability to help reduce biases.
Traditional requirements engineering practices of elicitation, analysis, specification, verification&validation can be adapted in order to allow for AI components to be more easily maintained [64]. Data scientists already have the skill-set to cover most of these needs, but it is important for an organized approach to help guide these efforts, and that can be achieved by following requirement engineering practices to convey their needs and expectations to relevant stakeholders [4].

Open Data Stakeholders

In the open data pipelines, there are different stakeholders present in different stages. For this study, the following ones were investigated:
Data portal maintainers, who are responsible for providing platforms that can be used by the other stakeholders to share datasets and enable engagement with each other [30].
Dataset providers, who are responsible for preparing the datasets that are accessible in the portals. This is usually done through research grants which require the datasets to become accessible afterwards, or through the drive of different science groups to improve the shared knowledge. Such endeavours are especially evident on environmental research, where some providers are even providing platforms for public access.
Dataset consumers, who use datasets collected from the portals to build their solutions. In the context of environmental research, such solutions include climate

models. Consumers have a more distant view of the data procurement and usually can't easily distinguish the nuances of quality relating to the software or hardware needed to provide them their data [30, 46]. As such, they need the Portals to enable them to gather this information before using the datasets in their models.

FAIR Principles

The FAIR Principles stand for (1) Findability, (2) Accessibility, (3) Interoperability and (4) Reusability. These will be treated as DQAs later in the study. They were first published in 2016 [66] as a collaborative effort of many researchers. These principles provide a set of guidelines to address challenges with managing and sharing data, aiming to benefit the scientific community as a whole. Since then, they have become the guiding principles of many data portals.

Bayesian Data Analysis

For the analysis of the quantitative data that were collected from the surveys, Bayesian data analysis was used. With this approach, a model is defined which tries to simulate the distribution probabilities of different variables. These variables include *independent variables*, that are changed by the model, *dependent variables*, that change due to the independent variables and *control variables*, that remain constant in the model.
To understand the relationship of these variables, it is necessary to do causal inference. This is done in order to identify confounders that might affect how different variables will interact with each other if present in the model. Usually this inference is presented in the form of directed acyclic graphs (DAG) An example of such a DAG can be seen in the next chapter in Figure 2.1.

1.2 Statement of the Problem

Data is an important aspect in ML, as it enables the scientists to get an accurate picture of the environment where their model is expected to be used. Unfortunately, the data aspect is often undervalued in favour of model performance and architecture [53]. It is important to understand how crucial data are, as they affect performance, robustness, and reliability [25, 47, 53, 60]. Shortcomings in the data are not always identified by the models, or in worse cases they can lead models to follow "spurious cues" resulting in unreliable performance [47].
ML models are opaque constructs that are difficult to explain intrinsically [7, 24, 26, 64]. ML models are also data-intensive [43] and developed in an iterative and exploratory fashion [46]. Data requires collection, annotation and potentially significant manual effort [23, 33, 46, 60]. Data pipelines can reduce manual effort in data handling, but they can be very costly if not handled correctly [7, 68]. Also, not all ML models are the same and require different DQAs to improve performance [10].
These impediments become more noticeable in datasets prepared and provided from public sources. Collected data have to be used-as-is in most cases, without the option for further enrichment from the original sensor. Also, publicly sourced data

might have been initially collected with different requirements in mind. This can make the use of such data on other tasks ultimately unreliable due to divergent requirements, leading to inconsistencies and errors [46]. This is especially problematic as nowadays large distributed teams are expected to work on the same model and good architecture is needed [13].
Such difficulties with managing data quality in ML development can be associated with a number of existing software engineering practices such as *software requirements*, *software design*, *software engineering process* or *software engineering professional practices* [8]. Software engineering has made major efforts in building these different practices and has provided solutions that can enable software engineers to perform their tasks more effectively. Environmental researchers can learn from such practices. This is especially relevant with ML, being a new and popular field in the last decade [4]. Its boom has led to ever-increasing procurement efforts. Ongoing research in software engineering attempts to address the challenges when it comes to using the currently available data [33], as well as identifying the scientist's needs [23].

1.3 Research Questions

For this study, the following research questions are proposed:

RQ 1: **How is Data Quality reported in environmental research by the data portals?**
The purpose of this RQ is to obtain a comprehensive understanding of how data quality is defined and communicated through data portals to the other stakeholders. The insights from this RQ will be quite valuable moving forward in this study, showing the different practices present in the portals.

RQ 2: **Which aspects of data quality are important from the perspective of different stakeholders associated with open data in environmental research?**
The research objective of this question is to understand if the expectations of the different data stakeholders in environmental research are aligned on their views surrounding quality reporting. The insights created can provide a more holistic view of data quality surrounding open data in environmental research and provide insights on how software engineering practices can potentially be used to improve the engagement of these stakeholders.

1.4 Purpose of the study

This study attempts to identify what are the important DQAs for environmental research and how they are identified and handled by the relevant stakeholders. To achieve this understanding, the problem will be viewed from different angles through the lens of three stakeholders. These perspectives, also known as actors [57], were picked considering their relevance to the different aspects of the open data used in environmental research:

- **Data Portal Maintainers** - who are the maintainers of platforms to share data
- **Dataset Providers** - who provide datasets to the portals
- **Dataset Consumers** - who use the data from the portals

Currently, these stakeholders are facing difficulties in their interactions due to different understanding and knowledge regarding data [46]. Dataset providers might have specific requirements when procuring data, and their collections potentially can't be used for other tasks [46]. Dataset consumers might not be equipped to understand the caveats of the data they want to use in their models [25, 33, 53]. Data portal maintainers might not provide appropriate mechanisms to enable the engagement of these stakeholders.
Simply conforming to the quality specifications is not enough from the perspective of consumers [30]. The aim of this study is to identify best practices and gaps currently present which can help make more educated decisions when selecting datasets, as well as help strengthen the findings of climate research initiatives. This can help integrate software engineering more closely to the environmental research pipelines [8]. Ultimately, the approaches followed in this study can potentially be used in other data pipelines as well.

1.5 Outline of the study

This report is separated into the following sections:

Section 2 - Related Work: Includes related literature to the methodology and research objectives of this study.

Section 3 - Methods: Provides the outline of the methodologies followed during the study. In particular:

- The general methodology of the study is presented.
- The plan for collecting the different aspects of data quality along with the thematic analysis.
- The sample survey process.
- The methodology along with the prior checks for the Bayesian data analysis.

Section 4 - Results: States the output of the executed study along with the different DQAs that were collected. It also includes the Bayesian data analysis of the survey responses along with the thematic analysis of the open-ended questions.

Section 5 - Conclusion: Presents the output of the study and derives insights from the collected data, connecting them to the research questions set by the study and software engineering at large. Also, provides suggestions for future research and validity threats are presented.

2
Related Work

This section will discuss the studies relating to the identification of DQAs and the difficulties of reporting them to different stakeholders.

In a study on marine life by Nguyen et al. [46], the authors identified DQAs around environmental data focusing on correctness, currentness, completeness, consistency. They noted factors that can degrade quality including slow data transmissions, sensor drifts, changes in the environment and biofouling. Biofouling refers to the degradation of equipment due to damages introduced from organisms in the deployed environment. Due to the high cost of collecting data in these conditions, the authors have raised the need for reusability, which is a view is supported by other sources as well [3, 23, 31]. The authors have also identified limitations in different areas of the data flow, and they provide some recommendations, like putting emphasis on enabling shareability when preparing requirements.

Sovacool et al., investigated the trends around data quality for the past 30 years regarding climate change regarding publicly funded projects [56]. They identified that the quality of information in these projects can be limited and inconsistent. They argue that there is a need for data managers and research councils to step up and focus on the quality and accuracy surrounding data.

The LoV-IoT project, aimed to improve the quality of the data produced from the water[1] and air quality sensors[2]. In this effort, due to the lack of funding, the climate scientists had to rely on open-source solutions to access the data from the sensors. Due to the lack of transparency in the process, they had to trust that the used solutions would not tamper the data coming from the sensors.

To help assess quality, different frameworks have been proposed. For example, the ORME-DQ model is proposed based on understanding the risk associated with the data [2]. Similarly, the NEAT framework tries to put the data quality in the spotlight as an internal problem instead of an external one that can be evaluated based on certain steps to be guided in the right direction [6].

Wang et al. [65], performed a two stage survey study in order to identify the DQAs that are important from the perspective of the consumers of the data. To achieve this they created a "Conceptual Framework of Data Quality" splitting the DQAs into: intrinsic data quality, contextual data quality, representational data quality and accessibility data quality.

[1]link: The LoV-IoT project: Water Quality Report
[2]link: The LoV-IoT project: Air Quality Report

The concept of *fitness for use* has been considered a typical way to define data quality [30, 60, 65]. Data quality is reliant on the context of use and should not be treated as an absolute concept, which is a view followed by other sources as well [5, 17, 22, 31, 40, 46]. However, it has been challenged by Kahn at al. [30] who argue that fitness changes over time, and it is difficult to measure. They have proposed the PSP/IQ model containing different dimensions with which data quality can be viewed, including which DQAs to be prioritised. In particular, they propose that presented information can be: sound, dependable, useful or usable. Depending on the need, different DQAs need to be accommodated even if they might be relating with each other.

Sambasivan et al. [53] also challenge the concept of *Fitness For Use*. They argue that "goodness-of-data" should be prioritized, meaning that instead of evaluating models based on metrics from existing data, the quality of the data themselves should be prioritised.

The data volume growth rate has been increasing over the past years. Tien argues that the gaps in quality can be tackled by increasing quantity [59]. However, others argue that just quantity is not enough and data quality is a fundamental component that can unleash the true potential of the data [17, 39]. This is especially true when considering the costs of time, collection and processing, or when handling challenges around robustness, security, or privacy [24, 31, 68].

Like environmental research, the automotive industry is transitioning from traditional software engineering practices. Such organizations are investigating AI Engineering approaches [7]. In this industry, the models rely on using data from multiple sensors that require good specification to provide correct information [24]. Multiple such organizations are in a similar situation, and it is necessary for practitioners to understand what data they provide to their models, especially when relying on open data to jump-start their efforts [44].

Currently, there are multiple challenges identified with creating clear specifications around annotations and data. Specifically: data collection, data quality and processes&ways of working. There is a need to follow specific sourcing and certification processes. Currently, when creating data requirements they usually do not follow defined standards, leading to the software products being considered unreliable [23].

Data requirements are often overlooked when looking into software requirements [23]. In some cases, it is perceived as the "Wild West" by practitioners with no defined processes to handle the data [25, 32]. Vogelsang et al. [64] argue that the traditional requirements engineering practices of elicitation, analysis, specification, verification&validation can be adapted in order to allow for AI components to be more easily interpreted. Data scientists should enable the requirement engineers in these tasks.

Some sources argue that data specifications should be created iteratively as more knowledge is created from the data [23, 60]. Still, there are some DQAs that are considered to be "inherent" in the data [46] originating, for example from the

ISO/IEC 25012 [20] framework. In environmental research there have also been attempts to standardize with the FAIR principles of findability, accessibility, interoperability, and reliability [66]. These scientists are not commercial competitors [46] and aim to create "global data" that follow high quality standards [15].

Among the DQAs identified in the literature, one that stands out is missing values in the data [3, 32, 33, 39, 41, 28]. In a course by McElreath [38], the instructor argues about the impact of such issues and how they are a quality aspect that is difficult to examine. This is because missing data can introduce bias in unexpected ways. This can affect the performance, leading to the model behaving incorrectly after deployment.

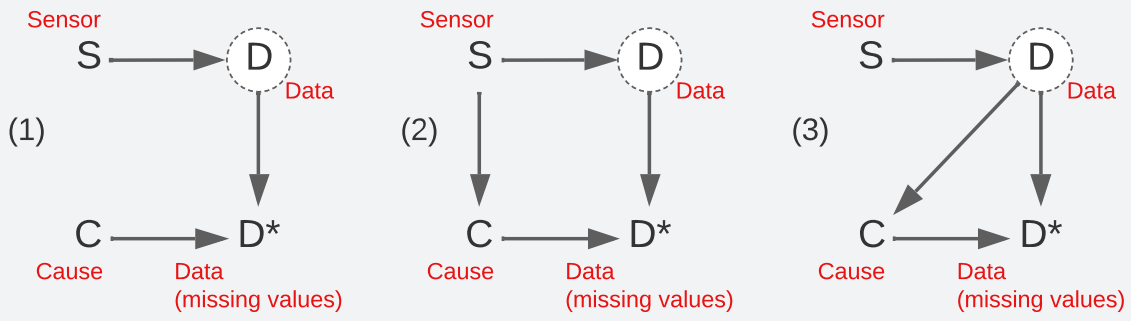

Figure 2.1: Recreation of example missing cases from McElreath [38].

In the recreation shown in Figure 2.1, some example cases of missing data and potential causal relationships are shown as argued by the instructor. In scenario (1) removing incomplete cases reduces efficiency, but it should not be detrimental. In scenario (2) as long as the cause is taken into account then it should be fine to proceed. Finally, in scenario (3) it can be really difficult to reach valid conclusions as the cause (C) can directly affect the missing values.

Easterbrook mentions the "Swiss cheese model" in his book [13]. He argues that different practices "placed on top of each other" increase reliability, as different "layers" can capture issues missed by earlier ones. Introducing multiple layers of protection can help to prevent such mistakes so even if "holes" exist, other layers can catch them. From his perspective, climate scientists have a remarkable set of test design practices which could only be achieved with the global collaboration and transparency.

The same author has challenged the view of "If the models and data disagree, believe the data". He mentions multiple cautionary tales that in history where the lack of understanding the data had an impact. Some examples include:

1) The untracked issues of global surface temperature measurements. This was done by collecting water from the ocean surface. However, different ships followed different practices. Some used insulated buckets, while others did not. The practices also changed over the years when instead of measuring water collected in buckets, measurements were made based on the water entering the engine rooms.

2) Data from satellites being treated as ground truth but without considering that they lose altitude which affected the measurements. This led the model future predictions to be considered untrustworthy by government bodies.
3) In the 2010s, it was perceived that global warming had stopped. However, what was not presented at the time was that a single dataset was used for making this decision, which coincidentally started its trend analysis during a particularly warm year. Certain areas of the planet were also underrepresented, like the arctics, due to the lack of sensor coverage. The data were collected by meteorologists who did not focus on long-term data consistency. Finally, the collection was performed by different teams without standardised approaches on handling missing values. When all these underlying issues were identified, the dataset was re-evaluated and the initial assumption was disproved.

As can be seen from the collected literature, there is a need for vigilance surrounding data, and it is important to promote transparency between their stakeholders [39]. Concerns around the quality of data can introduce scepticism, which can lead to a wait-and-see attitude, losing valuable time when attempting to understand the impact of climate change [13].

3
Methods

3.1 Study Methodology Overview

This study was exploratory in nature, as the aim was to understand what is happening in the target populations of environmental researchers and provide insights regarding using, handling, and working with DQAs. This also made it observational, following the definitions from Runeson and Höst [52]. To achieve this, a sample study approach was followed as suggested by Stol et al. [57] using input from the different stakeholders investigated in the study. The stakeholders considered in this study are the data portal maintainers, the dataset providers and the dataset consumers.

The study started with the collection of DQAs from the literature and data portals. Based on the literature findings, a questionnaire was created and shared to the investigated populations. With the collected information, the study could proceed to derive insights which, when viewed from the software engineering perspective, as defined by SWEBOK [8], can help identify gaps and provide suggestions for future efforts.

The initial collection of information for this study was performed using document analysis, which is a third-degree technique, on the portal website documentation. The sample study that followed was based on a survey, which is a first-degree technique following a mixed approach with qualitative (open-ended) and quantitative (close-ended) questions [54, 57]. These approaches were followed to enable triangulation [52, 54] which can increase reliability.

3.2 DQA Collection

The collection of the DQAs for this study started by performing a literature review. The gained knowledge was then used to create semantic themes which were evaluated by an expert to increase reliability. This individual has a PhD in machine learning data pipelines and two years of industrial experience as a data engineer.

These themes were then used to create questions to enable the document analysis of the collected data portals. These themes were updated as new DQAs were identified in the portals. The final set of DQAs and Themes was combined and used to build the survey instrument.

DQA Collection from Literature

Following the guidelines from Kitchenham and Charters, a literature review with systematic elements was followed to gather literature reporting on DQAs in Software Engineering [35]. It was performed with a limited scope due to resource constraints and the single researcher involved in the study. In particular, quality assessment of the collected literature was not performed and there was only a limited synthesis performed including the collected DQAs.

Literature Review

The literature review started with an initial set of articles suggested by the supervisor in order to get an understanding of the research field. For this initial set, snowballing was followed in order to identify more relevant articles.
Next, the literature analysis followed, which included systematic elements. Articles were collected using ACM Digital Library[1], IEEE Xplore[2], and SpringerLink[3] as sources. To help with finding all relevant articles around DQAs there was no limit introduced in the publication year or theme.
The investigation was performed using the following query with the aim of collecting all the relevant DQAs that currently exist in the platform:

Query 1: "Data Quality Aspects"

To help filter the evaluated articles before doing the full review, the following filters were used:

Filter 1: Only articles in English will be included in this study

Filter 2: Title or Keywords relate to data quality

Filter 3: Study explicitly states and relates DQAs to the conducted study

After filtering, the remaining articles were combined with the initial set used at the start of the study. The duplicates were removed between the different sources, and then the DQAs that were explicitly stated in the article were collected.
It was quickly identified during data extraction that many DQAs found in the literature were derivatives of each other (for example, *relevancy — relevance*) or syntactically equivalent (Free of Error — Free-of-Error) and were merged. Next, the Merriam-Webster thesaurus[4] was used in order to find semantically similar words. Finally, the remaining DQAs were run through a Generative AI[5] in order to identify the remaining semantically similar terms in the context of data quality.

DQA Collection from Data Portals

Using the DQAs identified in the literature, a thematic analysis as outlined by Braun and Clarke was followed to identify which DQAs were closely connected to each other, with the support of the aforementioned expert [9]. The set of created

[1]link: ACM Digital Library
[2]link: IEEE Xplore
[3]link: SpringerLink
[4]link: Merriam-Webster
[5]link: ChatGPT

themes was then used to define questions to enable the analysis of the data portals from the presented documentation on the websites.

Thematic Analysis

In this study the analysis was *semantic*, meaning looking into the surface of the data, and *theoretical*, meaning the themes were based (and expanded) upon an initial framework. This framework was based on the DQAs listed in ISO/IEC 25012 [20]. These were expanded as needed, which introduced an *editing approach* in the coding process with an initial set of a priori codes [52]. This approach was chosen to reduce the bias that can be introduced by a single researcher, which is an important consideration as identified by Eisenhardt [14] and Runeson and Höst [52]. This was also the reason for the expert input for this step.
As the set of DQAs was too large for creating the themes, only DQAs that appeared at least twice after the processing, were included. To perform the analysis, Microsoft Visio[6] was used. The diagram was passed to the expert to be reviewed. More details on this effort will be discussed in the results chapter.
The agreed themes were used to analyse the portals. The identified DQAs were combined with initial ones and used in order to create the questionnaire used for the rest of the study.

Portal collection and ordering

An initial set of data portals was collected based on non-probabilistic convenience sampling. This means that samples were picked in a non-random order based on accessibility [19]. The collection was enabled with input from an expert working at the IVL Swedish Environmental Research Institute[7], online searches, literature and sources from climate videos[8].
Since the collected portals can contain datasets from multiple subjects besides environmental research, it was important to find an efficient way to order them based on their relevance to the study. The first step was to get an understanding of what constitutes a dataset relevant to environmental research. To achieve this, different environmental glossaries from established organizations were used. In detail, a glossary from the US [1], EU [16] and UN [62] were used in order to cover a broad spectrum of terms related to environmental research since the data portals evaluated could originate from all over the world and in many disciplines of environmental research and other scientific fields.
In order to use the glossaries, they had to be processed to become machine-readable. This was tackled by creating custom python scripts to process the documents presenting the glossaries, and find common terms. All scripts and findings of this process can be found in the GitHub repository associated with this study[9].
In detail, after collecting the website content in *TXT* files directly from the online sources, they were processed into *CSV* files containing the term and an optional

[6]link: Microsoft Visio
[7]link: IVL Swedish Environmental Research
[8]link: Climate Videos Source
[9]link: Thesis GitHub Repository

definition. Next, a follow-up script was used to investigate which terms are common between the glossaries and a query was created that can be used for *advanced searches* in the data portals. Each term was encapsulated in quotation marks (“) and separated with an *OR* statement. In said script, a flag was added to ignore glossaries that contain less than 200 terms. The reason for this decision was the lack of terms present in one of the collected glossaries, which lead to a very small number of common terms. This will be explained more in the results chapter.
With the terms collected, the portals were ordered and evaluated based on the inclusion of the *datasets relating to environmental research* and the *total number of datasets* available in the portal. To do this, a *harmonic mean* metric was used. It assumed equal weight between these variables, as to not introduce bias towards any direction between them. Thus, the following formula was used to order the portals to be investigated:

$$HarmonicMean = \frac{2 \cdot RelevantDatasetsInPortal \cdot TotalDatasetsInPortal}{RelevantDatasetsInPortal + TotalDatasetsInPortal} \tag{3.1}$$

Unfortunately, not all portals supported advanced search or required some manual actions in order to collect the number of relevant datasets. In some cases, the portal was allowing the user to dynamically create their dataset based on different filters. As such, the ordering had to be redesigned. The proposed order was as follows:

- Portals that could not be sorted but are specifically dedicated to environmental research (ordered by dataset count, if accessible)
- Portals that could be sorted (ordered by harmonic mean)
- Remaining Portals

With the portals ordered, the last step was to introduce a cut-off point based on the mentioned ordering, this was set to a score that allowed to investigate a representative set of portals. The exact number will be explained in the results chapter.

Portal analysis

The document analysis followed, which was based on the available documentation in the data portals. The final aim of this phase was to expand the DQAs that were already in consideration. To do this, the themes identified before were used as a priori codes to help guide the investigation.
When a new DQA was identified it was added to the existing collection and then the previously checked portals were evaluated again to investigate for its potentially missed presence. The identified list of the new DQAs was then added to the original list of DQAs and a final list was created which was used to create the questionnaire. After the investigation, the collected information was used to update the survey instrument and answer the first research question set for this study.

3.3 Sample Survey Process

With the collection of the DQAs from the literature and data portals, the sample survey design could begin. The instrument used was a questionnaire containing

both open-ended and closed-ended questions [19]. It was distributed online using Microsoft Forms[10]. To introduce structure in the design, the approach defined by Ghazi et al. was used [19]. The relevant steps will be explained in the following subsections.

Research Objective of the Survey

The objective of this survey was to answer the second research question set at the beginning of the study:

> RQ2: Which aspects of data quality are important from the perspective of different stakeholders associated with open data in environmental research?

Population to be Studied

As defined by the research objective, the populations to be studied were based on the stakeholder group they belong in. Namely, the maintainers of data portals, dataset providers and dataset consumers.

Sample Plan

Since it was not possible to contact every portal, in this study a different approach was followed. Due to the nature of the task, non-probabilistic convenience sampling was used in order to identify which participants should be contacted [19]. This means that the subjects were selected based on accessibility. In particular, the data portals identified when collecting the relevant DQAs mentioned in Section 3.2 were used. The contacted respondents were also prompted to share the instrument with other subjects in order to promote snowball sampling, as suggested by Ghazi et al. [19], and Kitchenham and Pfleeger [34].

To identify the dataset providers, the contact information from the portals was used. In some cases, portals contained a significant number of providers. In such cases, a few of them were randomly picked after applying the advanced search query mentioned in Section 3.2, when applicable. To ensure randomness, a dedicated script was created. To allow traceability, the algorithm used can be found in the repository[11] associated with the thesis.

For contacting the dataset consumers, non-probabilistic convenience sampling [19] was used as well. The choice of the populations was based on their relevance to environmental research and open data. In particular, individuals were picked from the circles of contacted experts as well as researchers in various university staff lists with accessible contact information. Originally researchers from the IVL institute were expected to be contacted, however as it will be discussed in Section 5.2 it was avoided as a separate similar investigation already had taken place there, and it would introduce a potential bias in the results.

[10] link: Microsoft Forms

[11] link: random number picker script

Mode of Data Collection

In order to contact the potential respondents, an online questionnaire [49] was created using *Microsoft Forms*. For the participants, questionnaires require lower effort than physical surveys to access, and the process of completing them can be easier, as options can be more conveniently presented. For the researcher, the data collected are directly accessible for analysis, and it is easier to manage the response rate of the participants [49]. This is especially useful as a single researcher performed this study. Also, online questionnaires allow accessing real-time updates on the responses. Since the populations to be contacted are distributed all over the world, using online questionnaires also reduces the lead time to get the results to a minimum. As such, if it was discovered that a particular population was not covered in the current sample, it would be possible to share the instrument to more populations in a relatively short amount of time. Finally, as Chalmers is hosting an own instance of Microsoft Forms, the data could be kept on local servers.

Construction of Survey Instrument

The questionnaire was designed with the aim of being: (1) *descriptive*, as the study aimed to explain the traits of the investigate population, (2) *cross-sectional*, meaning the information collected was based on a specific point in time, and finally (3) *unsupervised*, since it was distributed online. The instrument made use of open-ended (*qualitative*) and close-ended (*quantitative*) questions [19].
With this context in mind, the questionnaire built was separated in two main sections, which from the perspective of the analysis can be understood as the independent variables section and the dependent variables section.
In practice, in the first section, the questions aim to help identify the nature of the respondent and by extension categorize them during the analysis. On this section, the questions were opted to be close-ended to more easily manage the independent variables during analysis. These questions aimed to understand the role and experience of the respondent. A subset of the investigated questions in the instrument included:

- Please select the category that best describes your role in Open Data.
- How many years have you been involved in Environmental Research?
- How many years of experience do you have in Data Management, such as Databases?

For the second section, inspiration was taken from the questionnaire from a previously used instrument for evaluating DQAs by Wang and Strong [65]. In that study, close-ended Likert scale questions were included with the aim of gathering the importance rating (1-9) of identified DQAs. In the study performed in this report, some changes were made. In order to promote the respondents to pick a side in the collected DQAs it was opted to ignore the "middle-ground" stance and the number of options was reduced leading to a scale between 1-6. Also, to avoid potential misunderstanding introduced by using numbers in a Likert scale, the numbers were presented as stars considering that they are a more universal term for identifying importance.
Another change was that in order to ensure that the respondents understand the

quality aspect they are tasked to comment on, a definition was provided from one of the sources where the quality aspect was collected along with similar DQAs as identified in Section 3.2. For allowing traceability for the respondent, a reference *digital object identifier* (DOI), or the *American Psychological Association* (APA) reference was provided as well.
Since multiple definitions exist in the literature, the definition that was most descriptive from the perspective of the researcher was picked. This of course can introduce bias in the study. To help mitigate it, the respondents were also prompted with a follow-up question to provide their thoughts on the provided definition and similar DQAs. The questions in this section follow the structure presented in the example in Figure 3.1 in order to streamline the process for the respondent and help increase the response speed.

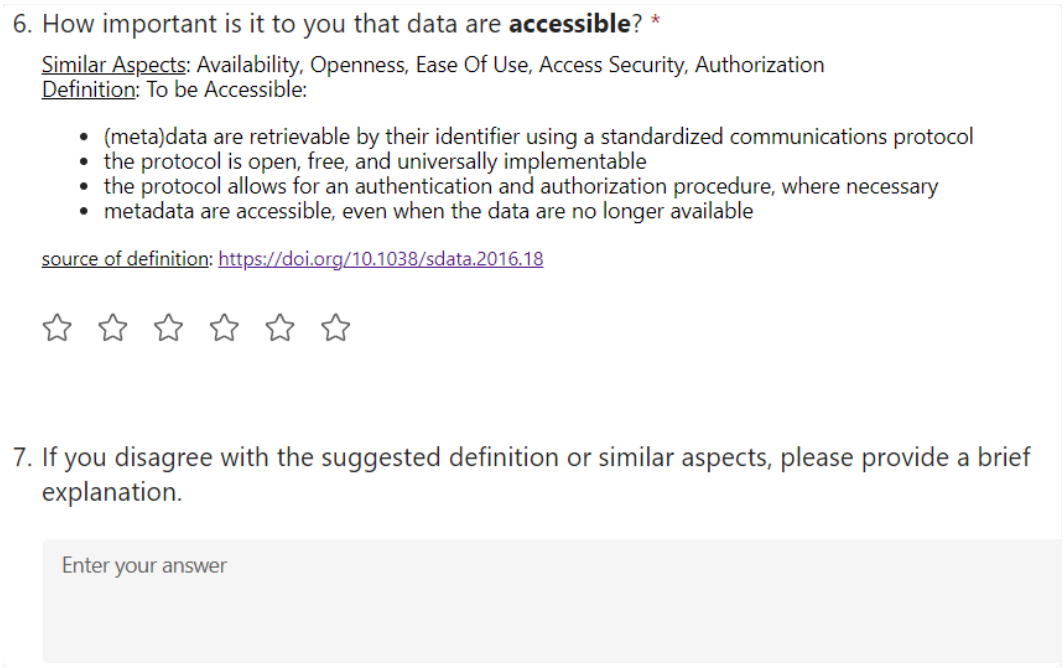

Figure 3.1: Example question in questionnaire for DQAs

Since there were multiple DQAs found during the collection phase of the study, not all of them were included in the instrument to not overwhelm the respondents. There were two cuts introduced. The first cut, which was also a cut-off point for the instrument, included the collected DQAs that appeared at least twice in the literature. Anything less than that was not included. In the second cut, the ones that appeared at least ten times were included with the structured mentioned in Figure 3.1. The second set was presented in a list to the respondents, where they were prompted to pick five of the aspects they considered to be most and least relevant. These cuts were chosen as they allowed for the best balance between the DQAs to be investigated in-depth and not asking too much time from the respondents.
Finally, the survey included an open-ended question prompting the respondent to provide the DQAs that they consider were missed by the current questionnaire. The full questionnaire in text form along with the reasoning behind each question can be found in the Appendix Section A.

Evaluation of Survey Instrument

As suggested by Ghazi et al. and Linåker et al. it is important to evaluate the created instrument before distribution [19, 36]. For this thesis, two pilot studies were performed. The first included two data engineering experts. Both had at least 3 years of experience in data management. One had experience of 3–5 years in data science and environmental research, while the second one had less than 1 year of experience in data science and no experience in environmental research. These experts were picked in order to evaluate the instrument even though they are not actively part of an environmental research effort, in order to avoid exhausting the potentially limited number of potential respondents in the study.
The second pilot study was held before the release of the instrument with two colleagues from the university with less experience around the investigated topics. Their input was quite valuable to help cover a wider spectrum of expertise around the instrument.

Data Analysis and Interpretation

Before the analysis could begin, the data were validated on their completeness and consistency. This was done by inspecting the report provided by Microsoft Forms. The responses were also partitioned based on the independent variables to help the analysis. Finally, descriptive analysis of the dependent and independent variables was performed, which will be presented in the results chapter.
For the analysis of the collected data, different approaches were followed for the open-ended and close-ended questions. The open-ended questions were coded based on thematic analysis with the aim of gathering views on the presented definitions for the suggested DQAs or gain insight on potentially missed DQAs [9]. These questions were treated as being *semantic*, meaning looking into the surface of the data. Also, an *inductive* approach was used, meaning basing the created themes on the data themselves. These approaches were chosen to reduce the bias that can be introduced by a single researcher. after removing sensitive information, the created themes were evaluated by the colleagues of the first pilot study who evaluated the instrument.
For the close-ended questions, since they are based on Likert scales, quantitative analysis was used. With the information collected and coded, statistical analysis was performed on the different variables [36]. Following that, since a few samples were collected, as will be presented in Section 4.2, Bayesian data analysis was used to get more information from the data. the approach was inspired by the Bayesian Workflow by Gelman et al. [18], as well as the work from McElreath [38]. More details on this process can be seen in Section 3.4.

3.4 Bayesian Data Analysis Methodology

Bayesian data analysis was used for this project due to the small number of samples and the predominant non-probabilistic convenience sampling collection [38].
For this analysis, the research goal was defined as follows: “How does the role and

experience of a stakeholder affects their view on different DQAs?". To satisfy this goal, different models were built based on the two types of questions present in the survey instrument with different prior and posterior predictive checks to evaluate the proposed priors.

Model Variables

For the analysis, the aim was to get an understanding of how different roles and backgrounds can affect the view regarding the different DQAs identified in the study. These were treated as independent variables and included in the final model for each DQA. The predictor was defined based on the ratings and picks from the respondents.

For each experience variable, the values were updated from the original survey during the analysis to the following options: no experience, up to 5 years of experience, 5 or more years of experience. This was done in order to reduce the parameters that the model had to learn, as the amount of experience gets more relevant in these larger intervals.

In the survey, the respondents could pick from three roles. They were also given the option to define their role. This was in consideration that a respondent might belong to multiple roles. With this in mind, the following possible values were included in the model:

1 - Dataset Provider
2 - Data Portal Maintainer
3 - Dataset Consumer
4 - 2 Roles: Dataset Provider and Dataset Consumer
5 - 2 Roles: Dataset Provider and Data Portal Maintainer
6 - 2 Roles: Dataset Consumer and Data Portal Maintainer

A proposed directed acyclic graph (DAG) outlining the causal relationships can be seen on Figures 3.2 and 3.6. When creating this DAG, the potential causal relationships were taken into consideration. Due to the nature of the sampling process, it is expected that many of the respondents will be associated with environmental research. This is because the sample, especially for the dataset providers, was based on random sampling from the available portals. This can decrease the potential for any unidentified confounders missed by the researcher.

Analysis Methodology of Part A: Ratings

The model proposed from this DAG contains the role as a *categorical variable*, which encodes all possible roles and combinations thereof, as explained before. It also includes three *ordered categorical* variables for each different type of experience. This is because it is expected that for a respondent to pick a higher level of experience it is only natural that they should cover the previous levels, especially since in the study it is represented as years. With this in mind, the proposed model is using an ordered logistic likelihood function as it aims to predict the pick from the rating scale. The proposed model is given in Equation 3.2.

