

Master Thesis

How to determine the **FAIRness** of Open Data by a Reference Model

Jorien van Ginkel BSc.

Master Business Informatics
Department of Information and Computing Sciences
Utrecht University



Utrecht University



Berenschot
Intellerts

ENABLING DATA DRIVEN
TRANSFORMATIONS

Author	Jorien van Ginkel BSc. 5766877 j.m.ginkel@students.uu.nl Utrecht University
First Supervisor	Dr. Marco Spruit m.r.spruit@uu.nl Department of Information and Computing Sciences Utrecht University
Second Supervisor	Armel Lefebvre MSc. A.E.J.Lefebvre@uu.nl Department of Information and Computing Sciences Utrecht University
External Supervisor	Glenn Elberse Data Scientist Berenschot Intellerts

I. Abstract

'Which requirements should a data repository meet to satisfy the FAIR principles?' is the main research question in this thesis. Whereas FAIR stands for: Findability, Accessibility, Interoperability and Reusability. This study is relevant since a data-driven economy and a digital mindset requires for companies and academia to extend their existing meta data management approaches. Regarding (open) data repositories it is no longer enough to take care of the quantitative part; the amount of data, sources and/or publications in the repository. Good data management becomes more and more important. To answer the main research question, we distinguish three main parts: A systematic Literature Research on the FAIR concepts (RQ1), the context where FAIR can be applied in (RQ2), and the requirements for measuring whether data repositories are FAIR (RQ3). The first research question results in four definitions on the FAIR concepts. Thereby, the main result of this literature research is that Reusability must be resultant of the other three, which implies: $(F + A + I)/3 = R$. Together with the result of research question 3, a set of Meta Data Attributes, these definitions and relation form the basis for the Reference Model (Artifact 2). The main goal of the model is to provide a step-by-step set of activities to ensure a data repository contains all aspects to measure FAIR, with the focus on comprehensibility, and accessibility in public and private sector. The context where this model can be applied in is the Data Scouting Process (Artifact 1). The main goal of this artifact is to determine the place of FAIR in the context of open data projects. We set up this artifact, as result of research question 2, based on a case study at Berenschot Intellerts. During this case study, we also applied the reference model into practice. The artifacts are evaluated based on iterative expert evaluation at Berenschot Intellerts (Artifact 1), interviews at CBS and Gemeente van Amsterdam (Artifact 2), and a survey among experts from business and academia (Artifact 2).

II. Acknowledgements

After ten months, this thesis is finally coming to an end. Many hours of hard work led to this document, that focuses on measuring the 'FAIRness' of open data. This thesis is the end of my student time, which started in 2012 with my bachelor Information Science at VU University and ends now, in September 2017, after the master Business Informatics in Utrecht.

I would like to thank my first supervisor, Dr. Marco Spruit, for his help and support during my research. I also want to thank my second supervisor Armel Lefebvre MSc for all his feedback, time, tips and support regarding this thesis. Additionally, I would like to thank my external supervisor, Glenn Elberse, and all the other colleagues at Berenschot Intellerts for the possibility they gave me to learn more about the 'open data world' in practice by guiding me in their work and organization. Finally, I want to thank my parents who always supported me and gave me the financial possibility to study.

Two years I studied under the motto of Utrecht University and it is my wish to pursue my life and career under the same device: *Sol Iustitiae, Illustra Nos.*

Jorien van Ginkel

Utrecht, September 2017.

III. Table of Content

1. Introduction	9
1.1 Problem Statement	11
1.2 Scope Thesis	13
1.3 Example Organization: Berenschot Intellerts.....	14
2. Research Approach	15
2.1 Research Questions.....	15
2.2 Research Method	16
2.3 Research Design	17
2.4 Academic relevance: open science	19
2.5 Business relevance: open innovation.....	19
3. Research Question 1: Systematic Literature Research.....	21
3.1 Data Quality in general.....	21
3.2 Findability	24
3.3 Accessibility	27
3.4 Interoperability.....	32
3.5 Reusability	36
3.6 Statistical conclusions and summary Literature Research	40
4. Research Question 2: Benefit of FAIR in Open Data Projects.....	42
4.1 Open Data Landscape.....	42
4.2 Roadmap Data Scouting Process	48
4.3 Process and iterative evaluation Data Scouting Process Roadmap	49
4.4 The place of FAIR in the Data Scouting Process	50
5. Case Study Project: Data Scouting Process in practice.....	52
5.1 Processes and Types of data Berenschot Intellerts.....	52
5.2 Step 1: Determine domain	53
5.3 Step 2: First search and identify objectives.....	55
5.4 Step 3: Determine KPIs.....	55
5.5 Step 4: Targeted Search	56
5.7 Step 5: Evaluate and ensure data quality.....	56
5.8 Step 6: Linking and Visualization	57
6. Research Question 3: Meta Data Management	60
6.1 Metadata management in general.....	60

6.2 Types and relevance of Metadata requirements	61
6.2 Realizations of metadata management	61
6.3 Basic Elements Metadata	62
6.4 FAIR Data Point Software Specification.....	62
6.5 DSA versus FAIR.....	63
6.6 Metric and Scoring Mechanism Approach	65
7. Model Design Reference Models	67
7.1 Key parts within FAIR definitions	68
7.2 Model 1: FAIR activities model.....	68
7.3 Relations between the concepts	70
7.4 Final Model.....	71
7.5 Interpretation Final Model	75
8. Evaluation.....	79
8.1 Survey	79
8.2 Main Findings Survey.....	79
8.3 Interviews	80
8.4 Main Findings Interviews.....	80
8.5 Is a data repository FAIR?.....	82
8.6 Motivation main improvements after evaluation	82
9. Overall Conclusion.....	86
9.1 Answering sub-questions	86
9.2 Answering main research question.....	88
References	91
Appendix A: Literature Research Set up	94
Appendix B: Process Systematic Literature Review	102
Appendix C: Survey Results	119
Appendix D: Interview Protocol	126

IV. List of Figures

Figure 1: Structure Thesis.....	10
Figure 2: Characteristics of the four FAIR principles as described by Wilkinson et al. (2016).	11
Figure 3: Characteristics of the five C's (as described by Sherman, 2014).....	12
Figure 4: Components Meta-Data dimension DAMA-DMBOK Framework (Mosley, 2008)	13
Figure 5: Visual representation of document structure.....	15
Figure 6: information Systems Research Framework applied to own research.	16
Figure 7: Design Science Research Process Model (Vaishnavi & Kuechler, 2004)	17
Figure 8: Seven key data quality research themes (Jayewardene et al., 2012)	22
Figure 9: Proposed configuration of data quality rules for information quality assessment	23
Figure 10: Results Survey regarding the concept Findability	27
Figure 11: Results Survey regarding the concept accessibility.....	31
Figure 12: Results Survey regarding the concept interoperability.....	35
Figure 13: Results Survey regarding the concept Reusability	39
Figure 14: Top-10 Open Data Barometer Project 2016	42
Figure 15: The Open Data status of the Netherlands compared to the rest of the world (Open Data Barometer, 2016).	43
Figure 16: Netherlands (green) versus UK (purple) – 2015.....	44
Figure 17: Radar chart of scaled sub-component scores. Comparison of UK and Europe average 2013.	45
Figure 18: Open data supply policy areas the Netherlands 2013-2015 (Algemene Rekenkamer)	45
Figure 19: supplier network of all the eleven ministries of the Dutch government	46
Figure 20: Place of FAIR in CRISP-DM.....	47
Figure 21: Place of FAIR in KDD Process	48
Figure 22: Data Scouting Process	49
Figure 23: Older set up Data Scouting Process	49
Figure 24: The Data Warehousing Process @ Berenschot Intellerts.	52
Figure 25: Taxonomy of Open Data Sources	54
Figure 26: Data Repository Finance	56
Figure 27: Data warehouse Structure Finance	58
Figure 28: Data Warehouse Structure Energy project	59
Figure 29: Where corporate knowledge within organizations is stored (Marco, 2000)	61
Figure 30: Overlap between DSA and FAIR requirements	65
Figure 31: Example FAIR Profile Dataset X	65

Figure 32: Visualization Model Construct	67
Figure 33: Model 1 FAIR Activities.....	69
Figure 34: Model mutual relations FAIR concepts	70
Figure 35: Final Reference Model 2	73
Figure 36: Draft Reference Model 2.....	74
Figure 37: Outer Circle Reference Model.....	75
Figure 38: Middle Part Final Reference Model.....	76
Figure 39: Example Data Repository (Based on the Case Study Project).....	78
Figure 40: 5 Star Model Open Data, Tim Berners-Lee.....	81
Figure 41: Repository Energy Project Berenschot Intellerts	83
Figure 42: Data Repository Gemeente Amsterdam	84

1. Introduction

In recent years, the digital transformation of business models and the enormous growth of data ensured that data became and becomes more and more the key for companies to enable new products and services. The economy becomes data-driven. We observe also in academia the desire for transparency, and therefore the need for data storage and publication, is increasing. The origin of more and more data repositories in both areas is then a logical consequence.

In many cases data repositories are for internal usage. However, the awareness increases that open data repositories, and the interoperability between different sources, can ensure faster developments and innovations. Well known open data repositories can be found at CBS (Statline), Kadaster, and KNMI, but there are many more initiatives. The next step, after making data open and available for external/public use, is taking care about the consistency and quality of data.

We state that quality of data roughly can be divided into two different aspects: the quality of the content, and the quality regarding the meta data. These two are not totally separate from each other; when the quality of meta data is high (i.e. information about the data is complete and accurate) it is easier to determine the content quality. Although in first instance the approach regarding data repositories is often a quantitative approach (the more publications, the better), the meta data management side, the more qualitative side, gets more and more attention.

Good meta data management is the key in determining quality of data. Thereby, it leads to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse (Wilkinson, Dumontier, Aalbersberg, Appleton, Axton, Baak & Bouwman, 2016). Wilkinson et al. (2016) provide in their paper four guiding principles for scientific data management and stewardship, the so called FAIR principles. These principles act as a guide to evaluate whether digital artifacts are Findable, Accessible, Interoperable, and Reusable. Thereby these principles help to maximize the added-value gained by contemporary, 'formal' digital publishing. The intention of the researchers is that the principles are not only applicable to 'data in conventional sense', but also to the algorithms, tools, and workflows that led to that data.

We observe in accordance to Doorn (2017) that there is a need for the translation of these theoretical principles into practice. Thereby, we assume that it is also relevant for business to use such guidelines in meta data management. Therefore, in Figure 1 (see next page) we provide the following structure of this research. We distinguish three main parts, which correspond to the research questions (discussed in next chapter): A systematic literature research on the FAIR concepts (RQ1), the context where FAIR can be applied in (RQ2), and the requirements for measuring whether data repositories are FAIR (RQ3). The first Research Question results in four 'new' definitions on the FAIR concepts. Together with the result of research question 3, a set of Meta Data Attributes, this forms the basis for a Reference Model (Artifact 2). The context where this model can be applied in is the Data Scouting Process (Artifact 1). We set up this artifact, as result of Research Question 2, based on a case study at Berenschot Intellerts. During this case study, we also applied the reference model into practice. The artifacts are evaluated based on iterative expert evaluation at Berenschot Intellerts (Artifact 1), interviews at CBS and Gemeente van Amsterdam (Artifact 2), and a survey among experts from business and academia (Artifact 2). The research questions will be discussed in logical order in this document, however it is good to keep in mind that the results from research question 2 and 3 have been elaborated parallel.