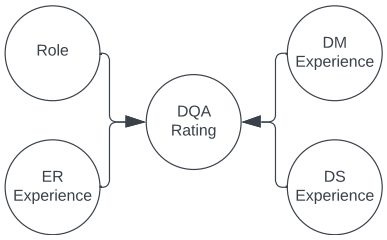

Figure 3.2: Proposed DAG for each DQA investigated in the ratings questions. The predictor includes the DQA itself. The role, data management experience (DM Experience), data science experience (DS Experience), environmental research experience (ER Experience), are treated as independent variables.

$$\text{rating}_i \sim \text{OrderedLogistic}(\phi_i, \alpha) \tag{3.2}$$

$$\phi_i = \beta_{role_i} + \beta_{e\_dm} \sum_{j=0}^{e\_dm_i - 1} \delta_j + \beta_{e\_ds} \sum_{j=0}^{e\_ds_i - 1} \delta_j + \beta_{e\_er} \sum_{j=0}^{e\_er_i - 1} \delta_j \tag{3.3}$$

$$\alpha_j \sim \text{Normal}(0, 1.5) \tag{3.4}$$

$$logit(\beta_{role_i}) \sim \text{Normal}(0, 1.5) \tag{3.5}$$

$$\beta\_ \sim \text{Normal}(0, 1) \tag{3.6}$$

$$\delta \sim \text{Dirichlet}(\alpha) \tag{3.7}$$

The linear model ϕ_i, as mentioned in Equation 3.3, is defining the thresholds for ratings dividing the continuous variable to observed ratings. The α_j in Equation 3.4, represents regression coefficients of the predictor variables splitting the model based on the cut-points. The β_{role_i}, shown in Equation 3.5, is used to measure the regression coefficients presented for the roles in the model and used in the model with a logit link function. Each experience, presented in Equation 3.6, is represented by a $\beta\_$ for the regression coefficients of the experience for data management, data science, and environmental research. Finally, the probability distribution of each experience is based on the accumulative sum defined using *Dirichlet* distribution. This can be weighted using an α, as shown in Equation 3.7. For this model, it was set to 1.0 in order to not bias the model towards any rating response and allow for high variability in the priors.

Prior Predictive Checks - Ratings

The proposed model contains multiple variables that require priors selected in an educated manner. It is important to take careful decisions as the choice of priors can influence the posterior distribution, as stated by McElreath [38]. Therefore, data were simulated for each of the variables and passed into the model. These sample data were created following a uniform distribution of the possible values in the dataset and included 200 simulated samples. Then, prior samples were selected from the model and density plots were created to investigate their distribution. Since each component in the model is following a normal distribution, the logistic function was used. The results of the picked model can be seen in Figure 3.3.

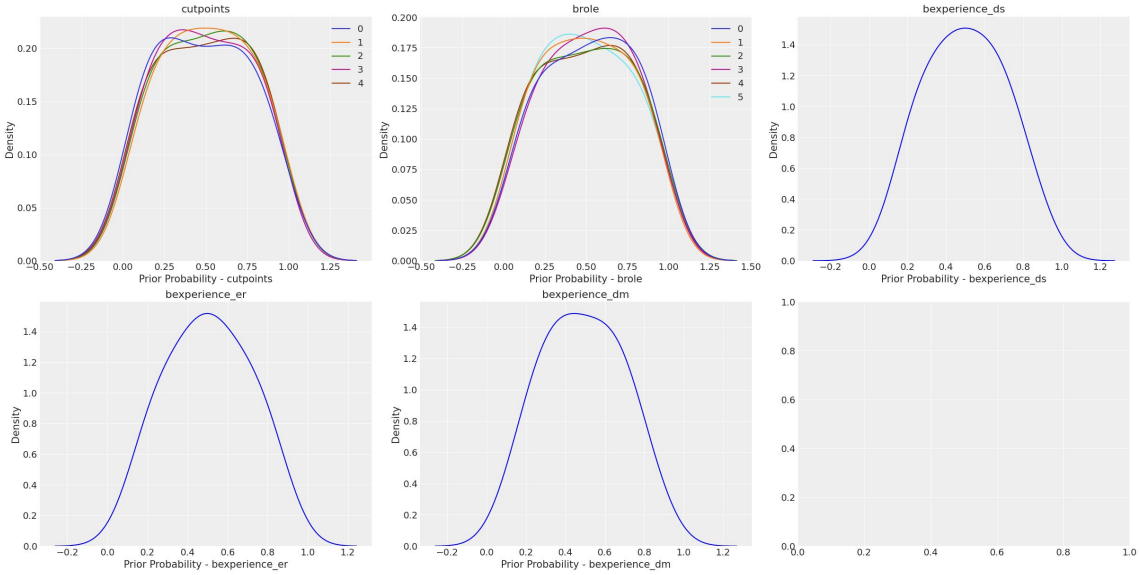

Figure 3.3: Prior predictive checks for Rating questions regardless of DQA. As shown, each curve aims to get as even of a distribution as possible between 0-1 in order to not bias the model towards any extremes.

Posterior Predictive Checks - Ratings

Finally, it is also good to perform posterior predictive checks with the proposed priors. The sampling process was performed with 2000 samples, 1000 for the draws and 1000 for the tuning. The output of this check can be seen in the posterior samples Figure 3.4 and the posterior densities Figure 3.5 for the ratings model.

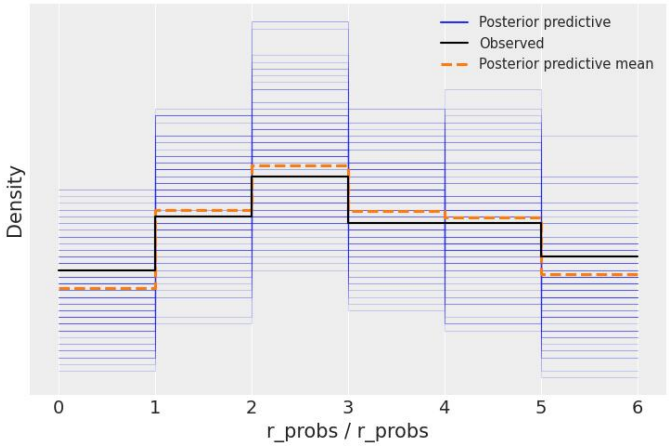

Figure 3.4: Posterior predictive samples along with observed data for Ratings model. There is no perfect alignment with the mock data. However, since this is the predictor it is good as it shows that the model is not overfit to the presented data.

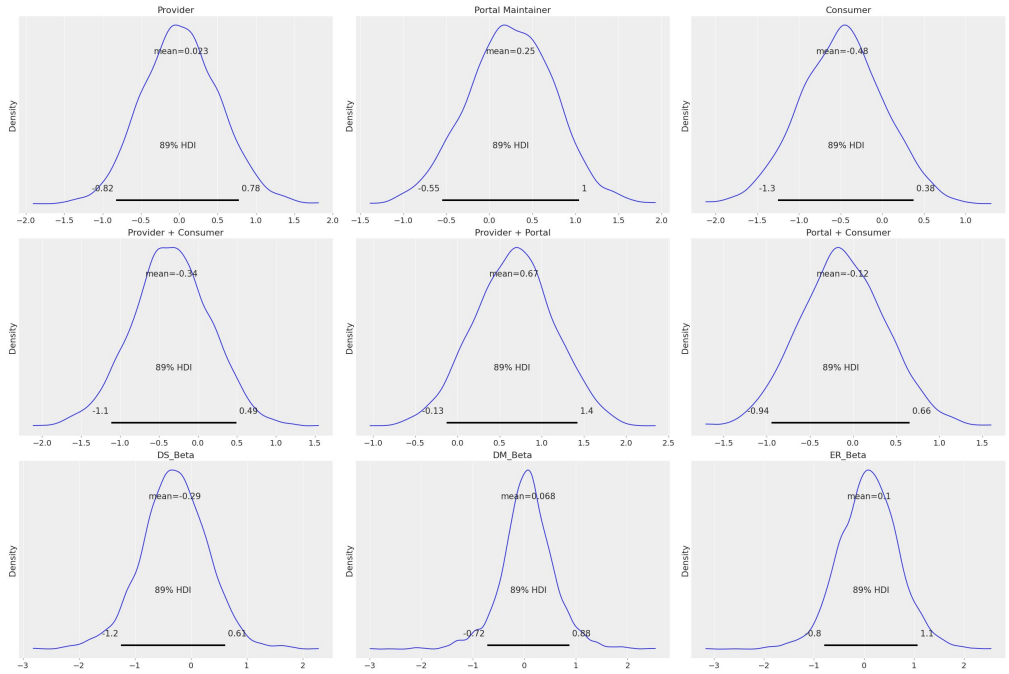

Figure 3.5: Posterior density plots for the Ratings model. All the plots are tending to 0 which is important for this model as it should not be biased towards any direction considering the uniform distribution used for all the simulated data.

Analysis Methodology of Part B: Pick-5

Similarly to the ratings model, the pick-5 model was based on including all the aforementioned variables. The only different being the likelihood function, as the question to be answered was different. Instead of aiming for ordered categories, the proposed model is looking into finding which categories are most likely to be picked by the respondents. As such, a multinomial distribution was used. Also, it is not visible in the equations themselves, but the possible values for the categorical options were limited to the maximum number of choices a respondent can make. In the case of this investigation, it was five. The proposed model for this investigation is given in Equation 3.8.

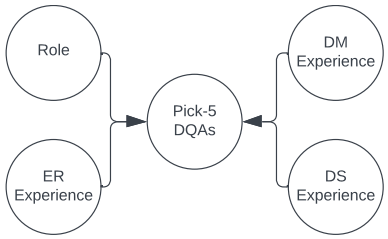

Figure 3.6: Proposed DAG for the Pick-5 most/least relevant DQAs question. The predictor includes the Pick-5 options. The Role, Data Management Experience (DM Experience), Data Science Experience (DS Experience), Environmental Research Experience (ER Experience), are treated as independent variables.

$$\text{pick\_5}_i \sim \text{Categorical}(\rho_i) \tag{3.8}$$

$$logit(\rho_i) = \beta_{role_i} + \beta_{e\_dm} \sum_{j=0}^{e\_dm_i - 1} \delta_j + \beta_{e\_ds} \sum_{j=0}^{e\_ds_i - 1} \delta_j + \beta_{e\_er} \sum_{j=0}^{e\_er_i - 1} \delta_j \tag{3.9}$$

$$logit(\beta_{role_i}) \sim \text{Normal}(0, 1.5) \tag{3.10}$$

$$\beta\_ \sim \text{Normal}(0, 1) \tag{3.11}$$

$$\delta \sim \text{Dirichlet}(\alpha) \tag{3.12}$$

In this model, ρ_i as defined in Equation 3.9, is used to present the variability across the different categories of the model. Since this is a multinomial model, it uses a logit link function. Like in the previous equation, the β_{role_i}, shown in Equation 3.10, defines the regression coefficients presented for the roles in the model and is also using a logit link function. The different experiences, presented in Equation 3.11, and represented by a $\beta\_$, are used as regression coefficients to model the experience for data management, data science, and environmental research. Lastly, as before, the *Dirichlet* distribution, shown in Equation 3.12, was used for the accumulative sum of the experience levels and the α was set to 1.0 in order to not bias the model towards any response and allow for high variability in the priors.

Prior Predictive Checks - Pick-5

As before, the priors need to be defined. A similar process as before was followed but for the new model. After evaluating different values for each prior, the ones currently presented were used. The prior predictive checks for these can be seen in Figure 3.7. Since the models are quite similar, the same priors can be used to provide adequate results.

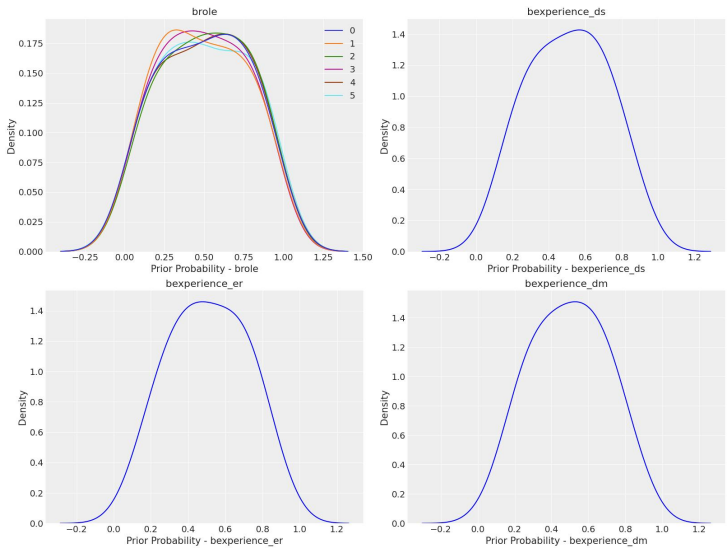

Figure 3.7: Prior predictive checks for rating questions regardless of DQA. As shown, each curve aims to get as even of a distribution as possible between 0-1 in order to not bias the model towards any extremes.

Posterior Predictive Checks - Pick-5

As for the previous model, the posterior predictive checks took place after the prior checks. The results can be seen in the posterior samples Figure 3.8 and the posterior densities Figure 3.9 for the pick-5 proposed model.

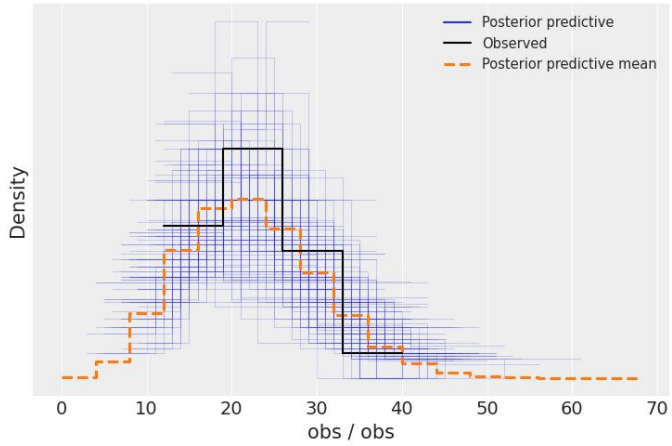

Figure 3.8: Posterior predictive samples along with observed data for the Pick-5 model. As before, the alignment is not perfect, but that is for the better.

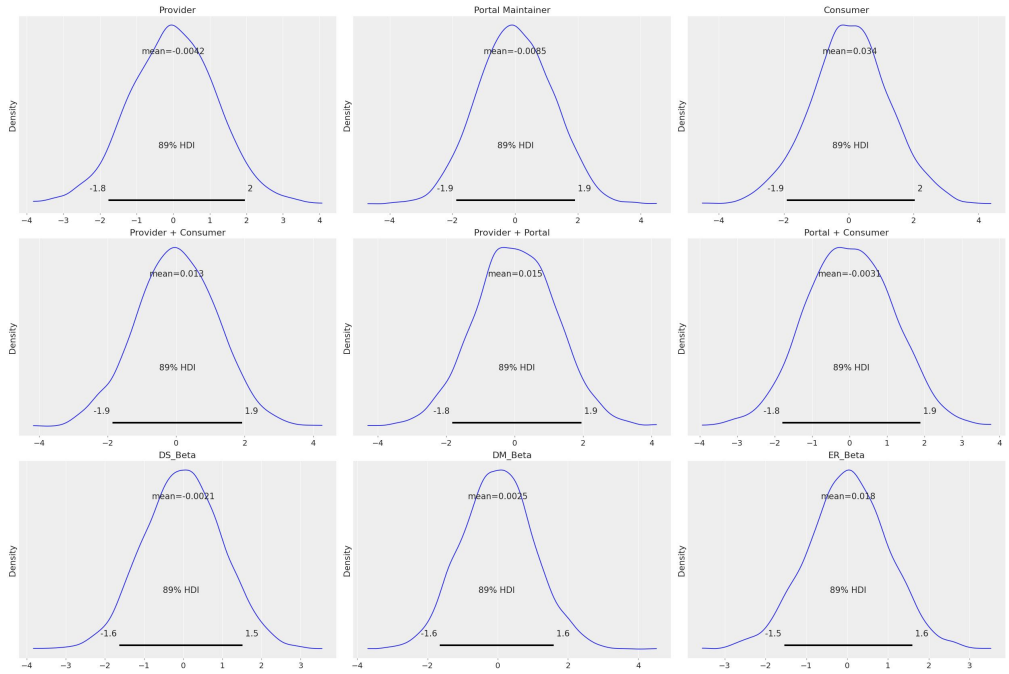

Figure 3.9: Posterior density plots for the Pick-5 model. As before, all the plots are tending to 0 which is the aim for this model.

4

Results

In the following sections, the results of this study will be presented. First, the collection of the DQAs, followed by the completion of the survey instrument and collection of the responses. The chapter will close with analysing the results collected from the respondents of the survey.

4.1 DQA Collection

DQA Collection from Literature

As mentioned in the previous chapter, for collecting the DQAs from the literature an initial set of articles was used which was expanded by snowballing. In this effort, **22** articles were included which contained DQAs relevant to this study. This "trail" of literature snowballing can be viewed in the Appendix: A.1.
This set of articles was also expanded with a literature review containing systematic elements [35]. This introduced **37** more articles collected with the approach shown in Table 4.1. The publication year of these articles is shown in Figure 4.1.

Table 4.1: Sources and filters applied during literature review

Source	Initial Search	Filter 1	Filter 2	Filter 3
IEEE	7	7	6	6
ACM	43	43	26	12
Springer Link	75	71	32	19
Total	125	122	64	**37**

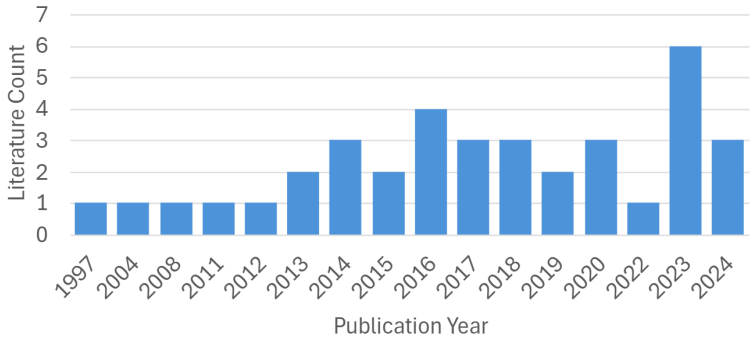

Figure 4.1: Publication year of literature selected for analysis.

Next, it was important to remove any duplicated entries, this effort included the initial **22** articles. In total **2** articles were removed. One appeared in two sources, and the second was included in the initial set of articles. In total **57** articles were included for the DQA collection from the literature.
At the end of the investigation of all the sources, a total of **130** unique DQAs were collected, with **404** instances of them appearing across the literature. These were too many to include in the questionnaire, so some steps were taken in order to group them and reduce the count. Initially, they were investigated for clear derivatives or structural similarities. In the end, **122** unique DQAs remained. Some example updates include:

- derivatives
  - ❖ relevancy — relevance
  - ❖ currency — currentness
  - ❖ Metadata — Metadata Update
  - ❖ Representative — Representatives
- syntactical similarities
  - ❖ Free of Error — Free-of-Error
  - ❖ Value added — Value-Added

The next step was to group them using the Merriam-webster thesaurus in order to identify semantically similar DQAs. During this effort, some created groups could be semantically connected with each other. This was performed as well in order to help reduce the created DQAs further. In the end, **98** DQAs remained. Some examples of these groupings include:

Table 4.2: Example DQAs grouped using Merriam-Webster

Kept DQA	**Similar DQAs**
Accessibility	Availability, Openness
Accuracy	Precision, Correctness
Relevance	Appropriateness, Validity, Usefulness, Fitness, Pertinence

The final step for finding semantic similarities was based on using a Generative AI[1]. In the tool, a list of the remaining DQAs was provided along with the following query: "Out of the following Data Quality Aspects, which ones are synonyms of each other: . . . ". This query was used three times in order to identify all the semantically relevant DQAs. In the end **72** DQAs remained which were used for the remainder of the study. A full list of the updated DQAs can be seen in the Appendix A.5. A list of counted instances after the update can be seen in Figure 4.2.

DQA Collection from Data Portals

To enable the investigation of the portals, thematic analysis was performed on the collected DQAs. This was done in order to derive potential directions for the document analysis. As it was discovered, most portals do not make explicit mentions of DQAs and therefore the themes were needed to introduce structure. The portals

[1]link: ChatGPT

DQA	Literature Count of DQA	DQA	Literature Count of DQA	DQA	Literature Count of DQA	DQA	Literature Count of DQA
Accuracy	50	Duplications	4	Findability	1	Statistical Processing	1
Completeness	44	Documentation	4	Biased	1	Frequency of collection	1
Consistency	37	Cost Effectiveness	3	Comparability	1	Synchronization	1
Timeliness	28	Presentation Quality	3	Access security	1	Geographical Coverage	1
Credibility	28	Compliance	3	Reusability	1	Temporal Coverage	1
Relevance	23	Skewness	3	Noise	1	Unit of Measure	1
Accessibility	19	Complexity	3	Statistical Presentation	1	Homogenous	1
Understandability	12	Flexibility	3	Cost	1	Comment	1
Interpretability	11	Interoperability	3	Technology Coverage	1	Uncertainty	1
Usability	11	Uniqueness	3	Age	1	Reference Period	1
Security	10	Volume	2	Feature Accuracy	1	Release Policy	1
Value-Added	10	Granularity	2	Alignment	1	Ontology	1
Appropriate amount of data	8	Provenance	2	Retrievability	1	Opportunity	1
Objectivity	8	Contactability	2	Popularity	1	Lineage	1
representation	6	Concise representation	2	Scalability	1	Verifiability	1
Conciseness	6	Recoverability	2	Error rate	1	Maintainability	1
Traceability	6	Resiliency	1	Definition	1	Meaningful	1
Unambiguous	5	Structure	1	Label Quality	1	Navigation	1

Figure 4.2: Count of Updated DQAs in the investigated literature. Ordered by count

were collected and ordered using the harmonic mean, when possible. Finally, the document analysis took place to identify current reporting practices and DQAs.

Thematic Analysis

To perform the thematic analysis, only DQAs that appeared at least twice after the clean-up were included. As shown in Figure 4.2, this included **33** DQAs. For the analysis, the process proposed by Braun and Clarke was followed [9]. First, it was important to familiarize the researcher with the codes. This was done by checking definitions from the literature. After an initial proposal was created based on ISO/IEC 25012 the themes were expanded based on the investigated DQAs. The result of this effort was evaluated by the expert as a Microsoft Visio diagram[2].
The themes were presented as large yellow circle shapes. In these circles the definition of the theme, based on ISO/IEC 25012, was added to provide an explanation of the relevance and scope of the theme. The DQAs were presented as smaller blue ellipsis shapes and were used as sub-codes for the created themes. For the DQAs the colour of the shape was changed based on the literature instances of the found DQA. The output of this investigation can be seen in the Appendix Figure A.3. These Themes were used to enable the documentation analysis performed on the portals. In the end, the following themes were created:

- Accuracy
- Completeness
- Consistency
- Cost-Effectiveness
- Credibility
- Transparency/Clarity

[2]Microsoft Visio

Portal Collection and Ordering

The collection of the data portals was enabled from inputs from different sources. As mentioned before, input was provided by an expert working in the IVL Swedish Environmental Research Institute, the investigated literature, climate videos and finally a number of sources included from online searches by the researcher. The distribution of said sources can be seen in Table 4.3. Using an ordering approach for these portals was ever more important considering that a good number of them were identified by the researcher, and as such it was important to reduce potential bias.

Table 4.3: Data Portal Sources distribution

Source	Proposed Portals
IVL Expert	20
Climate Videos	8
Online Searches	5
Literature	2

With the portals collected, and in order to introduce a structured way of ordering, a harmonic mean metric was introduced, which was presented before in the Equation 3.1. To create this metric, two variables were used: (1) the total number of datasets in the portals and (2) the total number of datasets relating to environmental research in the portal.
To identify the relevant datasets, keywords were created based on the collected glossaries. These included the UN Data Environmental Glossary [62], the European Environmental Agency Glossary [16] and the Arizona Department of Environmental Quality [1]. Initially, to get a US glossary, the United States Environmental Protection Agency [63] was preferred, but unfortunately their "Report on Environment" included only a few terms, while their full glossary was not possible to access. After performing the collection of the glossaries and the extraction of the terms, a total of **29** were common between them. These were then synthesized into creating the following query:

"absorption" OR "adsorption" OR "aquifer" OR "asbestos" OR "carcinogen" OR "catalytic converter" OR "confined aquifer" OR "ecosystem" OR "fly ash" OR "groundwater" OR "hazardous substance" OR "landfill" OR "lead" OR "nitrate" OR "nutrient" OR "ozone layer" OR "radioactive waste" OR "recycling" OR "scrubber" OR "secondary treatment" OR "septic tank" OR "sludge" OR "solid waste" OR "stationary source" OR "surface water" OR "tertiary treatment" OR "toxicity" OR "turbidity" OR "watershed"

With this information in mind, the result of the ordering can be seen in Table 4.4. Most of these portals get updated constantly with new datasets, so in order to introduce traceability the collection date has been added to the figure. Also, in the final column the harmonic mean can be seen, which defines the table order.
For this ordering, the criteria mentioned in the methodology were used. More explicitly, if the advanced search query was not possible and the portal stated envi-

ronmental data as their only source of information, then the ordering assumed an infinite amount of datasets. If the portal did not allow an advanced search query and contained other types of data besides environmental ones, then the harmonic mean was set to -1. The tool used to create this table was Microsoft Excel[3] and the calculation to create the harmonic mean was as follows:

=IF(EXACT(*Query Searchability*,"TRUE"), 2**Relevant Datasets in Portal***Total Datasets in Portal*/(*Relevant Datasets in Portal*+*Total Datasets in Portal*), IF(EXACT(*Environmental Portal*,"TRUE"), *Total Datasets in Portal*, -1))

Table 4.4: Data portals collection and initial evaluation after sorting based on the harmonic mean, which is the last column. Also, literature sources are from Thomer et al. [60]. The word "Environmental" has been shortened to "Env.", as well as "Searchability" to "Search." and "Percentage" to "Perc.".

Host name	Origin	Site	Data Collection Date	Total Datasets in Portal	Relevant Datasets in Portal	Query Search.	Relevant Perc.	Env. Portal	Harmonic Mean
OpenAQ	Expert IVL	https://explore.openaq.org	2024-03-02	9.00E+99	9.00E+99	FALSE	-1.00000	TRUE	∞
SMHI	Expert IVL	https://hypeweb.smhi.se	2024-03-02	9.00E+99	9.00E+99	FALSE	-1.00000	TRUE	∞
ENES	Expert IVL	https://enesdataspace.vm.fedcloud.eu	2024-03-02	9.00E+99	9.00E+99	FALSE	-1.00000	TRUE	∞
OceanDataFactory	Expert IVL	https://oceandatafactory.se	2024-03-02	9.00E+99	9.00E+99	FALSE	-1.00000	TRUE	∞
Climate Watch Data	Kurzgesagt	https://www.climatewatchdata.org	2024-03-02	9.00E+99	9.00E+99	FALSE	-1.00000	TRUE	∞
ICOS	Expert IVL	https://data.icos-cp.eu	2024-03-02	1,152,524	1,152,524	FALSE	-1.00000	TRUE	1,152,524
European data	Expert IVL	https://data.europa.eu	2024-03-02	1,671,917	345,459	TRUE	0.20662	FALSE	572,604
Zenodo	Expert IVL	https://zenodo.org	2024-03-02	3,418,881	79,413	TRUE	0.02323	FALSE	155,221
PANGEA	Expert IVL	https://www.pangaea.de	2024-03-02	279,303	78,936	TRUE	0.28262	TRUE	123,086
EOSC	Expert IVL	https://eosc-portal.eu	2024-03-02	1,253,656	47,463	TRUE	0.03786	FALSE	91,463
WDC Climate	Expert IVL	https://www.wdc-climate.de	2024-03-02	2,733,240	21,431	TRUE	0.00784	TRUE	42,529
Unep Data	Kurzgesagt	https://www.unep.org/data-resources	2024-03-02	20,213	20,213	FALSE	-1.00000	TRUE	20,213
EDI Data Portal	Expert IVL	https://portal.edirepository.org	2024-03-02	86,551	6,350	TRUE	0.07337	TRUE	11,832
ICPSR	Literture	https://www.icpsr.umich.edu	2024-03-02	19,708	6,524	TRUE	0.33103	FALSE	9,803
Nasa EarthData	Online Search	https://www.earthdata.nasa.gov	2024-03-02	9,263	9,263	FALSE	-1.00000	TRUE	9,263
CEDA	Online Search	https://catalogue.ceda.ac.uk	2024-03-02	8,990	8,990	FALSE	-1.00000	TRUE	8,990
Harvard Dataverse	Literature	https://dataverse.harvard.edu	2024-03-02	159,362	4,287	TRUE	0.02690	FALSE	8,349
EPA	Kurzgesagt	https://www.epa.gov/data	2024-03-02	5,855	5,855	FALSE	-1.00000	TRUE	5,855
Amazon	Expert IVL	https://aws.amazon.com/opendata	2024-03-02	4,221	1,666	TRUE	0.39469	FALSE	2,389
eLTER	Expert IVL	https://catalogue.lter-europe.net	2024-03-02	2,131	1,119	TRUE	0.52511	TRUE	1,467
NSIDC	Kurzgesagt	https://nsidc.org	2024-03-02	1,358	1,358	FALSE	-1.00000	TRUE	1,358
IPCC	Kurzgesagt	https://ipcc-browser.ipcc-data.org	2024-03-02	2,723	516	TRUE	0.18950	TRUE	868
voice of the ocean	Expert IVL	https://voiceoftheocean.org	2024-03-02	479	479	FALSE	-1.00000	TRUE	479
SIOS	Expert IVL	https://sios-svalbard.org	2024-03-02	505,776	177	TRUE	0.00035	TRUE	354
IMF Climate Data	Online Search	https://climatedata.imf.org	2024-03-02	230	230	FALSE	-1.00000	TRUE	230
European Env. Agency	Kurzgesagt	https://www.eea.europa.eu	2024-03-02	229	229	FALSE	-1.00000	TRUE	229
NEON	Expert IVL	https://data.neonscience.org	2024-03-02	182	182	TRUE	1.00000	TRUE	182
UC Irvine	Expert IVL	https://archive.ics.uci.edu	2024-03-02	9.00E+99	9.00E+99	FALSE	-1.00000	FALSE	-1
Google	Expert IVL	https://datasetsearch.research.google.com	2024-03-02	9.00E+99	9.00E+99	FALSE	-1.00000	FALSE	-1
huggingface	Expert IVL	https://huggingface.co	2024-03-02	111,535	111,535	FALSE	-1.00000	FALSE	-1
Our World In Data	Online Search	https://ourworldindata.org	2024-03-02	4,368	4,368	FALSE	-1.00000	FALSE	-1
World Resources Institute	Kurzgesagt	https://datasets.wri.org	2024-03-02	134	134	FALSE	-1.00000	FALSE	-1
USGS	Kurzgesagt	https://www.usgs.gov/products/data	2024-03-02	12,453	12,453	FALSE	-1.00000	FALSE	-1
UN Data	Online Search	https://data.un.org	2024-03-02	8,559	121	FALSE	-1.00000	FALSE	-1
DataCommons	Expert IVL	https://datacommons.org	2024-03-07	9.00E+99	9.00E+99	FALSE	-1.00000	FALSE	-1

[3]link: Microsoft Excel

With the portals ordered based on their harmonic mean, the final step was to introduce a cut-off point. This was decided to be the harmonic mean score of **300** as it allowed to check most of the portals and emphasize on the ones that are relevant to the scope of the study. This also means that portals that ranked a negative score were not evaluated.

Portal Analysis

As mentioned before, document analysis was used in order to investigate the available information [57]. During the investigation, it was difficult to specifically identify DQAs from the documentation present. To help introduce some structure in the process, the themes identified in Figure A.3 were used to define the following questions to help with the investigation:

- **Accuracy**: Does the portal have mechanisms to ensure accuracy in the data they provide?
- **Completeness**: Does the portal have mechanisms to ensure completeness in the data they provide?
- **Consistency**: Does the portal have mechanisms to ensure the provided data are consistent or harmonized with each other?
- **Currentness**: Does the portal inform the dataset consumer about the collection time of the data and give options on filtering based on collection time?
- **Cost-Effectiveness**: Does the portal consider the cost of collection/dissemination and make that information accessible to the dataset providers?
- **Credibility**: Does the portal provide information from credible sources, like international or state organizations?
- **Transparency**: Is the portal sharing information regarding the source, collection methods, processing, and metadata creation of the provided datasets?

Unfortunately, even with these guiding questions, as most portals made implicit statements about different DQAs and mainly through their legal documentation stating their absence of liability making identification difficult. Some notable exceptions are statements surrounding the **FAIR** guiding principles [66] which contain the DQAs of: (1) findability, (2) accessibility, (3) interoperability and (4) reusability. Also, a portal made a particular mention of the concept of contextuality[4], but since no other portal or investigated literature made specific mention of this DQA it was not included in the following steps of the study. The final output of the evaluation can be seen in the Appendix A.4. Snippets from the documentation that lead to these conclusions can be seen on the provided replication package that will be provided with this report.

Closing this investigation, some portals were not providing data from dataset providers directly but were acting as "meta-portals" meaning that they provided a common interface to multiple other portals. This shows that there are efforts to unify the existing solutions into a standardized framework that can bring a lot of benefits to future researchers. Examples like these included: SIOS[5], or the CEDA Archive[6].

[4]link: European Data - Contextuallity
[5]link: SIOS portal
[6]link: CEDA Archive portal

This of course meant that for such portals, among with some others as well, emphasis was placed on providing high quality metadata that can allow the user to find the data they need more effectively. In many such cases, the portals actively stated that they do not guarantee the quality of the data themselves. As will be seen later in the survey, this view is prompted by the mentality that the consumers should take note of the present metadata and decide if a dataset covers their needs.

DQAs Survey Inclusion

After investigating the literature and the portals, numerous DQAs were collected, but not all could be included in the survey as it would jeopardise response rate due to an excessive length of the survey. To address this, the DQAs that appeared at least **ten** times or were prominent in the portals, meaning the FAIR principles, were included in the Likert scale rating process as described in Figure 3.1. This included the 15 DQAs present in Table 4.5.

Table 4.5: DQAs included in the Likert Scale of the survey

Accessibility	Consistency	Interoperability	Reusability	Understandability
Accuracy	Credibility	Interpretability	Security	Usability
Completeness	Findability	Relevance	Timeliness	Value-Added

For these DQAs a definition was provided which can be seen in the Appendix A.2. The remaining DQAs were included in a separate set of questions as long as they appeared at least twice. For those, the respondents were tasked to tick the 5 ones they considered to be most and least important. These included **21** DQAs which are outlined in Table 4.6. The remaining **36** were not included in the survey.