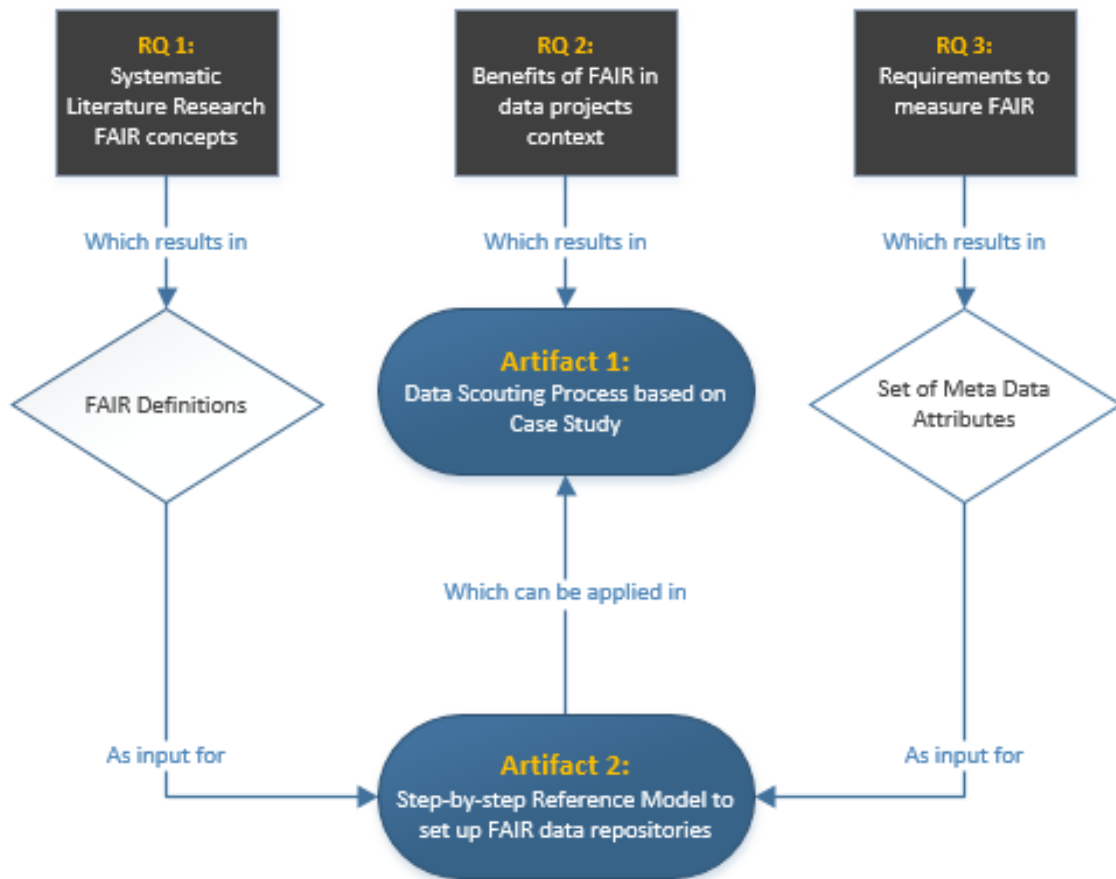


Figure 1: Structure Thesis

1.1 Problem Statement

In the main introduction, we already discussed a bit the relevance of the research topic. In this section, we will provide some more background information which results in the formal problem statement of this research.

The simplest definition of meta data is: data about data. A more elaborated version is: “Metadata is the description of the data as it created, transformed, stored, access, and consumed in the enterprise” (Sherman, 2014, p. 79). Meta data is from added value from business and IT perspective.

According to Sherman (2014) meta data management is not at the top of most BI teams or business people’s priority lists. However, handling metadata in a good way is essential to implement and manage technologies and products. Business people need to know how the data looks like which they want to use for their business analytics, and IT people need to know what happened to the data to provide consistent, comprehensive and clean data for (deeper) business analytics. Metadata management is the way to determine the quality of data in a structured way (Sherman, 2014).

Regarding the four principles of Wilkinson et al. (2016) one of the big challenges of data-intensive science is to improve knowledge discovery through both humans and their computational agents. These principles act like a guide to evaluate whether digital artifacts are Findable, Accessible, Interoperable, and Reusable. Also in their opinion, good data management is the key which leads to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse. Therefore, they distinguish multiple stakeholders (from business and academia) who deal with the same difficulties. For example, researchers who want to share, get credit, and reuse each other’s data and interpretations, but also software and tool-builders providing data analysis and processing services such as reusable workflows.

The definitions of Wilkinson on the four concepts can be found in Figure 2. Additionally, the term interoperability is in the paper interpreted as: “the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort” (Wilkinson et al., 2016).

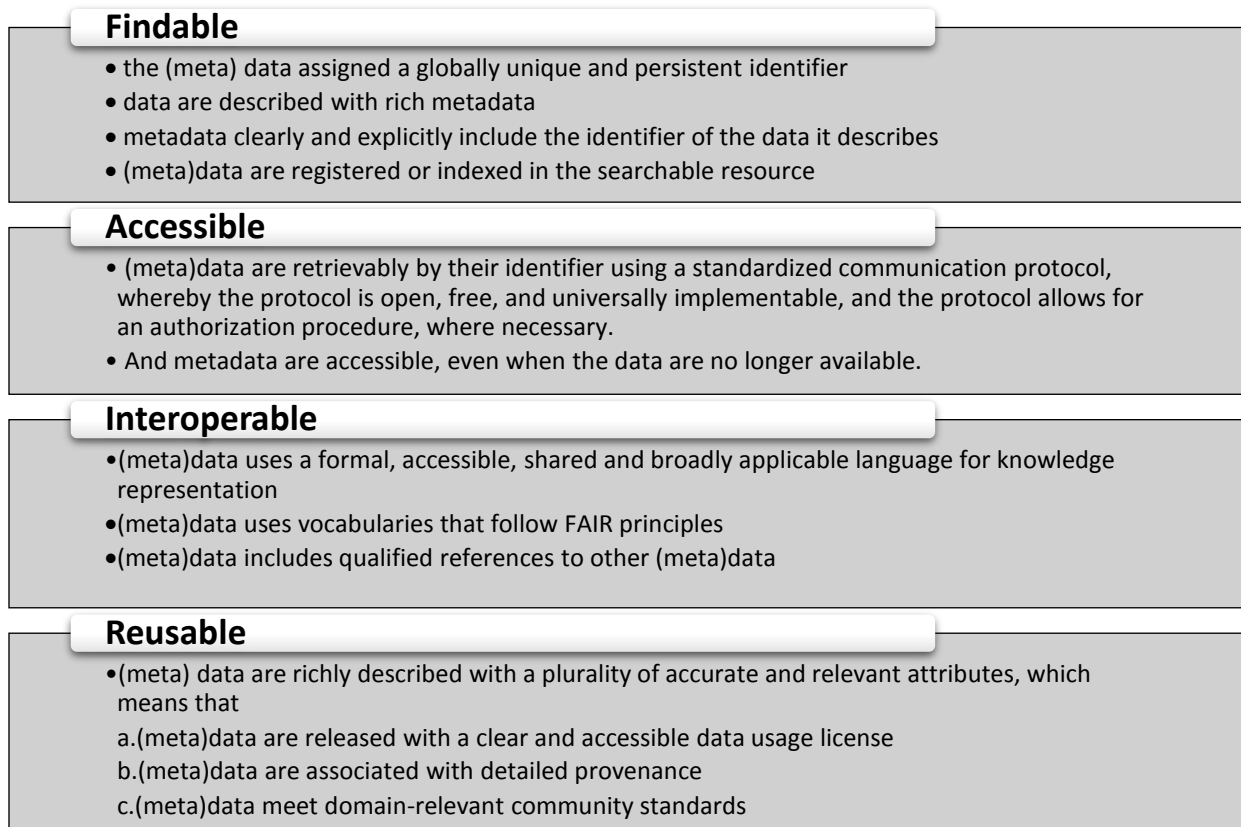


Figure 2: Characteristics of the four FAIR principles as described by Wilkinson et al. (2016).

As already said, business and academia deal with the same difficulties regarding good meta data management. The FAIR guidelines are provided in the academic area, however there is a comparable set of guidelines from business perspective; the five C's of Sherman (2014). According to him data should be 'whipped into shape', with the results that the data is: Clean, Consistent, Conformed, Current, and Comprehensive. Sherman (2014) describes the definitions provided in Figure 3.

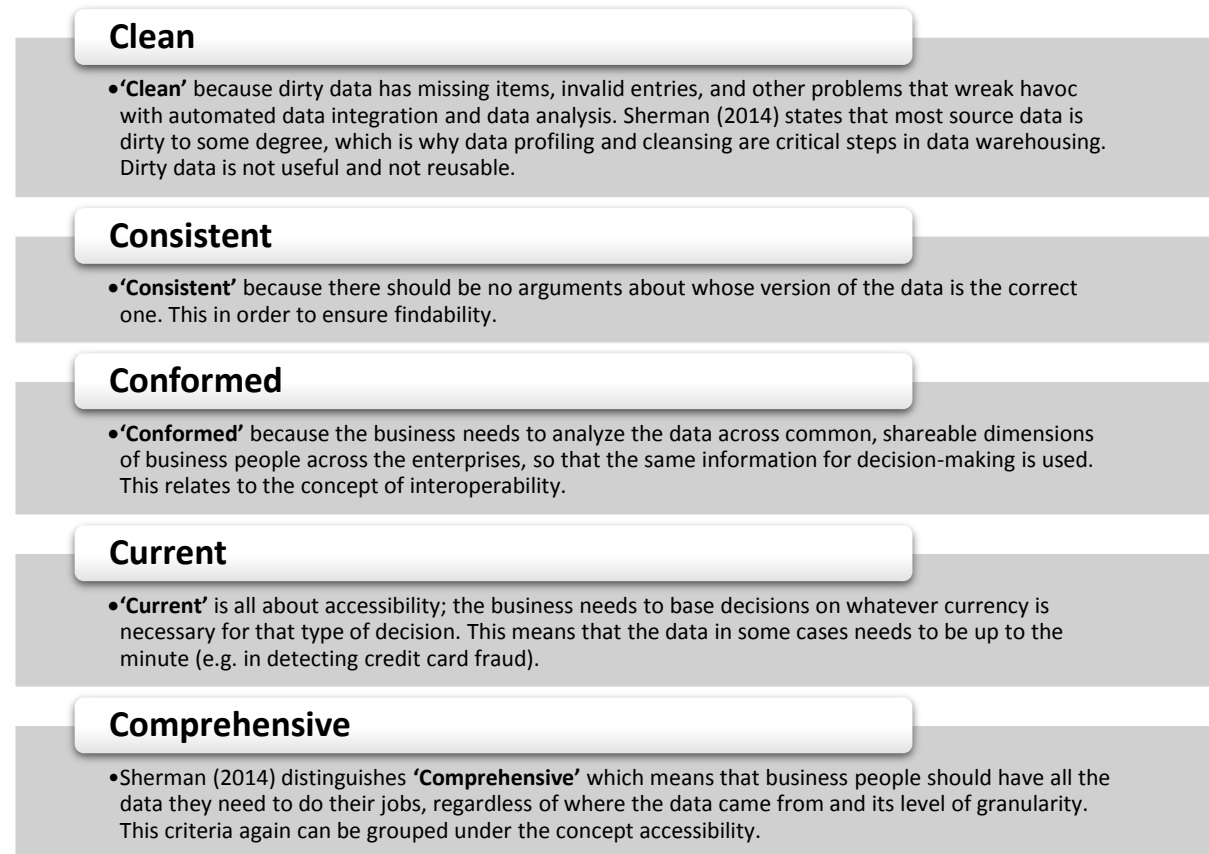


Figure 3: Characteristics of the five C's (as described by Sherman, 2014).