Table 4.6: DQAs included in the Pick-5 section of the survey

Appropriate Amount of Data	Documentation	Recoverability
Complexity	Duplications	Representation
Compliance	Flexibility	Skewness
Concise Representation	Granularity	Traceability
Conciseness	Objectivity	Unambiguous
Contactability	Presentation Quality	Uniqueness
Cost Effectiveness	Provenance	Volume

4.2 Survey Results

Survey Distribution

As mentioned before, the survey was distributed online to the stakeholders investigated in the study. Table 4.7 provides the numbers of contacted persons initially identified for the survey.

Table 4.7: Number of contacts for the different stakeholder groups

Stakeholder Group	e-mail contacts	online form contacts
Data Portals	16	7
Dataset Providers	57	17
Dataset Consumers	158	-
Total	**231**	**24**

From the collected contacts, three were duplicated and removed from the sample. In the end, **252** individuals or organizations were contacted. From them, three individuals, returned an automatic response that their e-mail is no longer valid as such they were excluded from the sample. So in total, **249** valid samples were initially and successfully contacted for the investigation.
The survey was available for three weeks and in the end a total of **34** responses were collected, with a reminder at the half-way point. The response rate covered about 13% of the contacted respondents by the end of the collection.
However, it is worth noting that since organizations were contacted as well, the questionnaire was shared with their employees and as such this response percentage is not representing the initial sample. As the survey was anonymous and snowballing was encouraged to the respondents in order to get more samples, it is not possible to get an accurate response rate from the initial sample.

Respondent Distribution

When the results were collected, some clean-up of the roles was needed. As mentioned before, the respondents were allowed to provide other roles than the provided ones. These were investigated, and it was possible to have them mapped to the values already expected by the model. As such, the following changes were made based on the roles defined in Section 3.4:

Table 4.8: Updates from the proposed roles of the respondents

Proposed Role	Respondent Statement
2 Roles: Dataset Provider and Data Portal Maintainer	I am connected with both providing data and hosting a platform to provide access to datasets.
Dataset Provider	Metrologist. I help data producers work out their uncertainties
Data Portal Maintainer	I manage a data centre so work with providers to deposit their data and maintain a 'portal' for users to access the data we hold
2 Roles: Dataset Provider and Dataset Consumer	dataset provider and consumer
Dataset Consumer	Data Scientist

With these changes implemented, the distributions of the roles can be seen in Figure 4.3. The different experiences investigated can be seen in Figures 4.4, 4.5, 4.6.

Survey Results — Open-Ended Questions

In the survey, two types of open-ended questions were provided. Firstly, an open-ended question was associated with each DQA rating to prompt the respondent

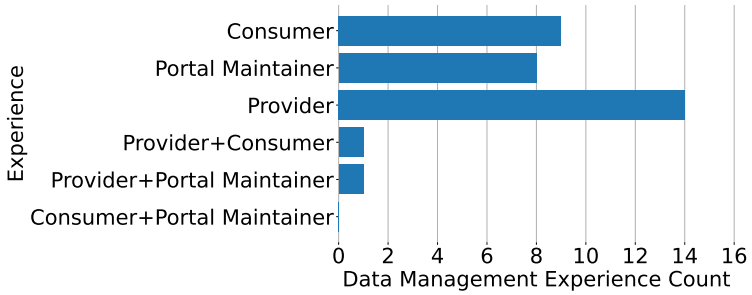

Figure 4.3: Distribution of roles in the results datasets. There is a similar amount of responses for the three main roles, but a minuscule presence of the double roles.

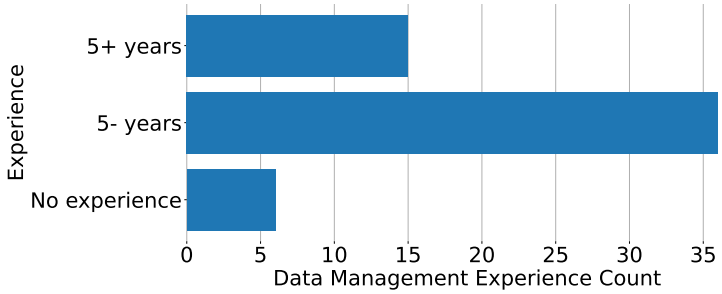

Figure 4.4: Distribution of the data management experience in the dataset.

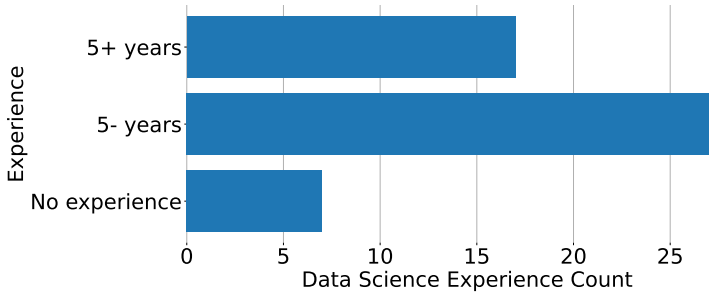

Figure 4.5: Distribution of the data science experience in the dataset.

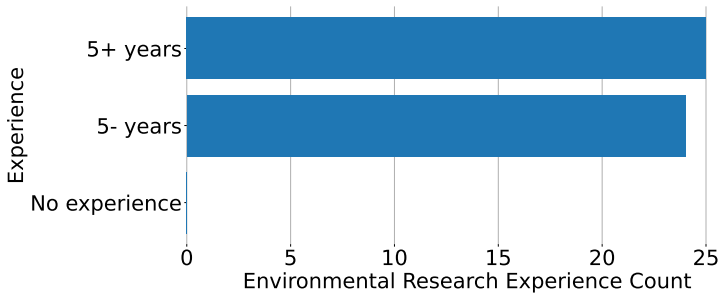

Figure 4.6: Distribution of the environmental research experience in the dataset. The lack of inexperienced samples is expected given the contacted populations.

to challenge the definition and similar DQAs if they so wished. The second was at the end of the survey to share any DQAs they believed were missed. Both were evaluated following thematic analysis [9]. After getting familiar with the data, initial codes were created and grouped into themes that were evaluated by the experts who supported the first pilot study.

Definition and Similar DQAs Evaluation Question

In the following tables, the comments separated by the DQA rating question will be presented. Also, the respondent ID is provided. It is not directly associated with the order or position of the respondent in the data. It is used in order to connect their background to their response, which can be seen in Table 4.9. The comments will be discussed and associated with the analysed ratings later in this section.
For this analysis the following themes were proposed:

- **Amendment** - Suggests a change to the existing definition.
- **Best Effort** - Suggest emphasis on best effort than absolute satisfaction.
- **Context** - Emphasizes the importance of context for this definition.
- **Other Definition** - Proposes a new definition.
- **Unclear** - Has a follow-up question or agrees with the definition

Table 4.9: Respondents background for rating question comments. Their **role**, data management experience (**exp_dm**), data science experience (**exp_ds**) and environmental research experience (**exp_er**) is presented.

ID	Role	exp_ds	exp_dm	exp_er
R1	Dataset Provider	3–5 years	1–2 years	5+ years
R2	Provider + Portal Maintainer	No experience	No experience	5+ years
R3	Dataset Consumer	No experience	5+ years	5+ years
R4	Dataset Provider	No experience	No experience	5+ years
R5	Data Portal Maintainer	5+ years	1–2 years	5+ years
R6	Dataset Provider	5+ years	5+ years	5+ years
R7	Data Portal Maintainer.	5+ years	5+ years	5+ years
R8	Dataset Provider	1–2 years	No experience	5+ years
R9	Data Portal Maintainer	5+ years	5+ years	5+ years
R10	Data Portal Maintainer	1–2 years	3–5 years	5+ years
R11	Dataset Provider	Less than 1 year	Less than 1 year	1–2 years
R12	Dataset Consumer	5+ years	3–5 years	5+ years

Table 4.10: Thematic Analysis of Respondent comments on Accessibility

Proposed Theme	Statement
Context	R1: It depends on the data. For example military and medical data require some level of protection. I can't think of a reason not to share environment data.
Unclear	R2: By "standardized communications protocol" do you mean something like a web site?
Amendment	R3: It is perfectly fine with data available upon request as well, as long as there is a guarantee that the function responsible is lasting.

Table 4.11: Thematic Analysis of Respondent comments on Accuracy

Proposed Theme	Statement
Amendment	R1: Having "correct" data could lead to data manipulation. Data providers may bias analysis while cleaning data. I think the better policy is to share the raw data and methods whenever possible.
Amendment	R2: You do not mention the importance of documentation of methodology and data quality control. Yes, it is very important that data are accurate, but this requires documentation that traces the methodology, standards, data quality examination, etc.
Other Definition	R4: That's a poor definition of accurate from the point of view of a metrologist (measurement scientist). We worked long and hard with GCOS[7]to talk about uncertainties rather than accuracy. More fundamentally, suitable accuracy will depend on the Application. We talk about assessing fitness for purpose and providing a robust indicator of quality (see [...]). Getting a robust uncertainty assessment allows users to judge fitness for purpose.
Amendment	R5: Sometimes even if data are not accurate, but precise, it offers a lot of information about the technology and corrections can still be made - so these data are still valuable.
Best Effort	R6: There is a proxy, data should be accurate enough, but in data sparse areas it may be better to release preliminary not QC data rather than no data
Amendment	R7: Fidelity of metadata should provide the means to ascertain the accuracy of data. For example, the metadata should describe the data type, bounds or limitations, missing values, and provenance, if derived.

Table 4.12: Thematic Analysis of Respondent comments on Completeness

Proposed Theme	Statement
Best Effort	R1: It's very important, but seldom realistic (especially in academic research).
Other Definition	R4: I like the way GCOS requirements describe spatial and temporal resolution and completeness.
Amendment	R6: Final data should be complete
Best Effort	R7: Complete as humanly and technologically possible. Data collections will inevitably be missing data points for various reasons.
Amendment	R8: Hard to define complete. Data is constant work and the data can be in different stages. It should be accuratly taget in what QC stage it is in.
Best Effort	R9: depending on the circumstances of collection, it may be impossible for some datasets to have values for all expected attributes. A different standard could be applied to metadata, however, and I would choose all 6 stars for completeness of metadata.
Amendment	R10: I do not disagree with this definition, but rather propose to extend the term completeness to data that had been measured in some point of time, but because deemed irrelevant for some specific purpose, e.g. the research question addressed in a specific manuscript, is never going to be published anywhere. And thus will probably never be relevant to any other research question by whoever.

Table 4.13: Thematic Analysis of Respondent comments on Consistency

Proposed Theme	Statement
Context	R1: Depends on the project, for experiments it's very important to have a balanced design and good quality. For field observations a complete dataset is generally unrealistic.
Other Definition	R4: As a metrologist I think of consistency as having traceability to SI
Best Effort	R6: For climate trend this is the target, not always possible.
Amendment	R7: Consistency is important. For example, dates should follow the same format (preferably ISO 8601) in the data collection.
Best Effort	R9: consistency across similar data is an ideal, difficult to achieve. Within a single data entity, 6 stars.
Other Definition	R11: As long as the data is accurate consistancy in the data isn't the most important thing. Consistancy in the data collection is important, that the time series has consistancy.

[7]GCOS - Global Climate Observing System

Table 4.14: Thematic Analysis of Respondent comments on Credibility

Proposed Theme	Statement
Other Definition	R4: Have a look at the [...] project and the concept of a maturity matrix. That captures what it means to be credible.
Amendment	R5: Sometimes data are not ground truth - but we can still learn from these "wrong data" - they offer insights into what could possibly be wrong out there. But to distinguish this, the metadata and associated information should be very clear.
Context	R7: Credible data is much harder to ensure within an open data repository. Trust in the data source provider is one approach to ensuring credible data, but is not 100% successful. We believe that the ultimate determination of fitness-of-use is between the provider and the consumer of the data, not necessarily the data repository.
Context	R9: It would depend on the context of use, whether credibility should be 6-star. The examples (origins, attributions, commitments) apply to metadata fields rather than data values.

Table 4.15: Thematic Analysis of Respondent comments on Interoperability

Proposed Theme	Statement
Other Definition	R4: This is a data science meaning of interoperable. We use the term in broader ways. Generally it's about the ability of systems to work together. That can be achieved in two ways - (1) by converting one data set to be more like another or (2) by providing information to allow both to be used in a process. From your viewpoint that describes things like data formats. But for us it's more about how we deal with the fact [...] have slightly different special response functions for their bands.
Other Definition	R5: In other earth science circles, interoperability was interpreted as interoperability of semantics
Best Effort	R9: Like consistency across similar datasets, interoperable is an ideal, rarely achieved.

Table 4.16: Thematic Analysis of Respondent comments on Relevance

Proposed Theme	Statement
Context	R1: This is very subjective to the end user.
Context	R3: Every piece of information stored may be of different relevance for different aspects.
Unclear	R4: This feels like an odd one - we wouldn't use data that weren't relevant.
Unclear	R5: At some point this will have to be truer and we all have to be more discerning of data to store because data storage also has its carbon footprint - and data storage is becoming more and more expensive!
Amendment	R6: What is relevant? As long as it is FAIR I belive it is up to users to find the relevance?
Context	R7: The relevancy of data is very subjective and may not become apparent for many years after archiving.
Unclear	R9: relevance depends on the intended use of the data. I think we can assume, no one will use irrelevant data (so perhaps this question is irrelevant?).
Context	R10: relevancy always depends on a specific perspective or fitness for a certain purpose - I'd second the definition, if "Every piece of information..." means "Every reliable, trustworthy piece of information..."
Context	R11: relevance is relative, what is irrelevant to me could be important for someone elses analysis
Amendment	R12: Just to clarify. It is easy to throw away data not relevant for a specific project. Thus, that there is irrelevant data in a dataset is not a problem, it might be important for someone else.

Table 4.17: Thematic Analysis of Respondent comments on Reusability

Proposed Theme	Statement
Amendment	R4: For me a key part of this is providing sufficient information about uncertainties that higher level processing algorithms for different algorithms can all use the information. It also is about what we call long term data preservation. Scientist of the future need to be able to reprocess our data sets to get climate records.
Unclear	R5: To me this is one of the main tenets of open data - if people want to reuse, repurpose and reanalyze the data - they should be able to do so.
Context	R7: The reusability of data is very subjective and may not become apparent for many years after archiving.
Amendment	R10: "(meta)data are associated with detailed provenance" would certainly be desirable, but is very rarely provided by authors of datasets, because there is no additional benefit for them beyond the DOI to be included in the publication, and no control by publishers if the full set of provenance data has not been provided.

Table 4.18: Thematic Analysis of Respondent comments on Security

Proposed Theme	Statement
Context	R1: Depends on the project. In my experience safety isn't very important as the data is of little commercial value.
Context	R4: Climate data should ALWAYS be completely public to anyone. But you don't want anyone to be able to change it. So to me secure is about being safe from being edited.
Context	R5: In terms of security and keeping sensitive data confidential (eg data that - if exposed - might endanger people's lives) - I think data should be kept secure. However, if data security is posed in the interest of commercial usage - this should be avoided and I think there should be more data openness.
Context	R8: Our data is open and you don't need to sign in/create an account to access it.
Context	R7: Public data means that data should be unfettered and available to anyone, including competitors. Data should, however, be secure from malicious or non-malicious tampering.
Other Definition	R9: I disagree with the definition. To me, Secure means avaialble for use far into the future, so stored reliably and backed up. Your definition of secure seems to make it inaccessible. I chose 1-star for your definition. With my definition, I would choose 6 stars.
Context	R11: The data should be open access to everyone even competitors, but secure so that no one can go in and change the raw data

Table 4.19: Thematic Analysis of Respondent comments on Timeliness

Proposed Theme	Statement
Context	R4: Well it should be "sufficiently up to date for the task in hand". But that will vary wildly. Someone monitoring forest fires or disaster zones or getting information for the shopping forecast needs data within hours. Climate modellers need better data within weeks or months. The space agencies distinguish "near real time" data from higher quality data with a latency.
Context	R5: timely data is useful, but historical data can still be useful. So even if it is "old" data, it would still be useful for trend analysis.
Context	R6: Timeliness needed for operational data in RD/NRT, for historical data timeliness is needed, but on another level
Context	R7: The timeliness of data is very subjective and may not become apparent for many years after archiving.
Context	R8: Depends on the scientific question if it is relevant or not
Context	R10: The concept currentness is in a field of tension to credibility and trustworthiness (or quality). "Quality takes time"

Table 4.20: Thematic Analysis of Respondent comments on Understandability

Proposed Theme	Statement
Amendment	R4: We usually have time to work it out ourselves in our readers. Should be well documented for humans. I don't need it to be AI-understandable. But that distinction is not clear - so you mean human or machine understandable?
Unclear	R5: this is another tenet of data openness. If data is not understandable and translated to "information" then its "openness" has failed.
Amendment	R6: What is understandable? Machine understandable, understandable for human eye?

Table 4.21: Thematic Analysis of Respondent comments on Usability. The comment from R5 is the same as the one in Table 4.20, so it was removed.

Proposed Theme	Statement
Amendment	R4: Odd definition that includes the word being defined!
~~Unclear~~	~~R5: this is another tenet of data openness. If data is not understandable and translated to "information" then its "openness" has failed.~~

Table 4.22: Thematic Analysis of Respondent comments on Value Added

Proposed Theme	Statement
Context	R1: This doesn't apply to my experience in academia
Context	R3: From the data user perspective the degree of imporatance here varies, I think.
Context	R4: Most of us are doing public good research. "Add value" is related to societal benefit vs taxpayer costs not competitive edge...
Unclear	R5: Data occupy space and take resources to collect and store - it is just fair to have value to data rather than to be just "numbers" being presented.
Context	R7: The added value of data is very subjective and may not become apparent for many years after archiving.
Unclear	R9: I would not pursue the project unless it added value. irrelevant question.
Context	R10: see my notion concerning relevancy. Also depends on purpose and perspective.

The comments show that some DQA definitions are insufficient to capture the respondents' views suggesting amendments and context consideration. As will be discussed in Section 5.3, a future study could create new definitions for these DQAs.

Missed DQAs Question

Like before, thematic analysis was performed on the comments. The respondents' background can be seen in Table 4.23 and the proposed themes are in Table 4.24. These themes were treated as new DQAs expanding the existing DQAs and will be discussed in Chapter 5.1.

Table 4.23: Respondents' background for missing DQAs. The ID is based on Table 4.24. Their **role**, data management experience (**exp_dm**), data science experience (**exp_ds**) and environmental research experience (**exp_er**) is presented.

ID	Role	exp_ds	exp_dm	exp_er
R13	Dataset Consumer	5+ years	5+ years	3-5 years
R14	Dataset Provider	1-2 years	No experience	5+ years
R15	Data Portal Maintainer	5+ years	5+ years	3-5 years
R16	Data Portal Maintainer	5+ years	5+ years	5+ years
R17	Data Portal Maintainer	1-2 years	3-5 years	5+ years

Table 4.24: Thematic Analysis of Missed DQAs based on Respondents. Availability is excluded as it was introduced in the similar DQAs around accessibility

Proposed Theme	Statement
~~Availability~~	~~R13: Availability~~
Documentation	R14: The importence of Meta data.
	R15: Metadata provided in iso format, so no abbreviations, units added etc. not sure if that is include in the word list (some words not clear to me what they mean).
-	R16: I would need to know how you define each of these terms before I can decide. which aspects are most relevant will depend on the project.
De-personalization	R17: Role of reviewing or de-personalizing data (e.g. by curation or following 4-eye-principles)
Rewarding Quality	R17: aspect of rewarding investments in data quality in the scientific value chain

Survey Results — Close-Ended Questions

For the close-ended questions, Bayesian data analysis was used, considering the small number of samples [38]. In order to reduce the amount of graphs to be presented in this section, only a single example will be explained in detail. The rest can be found in the relevant appendix sections. The only exclusion are the forest plots, as they will be directly used to explain the results of this study.
For the statistical analysis, the presentation of the scores was done using the independent variables, which were the first section of the questionnaire. These include the *role*, *data science experience*, *data management experience* and *environmental research experience*.
A box-plot representation of the data distribution can be seen in Figure 4.7. The median is shown with the orange line. The box extends between the first quartile (Q1) and third quartile (Q3). Following the definition from Tukey [61], the "whisker" of each plot occupies up to 1.5 times the range of the box. All other points, are treated as outliers (aka "fliers") and are presented as circles outside the whiskers. Individual box-plots of each identified combination of independent variables can be viewed in the Appendix Section A.

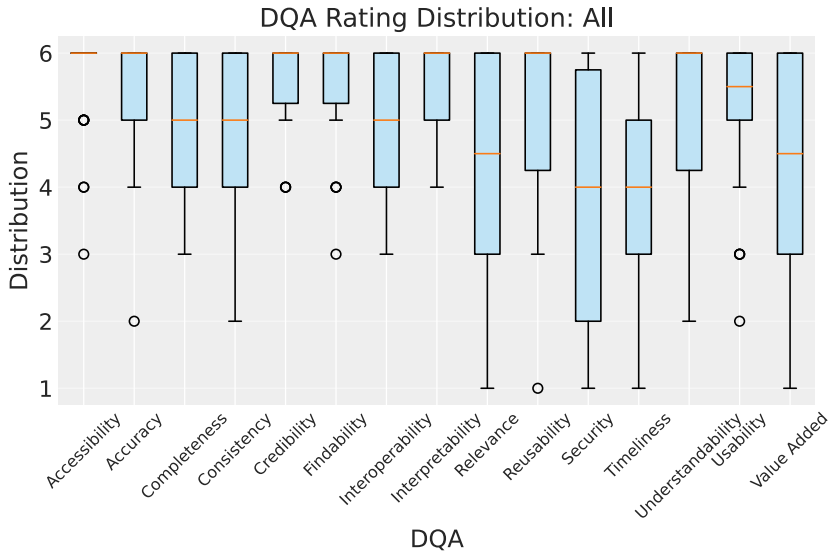

Figure 4.7: Distribution of the data of the closed-ended rating questions.

For the Pick-5 questions, collected picks can be seen in Figure 4.8 for the "most relevant" DQAs and in Figure 4.9 for the "least relevant" DQAs. Take note that the vertical axes of each graph is different. In general, there were a lot less responses on which DQAs are least relevant.

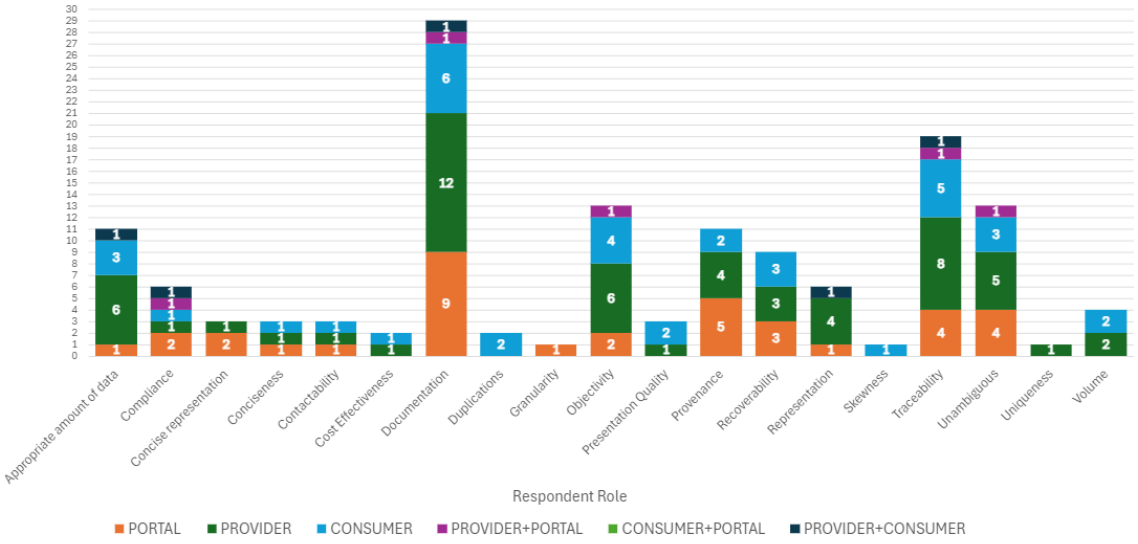

Figure 4.8: "Most Relevant" Pick-5 responses separated by role.

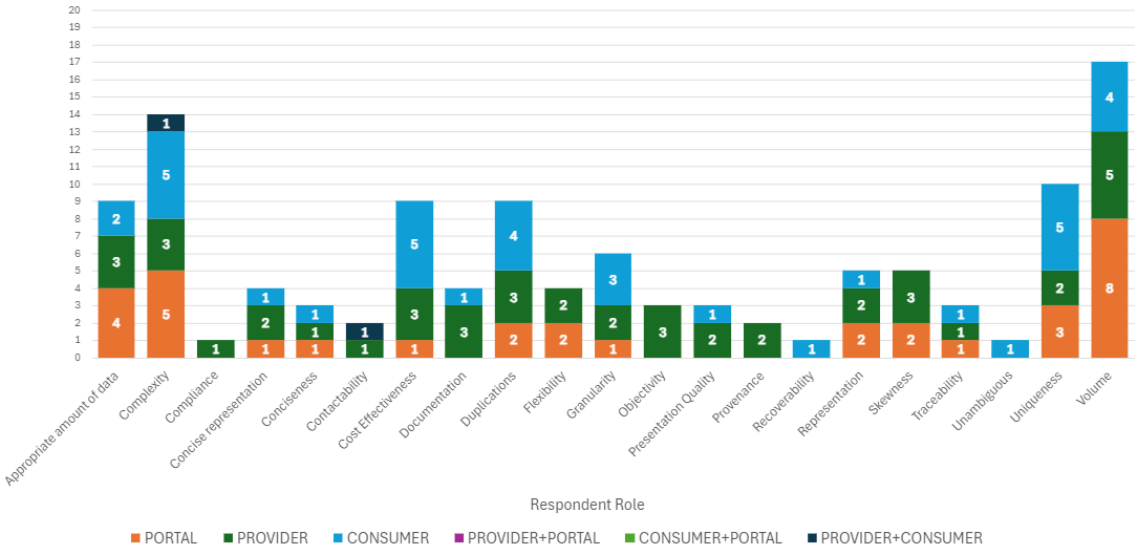

Figure 4.9: "Least Relevant" Pick-5 responses separated by role.

Bayesian data analysis of Part A: Ratings

The complete example case that will be presented in this section is the case of *accessibility*, as is the first one alphabetically. For the rest of the ratings, only the forest plots will be presented, but the other relevant plots and tables can be viewed in the corresponding appendix sections. Like in the checks with the mock data in

Section 3.4, all models were fit with 1000 samples and 1000 tuning samples. The hyperparameters proposed were not tuned further for each question investigated. Also, as shown in Figure 4.3 there was a very small number of samples collected for the combined roles. They were still included in the models, but more samples would be needed in order to derive reliable conclusions. This can be seen in the output of the models as well, with every such case having a very wide distribution.

Starting the analysis, each model was evaluated based on the posterior predictive check to ensure that the model fit the data correctly. For the case of accessibility, this can be seen in Figure 4.10. The rest of the ratings can be viewed in the Appendix Section A. As before, the aim was for the priors to closely resemble the model, but not directly align with it.

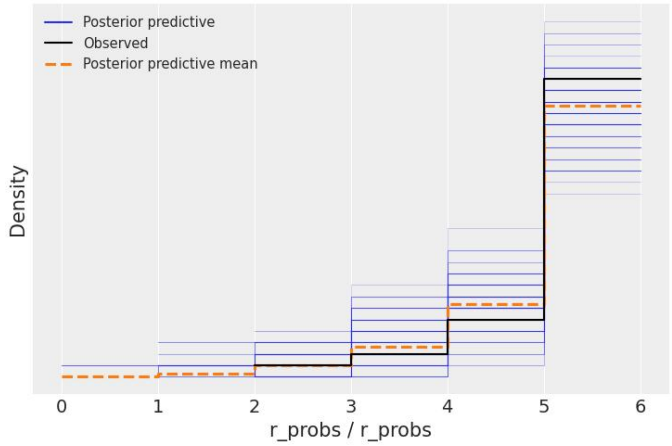

Figure 4.10: Posterior predictive check of the Accessibility model.

This was followed by performing the sampling of the model. The predicted betas (β) of the roles and experience can be viewed in the density plots 4.11. The plots for the rest of the ratings can be viewed in the Appendix Section A. Each sub-plot represents one parameter.

The trace used to create this density plot can be seen in 4.25 providing a more transparent view of the analysis. Like before, the trace of the remaining ratings can be seen in the Appendix Section A. In the trace, the regression coefficients are presented for the different single roles and double roles are presented. Also, the regression coefficients of the data science experience (DS_Beta), data management experience (DM_Beta) and environmental research experience (ER_Beta) are presented.

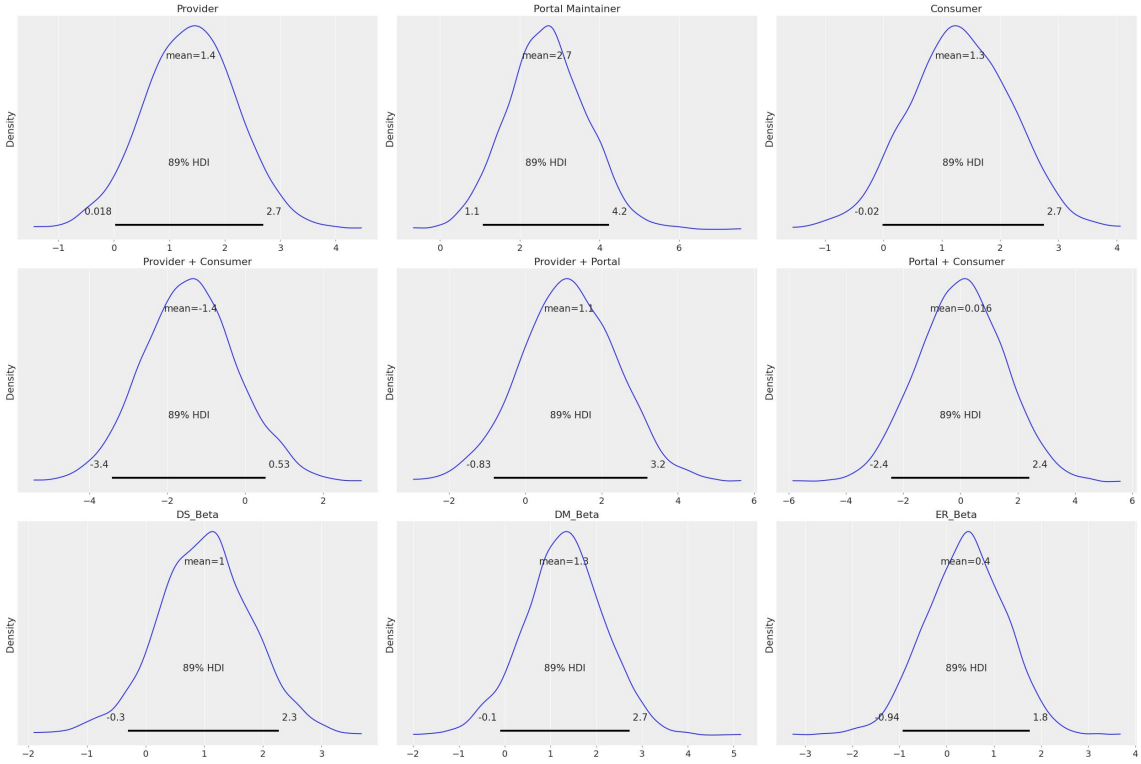

Figure 4.11: Density plots of betas for the Accessibility model.

Table 4.25: Trace representation of Accessibility model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	1.39	0.85	0.02	2.69	2574.47	1.0
Portal Maintainer	2.68	1.0	1.07	4.25	2906.66	1.0
Consumer	1.31	0.88	-0.02	2.75	2729.67	1.0
Provider + Consumer	-1.41	1.23	-3.43	0.53	2993.96	1.0
Provider + Portal	1.14	1.27	-0.83	3.2	3018.23	1.0
Portal + Consumer	0.02	1.52	-2.41	2.4	2947.69	1.0
DS_Beta	1.0	0.81	-0.3	2.27	2900.17	1.0
DM_Beta	1.29	0.88	-0.1	2.73	2434.15	1.0
ER_Beta	0.4	0.86	-0.94	1.76	2764.08	1.0

Trace plots have the reputation of being difficult to interpret [38]. With this in mind, forest plot representations of the trace of all ratings can be viewed in Figures 4.12-4.26 with the HDI set to 95%. Also, due to limitations in the chosen libraries, it was not possible to set a common horizontal axis range. Finally, in the following plots the double roles are presented as well, but due to the limited number of samples they will not be discussed.

In Figure 4.12 the DQA of accessibility is presented. This DQA is important for all roles, and especially for individuals identified as data portal maintainers with experience in data management and data science. Respondents with environmental research experience are not as positively associated with this DQA as other experiences.
Comments around this DQA, found in Table 4.10, mention that for environmental research, data should be openly accessible. They also make mention of the importance of long term accessibility.

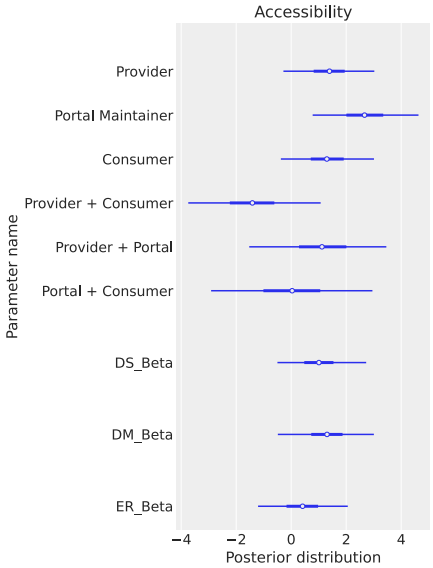

Figure 4.12: Forest plot representation of the Accessibility trace.

Accuracy, shown in Figure 4.13, is important for all roles but is slightly negatively associated with individuals with data science experience. An explanation can be found in the comments on accuracy provided in Table 4.11. These mention the higher importance of documentation and transparent methodology than accuracy, as it enables the consumer to check for uncertainty.
Individuals identified with data management experience are quite positively correlated to accuracy when compared to other experiences. Associated comments with such experience state the importance of accurate metadata instead of the data themselves.

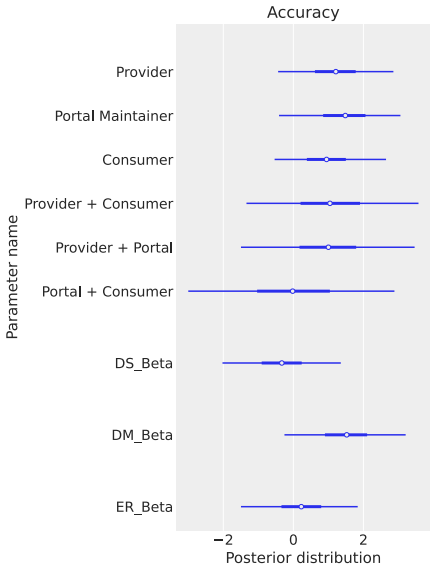

Figure 4.13: Forest plot representation of the Accuracy trace.

Completeness and consistency, found in Figures 4.14 and 4.15 respectively, are slightly relevant for everyone, but there are no outstanding correlations with any role or experience. Some comments around these DQAs make mention of "best effort" instead of aiming to completely satisfy them, as mentioned in the Tables 4.12 and 4.13.

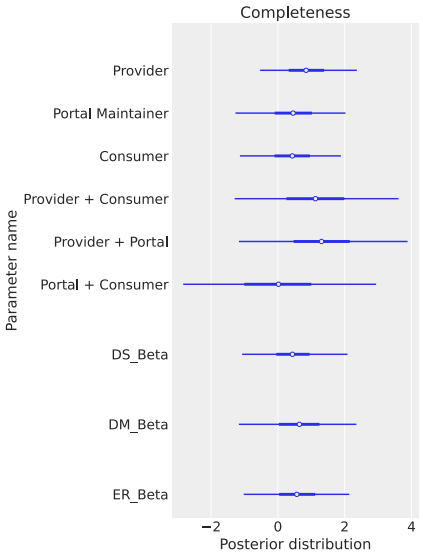

Figure 4.14: Forest plot representation of the Completeness trace.

Figure 4.15: Forest plot representation of the Consistency trace.

Credibility is valued by all roles and experiences, as seen in Figure 4.16. Notably, not so much by the data portal maintainers. Comments from such respondents in Table 4.14 state the importance of metadata and the need for other stakeholders to make best use of them in order to see if the dataset best fits their assumptions and needs. Also, individuals with data science experience do not seem to be interested in this DQA.

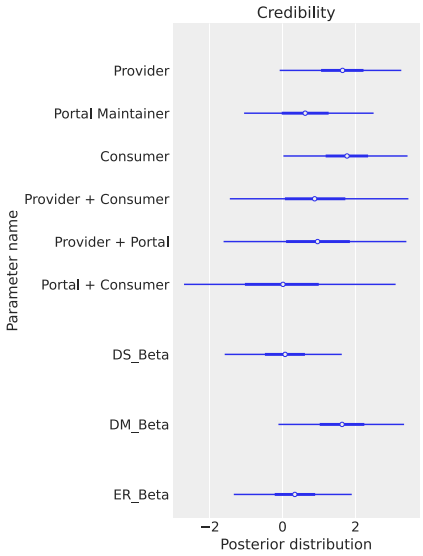

Figure 4.16: Forest plot representation of the Credibility trace.

A similar situation is seen with findability in Figure 4.17. However, data portal maintainers consider this DQA quite important, while the respondents who identified as consumers as less interested in this DQA. It is also worth mentioning that for this DQA as well as interpretability, no changes were suggested on the provided definitions.

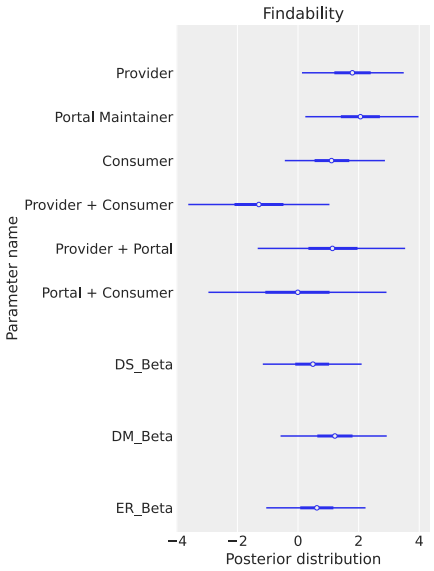

Figure 4.17: Forest plot representation of the Findability trace.

Interoperability and Interpretability, shown in Figures 4.18 and 4.19 respectively, are quite important for data portal maintainers with experience in data management and data science. Interestingly, environmental research experience seems to be not affecting these DQAs being at the zero point of both plots.
One noteworthy difference between these two DQAs is that consumers do not value interpretability as much as other roles. This contrasts Interoperability, where all roles appear to be similarly interested in this DQA.

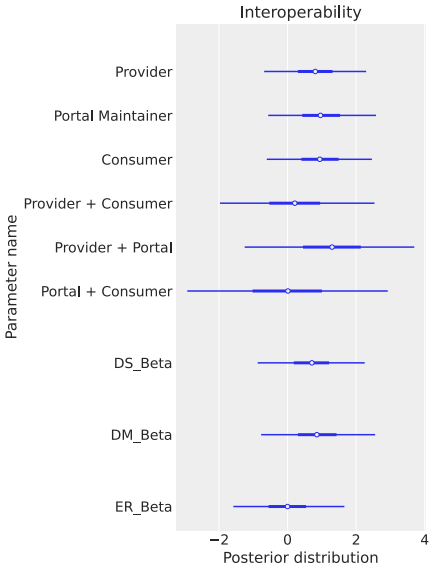

Figure 4.18: Forest plot representation of the Interoperability trace.

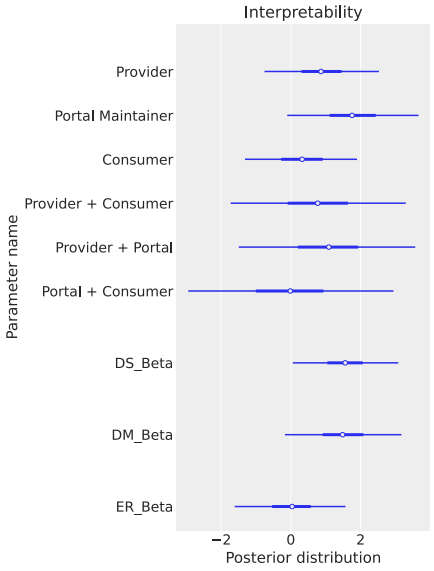

Figure 4.19: Forest plot representation of the Interpretability trace.

Relevance is not that strongly correlated with any experience or role. As shown in Figure 4.20 it is negatively impacted by data portals maintainers and dataset consumers with environmental research experience. While, it is slightly positively associated with respondents who identified as dataset providers or have experience in data science and data management. Quite a few of the comments in Table 4.16 raise the importance of context for defining the need of this DQA.

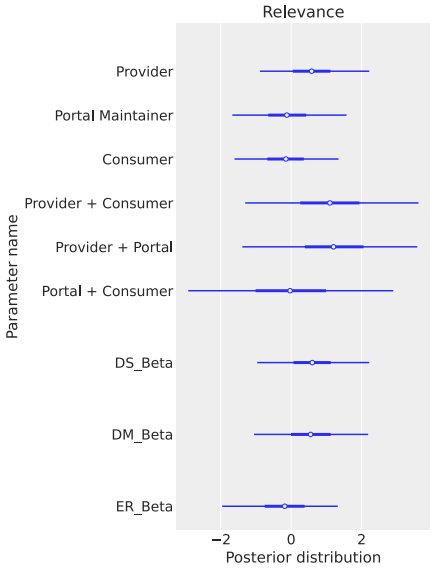

Figure 4.20: Forest plot representation of the Relevance trace.

Reusability, is highly important from the perspective of the dataset providers and data portal maintainers as shown in Figure 4.21. Interestingly, the respondent R10, who is a data portal maintainer, commented around reusability in Table 4.17, that from their perspective, it is not something that dataset providers actively adhere to due to the lack of incentives.
Another detail on reusability is that individuals identified with data science experience are slightly negatively correlated with this DQA. Dataset consumers as well do not view it as highly as other stakeholders.

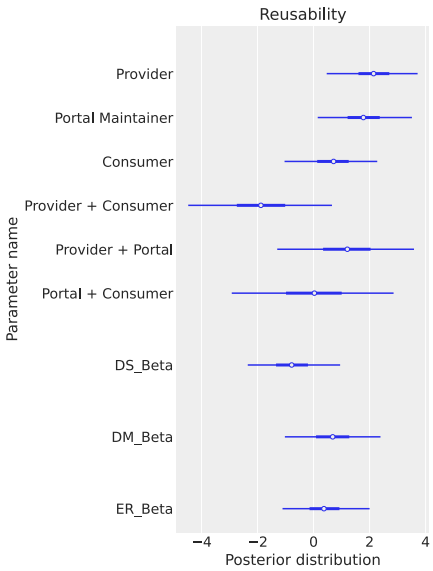

Figure 4.21: Forest plot representation of the Reusability trace.

Security, viewed in Figure 4.22 is not strongly correlated to any role or experience. It appears that individuals identified as data portal maintainers are negatively associated with this DQA. Another noteworthy negative association appears to be with environmental research experience. Quite a few comments, in Table 4.18, raise the importance of open data being accessible to everyone, many stating that there are no competitors as the proposed definition suggests. Still, a few comments also suggest the importance of protecting data from editing.

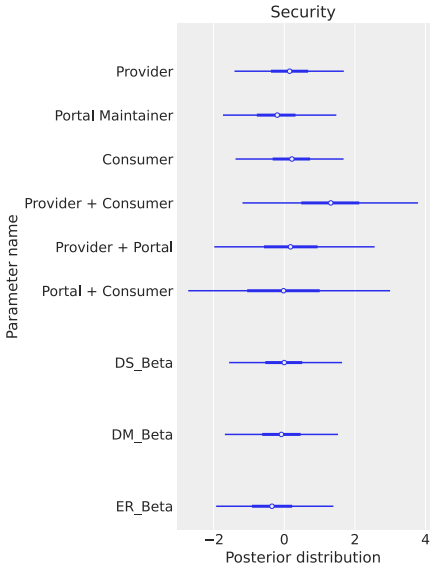

Figure 4.22: Forest plot representation of the Security trace.

Timeliness is important mainly for dataset consumers as well as respondents with experience in data science and environmental research, as shown in Figure 4.23. For the other parameters it is unimportant, as there appears to be a negative correlation with this DQA. The collected comments in Table 4.19 emphasize the importance of context for this DQA and as R10 puts it "quality takes time".

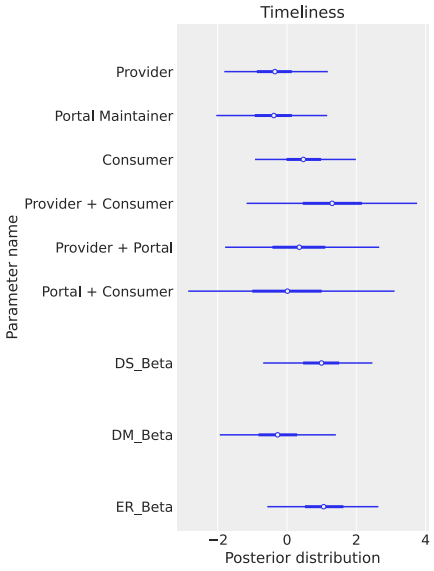

Figure 4.23: Forest plot representation of the Timeliness trace.

The DQAs of understandability, shown in Figure 4.24, and usability, shown in Figure 4.25 are showing similar results. All roles and experiences are positively associated, and especially the respondents identified as data portal maintainers. Dataset providers are slightly less but still positively associated with these DQAs. In Table 4.20 commenting around understandability, the respondents mention the need for defining if this DQA is related to machine or human understanding.

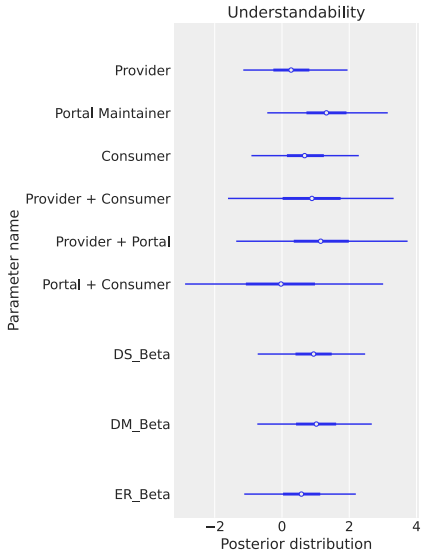

Figure 4.24: Forest plot representation of the Understandability trace.

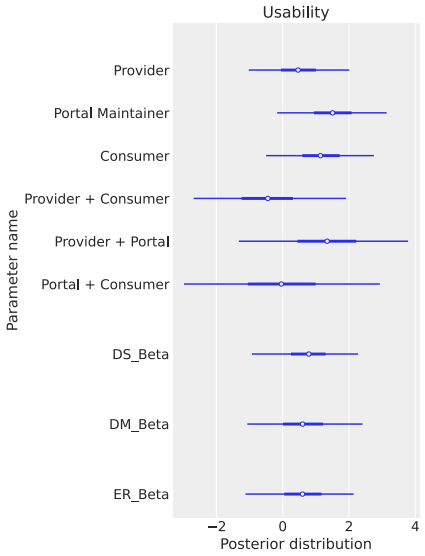

Figure 4.25: Forest plot representation of the Usability trace.

Looking into Figure 4.26 for the value added rating, an interesting situation is presented. Data portal maintainers, with experience in data management, are negatively associated with this DQA, while the other roles and experiences are positively associated. Most of the comments, collected in Table 4.22, raise about the importance of context. Most of the comments are from data portal maintainers. The definition makes mentions of competitive edge, which does not align with the view of open data.

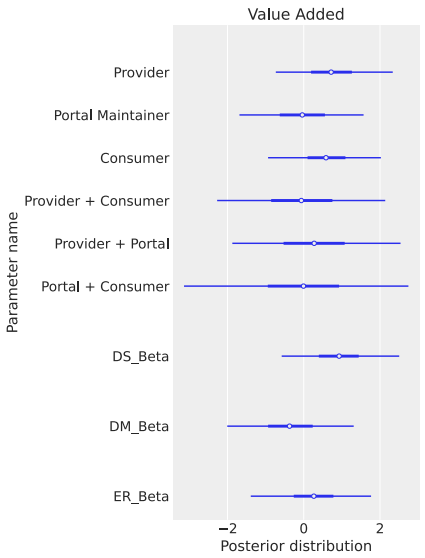

Figure 4.26: Forest plot representation of the Value Added trace.

Bayesian data analysis of Part B:Pick-5

Similar to the analysis of the ratings, for the Pick-5 questions the same procedures took place. The forest plots created for the analysis can be seen in Figure 4.27 for the most relevant DQAs, and Figure 4.28 for the least relevant DQAs. For this investigation the results were not separated by role or experience.
The rest of analysis plots and trace can be seen in corresponding appendix section. Posterior predictive checks can be seen in the Appendix Section A. The density plots of each model can be seen in the Appendix Section A. Finally, the trace of each model can be viewed in the Appendix Section A.
Looking at the most relevant DQAs in Figure A.15 documentation appears quite important to all respondents overshadowing the rest. It is followed by traceability and unambiguity. This view aligns with various comments that have been seen in the first part of the survey surrounding the ratings. More specifically, Tables 4.11, 4.12, 4.14. A noteworthy mention is also the comment from R14 and R15 in Table 4.24. All of them argue about the importance of providing metadata and allowing the consumers to decide on the usefulness of a dataset based on those.
Data volume, uniqueness, and duplications are considered by many stakeholders to be least relevant, as seen in Figure 4.28. This raises some potential insights around data quantity, which will be discussed more in the conclusion chapter.

A controversial point in both figures is the DQA around the appropriate amount of data. There is a similar number of responses arguing for both directions. Consulting Figures 4.8 and 4.9 it can be seen that mainly data portal maintainers are the ones who consider it to be least relevant, while dataset providers and dataset consumers consider it to be more relevant. This aligns with views presented in the comments emphasizing the higher importance of metadata from the perspective of the portals.

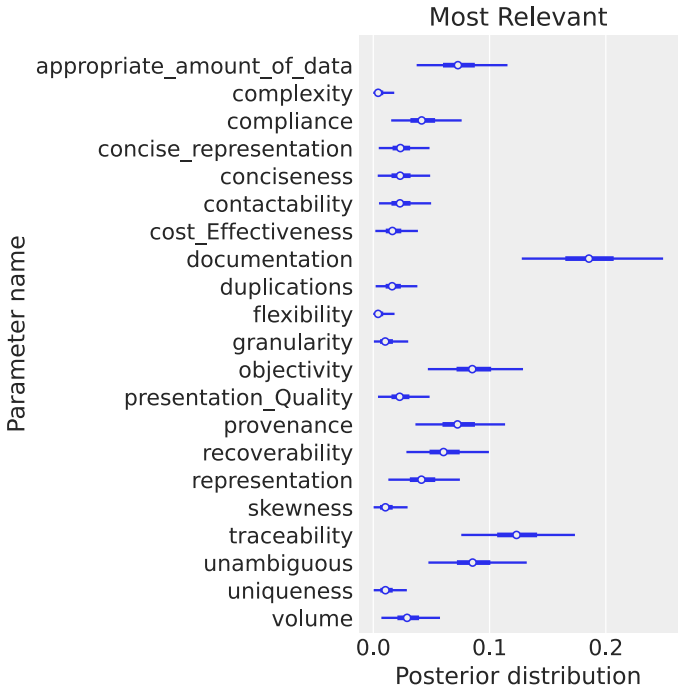

Figure 4.27: Forest plot representation of the "Most Relevant" trace.

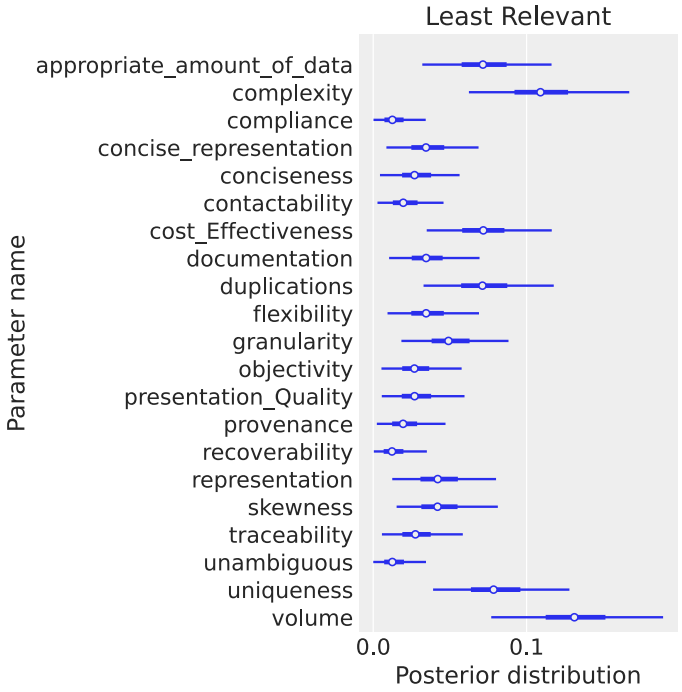

Figure 4.28: Forest plot representation of the "Least Relevant" trace.

5
Conclusion

5.1 Discussion

With the completion of the study, some noteworthy results should be discussed and associated with the research questions set at the beginning of the study.
Looking into **RQ1**, it was identified that currently there are multiple portals that provide data for environmental research and are approaching data quality reporting differently. Some noteworthy findings are the following:

Portal maturity) Not all portals are on the same level of maturity when it comes to ensuring and communicating the quality of the provided data. Some portals have defined procedures for assessing the quality of a dataset before it is accepted in their platform. In other cases, quality information was mostly absent with statements only accessible in legal documents about liability, informing consumers that they are not responsible for the data and are only available "as-is".

Portals and metadata) In many cases, portals only took responsibility for ensuring the quality of the metadata associated with the datasets. This view aligns with many of the comments collected during the survey. Thus, the data quality aspects that were investigated, were not straightforward to identity in the present documentation.

Meta-Portals) There are portals that operate as "meta-portals" allowing to access data from other portals. In these cases, similar procedures for data quality are defined for all relating portals. This shows the importance of open data communities working together in order to ensure interoperability. To this end, such portals made use of the FAIR framework [66], which defines best practices on handling data and metadata that are present in the portals.

Moving to **RQ2**, some noteworthy conclusions can be derived from the analysis of the questionnaires.

FAIR principles) Data portal maintainers value the FAIR principles highly. This can be seen by the positive trend of this role as identified in Figures 4.12, 4.17, 4.18, 4.21 compared to other roles. There is a positive association with the other roles as well, but not as noticeable. One interesting case is *interoperability*. It is viewed a little less highly than the rest. This could be attributed to different

definitions existing for different contexts, as seen in Table 4.15.
The FAIR principles were introduced in 2016 [66] and since then they have been integrated in a number of portals. Since they are valued so highly by the data portal maintainers investigated in the study, perhaps it is an opportunity for other portals as well to integrate these principles in their pipelines.

Perception of Reusability) As identified on reusability, in the Table 4.17, there appears to be a different perception on what constitutes good reusability between the data portal maintainers and dataset providers. Looking at *long term data preservation*[1] it appears that this initiative's targets are aligned with the FAIR principles and there potentially be an opportunity for combining these different guidelines to a common framework that data providers and data portal maintainers could use to communicate.

Reusability and ease of access) In Figure 4.21 it was identified that for individuals identified as dataset consumers and with data science experience, there is not a strong association with reusability. This can potentially be attributed to the ease of accessing new datasets in environmental research, thanks to the large number of open data portals, making it easier for data scientists to collect relevant datasets as needed. Also, consumers of open data are not as close to the procurement efforts as the other stakeholders, as such they are not exposed in the same way to the difficulties associated with collecting data as identified by Kahn et al. [30] and Nguyen et al. [46]. This also aligns with the view from Kim et al. [33] who criticized the "goodness-of-fit" mentality and proposing to focus on the "goodness-of-data".

Completeness and Consistency) Completeness and consistency, mentioned in Figures 4.14 and 4.15 respectively, along with their comments in Tables 4.12 and 4.13 show that these DQAs are not as impactful and that "best effort" should be the aim for these DQAs. Still, it is worth mentioning that McElreath warns about the handling of missing data [38], as discussed in Figure 2.1. Such activities require careful consideration, and information about uncertainties should be shared with consumers, preferably through documentation or metadata as mentioned in the comments.

Credibility) As identified in Figure 4.16, data portal maintainers do not value this DQA as much as other roles. A respondent identified with this role argued that relevant metadata should be available to allow stakeholders to make the decision on which data to use. Considering how important this DQA is from the perspective of dataset providers and dataset consumers, perhaps just providing metadata should not be enough. As some portals do, acceptance tests for the data should be introduced to make sure that the source of the data is credible.

Interpretability) Interpretability, discussed in Figure 4.19, is quite important for data portal maintainers with experience in data management and data science. This is reasonable as it reduces the manual labour of deciphering what information is

[1] link: Long-Term Data Preservation

present in a dataset. However, environmental research experience seems to be not affecting this DQA. This could be due to established standards surrounding environmental research. As identified by Easterbrook, the environmental research community is global and has made huge efforts to establish itself over many years [13]. This aligns with the comment from respondent R4 for a different DQA, in Table 4.11. R4 makes mention of the efforts taken with a global organization to define different aspects of quality that should be followed, especially in environmental research.

Timeliness) Timeliness, as discussed in Figure 4.23 is important for the consumers of the data but not as important for the dataset providers and data portal maintainers. Considering the importance for the users of this data, it could be argued that if it is not always possible to provide data in a timely manner, it might be preferable to provide transparent documentation as to why this is happening. It could also be seen as an opportunity to look into software engineering practices [8]. For example, in the Software Quality knowledge area there is a clear importance of understanding the trade-offs between quality, costs, and timeliness. Recognizing the need to for verification and validation practices to improve the quality of the provided product.

Value added) In Figure 4.26 about the value added DQA, it can be seen that portal maintainers with data management experience do not value this DQA highly, aligning with a view shared around the importance of metadata instead of the data. This is not the case for the dataset providers and dataset consumers, and especially for individuals with data science experience. This positive association most likely aligns with the concerns raised in the literature around "garbage-in, garbage-out" [39, 64]. This shows that there should be a shift in this mentality and mechanisms introduced to help consumers and providers identify how a dataset can best provide value.

Appropriate amount of data) As mentioned before, the appropriate amount of data is a controversial point. It appears it is not as important for data portal maintainers, while the other roles do consider it important. Considering the importance for other stakeholders, perhaps changes can be introduced to the platforms to allow providers to share with consumers what they consider to be an appropriate amount of data for the dataset they provide. An example use case can be seen in the cautionary tales shared in Section 2, where one of the reasons global warming was expected to have stopped was due to the lack of sensor coverage in certain geographical areas.

Data quantity) As mentioned before, data volume, uniqueness, and duplications are considered by many stakeholders to be least relevant. This is quite an interesting set of DQAs and perhaps align with the views presented by Tien [59] who argues that quality gaps can be covered with data quantity, especially considering the continuous improvements in performance and storage of computers.

Cost-effectiveness) An interesting case can be seen with cost-effectiveness. Most respondents and especially consumers consider it to be least relevant, showing that perhaps with open data in environmental research costs, are not as impactful as initially perceived.

De-personalization) De-personalization is an aspect that has not a DQA identified in the literature, but could be a valuable direction to look into in the future. This is especially relevant for other contexts, for example relating to the GDPR [50].

Quality rewarding) Rewarding quality is a DQA that has also not been identified in the literature. Quite a few articles raise that data quality is usually deglamoured [33], hidden and unappreciated [60]. Making efforts to reward these tasks can help swift this mindset and show the importance of data quality.

Lessons from Software Engineering

Looking into the knowledge areas of software engineering as defined by SWEBOK can help define strategies in order for these portals to become more reliable and provide a better service for the dataset consumers as well as the dataset providers [8]. When it comes to collecting requirement, there are multiple well established techniques around *quantifiable requirements*, *functional requirements* and *non-functional requirements*. As identified in the literature, these techniques can help provide solutions, especially now that environmental research is moving closer to using ML approaches to build their solutions [26].
Data portals should also make endeavours to allow for the communication of between the dataset providers and dataset consumers. By treating them as *process actors* in the *software process*, the potential for treating datasets as *process models* arises, which emphasizes that like software, data require refinement and need to be iterated upon in order to maximize their value. This would require *requirements elicitation* and *requirements analysis* in order to be enabled by the portals. Software engineering can be used in order to provide such avenues. If that is not possible, then more portals should promote acceptance tests to improve the standards of the available data.
Interestingly, the complexity surrounding data is not considered to be an important DQA from the perspective of the respondents, so looking into promoting *conceptual models* or *prototypes* using data is not a currently needed direction. However, many consider objectivity, traceability, and documentation to be of high importance. Following the themes proposed in this study, many stakeholders put emphasis in transparency and clarity and these should be promoted by the portals. In software engineering this can be achieved by looking into the *software design principles* that promote sufficiency, completeness, security and data persistence.
Dataset providers can also learn from software engineering practices. *Data structure-centered Design* aims to prompt engineers to focus on the input and output data structure and develop their solutions based on those. By coming in contact with their consumers, preferably via forums provided by the data portals, these solutions can follow existing Software Engineering approaches like requirement elicitation, requirement negotiation, requirement specification and requirement validation.
Also, currently there are technologies in software engineering for addressing issues with constructions of software. These approaches are looking into promoting robustness, completeness, and other non-functional requirements. Dataset providers can

get inspired by such approaches in order to provide metadata around Error Handling, Exception Handling, and Fault Tolerance of the information they provide.

Lessons for Software Engineering

Easterbrook in his work argues that there are many benefits of environmental research scientists over their commercial counterparts [13]. Due to the open nature of their work, they are proof of the importance of a global community, which can promote the creation of high quality software. With different disciplines involved, as well as openness about errors, practitioners can define robust standards to identify and fix issues. This attitude towards software development promotes higher work ethics. At the same time, there are organizations like the World Climate Research Programme[2], which create and provide standardized experiments that allow scientists to test their models.
This view surrounding openness and accessibility has been identified in this study as well. Especially from the perspective of the data portal maintainers, who value the FAIR principles [66]. This can also be seen in most of the comments mentioning the importance of data being accessible to all. These principles with some considerations can be applied to research software as well [37]. Software engineering can investigate the benefits of promoting such a mindset to the practitioners so that new technologies can become more easily accessible and integrated into the resulting software. Promoting transparency and learning from mistakes can also introduce higher standards of work, as practitioners can be held accountable by a larger community. This can be difficult to achieve in certain commercial cases, so different forms of transparency can be promoted, as identified by Creel [12]. Also, for tasks like ML, providing more standardized and internationally recognized datasets for evaluating performance of networks can help increase reliability of created products. This is especially relevant for safety critical tasks, like in the automotive industry, where regulations enforce compliance.
Another aspect that is identified in this study is the importance of rewarding data quality. When developing software, it is often the case that priority is given to satisfying metrics on the output of the project. However, it would be good to look into exploring the relevant aspects of the input of these products as well and considering how they can drive development and initiatives [27]. Data users value many DQAs identified in this study like timeliness and usability, which are not necessarily prioritised currently. The value of such aspects might not be immediately apparent, but by promoting them, the created products can also become more robust and flexible to the ever-changing needs of the environment where the software is released.

5.2 Threats to Validity

In the following sections, the threats to the validity of this study are presented, along with how they have been attempted to be addressed. In particular, the threats are

[2]link: World Climate Research Programme

separated in the four followed types as defined by Wohlin et al. [67].

Internal Validity

Threats to internal validity refer to the potential causal relationship between the independent variable and outcome, without the researcher's knowledge.
When considering the survey instrument, major efforts were taken in order to make it quick and concise to help reduce respondent attrition and by extension address concerns around *maturation*. The questions could be split into "blocks" dedicated to a separate DQA ordered alphabetically to reduce the potential *learning effects* in the final result. Finally, all the samples in the population were contacted with the same email in order not bias their response rate. *History* is another potential internal validity threat. For this study only one survey was sent, and the respondents were expected to respond once, so this bias should not cause concerns. The instrument was tested with pilot studies that were not included in the final output of the study, mitigating the concerns around *testing* bias.
One concern for the respondents is the potential *selection* bias introduced as the survey was based on voluntary responses. This unfortunately could not be mitigated completely except by introducing random selection of the population to be investigated. For this study this was only performed on the dataset providers due to the small number of identified populations to be contacted from the other stakeholders. Finally, for the analysis of the results from the survey, a concern is raised around ambiguity about direction of causal influence. To address this, different models were tested using the collected data to identify the causal relationships between the identified variables.
This study was performed by a single researcher, this can be quite problematic as it can introduce researcher bias [14, 52]. To handle this threat, extra care was taken to make the investigation traceable and automated wherever possible. For example, for the random selection of data providers from the data portals, a dedicated script was made. When not possible, expert suggestions and pilot studies were conducted in order to ensure the validity of the results. It is important to mention that the investigation of the portals did require judgement calls from the researcher. This is why when constructing the instrument, extra care was taken to only include DQAs that could be quantitatively identified in the literature and portals. Also, the overview of the analysis of the portals can be seen in Appendix A.4 and I welcome any scrutiny towards the selected methodology. Finally, the supervisor of this thesis, a senior researcher, regularly convened with the researcher and discussed aspects of data collection, analysis, and interpretation to mitigate this threat.

External Validity

External validity threats relate to limitations in the generalizability of the results.
One common external validity threat is the *interaction of selection and treatment* that can be introduced from a non-representative sample, especially with non-probabilistic convenience sampling as used in this study. To mitigate this threat, great emphasis was given to receiving expert input for identifying the data portals

to be investigated, but also from literature and online sources, enabling triangulation. The dataset providers from the portals were picked using a random algorithm and for the dataset consumers, individuals were contacted from universities based on their relevance to environmental research and open data. This cannot mitigate the potential bias completely and as such, snowballing was also implemented by asking respondents to share the instrument with colleagues they consider to be relevant to the study. Another aspect to consider is that even though the models used included the double roles, they were not included in the findings as they were underrepresented with only a single sample.
Considering the *interaction of setting and treatment*, the study was based on an online form, which is a standard for sharing surveys to large distributed populations. Also, the DQAs used in the instrument were taken from all available literature from three different sources. As seen in Figure 4.1 most of the included articles are from the past ten years which should make them relevant and increase generalizability due to using more relevant DQAs in the instrument. With these in mind, the created instrument should be able to generalize to other populations, but not necessarily the findings of this study.

Construct Validity

Construct validity refers to the generalizability of the result of the experiment concerning the theory it is built from.
Starting with *inadequate proportional explication of constructs*, in this study Bayesian data analysis was used and before deciding on the final model extensive prior and posterior predictive checks were performed in order to evaluate the potential output of the model. This can also help address the potential *confounding constructs* as the mock data relied on uniform distributions which were possible to check on the proposed models. For addressing *mono-operation bias*, multiple independent variables were used in the model, including the role and background aspects of the respondents. The instrument also included open-ended and close-ended questions.
Since the survey was constructed using information from the literature review, the respondents were also prompted to share any other DQAs they consider were missed in the instrument in order to help reduce the potential *mono-method bias*. Also, to reduce *mono-operational bias*, all investigated data portals were contacted to respond to the survey and dataset providers were randomly picked from the portals. However, it is important to note that since the study was cross-sectional, this bias could not be covered completely.
Looking into *interaction of different treatments*, originally this study was expected to be shared on populations that were close to contacted experts, however after identifying that a similar study was performed then these populations were not contacted in order to reduce this potential bias. For *interaction of testing and treatment*, as mentioned before, the pilot studies used for testing the instrument were not included in the final treatment sample, which can help mitigate this bias.
For the investigated DQAs the concern for *restricted generalizability across constructs* should be considered. As this study was based on environmental research and open data, it is not reasonable to make assumptions for these DQAs in other

contexts. However, the instrument was based on DQAs identified in the literature without considering this dimension and as such the instrument itself can potentially be used in other contexts. Since multiple different DQAs were identified during the analysis, in order to reduce them and understand their relationships an established thesaurus was used along with Generative AI. This helped combine similar DQAs and only keep the most prominent ones to be shared in the instrument.
Respondents might have different interpretations and terminologies of the DQAs which can lead to misunderstanding on what is being asked of them to comment on. This is especially important for the sample study phase, where the researcher is expected to use the collected data as is [57]. To address this, a definition was provided for the main DQAs identified in the first phases of the study that were included in the survey instrument. To ensure the respondents of their validity, the source of the definition was provided as well. Also, since multiple definitions exist, as mentioned before, the respondents were prompted to challenge the definition in subsequent questions. The instrument created was also reviewed by two different pilot studies with different levels of experience to ensure the validity of the instrument created.