We can conclude that stakeholders, both from academia and business, want to be transparent and create openness (Wilkinson et al., 2016), and know good meta data management is essential to indicate the data quality (Sherman, 2014, Batini, Cappiello, Francalanci & Maurino, 2009). Two examples of guidelines are elaborated: the FAIR principles and the five C's of Sherman. However still it seems that meta data management has low priority. Therefore, there is a need for clarity how to apply these theoretical guidelines into practice. The problem statement this research addresses can be summarized as follows:

Given the data-driven economy there is a need for openness and transparency in business and in academy. The awareness that good meta data management is essential to indicate data quality is increasing, however, meta data management still has low priority in many organizations. To bridge this gap (between theory and practice) there is a need for a set of practical steps in how to deal with meta data management in (open) data repositories.

1.2 Scope Thesis

Scope systematic literature research (RQ1)

In this research, we provide a systematic literature research, this means that the scope of the literature part is: all papers about (one of) the FAIR concepts. In Appendix B, we provide the steps in the process of determining the useful papers for this research. Based on reading abstracts in first instance 47 papers seemed to be useful for this research. We import these papers in Nvivo, and marked relevant parts per topic. In the end, we used 26 of those papers to set up the concept tables provided in Appendix A.

Scope Meta Data Management and Case Study (RQ 2 + RQ 3)

The broader scope of this research is data governance. According to the DAMA DMBOK2 Framework (Cupola, Earley, & Henderson, 2014) data governance can be divided into ten components: Data Governance, Data Architecture, Data Modelling and Development, Data storage and Operations, Data Security, Documents and Content, Reference and Master Data, Data Warehousing and BI, Meta-Data, and Data Quality. As described in the problem statement, there is in business and academia a need for good meta data management. Which means structured meta data description to indicate quality of data. Therefore, the scope for this research is Meta-Data. We provide the underlying components of the Meta-Data dimension based on the DAMA-DMBOK Functional Framework (Mosley, 2008) in Figure 4. The final reference model is a model which ‘supports meta data reporting and analysis’ (see Figure 4 final component). The other components are discussed in the theoretical part, or during the case study. For example, we set up a FAIR repository for Berenschot Intellerts which means we ‘Implement a Managed Meta Data Environment’. We do not to follow the components in Figure 4 strictly, but it provides a clear guidance. Additionally, regarding research question 2 (about the context of FAIR) the scope is open data projects in the Netherlands. And the Data Scouting Process (Artifact 1) we provide is based on open data projects at Berenschot Intellerts, however, it is general applicable. This also applies for the open data taxonomy we provide. The scope for the Case Study we carried out at Berenschot Intellerts (i.e. application of our reference model (Artifact 2)) comprises the steps of our own Data Scouting Process (Artifact 1).

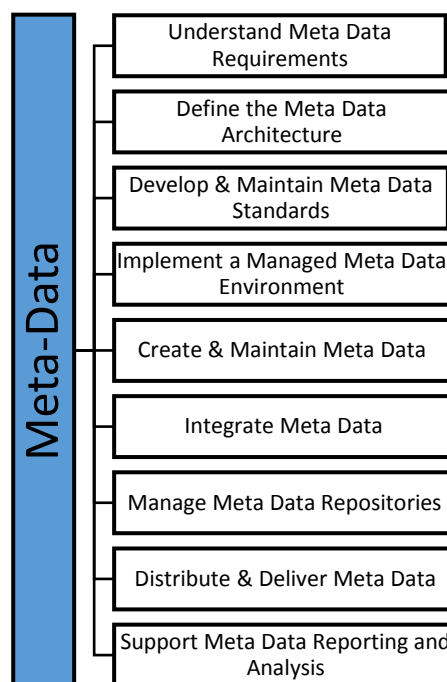


Figure 4: Components Meta-Data dimension DAMA-DMBOK Framework (Mosley, 2008)

1.3 Example Organization: Berenschot Intellerts

A few times we mentioned already the company Berenschot Intellerts. Berenschot Intellerts is the example organization in this research. Berenschot Intellerts is a separate B.V. which operates under the flag of Berenschot. Berenschot is a consultancy company in Utrecht which has almost 80 years of experience in organizational development and consultancy. They provide advice for public and private organizations all over the world to solve problems, realize new strategies, to increase achievements and to develop human capital. Berenschot Intellerts B.V. combines this experience with artificial intelligence solutions. With their expertise in big data, machine learning, and software engineering they deliver data driven solutions for clients. Based on the data they get from their clients, in combination with open data and data from partners, they provide interactive dashboards. These dashboards contain a heavy data component from a width array of sources, which are combined by Berenschot Intellerts as well. The Business people of Intellerts (including data scientists) are situated in Utrecht, the Netherlands, and the IT people in Kaunas, Lithuania. During this research, they provided expert evaluation (from data scientist perspective). In chapter 4 and 5 we elaborate more detailed over why their expertise is relevant for this research, and why this research is relevant for them. Additionally, their website provides some more background information: <http://intellerts.com/> .



2. Research Approach

2.1 Research Questions

The main research question is:

Which requirements should a data repository meet to satisfy the FAIR principles?

Three sub-questions are formulated:

1. What are the definitions and interpretations of the FAIR concepts in literature?
2. What is the benefit of FAIR in the context of open data projects?
3. What are the meta data requirements that satisfy the FAIR principles?

In Figure 5 we provide, corresponding with the structure of the thesis (Figure 1), the structure of this document. The three main parts correspond with the three research questions described above, and the links are the same as explained in the Introduction. Additionally, the main goal of Artifact 1 is to determine the place of FAIR in the context of open data projects. The main goal of Artifact 2 is to provide a step-by-step set of activities to ensure a data repository contains all aspects to measure FAIR, with the focus on comprehensibility, and accessibility in public and private sector.

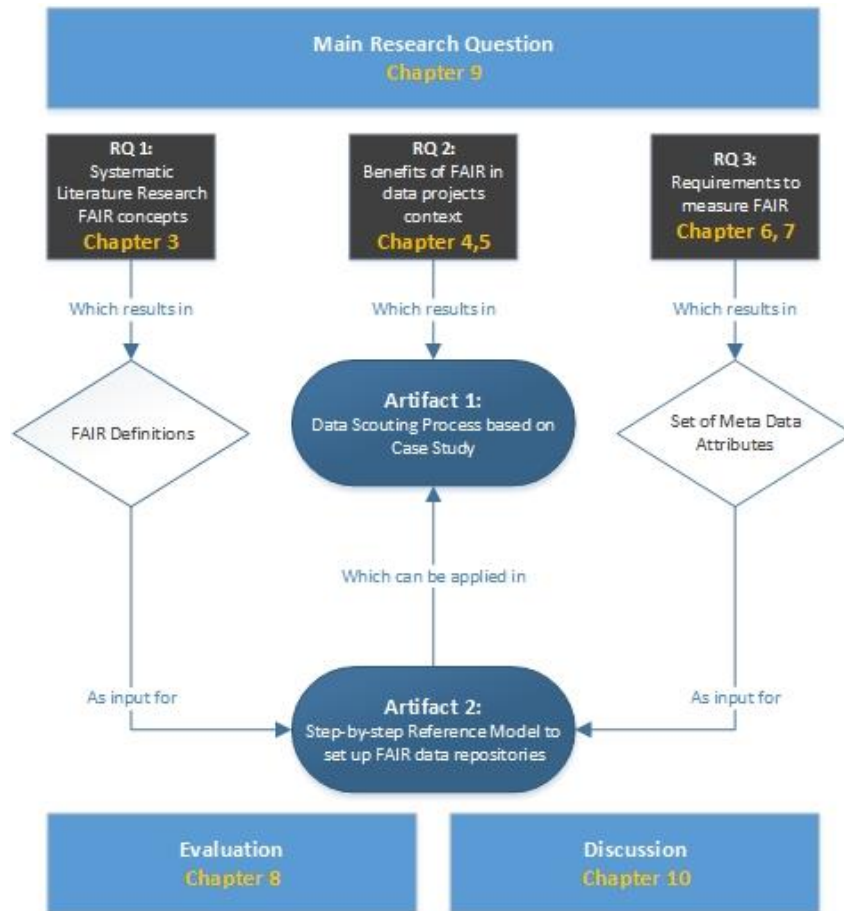


Figure 5: Visual representation of document structure.

2.2 Research Method

We determine Design Science Research is the most appropriate method for this research. This because “Design science addresses research through the building and evaluation of artifacts designed to meet the identified business need” (Hevner, March, Park & Ram, 2004, p.79). In our case there is a clear business need in companies and in academia as described in the problem statement. Also, the result of design science research, “a purposeful IT artifact created to address an important organizational problem” (Hevner et al., 2004, p. 82) applies. Finally, Design Science Research is chosen because it has a clear structure/deliverables and evaluation/consultation is very important.

Hevner et al. (2004) presented an Information Systems Research Framework. Two key components are the Environment and the Knowledge Base. The environment defines the problem space in which reside the phenomena of interest. And the knowledge base provides the raw materials from and through which IS research is accomplished. The main foundation for this research are the FAIR principles.

In Figure 6 the Information Systems Research Framework is applied to this research. We present two main artifacts: Data Scouting Process and FAIR Reference model. And we evaluate those by conducting interviews, a survey, a case study, and via iterative expert evaluation.

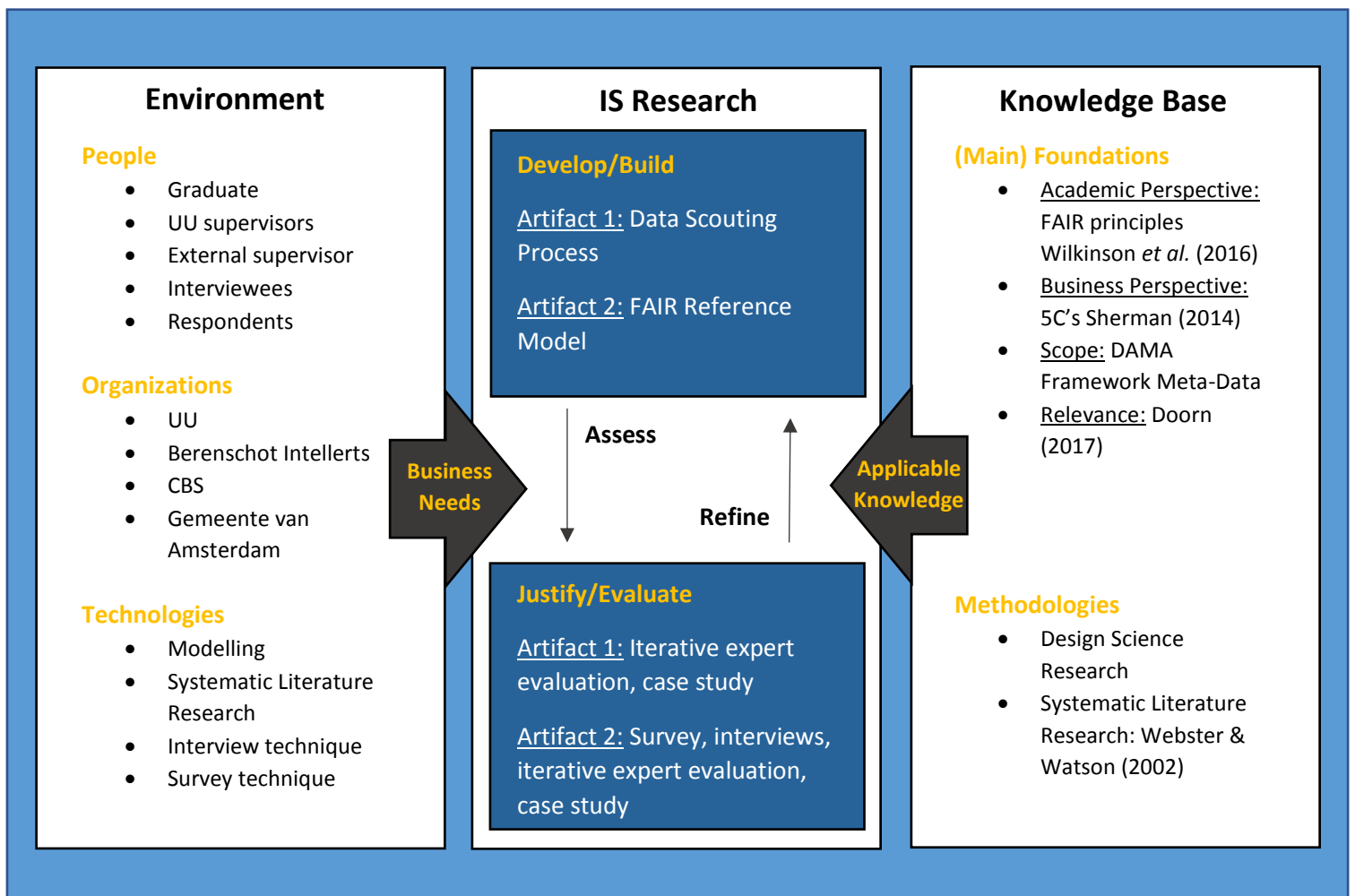


Figure 6: information Systems Research Framework applied to own research.

2.3 Research Design

The research design is an implementation of the research method. For this design, the DSR Cycle of Vaishnavi & Kechler (2004) is used. See Figure 7 a visualization of the Cycle applied to this study. The Cycle consists of five steps:

1. Awareness of problem
2. Suggestion
3. Development
4. Evaluation
5. Conclusion

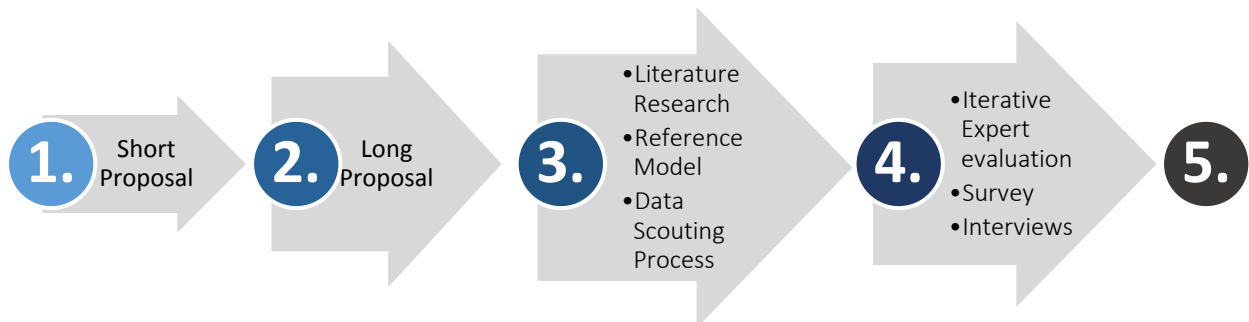


Figure 7: Design Science Research Process Model (Vaishnavi & Kuechler, 2004)

2.3.1 Literature Research Protocol

For the literature research, we follow the approach for a systematic literature research of Webster & Watson (2002). They provide suggestions on how to execute a review. Thereby, the study of Cram, Brohman & Gallupe (2016) is used as an example of a paper which is also based on the Webster & Watson's approach.

For this literature review 47 papers are systematically determined and categorized per theme (see Appendix B). These are further annotated in Nvivo per concept. There are four concepts: Findability, Accessibility, Interoperability, and Reusable, the so called FAIR principles. Based on the results of this analysis two concept tables will be provided. Table 1 describes which concepts can be found in which papers, and in table 2 the descriptions/definitions per concept per paper can be found. Additionally, we provide per concept a table like table 3: the main characteristics per concept with the papers where they are described. The concept tables (like template table 1 +2) can be found in Appendix A, and the concept table like template table 3 we provide in the literature research after each concept.

Authors	Concept1: Findability	Concept2: Accessibility	Concept3: Interoperability	Concept4: Reusability
Paper 1	x			X
Paper 2	x	x	X	X
Paper 3	x	x	X	

Table 1: Template table concepts per paper

Authors/ Concept definitions	Findability	Accessibility	Interoperability	Reusability
Paper 1	Definition/ Description			Definition/ Description
Paper 2	Definition/ Description	Definition/ Description	Definition/ Description	Definition/ Description
Paper 3	Definition/ Description	Definition/ Description	Definition/ Description	

Table 2: Template table definitions/descriptions per paper

Authors/ Characteristics concept	Characteristic 1	Characteristic 2	Characteristic 3	Characteristic 4
Paper 1	x			x
Paper 2	x	x	x	x
Paper 3	x	x	x	

Table 3: Template table characteristics concepts per paper

2.3.3 Evaluation Design

The evaluation phase is in Design Science very important. Therefore, evaluation and validation consist of three components: iterative expert evaluation, a survey, and interviews. The iterative expert evaluation is done by the external supervisor (data scientist) from Berenschot Intellerts. Especially for the two artifacts he provided his feedback. Second, the definitions and the input for the model is validated via a survey among IT professionals and FAIR experts. The total number of respondents is 21. The questions and results of this survey can be found in Appendix C. Finally, the model is evaluated during interviews with three interviews from data supplier perspective (Gemeente Amsterdam and CBS). We discuss the evaluation set up and results more detailed in chapter 8.

Goals iterative expert evaluation

The main goal of the iterative expert evaluation is to get feedback from business perspective. The expert is my external supervisor, a data scientist at Berenschot Intellerts. During this evaluation, the whole thesis is evaluated, and especially artifact 1.

Goals survey

Regarding the survey we distinguish two main goals.

First, during the survey we will evaluate the FAIR definitions, as result from our systematic literature research. These results will be immediately provided after each principle in chapter 3, the literature research.

Second, the survey is used to rank a set of meta data attributes, and to evaluate whether the original set of attributes is complete. This results in a set of numbers (0-5) which provides us the relevance per attribute due to the respondents. These results are provided in chapter 7 and 8. The respondents are FAIR experts from academia, and experts on meta data management from business.

The definitions and the data regarding the attributes is the main input for the reference model, artifact 2.

Goals interviews

The main goal of the interviews is to evaluate the definitions, the set of meta data attributes and the models from the open data supplier perspective, since this are the real experts regarding open data and therefore their feedback is essential in this research. The results of the interviews are provided in chapter 8.

2.4 Academic relevance: open science

In their paper McKiernan, Bourne, Brown, Buck, Kenall, Lin, ... & Spies (2016) state that the benefits of open science outweigh the potential costs. They motivate this by elaborating on benefits like: open publications gain more citations, open publications get more media coverage, new projects and collaborators are easier to find, which indirectly leads also to better job and funding opportunities. Not everyone in academia is as enthusiastic about open science as McKiernan et al. (2016). It raises especially critical questions about how to formalize such openness. How can we ensure things are not only available, but also useful? By whom should openness be practiced; early career researchers, established professors, everyone? There is also disagreement on the moment research should be open; at the very beginning, before publication, or only after? (Levin, 2015). However, overall researchers realize that this age asks to do something with the possibilities to ensure transparency and to prevent and avoid abuse of science. Open science is therefore widely promoted as a key component of modern society such as The WellcomeTrust, Royal Society, National Institute of Health, Center for Open Science, Open Knowledge Foundation and much more. Thereby, in the United Kingdom and the United States it already becomes more and more an important theme in science policy (Levin, 2015). Yet it still appears to be a complex change. Nosek et al. (2015) describe the current situation as a "classic collective action problem". There is a lack of strong initiatives by individual researchers to be more transparent. Thereby, they conclude that there is unfortunately no centralized means of aligning individual and communal incentives via universal scientific policies and procedures.

The FAIR principles are such an initiative. However, we observe that due to the FAIR principles also clarity is needed; clarity and agreement on definitions and interpretations of the concepts. Formalization on how to 'create' open science is desirable, to ensure things are not only available but also useful. Thereby the recommendation of Doorn (2017) is the starting point. According to him the first aim is to get agreement on the FAIR principles, and in addition there is a need to make the transition from theory to practice. Therefore, this research is also about the tangibility of these guidelines, because we assume that in that case FAIR can also become an indicator for data quality.

2.5 Business relevance: open innovation

In the first place, these principles are set up by Wilkinson et al. (2016) for science (especially life science), but we can assume that on a meta-data level business deals with similar problems as science. From business point of view it is about open data innovation with the goal to understand how open data can be harnessed to provide unique and valuable insights. Data innovation can be divided into three open innovation approaches: inbound open innovation, outbound open innovation, and coupled processes (Gassmann and Enkel cited in Cui, Teo & Li (2015). Cui et al. (2015) recognize that advances in IT have enabled firms to increasingly rely on open innovation. However, there is a lack of theoretically driven research on how IT impacts organizational open innovation performance. In their research Cui et al. (2015) assume that the effect of search openness on organizational innovation performance will depend on its alignment with organizational IT strategies. Two main important IT strategies are IT integration and IT flexibility. IT flexibility enables firms to quickly and economically adapt IT applications to support evolving knowledge sharing requirements with external sources. And IT integration facilitates the timely and idiosyncratic exchange of knowledge with collaborative partners. And this is also the matter where the guidelines are about. Cui et al. (2015) conclude that the alignment between IT flexibility and search openness enhances innovation radicalness and innovation volume, and the alignment between IT integration and search openness positively affects innovation volume. This means business has benefit from a smooth alignment between the IT strategies and search openness. Therefore, from business perspective this research is relevant to explore how among other things this search openness can be accomplished.