Conclusion Validity

Conclusion validity threats are related to issues regarding drawing correct conclusions between the treatment and outcome of the experiment.
One common conclusion validity threat is the *low statistical power*, which is a relevant concern with this study considering the low number of responses. To address this, Bayesian data analysis was used in order to get as much information as possible from the collected data.
Another concern is the *reliability of treatment implementation*. In this study, the same survey was sent to the investigated populations, which were contacted via an online form. This ensures that the same treatment was applied. Similarly, *random irrelevancies in experimental setting* should also not be a concern considering that the respondents could respond whenever they considered it to be most appropriate for them, however it is not possible to guarantee that it had no effect in their responses.
Looking into *random heterogeneity* of subjects, the respondents were specifically contacted based on their relevance to the study and as can be seen by the resulting distributions there is a good spread of roles and experiences in the study. That was not the case for environmental research experience, but that is expected considering the populations contacted.

5.3 Suggestion For Future Work

This study was mainly based on non-probabilistic convenience sampling in order to identify the portals to be investigated relating to environmental research. A future endeavour can look into portals focusing on other types of research or try to perform a more general investigation on a larger number of portals to identify best practices or compare them with the ones identified in this study.

Currently, there are numerous DQAs as discovered in the literature. Many of them are similar to each other and depending on the professional circle some can even have different meanings or be split into sub-DQAs with different facets of the main DQA being prioritised. One potential future investigation could be to gather the present DQAs and aim to define a set of guidelines on DQAs specifically for environmental research in order for these ambiguities to be removed.
Finally, for this study only a few samples were available for the statistical analysis, also the final model used in this study was based on making use of the different experience levels from the respondents. There are other aspects that might be hidden in the collected dataset, so I welcome any future researcher to make use of the collected data in order to find more interesting results. Also, in the Appendix Section A the full survey can be seen along with the reasoning of the questions. This should allow the replication and adaptation of the questionnaire and by extension the collection of more data in the context of environmental research or other contexts.

5.4 Conclusion

In this study, multiple data portals were investigated and stakeholders of open data were contacted in order to gather their views surrounding data quality. Like the studies performed before by Haug [22], Kahn et al. [31] and Meng [40], this study shows that there are different aspects of data quality that are prioritized based on the needs of the task. There is a need for data portals to enable and promote transparent communication between the dataset providers and dataset consumers so that educated decisions can be taken around which data to use.
Interestingly, some data portals are closer to following software engineering principles than others. These principles can help increase the reliability and validity of these platforms and promote higher standards between the stakeholders. Many stress the need for metadata to enable consumers and providers to transparently communicate. This is an important step and it should be followed by others as well. Portals can also consider providing solutions for performing validation and verification on the data they are accepting. As some portals do, before accepting a dataset it should be validated with well-defined metrics or statistical analysis to confirm the validity of the data with the associated metadata.
In general, documentation is an important aspect for the investigated stakeholders. It is also a well-defined software engineering principle, and more portals should look into providing the platforms and guidelines for the providers and consumers to make best use of it. This can be seen in the responses of the survey as well, with documentation being considered the most relevant DQA for the respondents from the Pick-5 Section.
Dataset providers should also look into accreditation and certification in order to present their results as more trustworthy. At the same time, they should promote data portals to allow for ways to gather feedback on the datasets they provide and should try to keep their data maintained or make efforts for long-term data preservation. Even historic data have their use as identified by respondents in the survey, so providers should make an effort to maintain the data they provide and have transparent and traceable solutions for managing them.

Dataset consumers should also look into data quality in a more positive light. Looking into defining appropriate ways to reward such endeavours can help improve the quality of data introduced in the models, which can by extension help improve the quality of the created output.
This study has made an effort to investigate the gaps present in the interactions between stakeholders surrounding data quality. Without understanding these needs, decisions can be taken that are not transparently shared between them, which can lead to creating solutions that might not be reliable. Hopefully this study has shown that there have been great steps already taken to improved on this front but more can be done in order to best help the climate scientists to create accurate models of the environment.

Bibliography

[1] Arizona Department of Environmental Quality. (n.d.). Glossary of Environmental Terms. ADEQ Arizona Department of Environmental Quality.https://legacy.azdeq.gov/function/help/glossary.html

[2] Batini, C., Barone, D., Mastrella, M., Maurino, A.,& Ruffini, C. (n.d.). A Framework And A Methodology For Data Quality Assessment And Monitoring. 333-346.

[3] Bhardwaj, E., & Khaiter, P. (2024). Is the climate getting WARMer? A framework and tool for climate data comparison. Environmental Modelling & Software, 171, 105879. https://doi.org/10.1016/j.envsoft.2023.105879

[4] Beyond Unicorns: Educating, Classifying, and Certifying Business Data Scientists. (2020). Harvard Data Science Review. https://doi.org/10.1162/99608f92.55546b4a

[5] Bobrowski, M., Marré, M.,& Yankelevich, D. (n.d.). A Software Engineering View of Data Quality.

[6] Bobrowski, M., Marré, M., & Yankelevich, D. (2002). A NEAT Approach for Data Quality Assessment. In M. G. Piattini, C. Calero, & M. Genero (Eds.), Information and Database Quality (Vol. 25, pp. 135–162). Springer US. https://doi.org/10.1007/978-1-4615-0831-1_7

[7] Bosch, J., Crnkovic, I., & Olsson, H. H. (2020). Engineering AI Systems: A Research Agenda. https://doi.org/10.48550/ARXIV.2001.07522

[8] Bourque, P. & Fairley, R. E. (eds.) (2014). SWEBOK: Guide to the Software Engineering Body of Knowledge. Los Alamitos, CA: IEEE Computer Society. ISBN: 978-0-7695-5166-1

[9] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

[10] Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2022). The Effects of Data Quality on Machine Learning Performance. https://doi.org/10.48550/ARXIV.2207.14529

[11] Christen, P. (2019). Data Linkage: The Big Picture. Harvard Data Science Review. https://doi.org/10.1162/99608f92.84deb5c4

[12] Creel, K. A. (2020). Transparency in Complex Computational Systems. Philosophy of Science, 87(4), 568–589. doi:10.1086/709729

[13] Easterbrook, S. M. (2023). Computing the Climate: How We Know What We Know About Climate Change (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781316459768

[14] Eisenhardt, K. M. (1989). Building Theories from Case Study Research. The Academy of Management Review, 14(4), 532. https://doi.org/10.2307/258557

[15] Edwards, P. N. (2013). A vast machine: Computer models, climate data, and the politics of global warming (First paperback edition). The MIT Press.

[16] European Environmental Agency. (n.d.). EEA Glossary.European Environmental Agency. https://www.eea.europa.eu/help/glossary/eea-glossary

[17] Fiore, Sandro& Elia, Donatello& Santos Pires, Carlos& Mestre, Demetrio& Cappiello, Cinzia& Vitali, Monica& Andrade, Nazareno& Braz, Tarciso& Lezzi, Daniele& Moraes, Regina& Basso, Tania& Kozievitch, Nádia& Fonseca, Keiko& Antunes, Nuno& Vieira, Marco& Palazzo, Cosimo& Blanquer, Ignacio& Meira Jr, Wagner& Aloisio, Giovanni. (2019). An Integrated Big and Fast Data Analytics smart_urban Platform for Smart Urban Transportation Management. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2936941.

[18] Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian Workflow. https://doi.org/10.48550/ARXIV.2011.01808

[19] Ghazi, A. N., Petersen, K., Reddy, S. S. V. R., & Nekkanti, H. (2019). Survey Research in Software Engineering: Problems and Mitigation Strategies. IEEE Access, 7, 24703–24718. https://doi.org/10.1109/ACCESS.2018.2881041

[20] Gualo, F., Rodríguez, M., Verdugo, J., Caballero, I., & Piattini, M. (2021). Data Quality Certification using ISO/IEC 25012: Industrial Experiences. https://doi.org/10.48550/ARXIV.2102.11527

[21] Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24(2), 8–12. https://doi.org/10.1109/MIS.2009.36

[22] Haug, A. (2021). Understanding the differences across data quality classifications: A literature review and guidelines for future research. Industrial Management& Data Systems, 121(12), 2651–2671. https://doi.org/10.1108/IMDS-12-2020-0756

[23] Heyn, H.-M., Habibullah, K. M., Knauss, E., Horkoff, J., Borg, M., Knauss, A., & Li, P. J. (2023). Automotive Perception Software Development: An Empirical Investigation into Data, Annotation, and Ecosystem Challenges. https://doi.org/10.48550/ARXIV.2303.05947

[24] Heyn, H.-M., Knauss, E., Muhammad, A. P., Eriksson, O., Linder, J., Subbiah, P., Pradhan, S. K.,& Tungal, S. (2021). Requirement Engineering Challenges for AI-intense Systems Development. 2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN), 89-96. 10.1109/WAIN52551.2021.00020

[25] Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M.,& Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–16. https://doi.org/10.1145/3290605.3300830

[26] Horkoff, J. (2019). Non-Functional Requirements for Machine Learning: Challenges and New Directions. 2019 IEEE 27th International Requirements Engineering Conference (RE), 386–391. https://doi.org/10.1109/RE.2019.00050

[27] "How the Wrong KPIs Doom Digital Transformation." MIT Sloan Management Review, Mar. 2022, p. 35. Gale Academic OneFile, link.gale.com/apps/doc/A734807133/AONE?u=anon 20a05955&sid=sitemap&xid=b5c2bf88. Accessed 11 June 2024.

[28] Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters, 14(12), 124007. https://doi.org/10.1088/1748-9326/ab4e55

[29] Jansen, H. (2010). The Logic of Qualitative Survey Research and its Position in the Field of Social Research Methods. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, Vol 11, No 2 (2010): Visualising Migration and Social Division: Insights From Social Sciences and the Visual Arts. https://doi.org/10.17169/FQS-11.2.1450

[30] Kahn, B. K., Strong, D. M.,& Wang, R. Y. (2002). Information quality benchmarks: Product and service performance. Communications of the ACM, 45(4ve). https://doi.org/10.1145/505999.506007

[31] Kahn, M. G., Brown, J. S., Chun, A. T., Davidson, B. N., Meeker, D., Ryan, P. B., Schilling, L. M., Weiskopf, N. G., Williams, A. E.,& Zozus, M. N. (2015). Transparent Reporting of Data Quality in Distributed Data Networks. eGEMs (Generating Evidence& Methods to Improve Patient Outcomes), 3(1), 7. https://doi.org/10.13063/2327-9214.1052

[32] Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Enterprise Data Analysis and Visualization: An Interview Study. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2917–2926. https://doi.org/10.1109/TVCG.2012.219

[33] Kim, M., Zimmermann, T., DeLine, R., & Begel, A. (2018). Data Scientists in Software Teams: State of the Art and Challenges. IEEE Transactions on Software Engineering, 44(11), 1024–1038. https://doi.org/10.1109/TSE.2017.2754374

[34] Kitchenham, B., & Pfleeger, S. L. (2002). Principles of survey research: Part 5: populations and samples. ACM SIGSOFT Software Engineering Notes, 27(5), 17–20. https://doi.org/10.1145/571681.571686

[35] Kitchenham, B., & Charters, S. M. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. 2.

[36] Linåker, J., Sulaman, S. M., Maiani de Mello, R., & Höst, M. (2015). Guidelines for Conducting Surveys in Software Engineering. Department of Computer Science, Lund University.

[37] Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., Van De Sandt, S., Ison, J., Martinez, P. A., McQuilton, P., Valencia, A., Harrow, J., Psomopoulos, F., Gelpi, J. Ll., Chue Hong, N., Goble, C., & Capella-Gutierrez, S. (2020). Towards FAIR principles for research software. Data Science, 3(1), 37–59. https://doi.org/10.3233/DS-190026

[38] McElreath, R., (2023). Statistical Rethinking Course for Jan-Mar 2023. GitHub https://github.com/rmcelreath/stat_rethinking_2023/raw/main/slides/Lecture_18-missing_data.pdf, 21

[39] Meng, X.-L. (2021). Enhancing (Publications on) Data Quality: Deeper Data Minding and Fuller Data Confession. Journal of the Royal Statistical Society Series A: Statistics in Society, 184(4), 1161–1175. https://doi.org/10.1111/rssa.12762

[40] Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics, 12(2). https://doi.org/10.1214/18-AOAS1161SF

[41] Meng, Xiao-Li. (2012). You want me to analyze data I don't have? Are you insane?. Shanghai archives of psychiatry. 24. 297-301. 10.3969/j.issn.1002-0829.2012.05.011.

[42] Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C., & Erickson, T. (2019). How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3290605.3300356

[43] Muslah, E., & Ghoul, S. (2019). Requirements variability specification for data intensive software. https://doi.org/10.48550/ARXIV.1904.12314

[44] Nascimento, E. D. S., Ahmed, I., Oliveira, E., Palheta, M. P., Steinmacher, I., & Conte, T. (2019). Understanding Development Process of Machine Learning Systems: Challenges and Solutions. 2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), 1–6. https://doi.org/10.1109/ESEM.2019.8870157

[45] Naumann, F. (2014). Data profiling revisited. ACM SIGMOD Record, 42(4), 40–49. https://doi.org/10.1145/2590989.2590995 Neill, C. J., & Laplante, P. A. (2003). Requirements engineering: The state of the practice. IEEE Software, 20(6), 40–45. https://doi.org/10.1109/MS.2003.1241365

[46] Nguyen, N.-T., Lima, K., Skålvik, A. M., Heldal, R., Knauss, E., Oyetoyan, T. D., Pelliccione, P., & Sætre, C. (2023). Synthesized Data Quality Requirements and Roadmap for Improving Reusability of In-Situ Marine Data. 2023 IEEE 31st International Requirements Engineering Conference (RE), 65–76. https://doi.org/10.1109/RE57278.2023.00016

[47] Paullada, A., Raji, I. D., Bender, E. M., Denton, E.,& Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. Patterns, 2(11), 100336. https://doi.org/10.1016/j.patter.2021.100336

[48] Phillips, C. J. (2019). The Bases of Data. Harvard Data Science Review. https://doi.org/10.1162/99608f92.5c483119

[49] Punter, T., Ciolkowski, M., Freimut, B., & John, I. (2003). Conducting on-line surveys in software engineering. 2003 International Symposium on Empirical Software Engineering, 2003. ISESE 2003. Proceedings., 80–88. https://doi.org/10.1109/ISESE.2003.1237967

[50] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance) (4 May 2016). http://data.europa.eu/eli/reg/2016/679/oj

[51] Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., ... Bengio, Y. (2019). Tackling Climate Change with Machine Learning. https://doi.org/10.48550/ARXIV.1906.05433

[52] Runeson, P., & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering, 14(2), 131–164. https://doi.org/10.1007/s10664-008-9102-8

[53] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3411764.3445518

[54] Seaman, C. B. (1999). Qualitative methods in empirical studies of software engineering. IEEE Transactions on Software Engineering, 25(4), 557–572. https://doi.org/10.1109/32.799955

[55] Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

[56] Sovacool, B. K., Daniels, C., & AbdulRafiu, A. (2022). Science for whom? Examining the data quality, themes, and trends in 30 years of public funding for global climate change and energy research. Energy Research & Social Science, 89, 102645. https://doi.org/10.1016/j.erss.2022.102645

[57] Stol, K.-J., & Fitzgerald, B. (2018). The ABC of Software Engineering Research. ACM Transactions on Software Engineering and Methodology, 27(3), 1–51. https://doi.org/10.1145/3241743

[58] Stol, K.-J., Caglayan, B., & Fitzgerald, B. (2019). Competition-Based Crowdsourcing Software Development: A Multi-Method Study from a Customer Perspective. IEEE Transactions on Software Engineering, 45(3), 237–260. https://doi.org/10.1109/TSE.2017.2774297

[59] Tien, J. M. (2013). Big Data: Unleashing information. Journal of Systems Science and Systems Engineering, 22(2), 127–151. https://doi.org/10.1007/s11518-013-5219-4

[60] Thomer, A. K., Akmon, D., York, J. J., Tyler, A. R. B., Polasek, F., Lafia, S., Hemphill, L., & Yakel, E. (2022). The Craft and Coordination of Data Curation: Complicating Workflow Views of Data Science. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW2), 1–29. https://doi.org/10.1145/3555139

[61] Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

[62] United Nations Statistics Division. (n.d.). UN data A world of information. UN Data.https://data.un.org/Glossary.aspx?q=datamart%5bEnvironmentG%5d

[63] U.S. Environmental Protection Agency.(2023, June 21). ROE Glossary. US EPA. https://www.epa.gov/report-environment/roe-glossary

[64] Vogelsang, A., & Borg, M. (2019). Requirements Engineering for Machine Learning: Perspectives from Data Scientists. 2019 IEEE 27th Inter-

national Requirements Engineering Conference Workshops (REW), 245–251. https://doi.org/10.1109/REW.2019.00050

[65] Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

[66] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

[67] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). Experimentation in Software Engineering. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-29044-2

[68] Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. Neurocomputing, 237, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026

A

Appendix 1

Figure A.1: Snowball trail

Quality Aspect	Definition	Source DOI/APA
Accessibility	To be Accessible: A1. (meta)data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary A2. metadata are accessible, even when the data are no longer available	https://doi.org/10.1145/3290605.3300830
Accuracy	The degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use	https://doi.org/10.1109/RE57278.2023.00016
Appropriate amount of Data	the extent to which the volume of information is appropriate for the task at hand	https://doi.org/10.1007/s40595-014-0030-9
Completeness	The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.	https://doi.org/10.1109/RE57278.2023.00016
Complexity	The sample complexity of a machine learning algorithm represents the number of training-samples that it needs in order to successfully learn a target function	https://doi.org/10.1145/3361242.3361260
Compliance	-	-
Concise Representation	The extent to which data is compactly represented	https://doi.org/10.1007/s40595-014-0030-9
Conciseness	The real world is represented with the minimum information required for the goal it is used for.	Bobrowski, M., Marré, M., & Yankelevich, D. (n.d.). A Software Engineering View of Data Quality.
Consistency	The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data regarding one entity and across similar data for comparable entities.	https://doi.org/10.1109/RE57278.2023.00016
Contactability	The extent to which the data publisher provide useful contact information	https://doi.org/10.1145/3209415.3209474
Cost Effectiveness	-	-
Credibility	The degree to which data has attributes that are regarded as true and believable by users in a specific context of use. Credibility includes the concept of authenticity (the truthfulness of origins, attributions, and commitments).	https://doi.org/10.1109/RE57278.2023.00016
Documentation	-	-
Duplications	duplicates or redundancy	https://doi.org/10.1145/3529190.3529222
Findability	To be Findable: F1. (meta)data are assigned a globally unique and persistent identifier F2. data are described with rich metadata (defined by R1 below) F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource	https://doi.org/10.1145/3290605.3300830
Flexibility	-	-
Granularity	-	-
Interoperability	To be Interoperable: I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles I3. (meta)data include qualified references to other (meta)data	https://doi.org/10.1145/3290605.3300830
Interpretability	the ease with which the user may understand and properly use and analyse the data	https://doi.org/10.1007/s11192-016-1914-5
Objectivity	Data is objective, i.e., it does not depend on the judgment, interpretation, or evaluation of people.	Bobrowski, M., Marré, M., & Yankelevich, D. (n.d.). A Software Engineering View of Data Quality.
Presentation Quality	-	-
Provenance	-	-
Recoverability	-	-
Relevance	Every piece of information stored is important in order to get a representation of the real world.	Bobrowski, M., Marré, M., & Yankelevich, D. (n.d.). A Software Engineering View of Data Quality.
Representation	Refers to the format and representation of data that should be the same in the dataset.	https://doi.org/10.1109/REW.2019.00050
Reusability	To be Reusable: R1. meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with detailed provenance R1.3. (meta)data meet domain-relevant community standards	https://doi.org/10.1145/3290605.3300830
Security	The extent to which access to data is restricted appropriately to maintain its security	https://doi.org/10.1007/s40595-014-0030-9
Skewness	Computes the distribution deviation of the observed data from a reference distribution.	https://doi.org/10.1145/3593434.3593445
Timeliness	the extent to which the information is sufficiently up to date for the task at hand	https://doi.org/10.1145/505999.506007
Traceability	-	-
Unambiguous	Each piece of data has a unique meaning.	Bobrowski, M., Marré, M., & Yankelevich, D. (n.d.). A Software Engineering View of Data Quality.
Understandability	the extent to which is easily comprehended	https://doi.org/10.1145/505999.506007
Uniqueness	Redundant data does not provide additional information to the ML-model for the training process. Thus, de-duplication is a common step in ML pipelines to avoid over tting	https://doi.org/10.48550/ARXIV.2207.14529
Usability	The stored information is usable by the organization.	Bobrowski, M., Marré, M., & Yankelevich, D. (n.d.). A Software Engineering View of Data Quality.
Value-Added	data give you a competitive edge, data add value to your operation	https://doi.org/10.1080/07421222.1996.11518099
Volume	-	-

Figure A.2: DQAs along with definitions from literature. Bold aspects were included in the full evaluation during the survey.

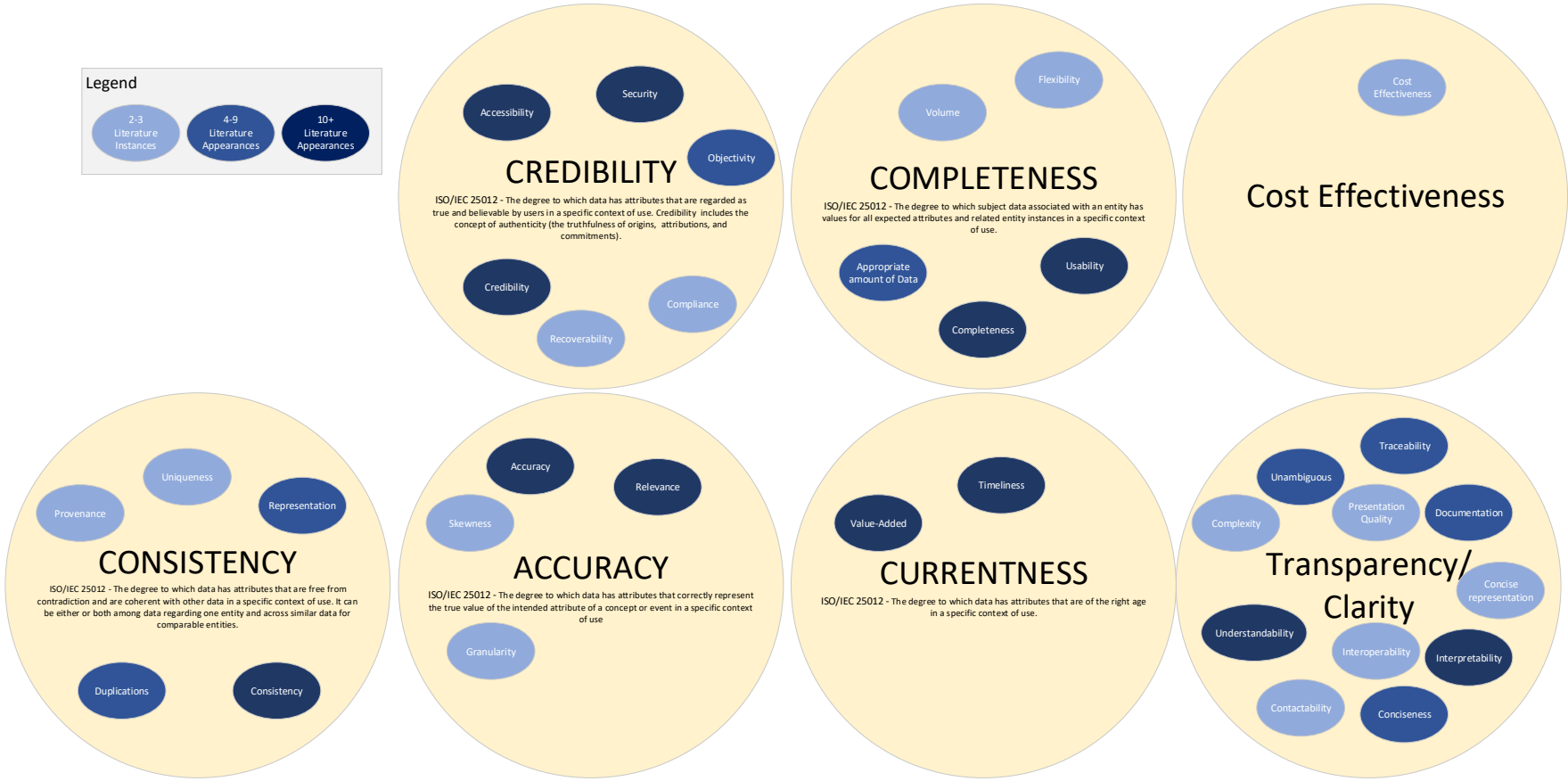

Figure A.3: Output of Thematic Analysis after the review from the expert.

Host name	Site	Harmonic Mean	ACCURACY	COMPLETENESS	CONSISTENCY	CURRENTNESS	COST-EFFECTIVENESS	CREDIBILITY	TRANSPARENCY	FAIR	FINDABLE	ACCESSIBLE	INTEROPERABLE	REUSABLE
OpenAQ	https://explore.openaq.org/	9E+99	O	O	E	E	X	O	E	X	X	X	X	X
SMHI	https://hypeweb.smhi.se/	9E+99	E	E	X	E	X	E	E	X	X	X	X	X
ENES	https://enesdataspace.vm.fedcloud.eu/home.html	9E+99	-	-	-	-	-	-	-	-	-	-	-	-
OceanDataFactory	https://oceandatafactory.se/	9E+99	X	X	X	X	X	X	X	X	X	X	X	X
Climate Watch Data	https://www.climatewatchdata.org/	9E+99	O	X	X	E	X	E	E	X	X	X	X	X
ICOS	https://data.icos-cp.eu/	1152524	V	V	V	E	V	V	V	V	V	V	V	V
European data	https://data.europa.eu/en	572603.9914	X	X	X	X	X	E	E	E	E	E	E	E
Zenodo	https://zenodo.org/	155220.5714	O	O	O	O	X	O	X	O	O	O	O	O
PANGEA	https://www.pangaea.de/	123085.7702	X	E	E	X	X	E	V	E	E	E	E	E
EOSC	https://eosc-portal.eu/	91463.23238	X	X	X	E	X	V	X	V	V	V	E	V
WDC Climate	https://www.wdc-climate.de/ui/	42528.53894	E	E	E	E	X	E	E	E	E	E	E	E
Unep Data	https://www.unep.org/data-resources	20213	X	X	X	V	X	V	X	X	X	X	X	X
EDI Data Portal	https://portal.edirepository.org/nis/home.jsp	11831.92538	O	O	E	E	X	E	E	X	X	X	X	X
ICPSR	https://www.icpsr.umich.edu/web/pages/index.html	9802.911863	V	V	E	E	V	E	E	V	V	V	V	V
Nasa EarthData	https://www.earthdata.nasa.gov/topics	9263	E	E	E	E	V	E	E	V	V	V	V	V
CEDA	https://catalogue.ceda.ac.uk/	8990	O	O	E	X	X	V	V	X	V	V	V	V
Harvard Dataverse	https://dataverse.harvard.edu/	8349.392835	O	O	O	O	X	O	O	V	V	V	V	V
EPA	https://www.epa.gov/data/where-find-epa-data	5855	V	V	V	X	V	V	V	X	V	X	X	V
Amazon	https://aws.amazon.com/opendata/	2389.055886	X	X	X	X	X	X	X	X	X	X	X	X
eLTER	https://catalogue.lter-europe.net/elter/documents	1467.439385	-	-	-	-	-	-	-	-	-	-	-	-
NSIDC	https://nsidc.org/data	1358	X	X	V	E	V	E	E	V	V	V	V	V
IPCC	https://ipcc-browser.ipcc-data.org/browser/search	867.5937018	V	X	V	E	X	E	E	V	V	V	V	V
voice of the ocean	https://voiceoftheocean.org/	479	O	O	E	E	X	X	V	X	X	X	X	X
SIOS	https://sios-svalbard.org/Data	353.8761585	O	O	E	E	X	E	E	E	E	E	E	E

LEGEND	Description
X	Missing DQA
O	Known DQA but not handled
V	Known DQA handled in statement
E	Known DQA handled in practice
-	Inaccesible Information

Lack Of Info	9	11	9	9	20	5	7	13	11	11	12	11
Present Info	16	14	16	16	5	20	18	12	14	14	13	14
Total	25	25	25	25	25	25	25	25	25	25	25	25

Figure A.4: Host evaluation descriptive results, interesting note on the CEDA portal where they seem to comply with the FAIR principles but are not actively part of the FAIR initiative. This visualisation is based on the judgements of the researcher and should not be considered factual but can be used to guide future investigations around DQAs in Portals.

Original DQA	Converted DQA
Access security	Access security
Accessibility	Accessibility
Accuracy	Accuracy
Age	Age
Alignment	Alignment
Amount of data	Appropriate amount of
Appropriate amount of	Appropriate amount of
Appropriateness	Relevance
Auditability	Traceability
Authorization	Accessibility
Availability	Accessibility
Believability	Credibility
Biased	Biased
Clarity	Understandability
Coherence	Consistency
Comment	Comment
Comparability	Comparability
Completeness	Completeness
Complexity	Complexity
Compliance	Compliance
Comprehensibility	Understandability
Concise	Concise
Conciseness	Conciseness
Confidentiality	Security
Consistency	Consistency
Contact	Contactability
Contactability	Contactability
Correctness	Correctness
Cost	Cost Effectiveness
Cost Effectiveness	Cost Effectiveness
Credibility	Credibility
Currency	Timeliness
Currentness	Timeliness

Original DQA	Converted DQA
Definition	Consistency
Distinctiveness	Unambiguous
Diversity	Complexity
Documentation	Documentation
Duplications	Duplications
Ease of Manipulation	Usability
Ease of Operation	Usability
Ease of	Usability
Ease of Use	Accessibility
Efficiency	Cost Effectiveness
Elasticity	Flexibility
Error rate	Error rate
Feature Accuracy	Feature Accuracy
Findability	Findability
Fitness	Relevance
Fitness for Use	Relevance
Flexibility	Flexibility
Free of Error	Accuracy
Free-Of-Error	Accuracy
Frequency	Frequency of
Frequency of	Frequency of
Geographic	Completeness
Geographical	Geographical
Granularity	Granularity
Homogenous	Homogenous
Institutional Mandate	Compliance
Interoperability	Interoperability
Interpretability	Interpretability
Label Quality	Label Quality
Learnability	Understandability
Lineage	Lineage
Maintainability	Maintainability
Meaningful	Meaningful

Original DQA	Converted DQA
Metadata	Documentation
Metadata Update	Documentation
Minimality	Completeness
Navigation	Navigation
Noise	Noise
Objectivity	Objectivity
Ontology	Ontology
Openness	Accessibility
Opportunity	Opportunity
Outdated Date	Timeliness
Pertinence	Relevance
Popularity	Popularity
Portability	Interoperability
Precision	Accuracy
Presentation	Consistency
Presentation Quality	Presentation Quality
Privacy	Security
Provenance	Provenance
Quality Management	Compliance
Readability	Understandability
Recoverability	Recoverability
Reference Period	Reference Period
Release Policy	Release Policy
Relevance	Relevance
Relevancy	Relevance
Reliability	Credibility
representation	Representation
Representational	Consistency
Representative	Completeness
Representativeness	Completeness
Reproducibility	Consistency
Reputation	Credibility
Resiliency	Resiliency

Original DQA	Converted DQA
Retrievability	Retrievability
Reusability	Reusability
Safety	Security
Scalability	Scalability
Security	Security
Skewness	Skewness
Statistical	Statistical
Statistical Processing	Statistical Processing
Structure	Structure
Synchronization	Synchronization
Target Accuracy	Accuracy
Target Class Balance	Consistency
Technological	Completeness
Technology Coverage	Technology Coverage
Temporal Coverage	Temporal Coverage
Timeliness	Timeliness
Time-related	Completeness
Traceability	Traceability
Trust	Popularity
Unambiguous	Unambiguous
Uncertainty	Uncertainty
Understandability	Understandability
Uniqueness	Uniqueness
Unit of Measure	Unit of Measure
Up-to-date	Timeliness
Usability	Usability
Usefulness	Relevance
Validity	Relevance
Value Added	Value-Added
Value-Added	Value-Added
Variety	Complexity
Verifiability	Verifiability
Volume	Volume

Figure A.5: Updated DQAs after clean up. Ordered by name

Survey Instrument

Section - Tell us some things about yourself

Q1: Please select the category that best describes your role in Open Data

Type: Categorical, mandatory

Options: 1) **Dataset Provider** - You work with collecting and preparing data to be used by Dataset Consumers. 2) **Dataset Hosting Portal Maintainer** - You work with providing the platforms to allow Data Consumers to access the datasets provided by the Data Providers. 3) **Dataset Consumer** - You work with building solutions based on datasets collected from different data portals.

Reasoning: Considering that the study aims to identify the view of different DQAs from the perspective of different stakeholder groups identifying their role is an important bedrock for the of the analysis.

Q2: How many years of experience do you have in Data Management such as Databases?

Type: Ordered Categorical, mandatory
Options: 1) No experience. 2) Less than 1 year. 3) 1-2 years 4) 3-5 years 5) 5+ years
Reasoning: Management of Data is important for any stakeholder working with them. Using this information can provide some useful insights about views on different DQAs.

Q3: How many years of experience do you have in Data Science?

Type: Ordered Categorical, mandatory
Options: 1) No experience. 2) Less than 1 year. 3) 1-2 years 4) 3-5 years 5) 5+ years
Reasoning: Data Science is closely connected with data and their usage. The level of experience of the respondent can provide useful insights about their familiarity with their caveats.

Q4: How many years have you been involved in Environmental Research?

Type: Ordered Categorical, mandatory
Options: 1) No experience. 2) Less than 1 year. 3) 1-2 years 4) 3-5 years 5) 5+ years
Reasoning: Environmental Research is a core component of the investigations in this thesis as most of the respondents were contacted due to their relationship to Open data portals relating to Environmental Research. Looking into their experience can provide useful insights about the view of different DQAs.

Q5: If you are interested in receiving a short summary of the results on this study please share your email with us.

Supplementary Text: Alternatively, you can send a request for the summary to this email: geoeft@chalmers.se
Type: Text, Optional
Reasoning: In order to thank the respondents for their time and also share the the knowledge created by the study the respondents can optionally add their email so that they can be contacted with a small summary of the survey.

Section - Tells us your thoughts around the following Data Quality Aspects surrounding Open Data

Q6: How important is it to you that data are accessible?

Supplementary Text:
Similar Aspects: Availability, Openness, Ease Of Use, Access Security, Authorization
Definition: To be Accessible:

- (meta)data are retrievable by their identifier using a standardized communications protocol
- the protocol is open, free, and universally implementable
- the protocol allows for an authentication and authorization procedure, where necessary
- metadata are accessible, even when the data are no longer available

source of definition: https://doi.org/10.1038/sdata.2016.18
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q7: If you disagree with the suggested definition or similar aspects, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q8: How important is it to you that data are accurate?