Also, governments can generate social and economic values by using data-driven innovation processes. Due innovation processes within governments there is many times the difference of countries in the level of open innovation maturity model of open data provision and usage is disregarded. This is a specific point of attention. Therefore, researching meta data management (within governments and companies) is interesting, to answer questions like: which international meta data standards are available and useful to overcome differences in approach between countries. As described already in the problem statement, unfortunately meta data management is often not at the top of business people's priority lists (Sherman, 2014). This is a pity, because handling meta data is essential to implement and manage technologies and products. Business people need to

know how the data looks like they want to use for their business analytics, and IT people need to know what happened to the data to provide consistent, comprehensive and clean data for (deeper) business analytics. Thereby it is also valuable for the clients, since it causes transparency in the process their data goes through (Sherman, 2010).

3. Research Question 1: Systematic Literature Research

WHAT ARE THE DEFINITIONS AND INTERPRETATIONS OF THE FAIR DATA CONCEPTS IN THE LITERATURE?

WHY Systematic Literature Research on the FAIR concepts?

In this first research question, we research the FAIR concepts – Findability, Accessibility, Interoperability, Reusability – separately. This is useful since the FAIR concepts are defined by Wilkinson et al. (2016). in the specific context of using them together, but it is also interesting to see how these concepts are defined separately in literature. The goal is not to define completely new definitions on the concepts, but to improve the existing definitions. Although when bigger changes are needed, because something cannot be confirmed with other research, we will do. The main goal of this research question is to get more agreement on the definitions, because they are grounded on a more thorough literature research. The steps which are followed are already described in section 1.2.

3.1 Data Quality in general

In the introduction, we distinguished already two aspects to make an indication on the quality of data, namely: the quality of the content, and the quality regarding the meta data. Thereby, we concluded there are many different stakeholders (from business and academia) when it comes to indicate data quality. This is also underlined by Chatfield, Akemi, Reddick & Al-Zubaidi (2015). They conclude that all (big) data users must have some focus on data quality. In practice, this means often that it is about huge amounts of data. And this makes selecting on quality complex; which data to dismiss, how to select the most appropriate data, and eventually how to evaluate the value of the data. Before we elaborate more on the FAIR concepts separately we give some insight in what the term data quality as whole comprises according to literature, and some initiatives to guide measuring data quality are mentioned. This to provide some background information.

Data Quality Research Themes

An example of an initiative on data quality management comes from Jayewardene, Sadiq & Indulska (2012). They distinguish based on a literature review seven key data quality research themes: Data Quality Assessment, Data Quality Framework, Data Modelling and Design, Data Integration and Linkage, Data Constraints and Rules, Data Lineage, and Data Acquisition and Presentation (see Figure 8). Batini et al. (2009) conclude in addition that the quality of data is a relevant performance issue of operating processes, decision-making activities and interorganizational cooperation requirements. By conducting a survey Jayewardene et al. (2012) concluded that seventy per cent of the participants (business people) indicated these factors as ‘high’ or ‘very high’ to achieve good data quality within organizations. However, only 20-30% of them was satisfied about how these factors are implemented. And therefore, was asked for the most significant factors. These were: Data Quality Assessment, Data Quality Frameworks, and Data Constraints and Rules. Nevertheless, we can conclude that these themes of Jayewardene et al. (2012) are still quite abstract from practical perspective.

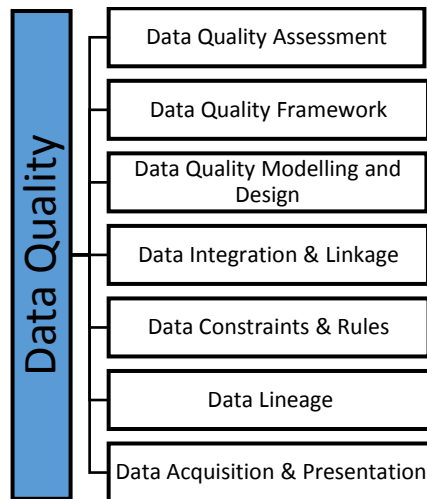


Figure 8: Seven key data quality research themes (Jayewardene et al., 2012)

Changing context

The issue of data quality is not only abstract, it has become also increasingly complex. This because of a fast-changing context; information systems have been migrating from a hierarchical to a network-based structure, where the set of potential data sources that an organization can have is increased in size and scope (Batini et al., 2009). Therefore Batini et al. (2009) refer in their paper to basic set of data quality dimensions of Scannapieco & Catarci (2002): accuracy, completeness, consistency, and timeliness. These factors are so important, especially when we realize that 40 per cent of the material on the net disappears within one year, another 40 per cent is modified, and only 20 per cent stands in original form, which shows again the relevance of determine data quality (Rao,2003 in Batini et al., 2009). The idea of Scannapieco & Catarci (2002) is in line with a more recent study of Fürber & Herp (2011). Their semantic web information quality assessment framework of is based on more or less the same factors, namely: accuracy, completeness, timeliness and uniqueness. And this reminds and is related in interpretation to the four principles: findability, accessibility, interoperability and reusability.

Performance Issue

The definition of Batini et al. (2009) also contributes to a better understanding of the concept Data Quality. According to them data quality can be seen mainly as a performance issue. A performance issue of: 1. Operating processes, 2. Decision-making activities, and 3. Interorganizational cooperation requirements. This is an interesting definition because it shows similarities with the FAIR concepts as well. First of all, interoperability comprises the issue 'operating processes'. Second, 'decision-making activities' means the information about what happened to the data, the so-called provenance of data. Provenance is very important in meta data management, and relates especially to the concept Findability. We provide more details in the next section on this. Finally, 'interorganizational cooperation requirements' is basically what we are looking for due to the main research question; interorganizational meta data requirements (in this research: which satisfy FAIR) for a data repository.

Quality Assessment & Quality Frameworks

Coming back on the framework provided in Figure 8, the dimensions Data Quality Assessment and Data Quality Frameworks are most important regarding this research. Because Data Quality Assessment refers to investigating and measuring data related problems within organizations with the aim to implement data quality improvement strategies, and Data Quality Frameworks are simply the practical translation on how to manage this within organizations (Jayewardene *et al.*, 2012).

A research where Quality Assessment and a Quality Framework comes together is in the paper of Fürber & Hepp (2011). In their paper, they provide a framework for information quality assessment of Semantic Web data called SWIQA. SWIQA employs data quality rule templates to express quality requirements which are automatically used to identify deficient data and calculate quality scores. They identified five dimensions (see Figure 9) that can be measured by applying the data quality rules they set up as well: syntactic accuracy, semantic accuracy,

completeness, timeliness, and uniqueness. SWIQA may be used by data owners to keep track of the quality of their data, and by data consumers to find high quality data sources.

For now, the main conclusion is that the dimensions mentioned by Jayewardene et al. (2012) and the dimensions of Fürber & Hepp (2011) show similarities mutually, but also with the FAIR concepts which is important for the foundation of the FAIR concepts. In the next sections, we discuss further on the concepts separately, and then we mention also important differences between the concepts and the interpretations in literature.

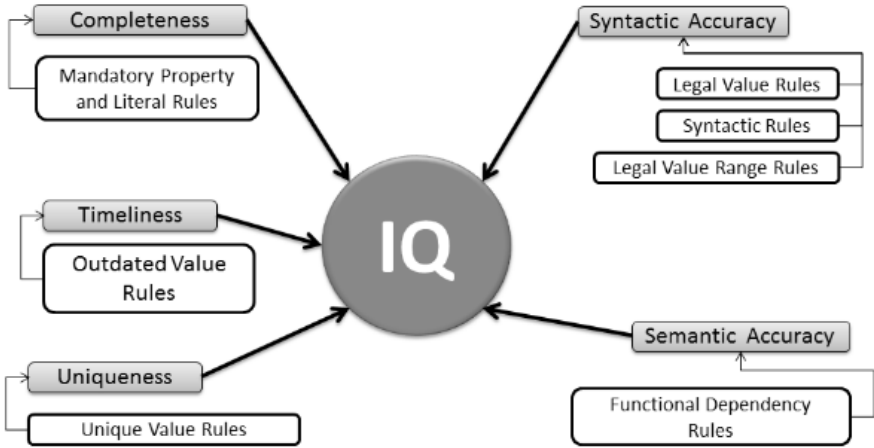


Figure 9: Proposed configuration of data quality rules for information quality assessment

3.2 Findability

According to Wilkinson et al. (2016) data is findable when the (meta)data 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, and (meta)data are registered or indexed in the searchable resource. We assume that in general the term findability is known as ‘searchability’. However, the definition of Wilkinson et al. (2016) is focused on the meta data perspective on findability, therefore this is in first instance the starting point of this research as well. In this section, we discuss three founded characteristics of findability, based on different literature studies: ‘Versioning’, ‘Provenance’, and ‘an Unique Identifier’.

Versioning

Sheridan & Tennison (2010) provide some guidelines to ensure a web of linked government data (data.gov.uk). These guidelines are to give data publishers a direction along the way, and have some overlap with the FAIR principles. Thereby, it is in our opinion a proper research and especially from added value because it explains the cohesion between different (in other studies separate) aspects of findability. Especially three of them are interesting regarding findability: ‘URI’, ‘Versioning’, and ‘Provenance’. Whereby versioning is “*the process of assigning unique names or unique version numbers to unique states*”¹. In practice, this means according to Sheridan & Tennison (2010) that while many sources may provide information about a given resource, only one should provide authoritative information about a property of that resource. These sources then can be combined to give slices of information at a point in time. Generally, consumers will be interested in the current state of the world, but policy makers will look back into the past and project into the future. Therefore, we conclude that versioning is especially important regarding government data. Because government data is a typical example of data consisting of different sets, updated and modified at different times, with potential overlap between different sources. This makes it difficult to distinguish which information is the most important (Sheridan & Tennison, 2010).

Provenance

Second, ‘provenance’ is a concept to focus on according to Sheridan & Tennison (2010). According to them provenance comprises the provenance of the data itself but also the way it is manipulated. But not only they mention this, the term is shard wider. Also, the definition of Wilkinson et al. (2016) comprises this concept by including the part ‘data are described with rich metadata’. At the same time, this explains the goal of provenance: rich meta data. We conclude that in general provenance of data is about “the tracking of historical information”, like metadata about what, when, where, how, by whom, and why a data set is created (Morau *et al.*, 2011). Since Data Quality is a subjective concept (data which is good for one organization can be bad for another), the context of data is essential to evaluate the quality of data (Malaverri, Mota & Medeiros, 2013), and thus this ‘contextual information’ must be easy to find for the data user. The question is then: how to associate provenance metadata with datasets? Chong, Skalka & Vaughan (2015) present in their paper an ambitious approach to realize this. They provide a scheme for embedding a provenance identifier in environmental datasets, that associates meta data with datasets in a manner that does not rely on external structure (like XML formats or database schema). The identifier could be for example an URL link to a data object with extensive provenance information. Chong *et al.* (2015) call it a kind of ‘watermarking scheme’, and call these datasets as being ‘self-identifying’.

Unique Identifier

This approach of Chong *et al.* (2015) brings us to the third aspect of findability which is indispensable: a unique identifier. In the approach of Chong *et al.* (2015) this is an URL link. Batini *et al.* (2009) also argues that the use of URL as unique identifier ensures better findability of data. They refer to a methodology of Cappiello, Francalanci & Pernici (2003) to support the preservation process of the entire life cycle of information, where in the publishing stage (new Web page replaces an old one) the volatility of old data is evaluated. And when the old data is still valid, it is associated with a new URL. According to the definition of Wilkinson et al. (2016) this aspect is described as ‘(meta) data assigned a globally unique and persistent identifier’ and ‘metadata clearly and explicitly include the identifier of the data it describes’. At the same time, we think this shows the main weakness of the current definition of Wilkinson et al. (2016): it is cautiously described that data needs a unique identifier to ensure the

¹ https://en.wikipedia.org/wiki/Software_versioning

findability. Therefore, when we in short translate the combination of the versioning process, the need for provenance, and a unique identifier into practice – in terms of attributes of meta data in a data repository - we propose that every dataset (not only data source!) contains a unique identifier (e.g. a series of numbers). As said, an identifier is crucial to avoid confusion about which version is the correct one, in other words: to determine whether the data is ‘consistent’ (5Cs of Sherman, 2014). Uniqueness is where it is all about regarding findability. Uniqueness is defined as ‘the degree to which data is free of redundancies’ (Fürber & Hepp, 2011). However, it seems findability itself is a term which is hard to find literally in IT related, data (quality) oriented literature. For this study, no literal definitions are founded (as shown in the concept tables in Appendix A). We think the main reason is because findability is a relative new term regarding meta data management, and should not be confused with ‘searchability’. A demarcation for this definition is therefore difficult, but because Kalliknikos, Aaltonen & Marton (2013) describe that ‘immediate findability’ (in other words ‘searchability’) and ‘the effects to findability provided by search engines’ are covered by the concept ‘accessibility’, therefore we decided to not explicitly include this distinction in the definition of findability.

3.2.1 Conclusion

We conclude that findability has all to do with uniqueness; using a unique key to make sure data is separated and findable. A guideline from literature is to make use of URI/URL with the ideal that a database becomes ‘self-identifying’. Also ‘versioning’ and ‘provenance’ are important terms regarding findability. Versioning is the process of assigning unique identifiers to data sources. And provenance is the tracking of historical information about the data, which are both processes to avoid data from redundancies, so to make sure data is unique and there is one source providing authoritative information. Data provenance makes sure that the context of data is considered. This is important since data quality is not always seen as a subjective concept.

Based on these findings we provide the following definition of findability:

FINDABILITY (of data) is about the uniqueness and provenance of data to ensure data sources are searchable, free of redundancies and the context is clear. This is achieved by versioning, associating (new) URLs or by embedding other kind of (provenance) identifiers.

3.2.2 Benefits new definition versus definition Wilkinson et al. (2016)

The new definition of findability is in line with the definition of Wilkinson et al. (2016). For both applies that ‘identifying’ is the keyword, but on several aspects our definition solves the following shortcomings of Wilkinson et al. (2016):

- The definition of Wilkinson et al. (2016) provides a set of requirements to reach findability, but does not include a description of the concept itself.
 - Therefore, our definition includes: ‘Findability (of data) is about the uniqueness and provenance of data’.
- Additionally, Wilkinson et al. (2016) describes how things can be realized, but does not mention relevant processes/terms.
 - Therefore, our definition includes: ‘This is achieved by versioning, associating (new) URLs or by embedding other kind of (provenance) identifiers.’
- The goal of findability is missing in the definition of Wilkinson et al. (2016).
 - Therefore, our definition includes: ‘to ensure data sources are searchable, free of redundancies and the context is clear’.

3.2.3 Concept Table Findability

We summarize the key concepts regarding findability as follows:

Findability					
	Versioning	Provenance	Unique Identifier	Uniqueness	Searchability
Batini, Capiello, Francalanci, Maurino (2009)			X		
Chong, Skalka and Vaughan (2015)		X	X		X
Fürber and Hepp, (2011)				X	
Kallinikos, Aaltonen, Marton (2013)					X
Malaverri, Mota, and Medeiros (2013)		X			
Moreau, Freire, Futrelle, Groth,..., Plale (2011)		X			
Sheridan and Tennison (2010)	X				
Sherman (2014)				X	

Table 4: Detailed Concept Table on Findability

3.2.4 Results Evaluation

According to the survey only 15% (see Figure 10) from the respondents chooses for our definition when we asked which definition describes findability best. The respondents could choose for the definition of Wilkinson, and one other random formulated definition, namely: 'Findability is about the ease with which information can be found using search engines'. This result is quite disappointing, especially when we compare it with the results on the other three fAIR concepts. And since especially the definition of Wilkinson is appreciated by the respondents (55%), a reformulation (see 3.2.5) seems in place to ensure the new definition is an improvement.

However, the 'random' definition is not random formulated. It seems the assumption we did in the beginning of this chapter that the term findability is in general known as 'searchability' is right. Because a notable part of the respondents (30%) interprets findability like this. Therefore, we conclude the concept of Findability is, as already mentioned before, relatively new as concept regarding meta data management.

Therefore, the main improvement on the definition of Findability is that we now explicitly mention the term 'rich meta data', like Wilkinson did. This must ensure that it is clearer for people that findability is not only about 'searchability', and must be read regarding this research context in the context of meta data management. The new definition is:

"FINDABILITY (of data) is about the uniqueness and provenance of data, which means data described with rich meta data with focus on contextual information. This is achieved by versioning, associating URIs or embedding other kind of identifiers"

We discuss the additional results of the evaluation, derived via the interviews, in chapter 8. The results do not differ in the core of what we have described above.

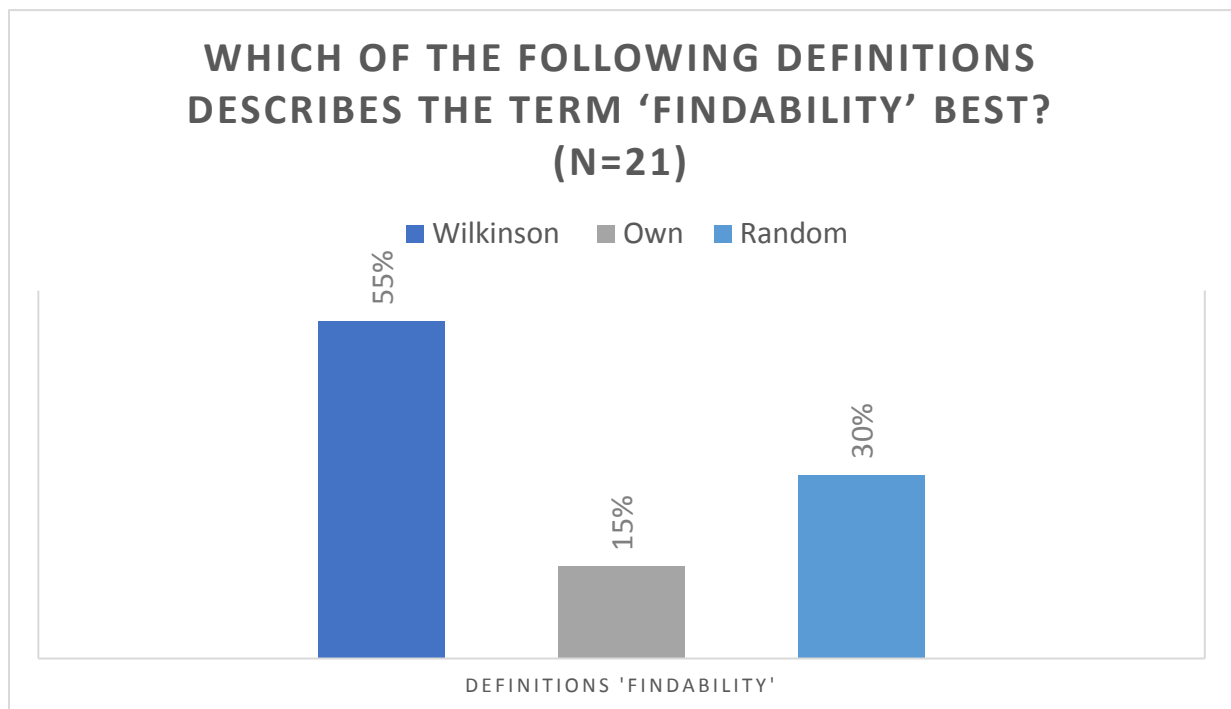


Figure 10: Results Survey regarding the concept Findability

3.3 Accessibility

According to Wilkinson et al. (2016) data is accessible when the metadata is retrievable by their identifier using a standardized communication protocol, whereby the protocol is open, free, and universally implementable, and whereby the protocol allows for an authorization procedure. Thereby, metadata must be accessible, even when the data are no longer available. Already at first sight we conclude that this definition comprises a very detailed, quite specific description of the term 'accessibility'. Are we right according to literature? In this section, we discuss the accessibility concept, especially two characteristics: accessibility from syntactic and from semantic perspective.

Unified Data Access

Matiaško, Záborská, & Záborský (2004) provide in their paper a Unified Data Access Framework. Their main aim is to allow unified data access on the international level for educational, commercial and security purposes. This because they recognize that data stored inside many database systems is under strong pressure to be accessible. For example, databases with different national requirements, habits and history have different structures for the same information. They mention the XML language as a general requirement. Because with XML it is possible to describe, verify, and to manipulate data in the heterogeneous environment. Thereby have XML documents a mechanism for self-description, so called DTD. This is in agreement with the definition of Kallinikos, Aaltonen & Marton (2013) as well: "Accessibility is the functional identity and innovativeness of generative technologies". Both imply that functional mechanisms can ensure that data is accessible. A second requirement recognized by Matiaško, Záborská, & Záborský (2004) to achieve unified data access is that information must be provided without dependency to the current representation, because it must be based on the facility that it can expand. Thus, to achieve accessibility a general language is important, as well as the possibilities for expansion.

Accessibility as semantic quality dimension

However, this view is maybe a bit simplistic. Accessibility might cover more aspects. Harn, Kim, Lee & Choi (2015) provide in their paper an open innovation maturity model for the government. They mention a few challenges regarding open data. And according to them the most identified barriers include lack of comprehensive data policies, lack of validity, completeness of datasets, lack of motivation within public sector, lack of technical and semantic interoperability, lack of technical ability within public and private sectors, and inaccessible datasets. So inaccessible datasets are a challenge. Why this is such a challenge can be explained by the fact that accessibility is more about the semantic side of data quality than that it is syntactical (Olbrich, 2010). This semantic nature can be explained by characteristics of data like 'current' and 'comprehensive'; two of the 5Cs from Sherman (2014). Data needs to be 'current' and 'comprehensive' to have up to date data and to make sure there is access to the data, regardless of where the data comes from and its level of granularity, to base decisions on (Sherman, 2014). Or regarding requirements to ensure accessible data in terms of 'interactivity' and 'data flexibility' (Philips-Wren, Iyer, Kulkarni & Arivachandra (2015). Two other keywords of Fürber & Hepp (2011) regarding data are also in line with these interpretations, namely: the 'completeness' and 'timeliness' of data is important according to them. This needs to be considered to ensure high data quality, and is about the extent to which data are of sufficient breadth, depth, and scope for the task at hand and reflects how up-to-date the data is. In short, we observe accessibility is important and at the same time difficult to realize because in terms of 'unified access' unified requirements, structures, and for example language is needed. Second, accessibility is a challenge because of the 'semantic nature'; accessible data needs to be current, comprehensive, interactive, complete, timely and flexible, which are all non-tangible features.