Supplementary Text:
Similar Aspects: Correctness, Precision, Free-Of-Error, Target Accuracy
Definition: The degree to which data has attributes that correctly represent the true value of the intended. Attribute of a concept or event in a specific context of use.
source of definition: https://doi.org/10.1145/505999.506007
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q9: If you disagree with the suggested definition or similar aspects, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q10: How important is it to you that data are complete?

Supplementary Text:
Similar Aspects: Minimality, Representativeness, Geographic representativeness, Technological representativeness, Time-related representativeness
Definition: The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.
source of definition: https://doi.org/10.1109/RE57278.2023.00016

Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q11: If you disagree with the suggested definition or similar aspects, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q12: How important is it to you that data are consistent?

Supplementary Text:
Similar Aspects: Coherence, Consistent Representation, Representational Consistency, Target Class Balance
Definition: The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data regarding one entity and across similar data for comparable entities.
source of definition: https://doi.org/10.1109/RE57278.2023.00016
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q13: If you disagree with the suggested definition or similar aspects, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q14: How important is it to you that data are credible?

Supplementary Text:
Similar Aspects: Believability, Reliability, Reputation, Trust
Definition: The degree to which data has attributes that are regarded as true and believable by users in a specific context of use. Credibility includes the concept of authenticity (the truthfulness of origins, attributions, and commitments).
source of definition: https://doi.org/10.1109/RE57278.2023.00016
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q15: If you disagree with the suggested definition or similar aspects, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q16: How important is it to you that data are findable?

Supplementary Text:
Definition: To be Findable:

- (meta)data are assigned a globally unique and persistent identifier
- data are described with rich metadata
- metadata clearly and explicitly include the identifier of the data it describes
- (meta)data are registered or indexed in a searchable resource

source of definition: https://doi.org/10.1038/sdata.2016.18
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q17: If you disagree with the suggested definition, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q18: How important is it to you that data are findable?

Supplementary Text:
Similar Aspects: Portability
Definition: To be Interoperable:

- (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- (meta)data use vocabularies that follow FAIR principles
- (meta)data include qualified references to other (meta)data

source of definition: https://doi.org/10.1038/sdata.2016.18
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q19: If you disagree with the suggested definition or similar aspect, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q20: How important is it to you that data are interpretable?

Supplementary Text:
Definition: The extent to which information is in appropriate languages, symbols, and units, and the definitions are clear.
source of definition: https://doi.org/10.1145/505999.506007
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q21: If you disagree with the suggested definition, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q22: How important is it to you that data are relevant?

Supplementary Text:
Similar Aspects: Appropriateness, Fitness, Pertinence, Usefulness, Validity
Definition: Every piece of information stored is important in order to get a representation of the real world.
source of definition: Bobrowski, M., Marré, M., & Yankelevich, D. (n.d.). A Software Engineering View of Data Quality.
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q23: If you disagree with the suggested definition or similar aspects, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q24: How important is it to you that data are reusable?

Supplementary Text:
Definition: To be Reusable:

- meta(data) are richly described with a plurality of accurate and relevant attributes.
- (meta)data are released with a clear and accessible data usage license
- (meta)data are associated with detailed provenance
- (meta)data meet domain-relevant community standards

source of definition: https://doi.org/10.1038/sdata.2016.18
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q25: If you disagree with the suggested definition, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q26: How important is it to you that data are secure?

Supplementary Text:
Similar Aspect: Safety
Definition: Data cannot be accessed by competitors, data are of a proprietary nature, access to data can be restricted, secure.
source of definition: https://doi.org/10.1080/07421222.1996.11518099
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q27: If you disagree with the suggested definition or similar aspect, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q28: How important is it to you that data are timely?

Supplementary Text:
Similar Aspects: Currentness, Frequency

Definition: The extent to which the information is sufficiently up to date for the task at hand.
source of definition: https://doi.org/10.1145/505999.506007
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q29: If you disagree with the suggested definition or similar aspects, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q30: How important is it to you that data are understandable?

Supplementary Text:
Similar Aspects: Clarity, Comprehensibility, Learnability, Readability
Definition: The extent to which is easily comprehended.
source of definition: https://doi.org/10.1145/505999.506007
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q31: If you disagree with the suggested definition or similar aspects, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q32: How important is it to you that data are usable?

Supplementary Text:
Similar Aspects: Ease of Manipulation, Ease of Operation, Ease of Understanding
Definition: The stored information is usable by the organization.
source of definition: Bobrowski, M., Marré, M., & Yankelevich, D. (n.d.). A Software Engineering View of Data Quality.
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q33: If you disagree with the suggested definition or similar aspects, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q34: How important is it to you that data add value?

Supplementary Text:
Definition: Data give you a competitive edge, data add value to your operation
source of definition: https://doi.org/10.1080/07421222.1996.11518099
Type: Ordered Categorical, Mandatory
Options: Likert Scale (1-6), presented as stars
Reasoning: This DQA was picked as it appeared many times in the literature and data portals.

Q35: If you disagree with the suggested definition, please provide a brief explanation.

Type: Text, Optional
Reasoning: In the literature a common mention is that different DQAs have different meaning for different individuals. This question is to allow the respondents to challenge the presented definition and share their view.

Q36: From the following Data Quality Aspects, please pick up to five quality aspects that you consider to be most relevant?

Type: Categorical, mandatory
Options: 1) Appropriate amount of data 2) Complexity 3) Compliance 4) Concise representation 5) Conciseness 6) Contactability 7) Cost Effectiveness 8) Documentation 9) Duplications 10) Flexibility 11) Granularity 12) Objectivity 13) Presentation Quality 14) Provenance 15) Recoverability 16) Representation 17) Skewness 18) Traceability 19) Unambiguous 20) Uniqueness 21) Volume
Reasoning: As it would be good to avoid overburdening the respondents with too many questions, the DQAs that appeared at least a minimum number of times in the literature and Open data portals were included. This was done in order to investigate which of the found DQAs are most important but were unfortunately not included in the Likert scales,

Q37: From the following Data Quality Aspects, please pick up to five quality aspects that you consider to be least relevant?

Type: Categorical, mandatory
Options: 1) Appropriate amount of data 2) Complexity 3) Compliance 4) Concise representation 5) Conciseness 6) Contactability 7) Cost Effectiveness 8) Documentation 9) Duplications 10) Flexibility 11) Granularity 12) Objectivity 13) Presen-

tation Quality 14) Provenance 15) Recoverability 16) Representation 17) Skewness 18) Traceability 19) Unambiguous 20) Uniqueness 21) Volume

Reasoning: As it would be good to avoid overburdening the respondents with too many questions, the DQAs that appeared at least a minimum number of times in the literature and Open data portals were included. This was done in order to investigate which of the found DQAs are most important but were unfortunately not included in the Likert scales.

Boxplots

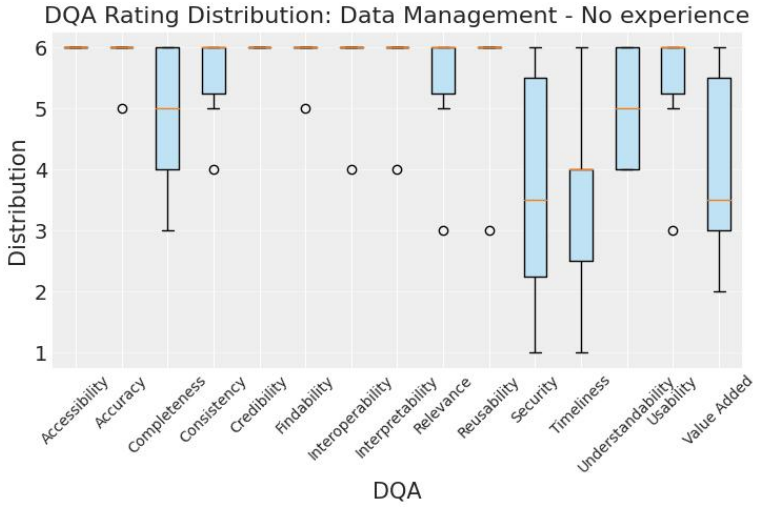

Figure A.6: Distribution of the data for Data Management - No Experience

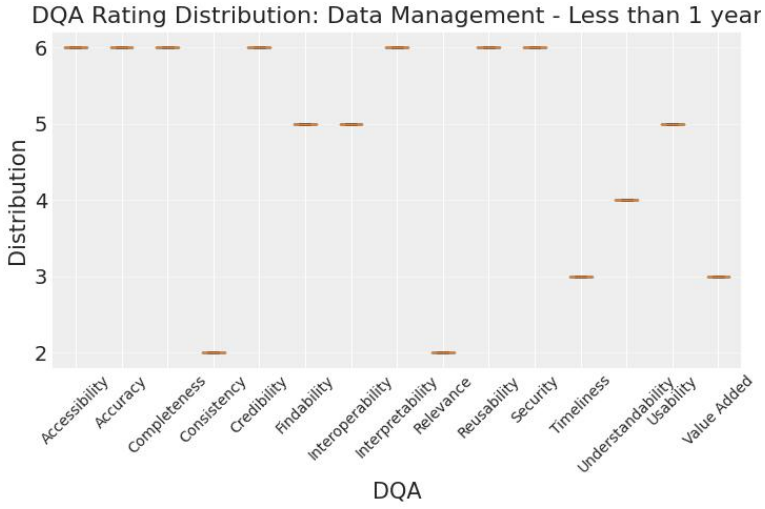

Figure A.7: Distribution of the data for Data Management - Less than 1 year

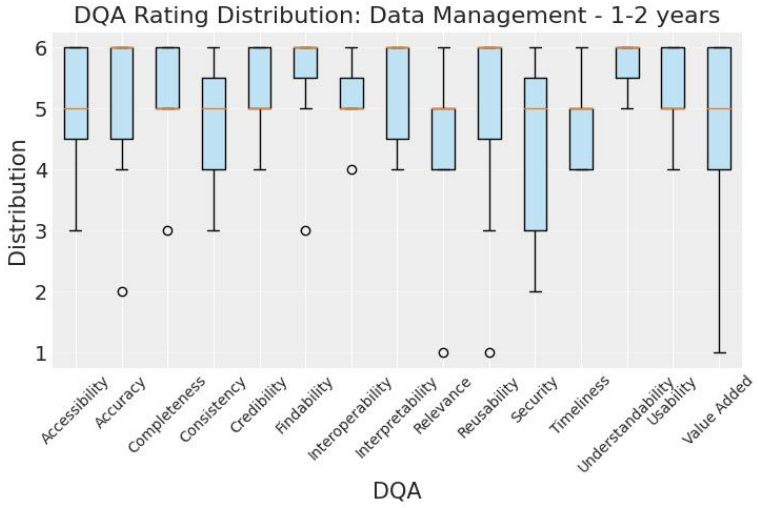

Figure A.8: Distribution of the data for Data Management - 1-2 years

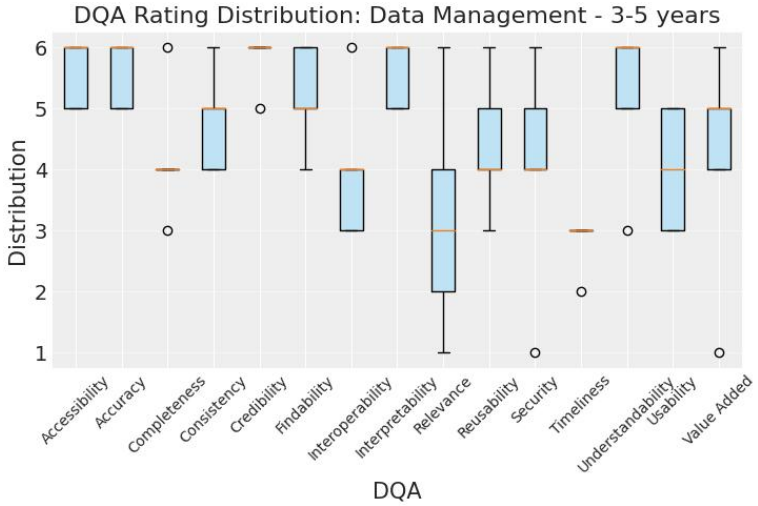

Figure A.9: Distribution of the data for Data Management - 3-5 years

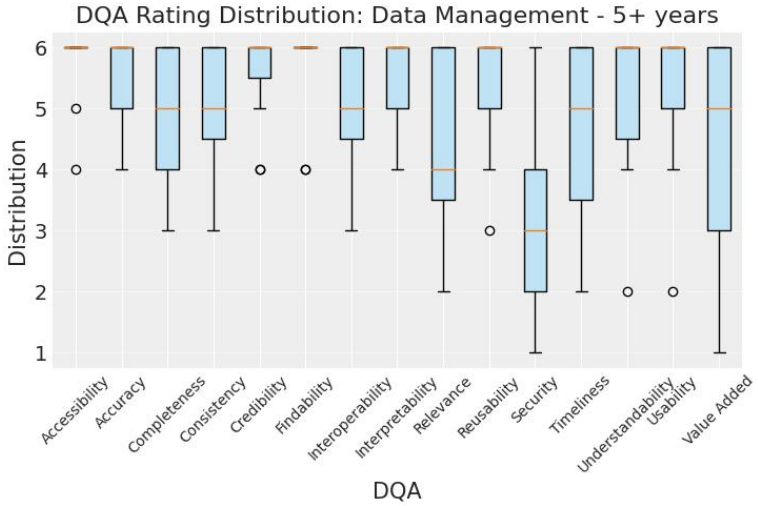

Figure A.10: Distribution of the data for Data Management - 5+ years

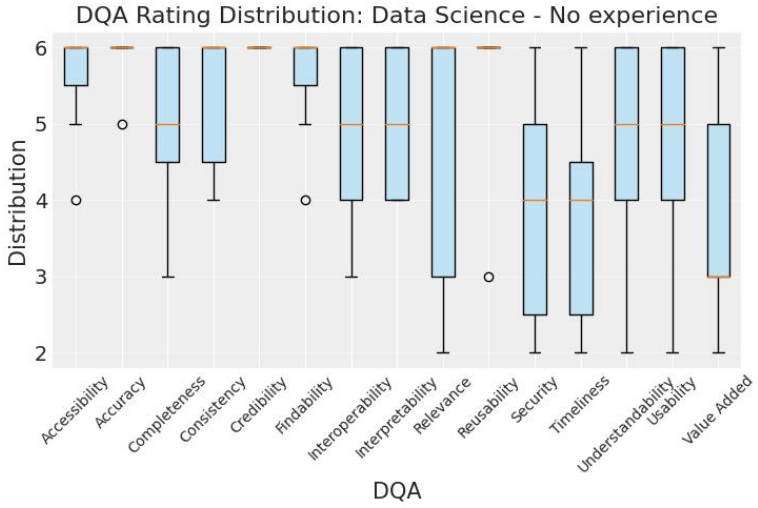

Figure A.11: Distribution of the data for Data Science - No Experience

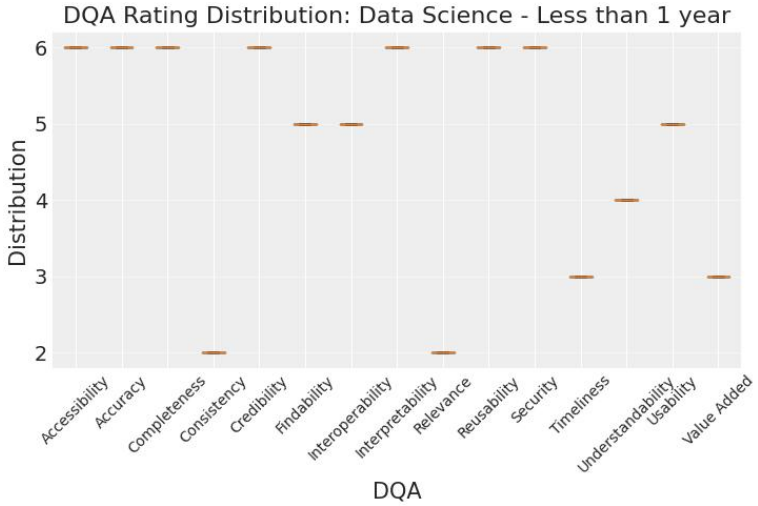

Figure A.12: Distribution of the data for Data Science - Less than 1 year

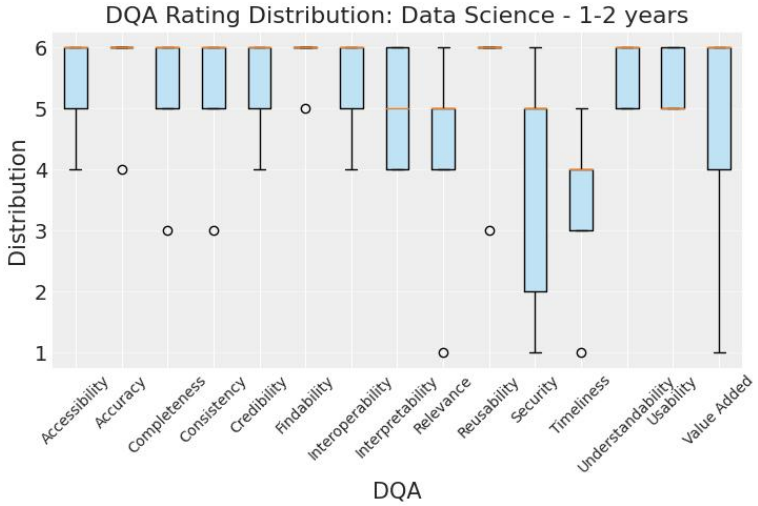

Figure A.13: Distribution of the data for Data Science - 1-2 years

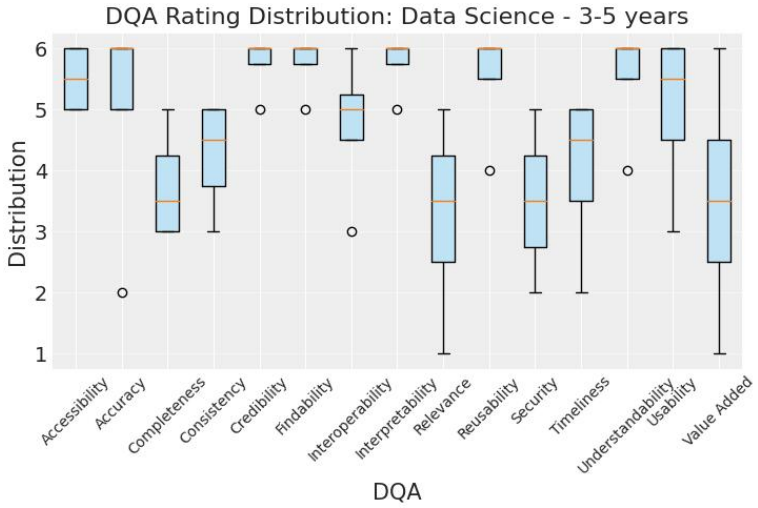

Figure A.14: Distribution of the data for Data Science - 3-5 years

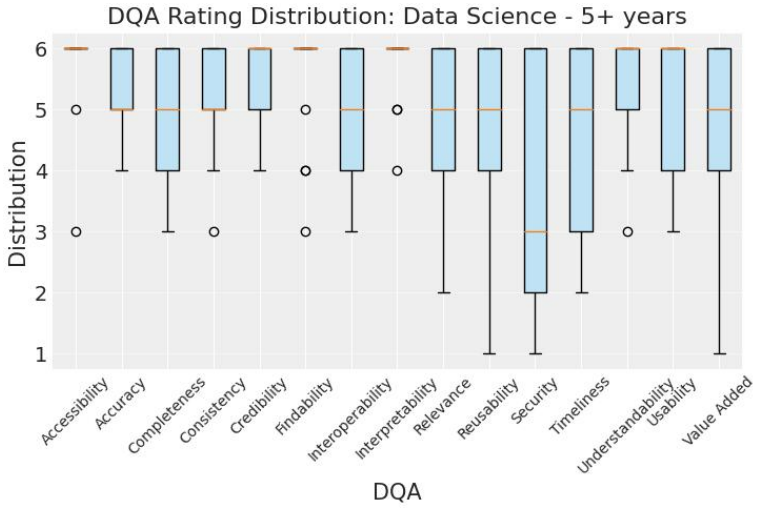

Figure A.15: Distribution of the data for Data Science - 5+ years

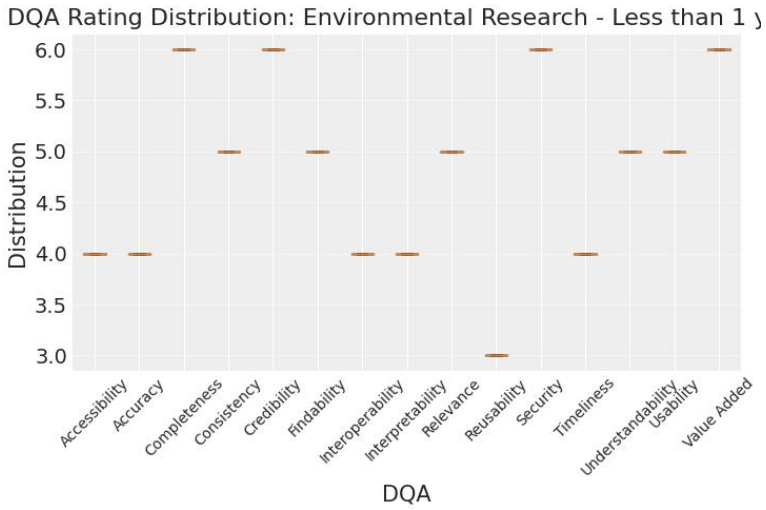

Figure A.16: Distribution of the data for Environmental Research - Less than 1 year

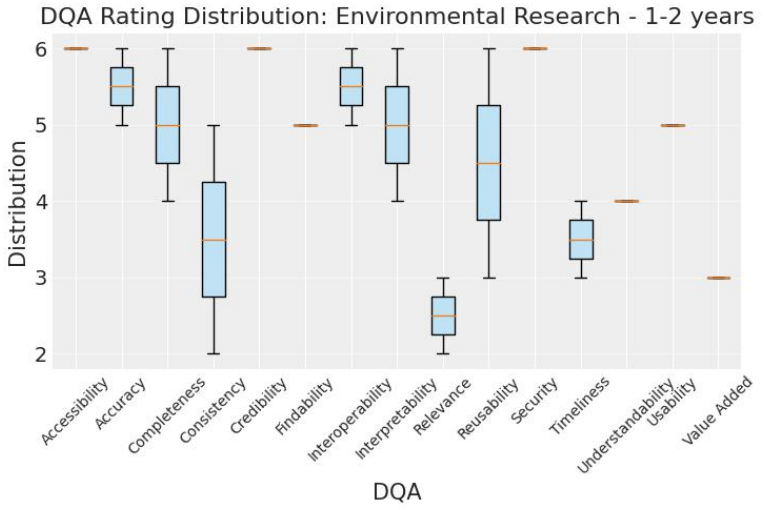

Figure A.17: Distribution of the data for Environmental Research - 1-2 years

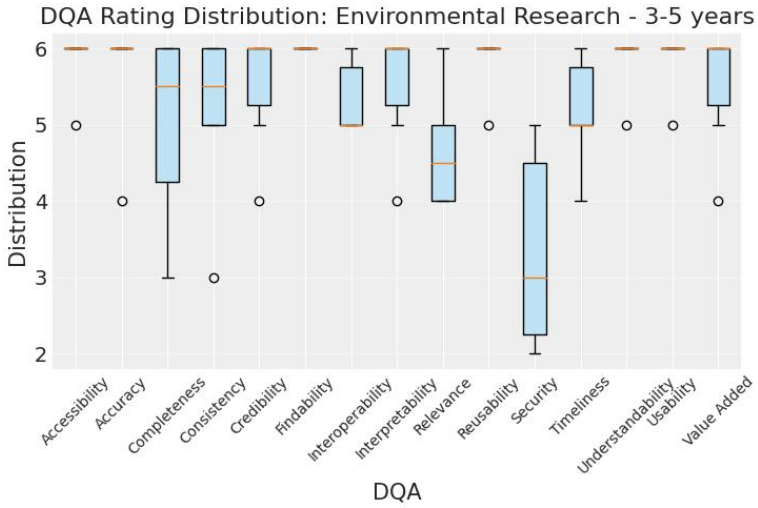

Figure A.18: Distribution of the data for Environmental Research - 3-5 years

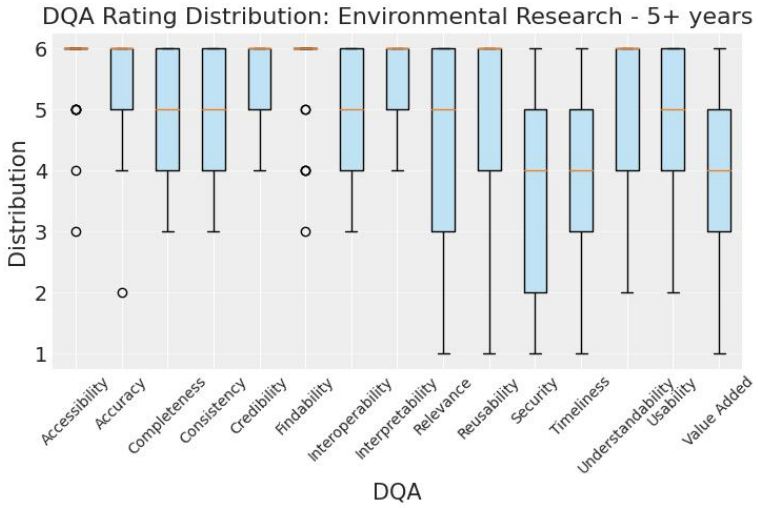

Figure A.19: Distribution of the data for Environmental Research - 5+ years

Ratings: Posterior Predictive Checks Plots

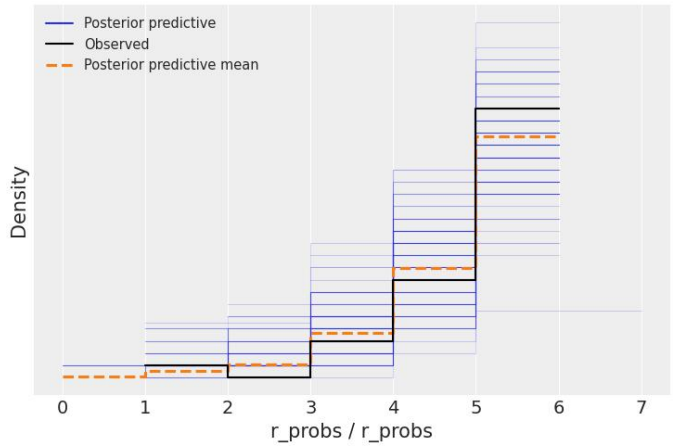

Figure A.20: Posterior predictive check of the Accuracy model.

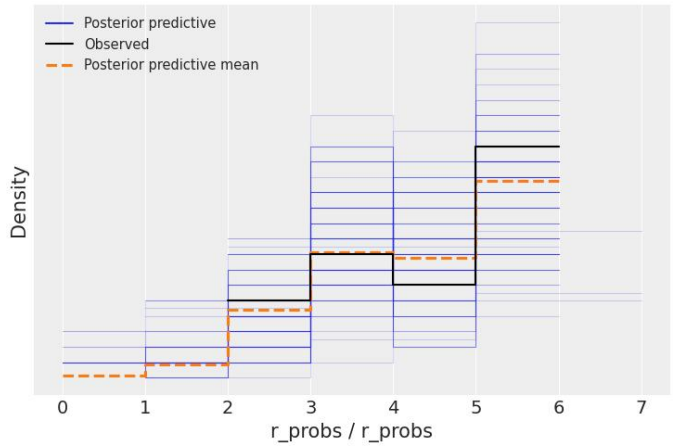

Figure A.21: Posterior predictive check of the Completeness model.

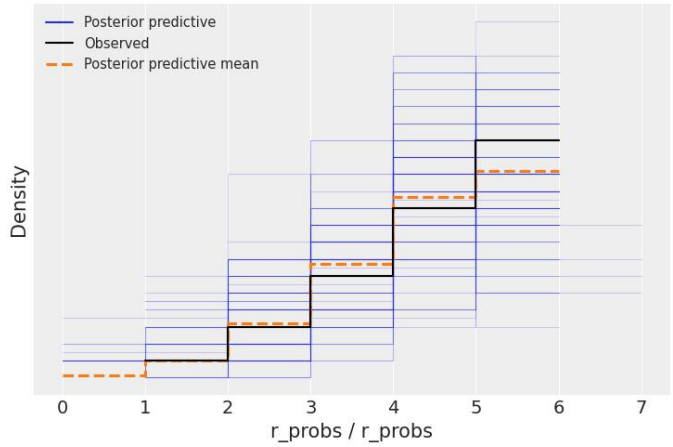

Figure A.22: Posterior predictive check of the Consistency model.

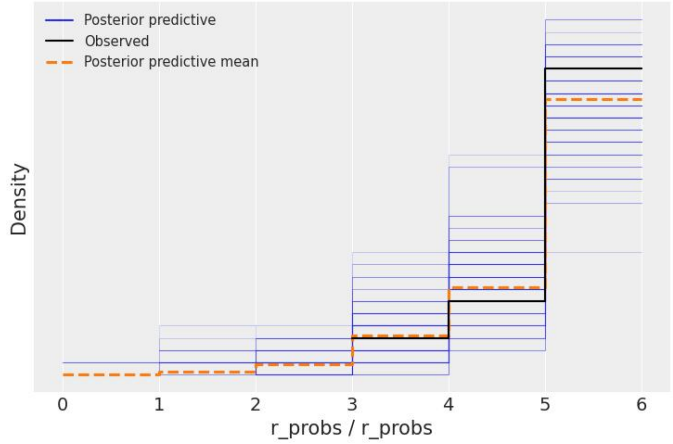

Figure A.23: Posterior predictive check of the Credibility model.

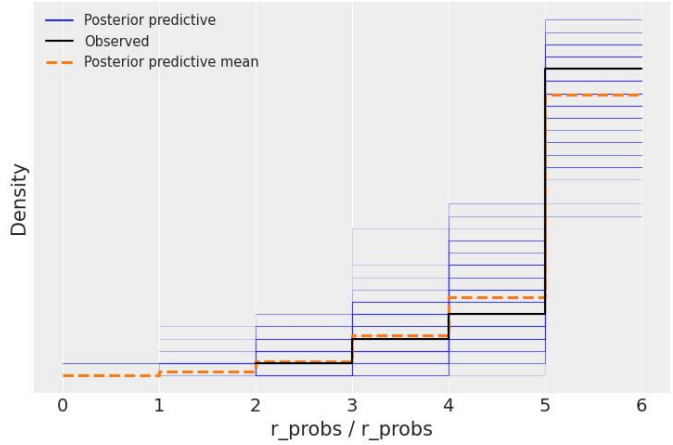

Figure A.24: Posterior predictive check of the Findability model.

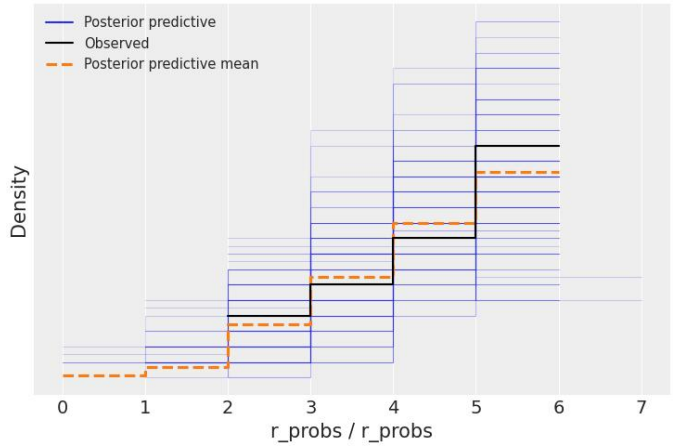

Figure A.25: Posterior predictive check of the Interoperability model.

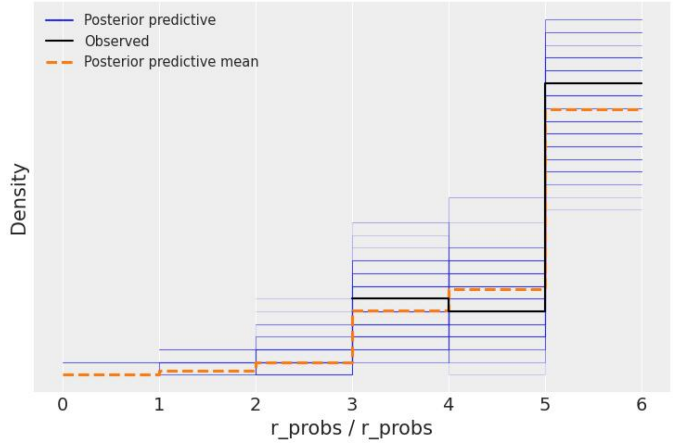

Figure A.26: Posterior predictive check of the Interpretability model.

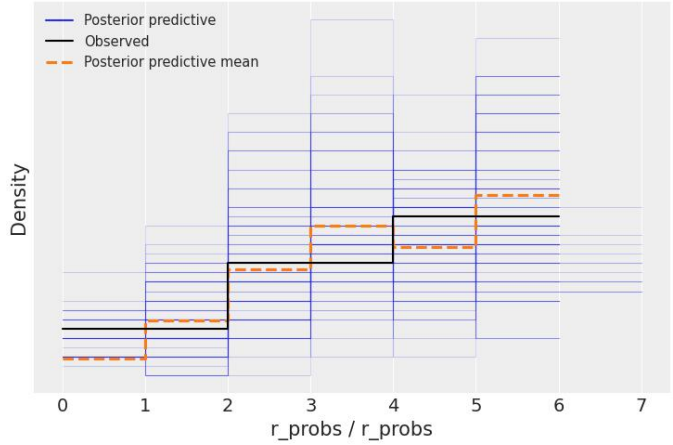

Figure A.27: Posterior predictive check of the Relevance model.

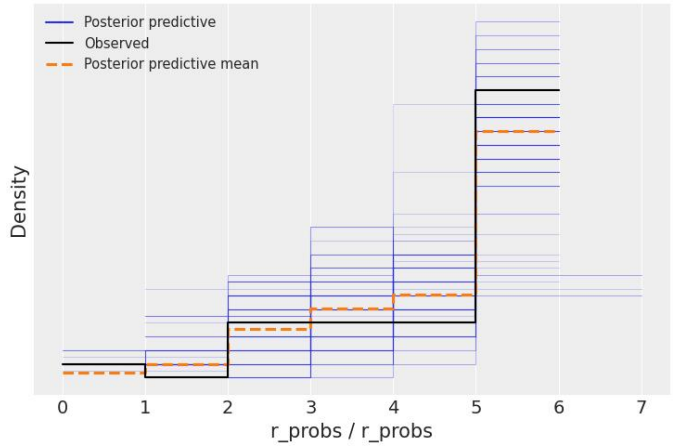

Figure A.28: Posterior predictive check of the Reusability model.