Nevertheless, accessibility is also explained from other perspectives in the literature. It is the most frequent found and defined term from the four FAIR concepts. And we conclude this is among other things because accessibility is often described as a data quality dimension. For example, Geisler, Weber & Jarke (2016), Jayewardene et al. (2012), Beebe & Walz (2005), and Tilly, Posegga, Fischbach & Schoder (2015) define accessibility as a quality dimension to measure (see Table 5, column 1), describe, and ensure data quality. This interpretation implies that accessibility is not so much a 'characteristic' or 'feature' of data, but accessibility is a dimension which covers multiple, different, 'tangible' metrics. Like accessibility can be measured by the accessibility of a SPARQL endpoint or of an RDF dump. So, accessibility as dimension and a metric as concrete quality measure for a concrete quality indicator usually associated with a measuring procedure (Debattista, Auer & Lange, 2016). However, this interpretation is not inconsistent with the semantic perspective on accessibility, therefore we decide to describe it as a: 'semantic quality dimension'.

Fitness for use

Accessibility is not only a term recognized by researchers, academics, and business professionals. The findings of the research of Wang & Strong in 1996 show that consumers also recognize its importance for quite a long time. This corresponds to the statement of Debattista et al. (2016) that "Accessibility comprises not only availability but also dimensions such as security or performance", since this definition implies that accessibility is a broad term with potential high risks for consumers (e.g. bad security). We conclude accessibility is, as dimension of a dataset, relevant for the consumer. Thereby, citizens itself are more and more a source for e.g. Open Government Data. Although information systems of government agencies are still the main sources, two more sources are gradually emerging, namely citizens and sensors (Charalabidis, Alexopoulos & Loukis, 2016). Therefore, because it is about their own data, their own personal information, citizens become more and more an important stakeholder whether they realize it or not. In their taxonomy of open government data research areas and topics Charalabidis et al. (2016) distinguish the research area 'Open Government Data Infrastructures'. This area includes topics concerning various important technological aspects of the ICT infrastructures developed by government agencies to make Open Government Data accessible to different groups of actors. In this context, the actors are like architectures, APIs provision, and personalization capabilities. Relating to this another research topic in this area is storage and long-term preservation of the data, and the use of cloud services in this domain. We can conclude that in this research accessibility is basically seen in the context of data infrastructures, and it relates to topics like data storage and long-term preservation, whereby it is important data is accessible to different groups of actors. And this is exactly what we also found back in other literature. We would say, accessibility more from a user perspective. Batini et al. (2009) describe it in their definition of accessibility as follows: "Accessibility measures the

ability of users to access data, given their culture, physical status and available technologies, and is important in cooperative and network-based information systems.” (Batini et al., 2009). We can also describe it as the more ‘social perspective’; users have right to access regardless their different context and background. And this is not only about whether access is possible (in terms of availability), it is also about how easy access can be established. Zhang, Jayawardene, Indulska, Sadiq & Zhou (2014) describe it as “ease of use, maintainability and control of the data from end users’ perspective”. Also, Yu (2016) writes about accessibility in terms of an ‘(End) User Acceptance Factor’. Summarized we say it is about ‘fitness for use’. Tilly et al. (2015) use this in their paper. Although they use it as definition of information quality which comprises for example the dimension ‘accessibility’, and their interpretation of ‘fitness for use’ is literally that information can be “easily perceived, interpreted, and applied”.

3.3.1 Conclusion

We can conclude that Accessibility is something which is relevant for researchers, academics, business professionals and consumers. We can conclude that Accessibility is a semantic quality dimension, which comprises especially important aspects for (end) users, like: availability, security, performance, interactivity, and flexibility. Thereby, accessibility is focused on providing access in heterogenous environments regardless context and background. Accessibility must be seen in the context of data infrastructures, and is one of the main challenges regarding open data.

Based on these findings we provide the following definition of Accessibility:

ACCESSIBILITY is a semantic quality dimension, which comprises availability, security, performance, interactivity, flexibility of data, to ensure data access to (end) users regardless their different context and background.

3.3.2 Benefits new definition versus definition Wilkinson et al. (2016)

The new definition of accessibility differs quite a lot with the definition of Wilkinson *et al.* (2016). The main reason for this is because the definition of Wilkinson is too small in comparison with how the concept is elaborated in literature, thereby:

- The guidelines for Accessibility mentioned by Wilkinson et al. (2016) are not found in other literature. Especially a communication protocol is not mentioned.
 - Therefore, we did not conclude this in the new definition.
- Accessibility must be viewed from higher level
 - Therefore, we include that accessibility is a ‘semantic quality dimension’
- Regarding accessibility the (end) User is very important
 - Therefore, we included ‘availability’, ‘security’, ‘performance’, ‘interactivity’, and ‘flexibility’, because this are all indicators for the (end) User to determine whether the data is accessible to them
- Accessibility must also be seen from ‘social perspective’
 - Therefore, we included: ‘regardless their different context and background’.

3.3.3 Concept Table Accessibility

We summarize the key concepts regarding accessibility as follows:

Accessibility					
	Quality Dimension	Semantic perspective	Syntactic perspective	Social perspective	Importance (End) User
Batini, Cappiello, Francalanci, Maurino (2009)				X	X

Charalabidis, Alexopoulos, Loukis (2016)				X	X
Debattista, Auer, Lange (2016)		X			
Fürber and Hepp, (2011)			X		
Geisler, Quix, Weber, Jarke (2016)	X				
Harn, Kim, Lee, Choi (2015)					X
Kallinikos, Aaltonen, Marton			X		
Beebe & Walz (2005)	X				
Matiasco, Zabovska, Zabovsky (2004)		X			
Olbrich (2010)		X			
Philips-Wren, Iyer, Kulkarni & Arivachandra (2015)		X			
Sherman (2014)			X		
Tilly, Posegga, Fischbach, Schoder	X				X
Wang and Strong (1996)					X
Yu (2016)					X
Zhang, Indulska, Jayawardene, Sadiq, Zhou (2014)					X

Table 5: Detailed Concept Table on Accessibility

3.3.4 Results Evaluation

According to the survey 60% (see Figure 11) from the respondents chooses for our definition when we asked which definition describes findability best. The respondents could choose for the definition of Wilkinson, and one other random formulated definition, namely: 'Accessibility is defined as a 'quality dimension' to 'measure', describe and ensure data quality'. We chose this definition as third option, because it only comprises the 'quality dimension' aspect, which occurs in literature the most. However, the results show that our final definition scores best, therefore we decide to not reformulate the definition on accessibility. Which does not imply this definition is perfect. We stimulate improvements, based on even more extensive research.

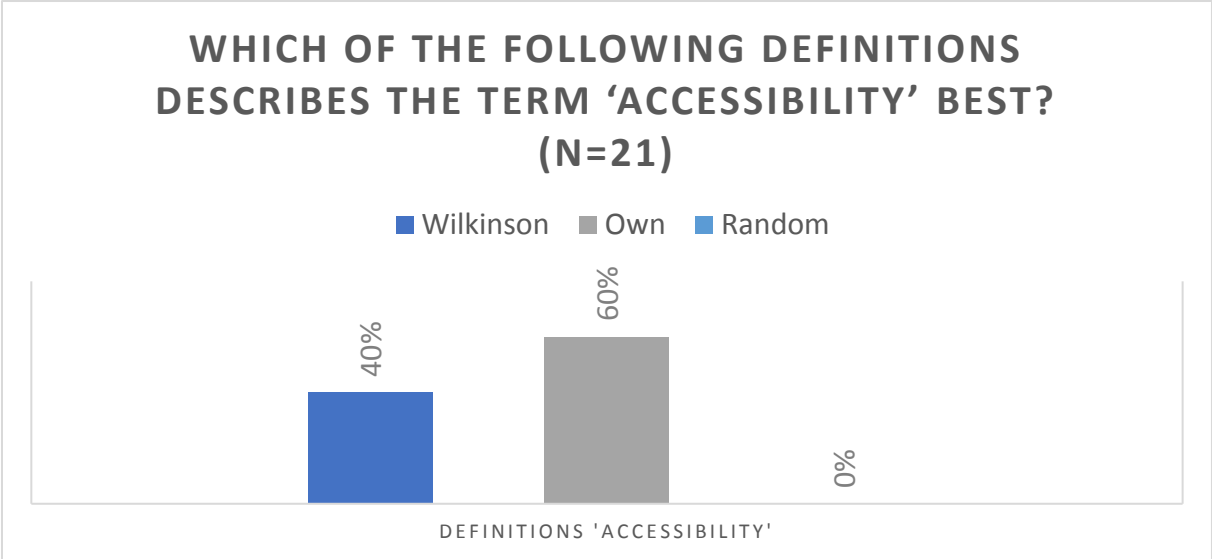


Figure 11: Results Survey regarding the concept accessibility

3.4 Interoperability

According to Wilkinson et al. (2016) interoperability is defined as “the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort”. And data is Interoperable when (meta)data uses a formal, accessible, shared and broadly applicable language for knowledge representation, (meta)data uses vocabularies that follow FAIR principles, (meta)data includes qualified references to other (meta)data. Is this in accordance with literature? In this section, we first discuss the need for interoperability, then related research topics, different types of interoperability, the goal, and the value for of (end) user.

Sharing among heterogeneous environments

Wang, Truptil & Benaben (2015) state that interoperability is one of the key competition factors for modern enterprises, since it describes the ability to establish partnership activities. But we would say not only partnership activities, interoperability has many aspects. These aspects are defined by Shukair, Loutas, Peristeras & Sklarß (2013) as technical, semantic and organizational in nature. According to them interoperability also becomes more and more important, because of the different interpretations of data, the lack of common metadata, and the absence of universal reference data. Thus, the goal and the strength of interoperability is to connect between different partners. However, this is at the same time the key problem of interoperability; sharing data among heterogeneous partners (Wang et al. 2015). Overall, we could say according to Malaverri et al. (2013) that the underlying hypothesis behind interoperability is that in heterogeneous environments there should be a set of common characteristics, such as a wide variety of data sources or the need to coordinate data-driven processes.

System functional feature

We mentioned that interoperability has organizational, semantic and technical aspects. We saw that from organizational perspective it is especially about sharing data between different partners to establish partnership activities. On the semantic perspective, we will elaborate more later in this section. When we look more detailed from technical perspective, we can define interoperability as the cooperation among processes by exchanging procedures within cooperative information systems which are not logically integrated, since they are stored in separate databases according to different schemas (Batini *et al.*, 2009). We can describe interoperability in this context as a ‘system functional feature’ (Yu, 2016).

A more concrete interpretation of these three aspects is provided by Charalabidis et al. (2016). In their paper, they provide a taxonomy of open government data research topics of the Open Government Data (OGD) domain. Their taxonomy includes four main research topics of the Open Government Data (OGD) domain, namely: management and policies, infrastructures, interoperability, and usage and value. And regarding interoperability they distinguish eight research topics, namely: 1. Metadata for OGD, 2. Multilingualism Issues, 3. Services Interoperability Standards, 4. Semantic Annotation, 5. OGD Ontologies, 6. Platform & Technical Interoperability, 7. Organizational interoperability, 8. Controlled Vocabularies/ Code lists Preservation. Just to give an indication what kind of themes we should think of regarding the term interoperability in literature. We observe that interoperability not only in business, but also in academia a widely used term is.