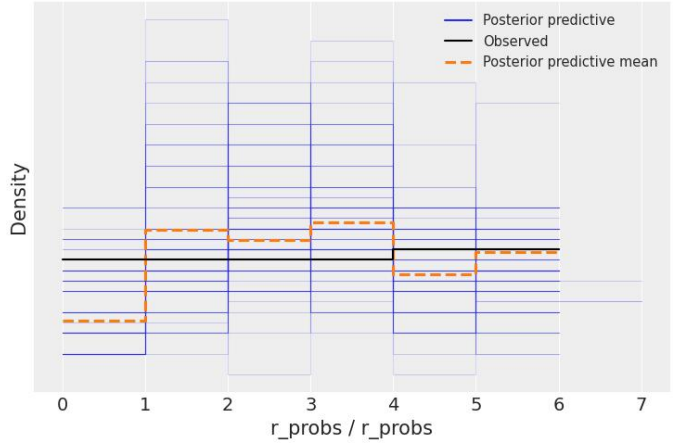

Figure A.29: Posterior predictive check of the Security model.

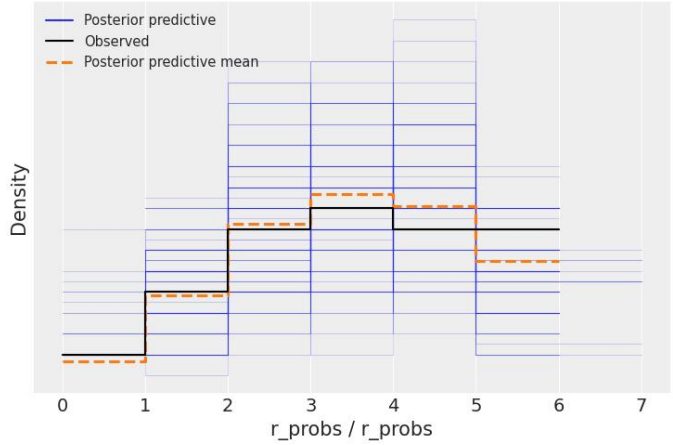

Figure A.30: Posterior predictive check of the Timeliness model.

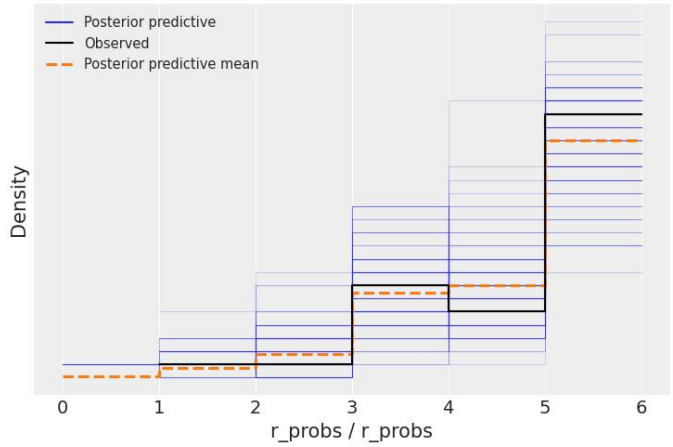

Figure A.31: Posterior predictive check of the Understandability model.

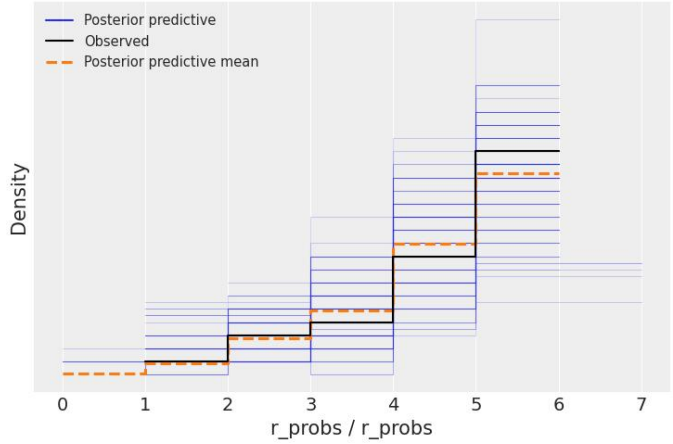

Figure A.32: Posterior predictive check of the Usability model.

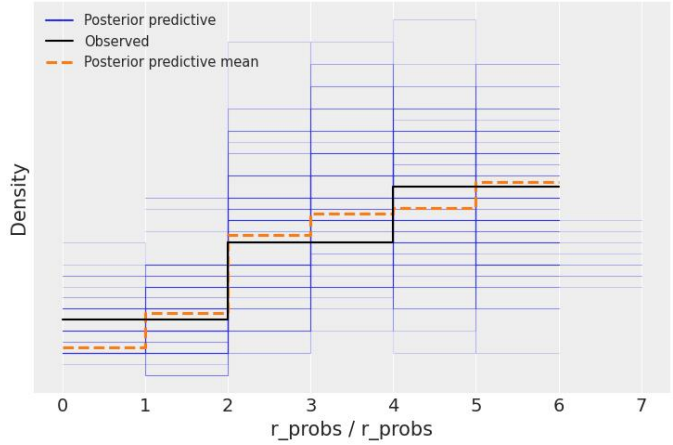

Figure A.33: Posterior predictive check of the Value-Added model.

Ratings: Beta Prediction Density Plots

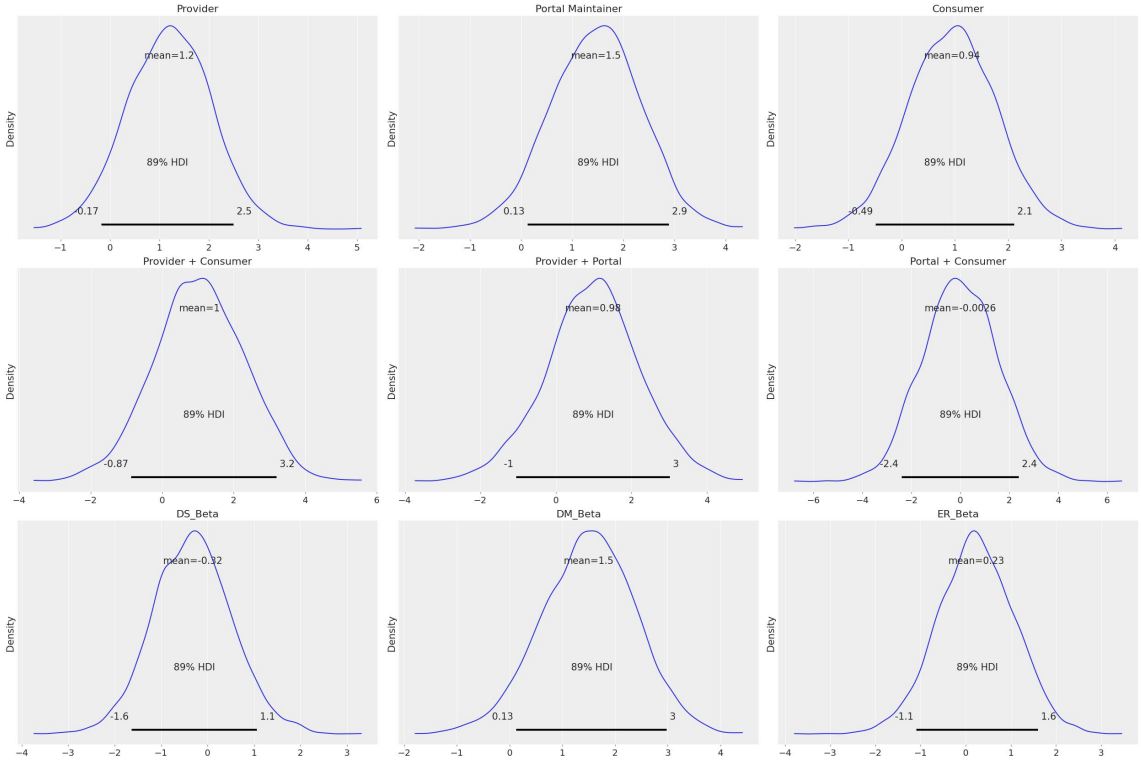

Figure A.34: Density plots of betas for the Accuracy model.

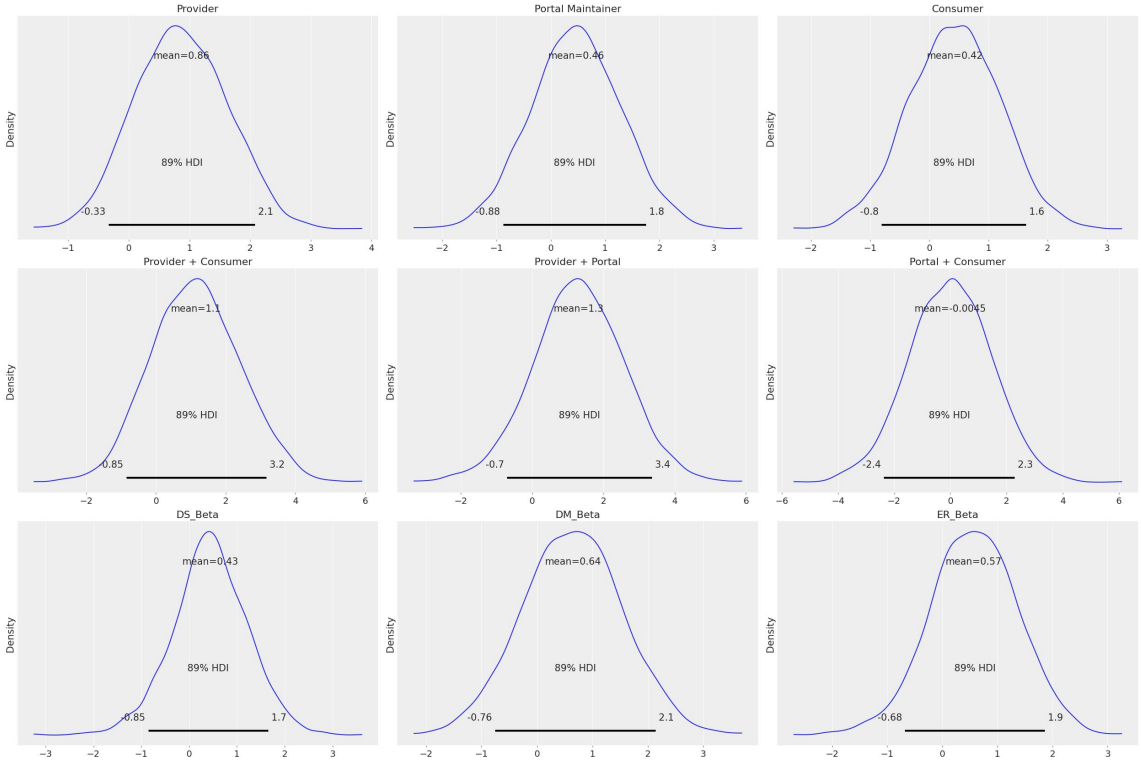

Figure A.35: Density plots of betas for the Completeness model.

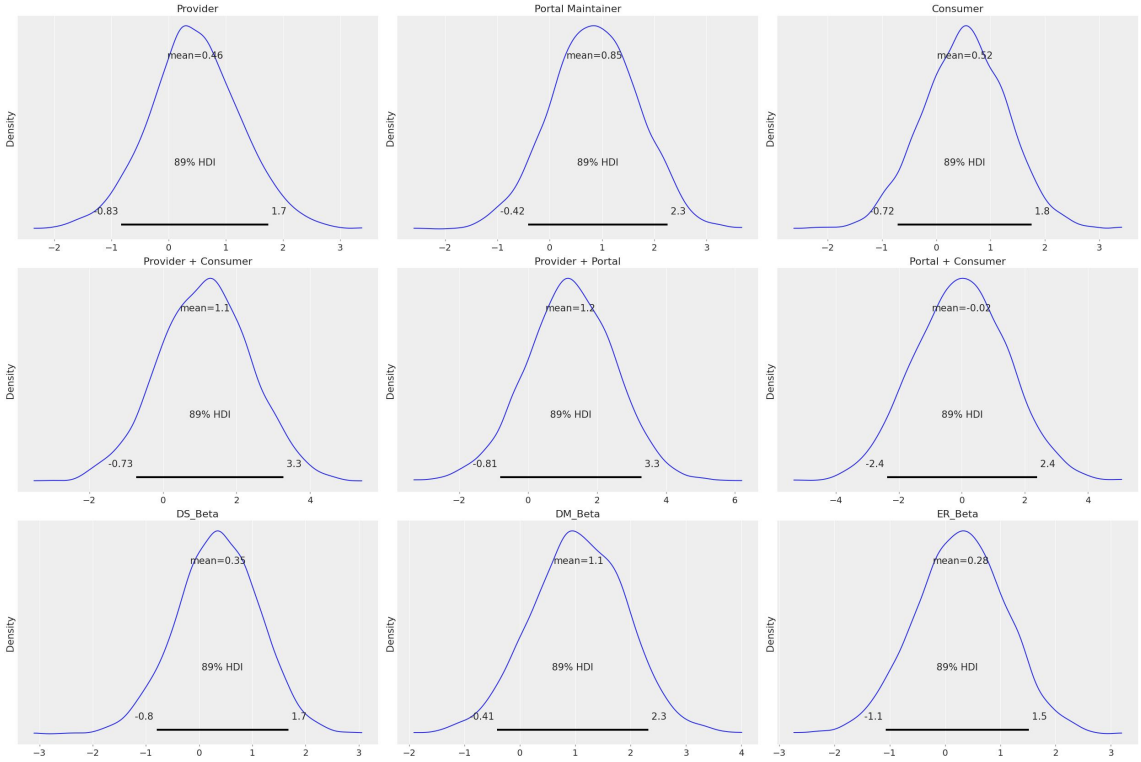

Figure A.36: Density plots of betas for the Consistency model.

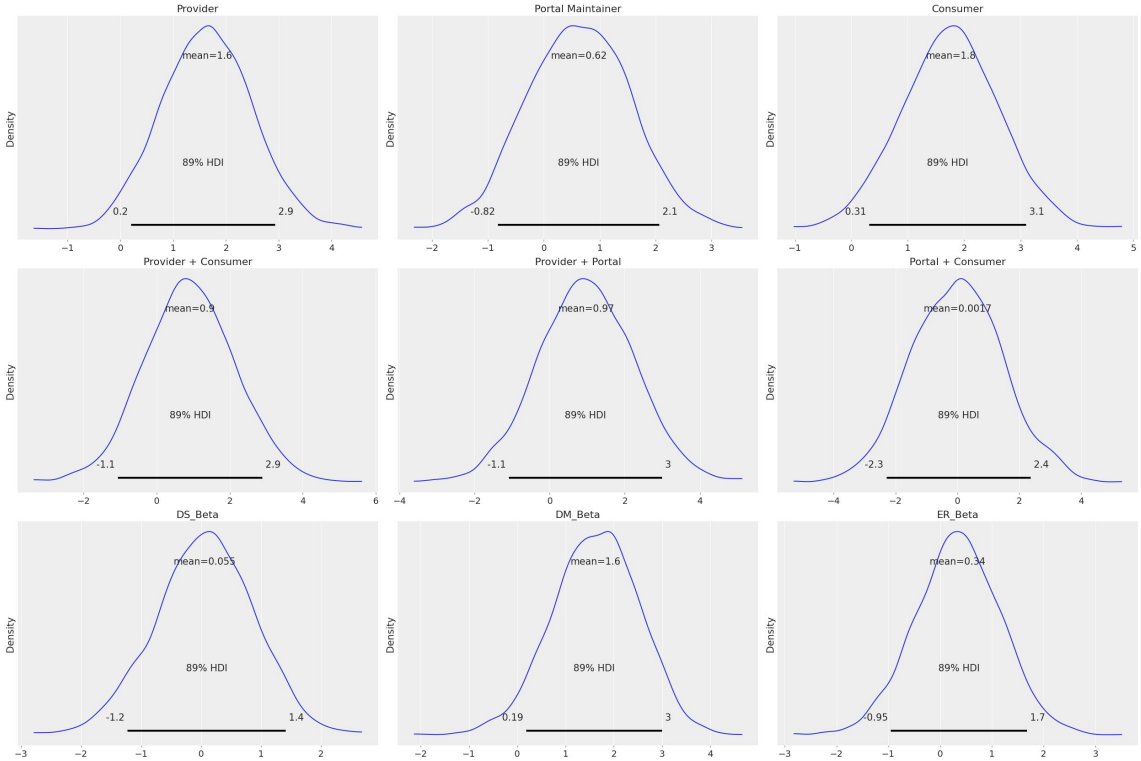

Figure A.37: Density plots of betas for the Credibility model.

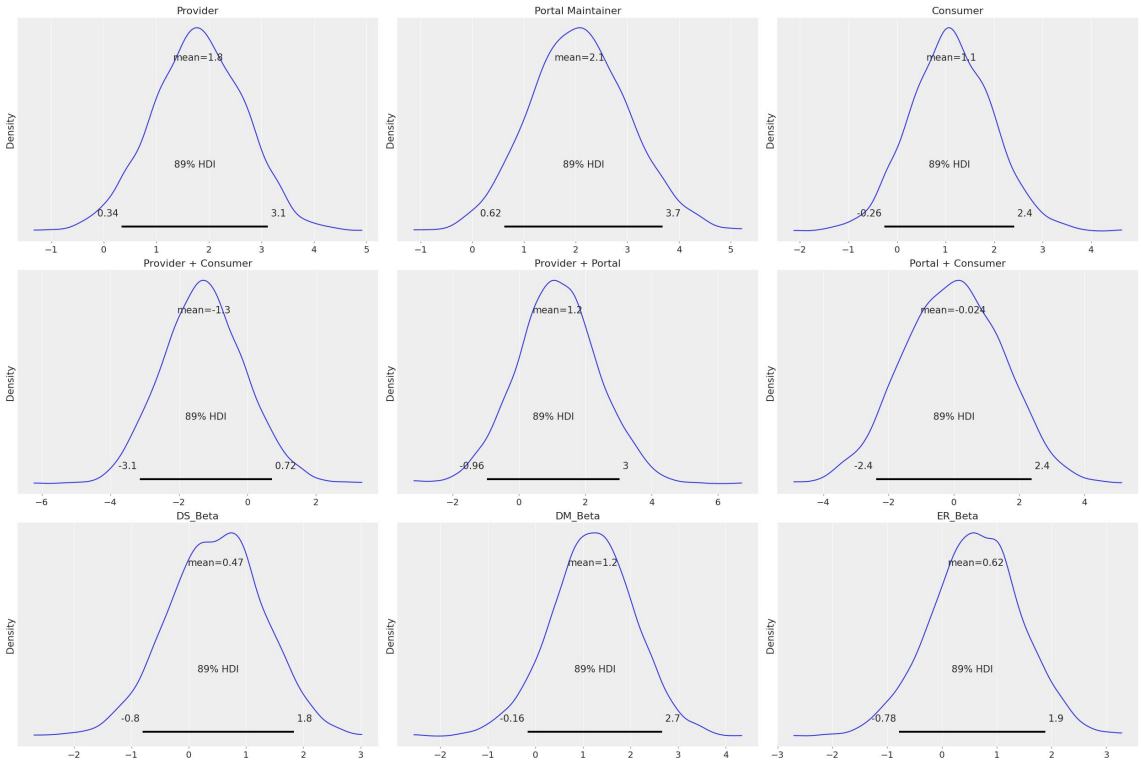

Figure A.38: Density plots of betas for the Findability model.

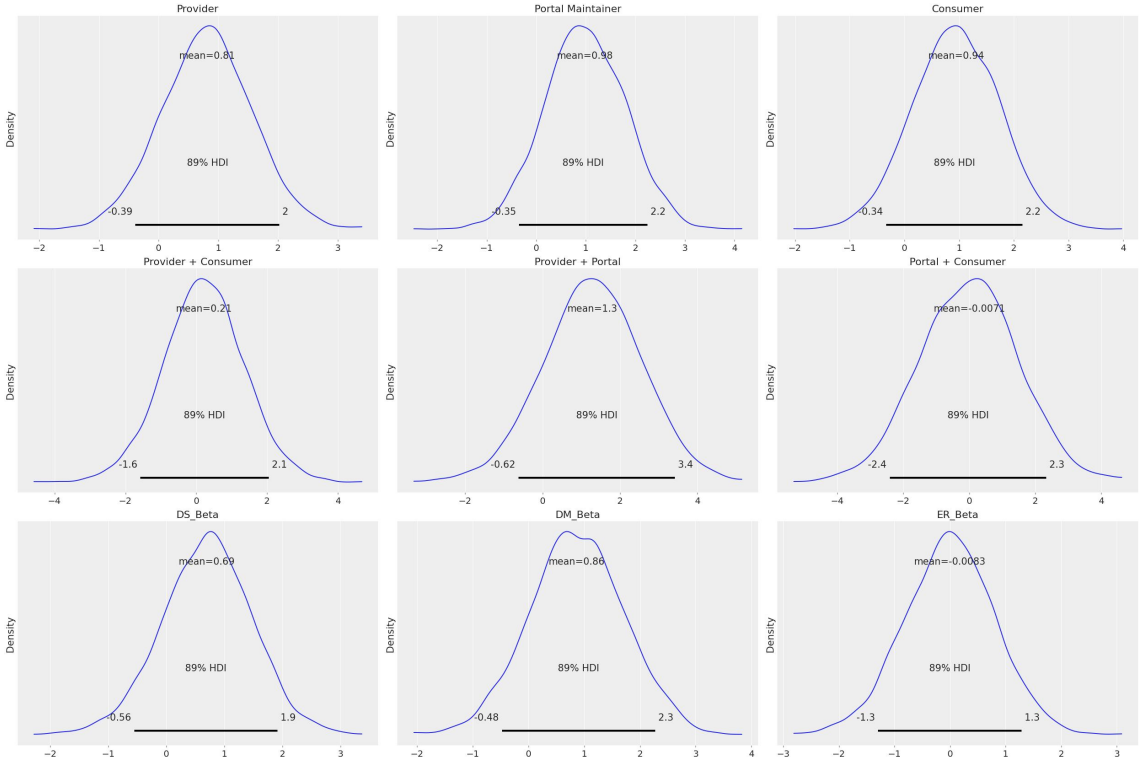

Figure A.39: Density plots of betas for the Interoperability model.

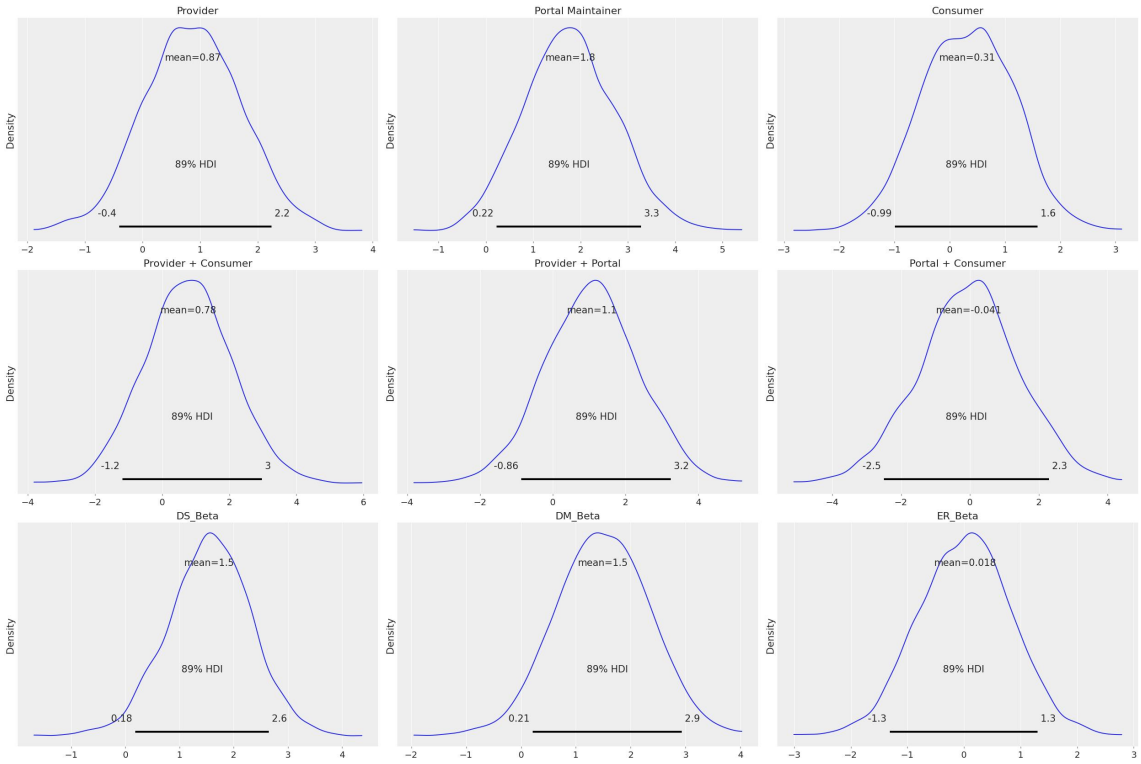

Figure A.40: Density plots of betas for the Interpretability model.

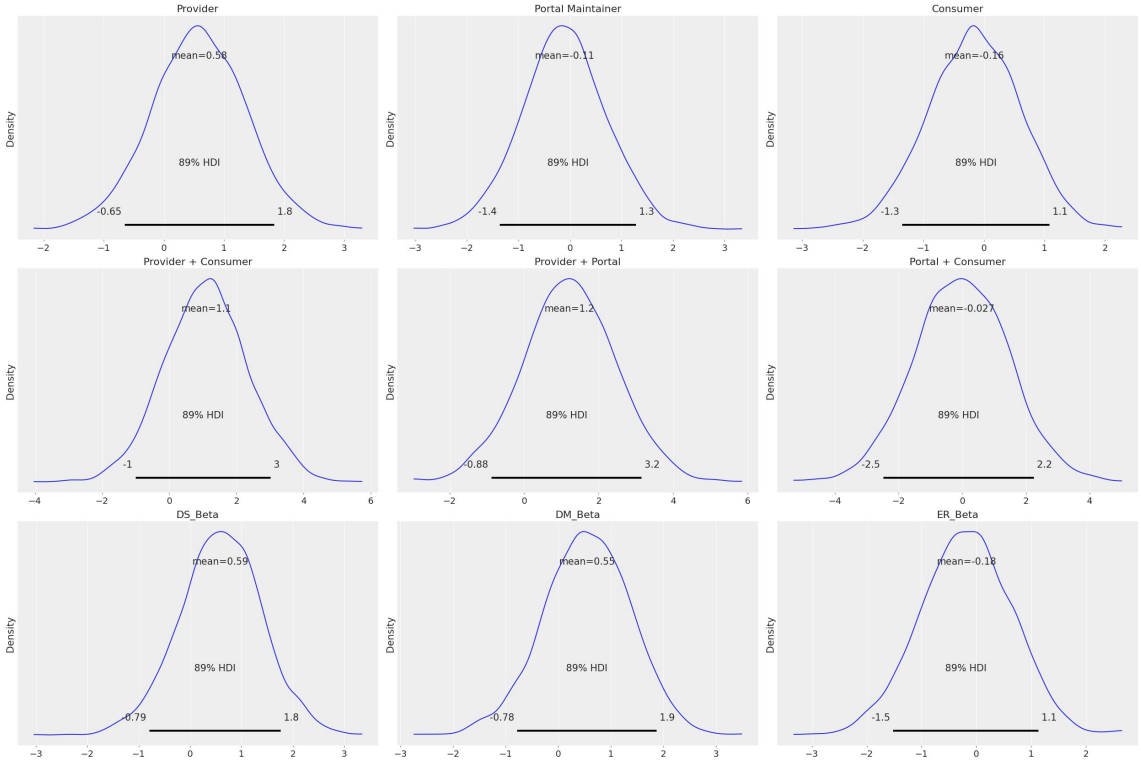

Figure A.41: Density plots of betas for the Relevance model.

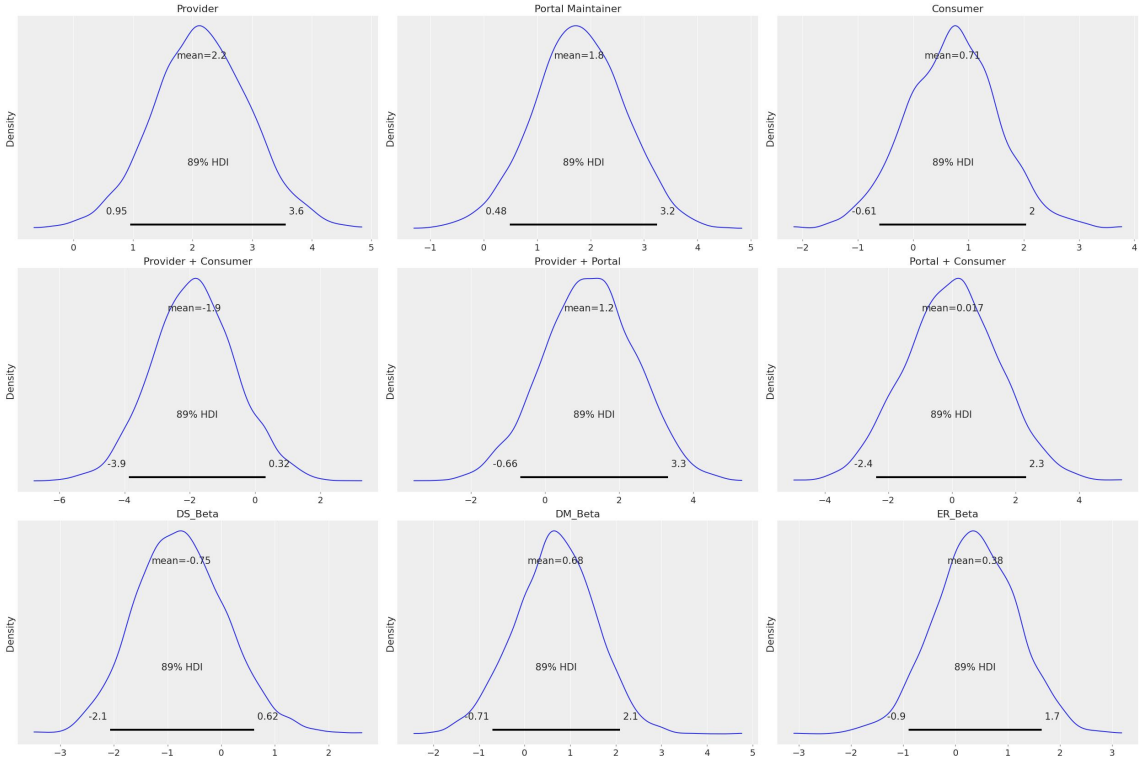

Figure A.42: Density plots of betas for the Reusability model.

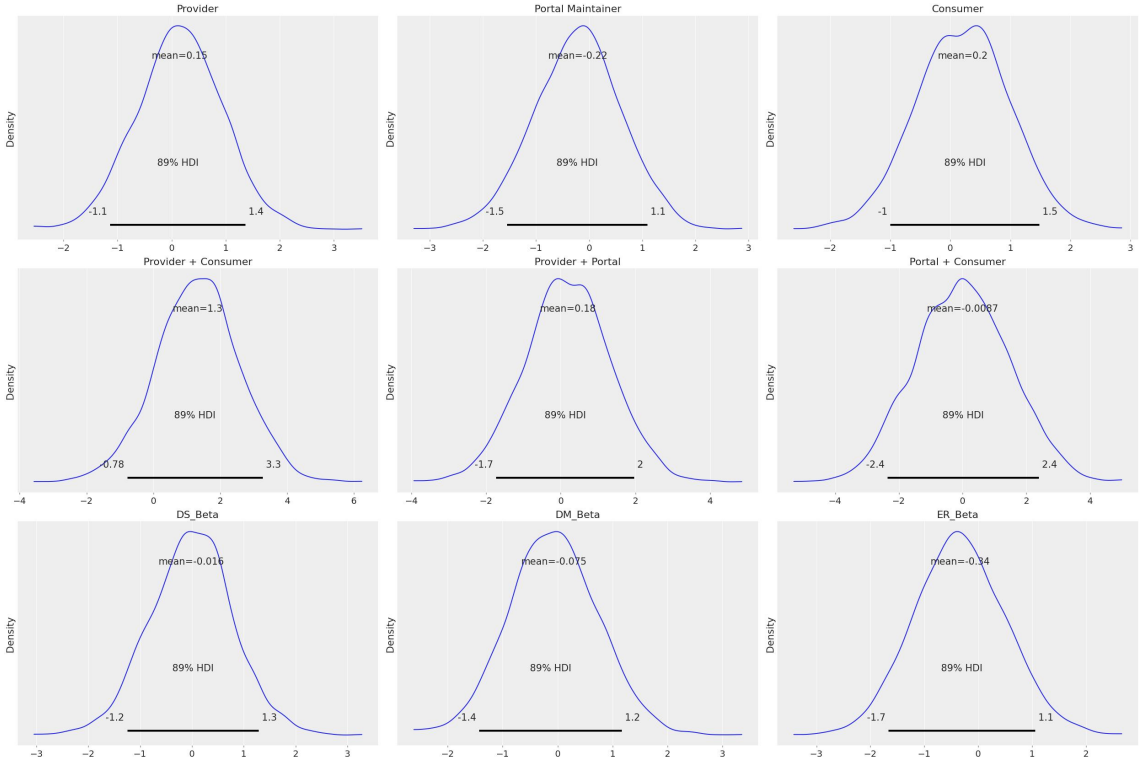

Figure A.43: Density plots of betas for the Security model.

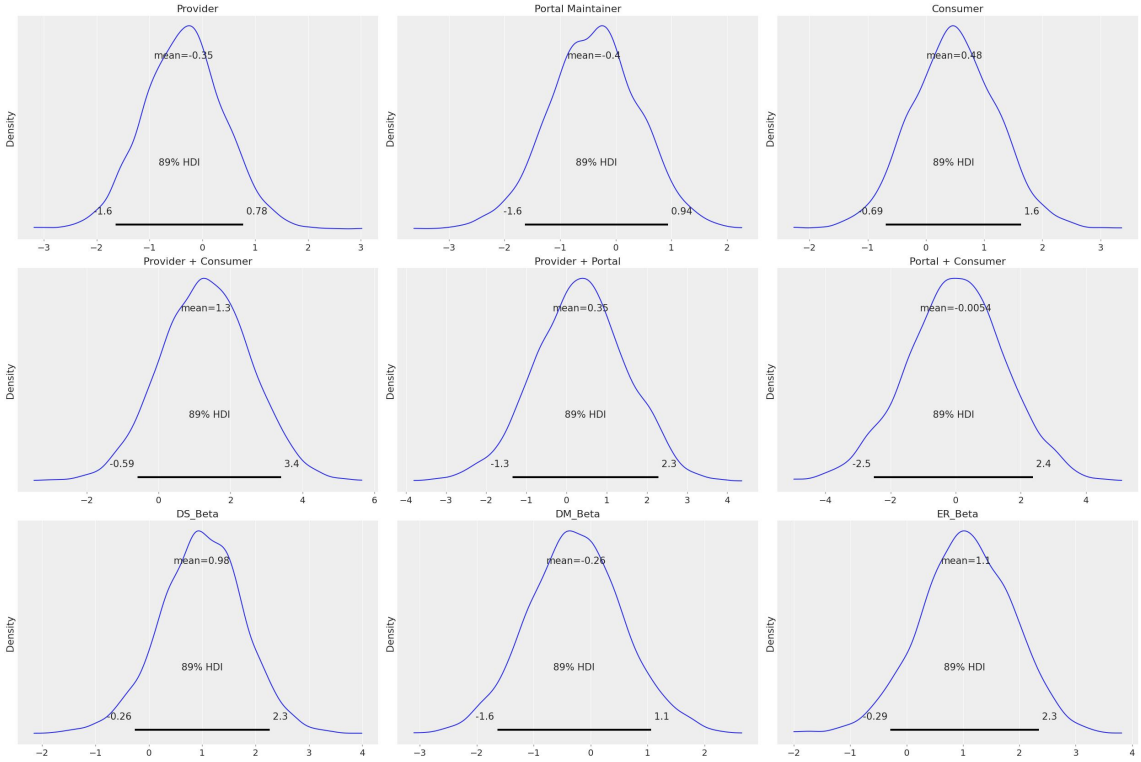

Figure A.44: Density plots of betas for the Timeliness model.

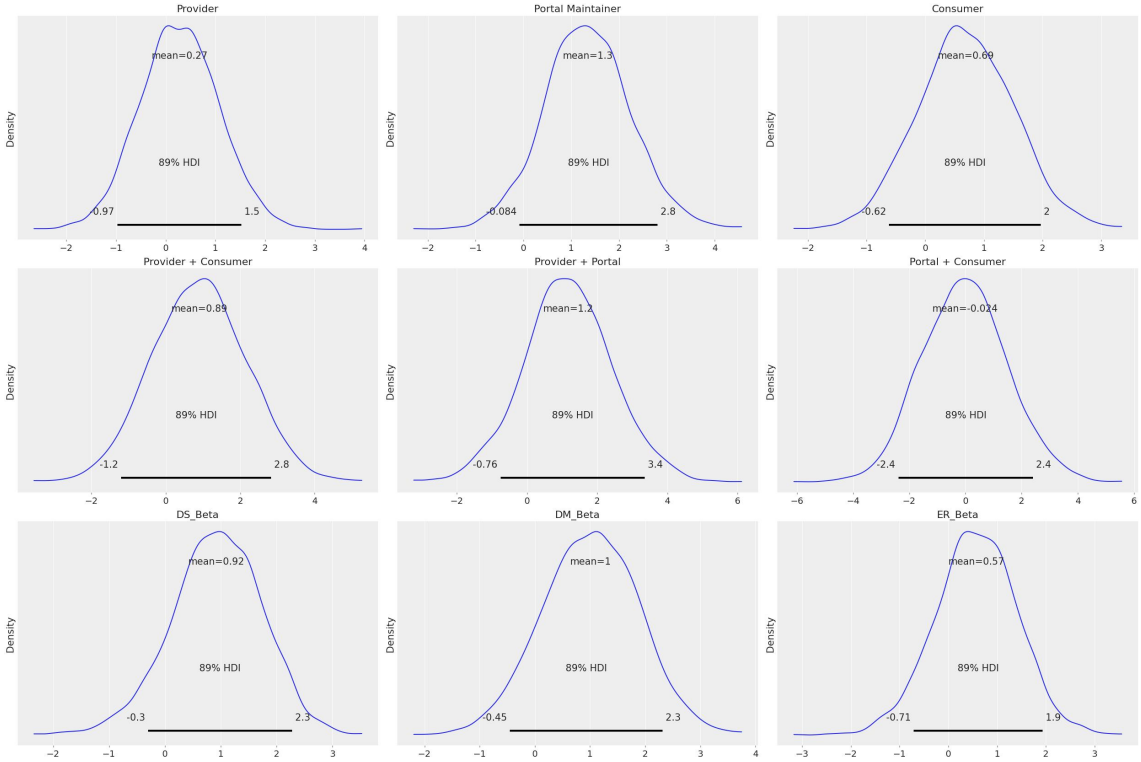

Figure A.45: Density plots of betas for the Understandability model.

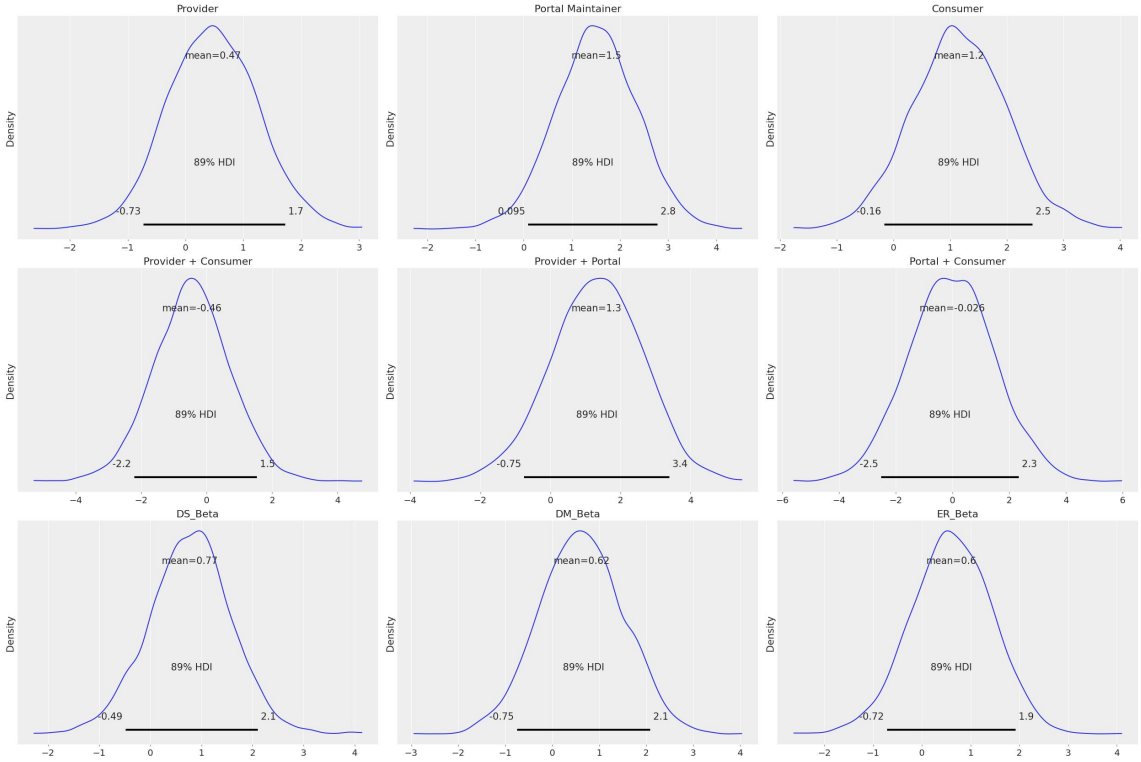

Figure A.46: Density plots of betas for the Usability model.

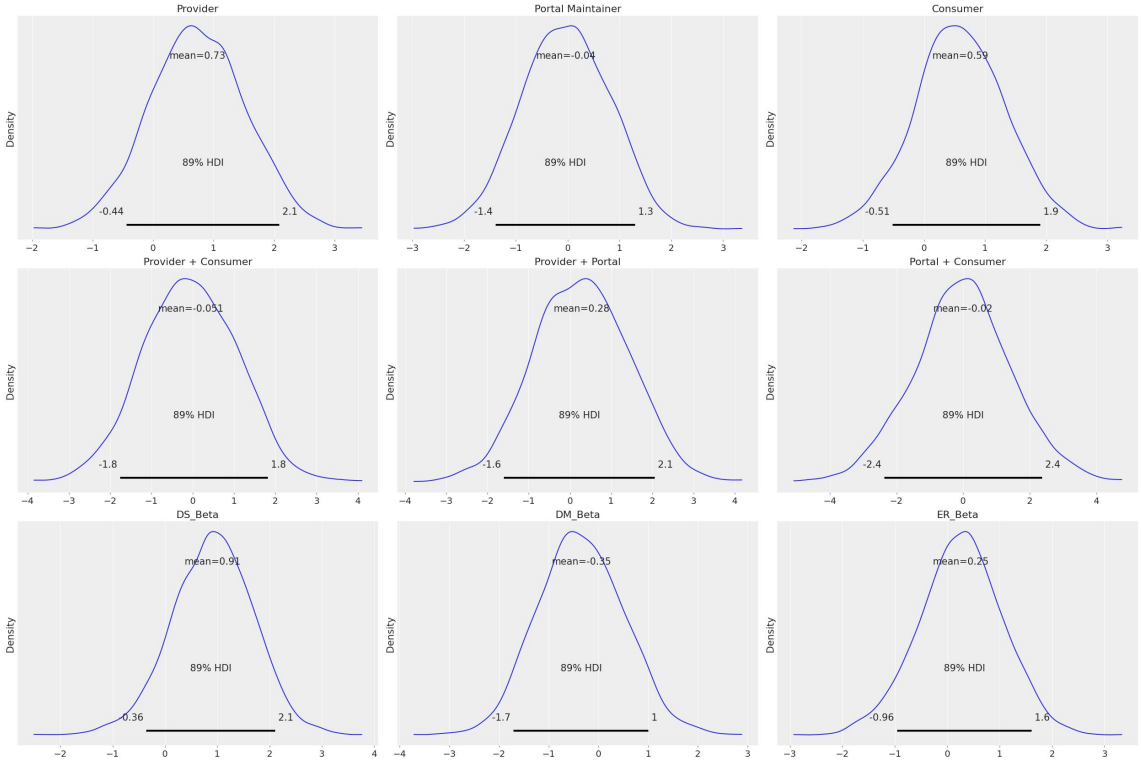

Figure A.47: Density plots of betas for the Value Added model.

Ratings: Trace Plots

Table A.1: Trace representation of Accuracy Model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	1.2	0.84	-0.17	2.5	2905.92	1.0
Portal Maintainer	1.46	0.89	0.13	2.89	2996.26	1.0
Consumer	0.94	0.82	-0.49	2.11	2885.98	1.0
Provider + Consumer	1.04	1.28	-0.87	3.19	2770.35	1.0
Provider + Portal	0.98	1.25	-1.05	3.02	2856.53	1.0
Portal + Consumer	-0.0	1.54	-2.4	2.4	3166.24	1.0
DS_Beta	-0.32	0.86	-1.63	1.06	3052.23	1.0
DM_Beta	1.5	0.89	0.13	2.98	2966.26	1.0
ER_Beta	0.23	0.85	-1.09	1.6	2897.20	1.0

Table A.2: Trace representation of Completeness model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	0.86	0.76	-0.33	2.08	2589.29	1.0
Portal Maintainer	0.46	0.84	-0.88	1.75	2539.95	1.0
Consumer	0.42	0.77	-0.8	1.63	3090.24	1.0
Provider + Consumer	1.14	1.27	-0.85	3.17	2981.43	1.0
Provider + Portal	1.3	1.28	-0.7	3.36	2804.63	1.0
Portal + Consumer	-0.0	1.47	-2.36	2.29	2818.14	1.0
DS_Beta	0.43	0.79	-0.85	1.65	2731.03	1.0
DM_Beta	0.64	0.9	-0.76	2.14	2811.95	1.0
ER_Beta	0.57	0.8	-0.68	1.86	2821.93	1.0

Table A.3: Trace representation of Consistency model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	0.46	0.8	-0.83	1.75	2259.39	1.0
Portal Maintainer	0.85	0.84	-0.42	2.25	2268.19	1.0
Consumer	0.52	0.78	-0.72	1.76	2464.37	1.0
Provider + Consumer	1.14	1.25	-0.73	3.27	3278.45	1.0
Provider + Portal	1.21	1.3	-0.81	3.29	2863.20	1.0
Portal + Consumer	-0.02	1.5	-2.37	2.39	3176.07	1.0
DS_Beta	0.35	0.78	-0.8	1.68	2922.75	1.0
DM_Beta	1.05	0.86	-0.41	2.32	2438.63	1.0
ER_Beta	0.28	0.83	-1.07	1.52	2809.51	1.0

Table A.4: Trace representation of Credibility model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	1.64	0.86	0.2	2.93	2702.42	1.0
Portal Maintainer	0.62	0.92	-0.82	2.07	2867.70	1.0
Consumer	1.76	0.87	0.31	3.1	3000.19	1.0
Provider + Consumer	0.9	1.25	-1.06	2.89	2975.16	1.0
Provider + Portal	0.97	1.28	-1.09	2.99	2775.21	1.0
Portal + Consumer	0.0	1.47	-2.29	2.36	2919.69	1.0
DS_Beta	0.06	0.82	-1.22	1.41	2870.30	1.0
DM_Beta	1.62	0.9	0.19	3.0	2845.10	1.0
ER_Beta	0.34	0.83	-0.95	1.68	3005.67	1.0

Table A.5: Trace representation of Findability model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	1.79	0.87	0.34	3.12	2756.30	1.0
Portal Maintainer	2.07	0.95	0.62	3.68	3151.90	1.0
Consumer	1.12	0.85	-0.26	2.42	3071.97	1.0
Provider + Consumer	-1.28	1.21	-3.14	0.72	3130.04	1.0
Provider + Portal	1.15	1.24	-0.96	3.03	3191.64	1.0
Portal + Consumer	-0.02	1.52	-2.39	2.38	3006.87	1.0
DS_Beta	0.47	0.83	-0.8	1.84	2684.83	1.0
DM_Beta	1.2	0.89	-0.16	2.66	2786.49	1.0
ER_Beta	0.62	0.83	-0.78	1.89	2857.81	1.0

Table A.6: Trace representation of Interoperability model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	0.81	0.76	-0.39	2.02	3156.37	1.0
Portal Maintainer	0.98	0.81	-0.35	2.24	3137.28	1.0
Consumer	0.94	0.79	-0.34	2.16	2618.52	1.0
Provider + Consumer	0.21	1.14	-1.58	2.06	3085.69	1.0
Provider + Portal	1.29	1.27	-0.62	3.41	2867.06	1.0
Portal + Consumer	-0.01	1.5	-2.41	2.33	3112.12	1.0
DS_Beta	0.69	0.79	-0.56	1.91	2936.66	1.0
DM_Beta	0.86	0.85	-0.48	2.28	3025.84	1.0
ER_Beta	-0.01	0.82	-1.3	1.29	2635.70	1.0

Table A.7: Trace representation of Interpretability model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	0.87	0.84	-0.4	2.24	2802.33	1.0
Portal Maintainer	1.78	0.97	0.22	3.28	2626.31	1.0
Consumer	0.31	0.84	-0.99	1.59	2917.84	1.0
Provider + Consumer	0.78	1.3	-1.18	2.98	2770.32	1.0
Provider + Portal	1.07	1.29	-0.86	3.25	2855.36	1.0
Portal + Consumer	-0.04	1.48	-2.49	2.29	3062.93	1.0
DS_Beta	1.54	0.78	0.18	2.65	2536.39	1.0
DM_Beta	1.49	0.86	0.21	2.93	2725.59	1.0
ER_Beta	0.02	0.83	-1.31	1.3	3162.29	1.0

Table A.8: Trace representation of Relevance model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	0.58	0.78	-0.65	1.84	2963.66	1.0
Portal Maintainer	-0.11	0.83	-1.36	1.28	3248.83	1.0
Consumer	-0.16	0.77	-1.35	1.08	3280.24	1.0
Provider + Consumer	1.1	1.25	-1	3.01	3027.69	1.0
Provider + Portal	1.21	1.25	-0.88	3.15	2459.04	1.0
Portal + Consumer	-0.03	1.49	-2.48	2.25	2769.42	1.0
DS_Beta	0.59	0.81	-0.79	1.77	3205.18	1.0
DM_Beta	0.55	0.83	-0.78	1.87	3054.90	1.0
ER_Beta	-0.18	0.84	-1.52	1.13	3385.06	1.0

Table A.9: Trace representation of Reusability model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	2.15	0.82	0.95	3.57	2939.66	1.0
Portal Maintainer	1.78	0.86	0.48	3.25	3199.97	1.0
Consumer	0.71	0.84	-0.61	2.05	2898.80	1.0
Provider + Consumer	-1.86	1.31	-3.87	0.32	2891.46	1.0
Provider + Portal	1.19	1.26	-0.66	3.32	2846.32	1.0
Portal + Consumer	0.02	1.49	-2.39	2.34	2138.23	1.0
DS_Beta	-0.75	0.85	-2.07	0.62	2727.92	1.0
DM_Beta	0.68	0.88	-0.71	2.1	2850.58	1.0
ER_Beta	0.38	0.8	-0.9	1.65	2416.36	1.0

Table A.10: Trace representation of Security model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	0.15	0.79	-1.14	1.36	3144.43	1.0
Portal Maintainer	-0.22	0.82	-1.55	1.1	3180.62	1.0
Consumer	0.2	0.78	-1.0	1.48	3004.20	1.0
Provider + Consumer	1.32	1.26	-0.78	3.27	2591.80	1.0
Provider + Portal	0.18	1.16	-1.73	1.97	3044.30	1.0
Portal + Consumer	-0.01	1.49	-2.35	2.4	3026.85	1.0
DS_Beta	-0.02	0.8	-1.25	1.28	2787.26	1.0
DM_Beta	-0.08	0.82	-1.43	1.17	3242.94	1.0
ER_Beta	-0.34	0.85	-1.66	1.06	3015.20	1.0

Table A.11: Trace representation of timeliness model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	-0.35	0.76	-1.64	0.78	3158.60	1.0
Portal Maintainer	-0.4	0.81	-1.63	0.94	2766.36	1.0
Consumer	0.48	0.74	-0.69	1.63	3381.92	1.0
Provider + Consumer	1.31	1.26	-0.59	3.41	3126.86	1.0
Provider + Portal	0.35	1.15	-1.35	2.28	2902.44	1.0
Portal + Consumer	-0.01	1.51	-2.51	2.37	2915.93	1.0
DS_Beta	0.98	0.8	-0.26	2.26	3116.90	1.0
DM_Beta	-0.26	0.84	-1.64	1.06	3092.75	1.0
ER_Beta	1.05	0.83	-0.29	2.34	2590.40	1.0

Table A.12: Trace representation of Understandability model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	0.27	0.79	-0.97	1.52	2675.76	1.0
Portal Maintainer	1.34	0.91	-0.08	2.81	2883.39	1.0
Consumer	0.69	0.82	-0.62	1.97	2831.14	1.0
Provider + Consumer	0.89	1.27	-1.2	2.83	2822.63	1.0
Provider + Portal	1.17	1.27	-0.76	3.35	2409.78	1.0
Portal + Consumer	-0.02	1.52	-2.39	2.41	2877.07	1.0
DS_Beta	0.92	0.82	-0.3	2.28	2150.92	1.0
DM_Beta	1.0	0.87	-0.45	2.31	2916.11	1.0
ER_Beta	0.57	0.84	-0.71	1.93	2678.80	1.0

Table A.13: Trace representation of Usability model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	0.47	0.78	-0.73	1.72	2979.54	1.0
Portal Maintainer	1.51	0.85	0.1	2.78	2987.52	1.0
Consumer	1.16	0.83	-0.16	2.46	2814.02	1.0
Provider + Consumer	-0.46	1.19	-2.22	1.53	2723.96	1.0
Provider + Portal	1.31	1.31	-0.75	3.4	2786.88	1.0
Portal + Consumer	-0.03	1.52	-2.52	2.34	2901.25	1.0
DS_Beta	0.77	0.82	-0.49	2.11	2424.34	1.0
DM_Beta	0.62	0.89	-0.75	2.08	2791.55	1.0
ER_Beta	0.6	0.83	-0.72	1.93	2746.93	1.0

Table A.14: Trace representation of Value Added model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	0.73	0.79	-0.44	2.09	3306.45	1.0
Portal Maintainer	-0.04	0.86	-1.39	1.3	3391.02	1.0
Consumer	0.59	0.75	-0.51	1.9	2832.09	1.0
Provider + Consumer	-0.05	1.15	-1.76	1.82	3054.92	1.0
Provider + Portal	0.28	1.16	-1.6	2.06	3172.11	1.0
Portal + Consumer	-0.02	1.47	-2.37	2.37	2556.19	1.0
DS_Beta	0.91	0.79	-0.36	2.11	3085.16	1.0
DM_Beta	-0.35	0.87	-1.71	1.01	3097.39	1.0
ER_Beta	0.25	0.8	-0.96	1.61	2527.64	1.0

Pick-5: Posterior Predictive Checks Plots

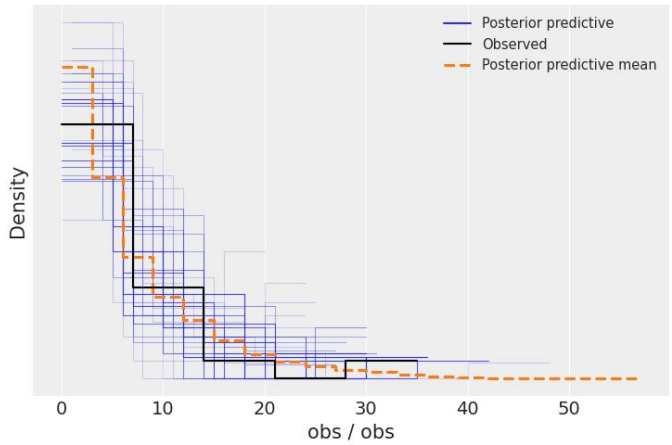

Figure A.48: Posterior predictive check of the Pick-5 "Most Relevant" model.

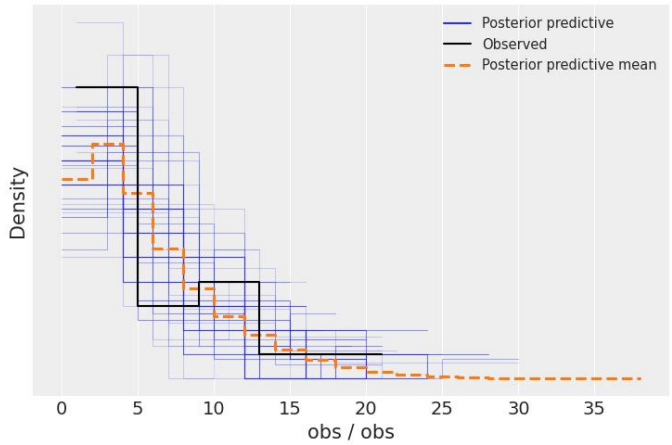

Figure A.49: Posterior predictive check of the Pick-5 "Least Relevant" model.

Pick-5: Beta Prediction Density Plots

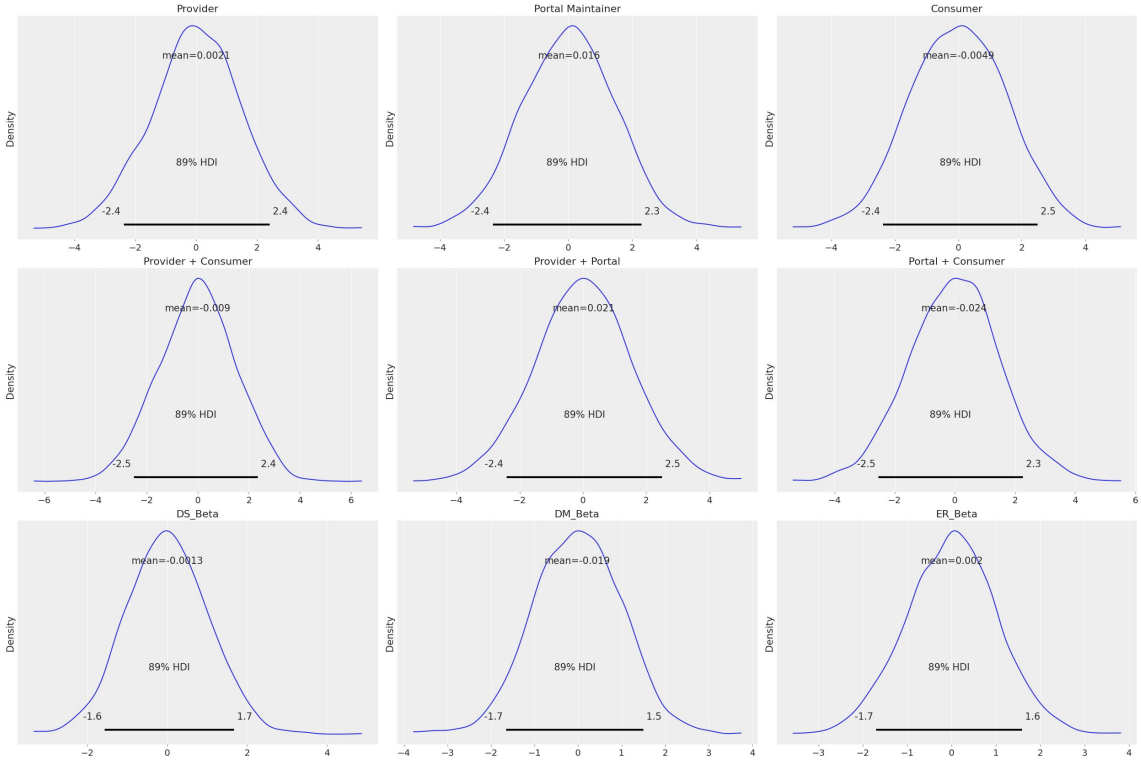

Figure A.50: Density plots of betas for the Pick-5 "Most Relevant" model.

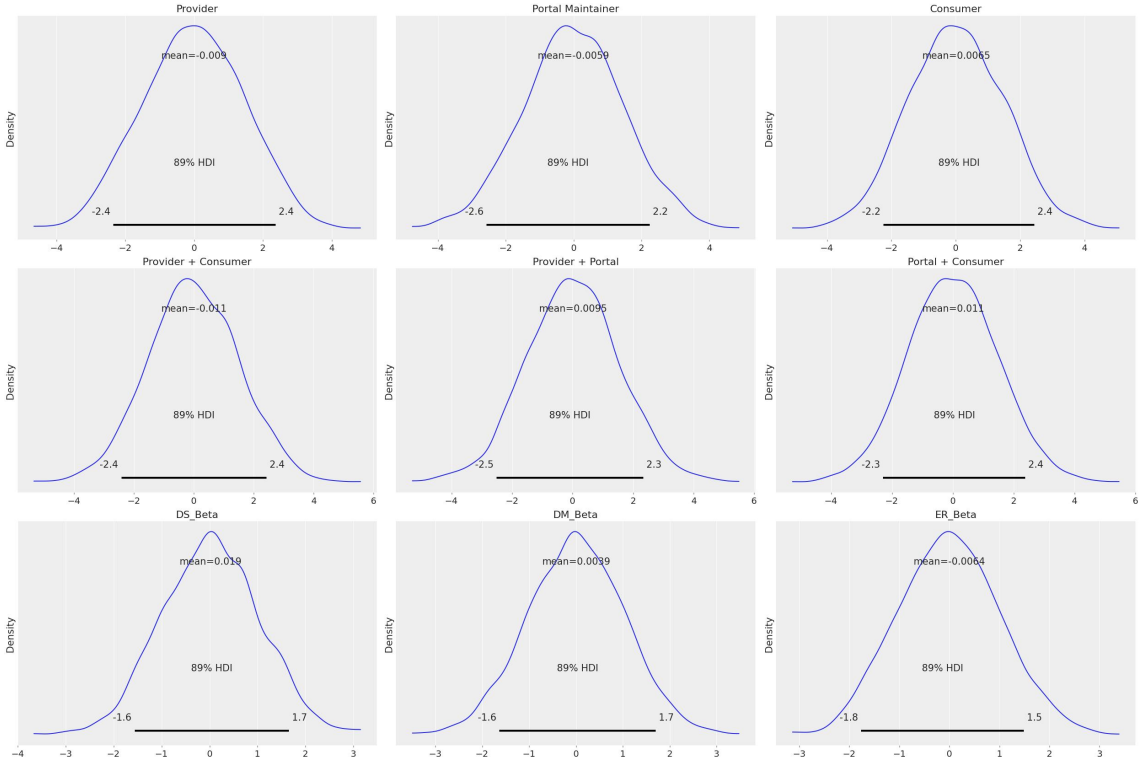

Figure A.51: Density plots of betas for the Pick-5 "Least Relevant" model.

Pick-5: Trace Plots

Table A.15: Trace representation of pick-5 "Most Relevant" model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	0.002	1.492	-2.381	2.412	2668	1.0
Portal Maintainer	0.016	1.475	-2.36	2.291	2555	1.0
Consumer	-0.005	1.529	-2.373	2.5	2968	1.0
Provider + Consumer	-0.009	1.515	-2.501	2.357	2887	1.0
Provider + Portal	0.021	1.53	-2.42	2.501	2769	1.0
Portal + Consumer	-0.024	1.508	-2.543	2.259	2866	1.0
beta_ds	-0.001	1.025	-1.566	1.675	2547	1.0
beta_er	-0.019	1.017	-1.658	1.504	3113	1.0
beta_dm	0.002	1.024	-1.701	1.586	2668	1.0
appropriate amount of data	0.075	0.021	0.043	0.106	2761	1.0
complexity	0.006	0.006	0.0	0.014	1501	1.0
compliance	0.044	0.016	0.019	0.069	2162	1.0
concise representation	0.025	0.012	0.007	0.042	2496	1.0
conciseness	0.025	0.012	0.006	0.043	2204	1.0
contactability	0.025	0.013	0.006	0.043	2098	1.0
cost effectiveness	0.018	0.01	0.002	0.032	1963	1.0
documentation	0.187	0.031	0.136	0.236	3081	1.0
duplications	0.018	0.01	0.003	0.032	2038	1.0
flexibility	0.006	0.006	0.0	0.013	1282	1.0
granularity	0.012	0.009	0.001	0.024	1916	1.0
objectivity	0.087	0.022	0.054	0.122	2383	1.0
presentation quality	0.025	0.012	0.007	0.042	2371	1.0
provenance	0.074	0.021	0.044	0.107	2854	1.0
recoverability	0.062	0.019	0.032	0.091	2960	1.0
representation	0.043	0.016	0.018	0.068	1954	1.0
skewness	0.012	0.009	0.001	0.024	2108	1.0
traceability	0.125	0.026	0.085	0.166	2639	1.0
unambiguous	0.087	0.022	0.053	0.122	2332	1.0
uniqueness	0.012	0.008	0.0	0.023	1182	1.0
volume	0.031	0.014	0.01	0.051	2904	1.0
ds_No Experience	0.327	0.235	0.0	0.662	2045	1.0
ds_5- Years	0.339	0.241	0.001	0.688	2106	1.0
ds_5+ Years	0.334	0.24	0.0	0.675	2159	1.0

Table A.16: Trace representation of pick-5 "Least Relevant" model prediction

Parameter	mean	sd	hdi_5.5%	hdi_94.5%	ess_tail	r_hat
Provider	-0.009	1.474	-2.353	2.376	3061	1.0
Portal Maintainer	-0.006	1.494	-2.561	2.243	2422	1.0
Consumer	0.007	1.491	-2.248	2.443	3088	1.0
Provider + Consumer	-0.011	1.506	-2.422	2.422	2862	1.0
Provider + Portal	0.01	1.542	-2.516	2.333	2679	1.0
Portal + Consumer	0.011	1.477	-2.291	2.386	2571	1.0
beta_ds	0.019	1.027	-1.564	1.657	2753	1.0
beta_er	0.004	1.042	-1.632	1.703	3031	1.0
beta_dm	-0.006	1.014	-1.763	1.491	2874	1.0
appropriate amount of data	0.073	0.022	0.04	0.108	2793	1.0
complexity	0.111	0.027	0.067	0.152	2675	1.0
compliance	0.015	0.01	0.002	0.029	2464	1.0
concise representation	0.037	0.016	0.013	0.061	2774	1.0
conciseness	0.029	0.014	0.008	0.05	2523	1.0
contactability	0.022	0.012	0.004	0.038	2666	1.0
cost effectiveness	0.073	0.021	0.041	0.107	2850	1.0
documentation	0.037	0.016	0.012	0.059	2459	1.0
duplications	0.073	0.022	0.039	0.108	2676	1.0
flexibility	0.037	0.016	0.011	0.059	2690	1.0
granularity	0.051	0.019	0.022	0.079	3056	1.0
objectivity	0.029	0.014	0.008	0.049	2550	1.0
presentation quality	0.029	0.015	0.007	0.05	2511	1.0
provenance	0.022	0.013	0.003	0.039	2002	1.0
recoverability	0.015	0.01	0.001	0.028	2180	1.0
representation	0.044	0.018	0.016	0.071	2842	1.0
skewness	0.044	0.018	0.017	0.07	2699	1.0
traceability	0.03	0.014	0.006	0.048	2047	1.0
unambiguous	0.015	0.01	0.0	0.028	2375	1.0
uniqueness	0.081	0.024	0.044	0.117	2577	1.0
volume	0.133	0.029	0.091	0.182	3205	1.0
ds_No Experience	0.332	0.232	0.0	0.655	2360	1.0
ds_5- Years	0.335	0.233	0.001	0.665	2267	1.0
ds_5+ Years	0.333	0.237	0.001	0.662	2235	1.0
er_No Experience	0.336	0.234	0.0	0.661	2154	1.0
er_5- Years	0.332	0.232	0.0	0.656	2208	1.0
er_5+ Years	0.332	0.233	0.0	0.664	2479	1.0
dm_No Experience	0.33	0.234	0.003	0.669	2836	1.0
dm_5- Years	0.334	0.234	0.0	0.674	2383	1.0
dm_5+ Years	0.336	0.235	0.001	0.676	2247	1.0