Requested interoperability

One type of interoperability we want to elaborate on more is ‘requested interoperability’. This type of interoperability relates most to the ‘Platform & Technical Interoperability’ topic from Charalabidis *et al.* (2016). It is about the ability for multiple software components to interact regardless of their implantation, programming language or hardware platform (Matiasko, Zabovska & Zabovsky, 2004). Thus, we conclude again that interoperability is about connecting and interacting between different environments.

Based on this interpretation Matiasko et al. (2004) distinguish several mechanisms for software interoperability, namely: data-type interoperability, specification-level interoperability, and semantic interoperability. Whereas data-type interoperability and specification-level interoperability is on the technical side, and semantic interoperability obvious is from semantic perspective. Data-type interoperability is about supporting structured exchange of information through APIS. Whereas specification-level interoperability is almost the same, but also encapsulates knowledge representation differences at the level of abstract data types (like Table Tree). Thus, specification-level interoperability enables programs to interact at a higher level of abstraction (e.g. Java Beans

fall in this category). Finally, regarding semantic interoperability Matiasco *et al.* state that it assumes different information sources store information on related issues but each may offer a different meaning (semantic) of it. We conclude in this context the semantic perspective has to do with the interpretation of the data. And therefore, the challenge of semantic interoperability can be described as the ability of the user to access consistently and coherently similar digital objects distributed across heterogeneous repositories. This is where accessibility and interoperability show similarities. In both cases finally the goal is that the (end)user can (re)use the information regardless his/her environment with related characteristics.

Data needs to be conformed

Different frameworks provide guidelines in how to approach the above described challenge. For example, in the Linked Data Quality Assessment framework – Luzzu – of Debattista *et al.* (2016) interoperability is accompanied by a set of ontologies for capturing quality-related information reuse, including quality measures, issues, and reports with the goal that it can be reused in other semantic frameworks and tools. Or more in business terms: data needs to be ‘Conformed’, because business wants to analyze data across common, sharable dimension, so that the same information is used (one of 5Cs of Sherman, 2014). And then the data user is central. Therefore, according to Kallinikos *et al.* (2013) interoperability is an important condition of the digital ecosystem.

We observed already that interoperability is about the interaction between heterogeneous environments/objects, and in the previous paragraphs we described that the (end) user is the main stakeholder in this. Wang *et al.* (2015), mentioned earlier in this section as authors who define interoperability also as a measure of the extent to which systems, organizations and individuals can cooperate, additionally mention an interesting requirement. They note that the cooperation between systems should require minimal efforts from their users to be qualified as interoperable. Again, this shows the importance of the (end)user, but thereby is it notable that this is the same what Wilkinson *et al.* (2016) define in their paper. Wang *et al.* (2015) conclude that to solve this problem a general model transformation methodology is required. Because traditional model transformation practices have several weaknesses, like low reusability (!), repetitive tasks, huge manual effort.

3.4.1 Conclusion

Regarding interoperability we conclude that with different resources (with common characteristics) interoperability is needed to connect them with minimal or no special effort from the user to realize data can be shared among heterogeneous partners. Interoperability is also a widely used term in academia. Charalabidis *et al.* (2016) distinguished about eight different research topics. The main goal of interoperability is to ensure that data can be shared, reused and exchanged.

Based on these findings we provide the following definition of interoperability:

INTEROPERABILITY is a system feature to connect and conform heterogeneous environments, to ensure sharing, reusing and exchanging of data between these environments without special effort from the (end) user.

3.4.2 Benefits new definition versus definition Wilkinson *et al.* (2016)

The concepts regarding Interoperability found in literature are quite in accordance to the guidelines of Wilkinson, especially that data must be interoperable with minimal effort (for (end) users). Therefore, no big changes are implemented, besides that:

- In the original description, a universal language and vocabularies that follow FAIR principles were explicitly mentioned. In our opinion, this is too specific for a general definition.
 - Therefore, we do not include this in the new definition.

3.4.3 Concept Table Interoperability

We summarize the key concepts regarding interoperability as follows:

Interoperability				
	Types of Interoperability	Need for interoperability	Goal of interoperability	Identity/ Minimal effort user
Batini, Capiello, Francalanci, Maurino (2009)	X			
Charalabidis, Alexopoulos, Loukis (2016)		X	X	
Debattista, Auer, Lange (2016)			X	
Kallinikos, Aaltonen, Marton (2013)		X		
Malaverri, Mota, and Medeiros (2013)			X	X
Matiasco, Zabovska, Zabovsky (2004)			X	X
Sherman (2014)		X		
Shukair et al (2013)	X	X		
Wang, Truptil & Benaben (2015)				X
Yu (2016)	X			X

Table 6: Detailed Concept Table on Interoperability

3.4.4 Results Evaluation

According to the survey 67% (see Figure 12) from the respondents chooses for our definition when we asked which definition describes findability best. The respondents could choose for the definition of Wilkinson, and one other random formulated definition, namely: 'Interoperability describes the ability to establish partnership activities in an environment of unstable market'. This definition is from Wang et al. (2015), found during this literature research, and is taken as third option in the survey because it comprises a quite different perspective as our final definition.

We decide to not reformulate the definition on interoperability based on the results of the survey. But just what applies to accessibility, it does not mean this definition is perfect. We stimulate improvements, based on even more extensive research.

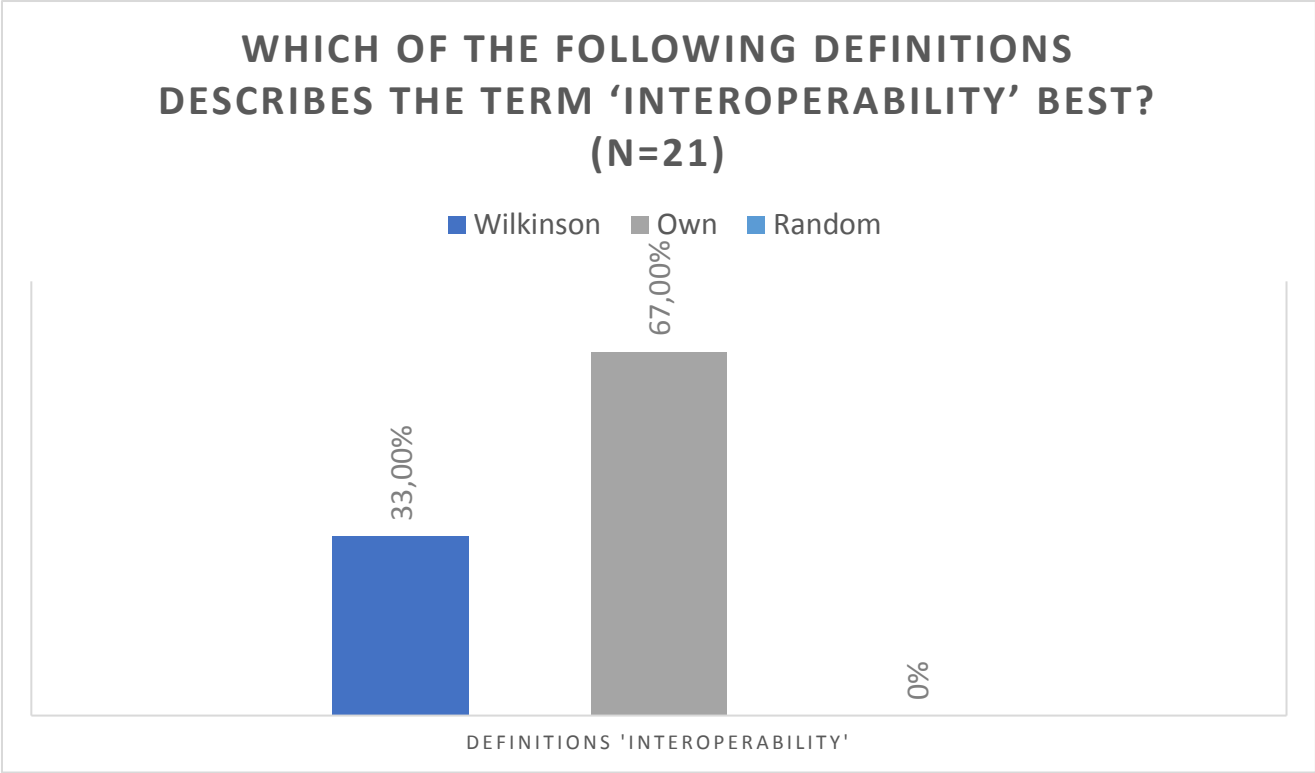


Figure 12: Results Survey regarding the concept interoperability

3.5 Reusability

The five Cs of Sherman (2014) came back a couple of times in the sections before. The last one 'clean' suits the reusability concept, since the final results of findable, accessible and interoperable data is that it can be reused. For reuse data needs to be clean, in terms of: no missing items or invalid entries and such (Sherman, 2014). According to Wang et al. (2015) reusability still is a point of attention, since low reusability it is one of the weaknesses of traditional model transformation practices. Notable is that Debattista et al. (2016) state that reusability helps to decrease the number of duplicate and redundant resources on the Web. It makes that the internet is currently evolving from the 'Web of Documents' into the 'Web of Data' (Fürber & Hepp, 2011). This is an interesting perspective, because from this perspective reusability is mentioned more as a cause for other quality indicators, than as a quality factor itself. This seems in contradiction with the definition according to Wilkinson et al. (2016) that data is reusable when (meta) data are richly described with a plurality of accurate and relevant attributes, which means that (meta)data are released with a clear and accessible data usage license, (meta)data are associated with detailed provenance, and (meta)data meet domain-relevant community standards. This implies that reusability can be achieved by fulfilling a set of specific requirements.

Linked Data

This approach seems to be in line with the approach of the UK government. The UK is a country with explicit policy on public data. They have public data principles which state that the government should make public data available in machine-readable formats, publishing using open standards and released under an open license. Of course, governments are worried that their data will be used in incorrect or misleading ways, but from their point of view linked data methods can prevent this. The term Linked Data refers to a set of best practices for publishing and connecting data on the web (Sheridan & Tennison, 2010). A lot of data today is available for us in the form of webpages, HTML documents which are linked to each other by hyperlinks. This occurs some problems, because the quality of structure and semantics of the data decreases, machines have difficulties with extracting any meaning, and these are not 'powerful' enough to enable entities described in a document to be connected to related entities. Linked data methods are so relevant for publishing open government data (especially for statistical and geo-spatial information), because linked data standards uniquely allow governments to publish data responsibly. It leaves the publisher in control of their data in a unique way, but also enables reusability.

Open Data Policy UK

In the Open source, Open standards and Re-use Government Action Plan of the UK government (2010) for reuse they provide among other things two interesting insights: 1. The government must enable straightforward reuse of data elsewhere in the public sector, and their own government data will be released on an open source basis. 2. Systems, designs, or architecture already owned by the public sector must be reused by the government, and supplier will be required to warrant that they have not developed or produced something comparable. Based on these points and some others (about open standards, and (non)-open source software) the UK government set up ten key actions to execute this policy. Three of them are interesting regarding reusability, namely: Reuse as a practical principle, Open Standards, Open Sources techniques and reuse within Government, and appropriate release of code. The idea of 'Reuse as a practical principle' is that where open source solutions are evaluated and approved by one part of Government, that evaluation should not be repeated but should be shared. This means that different departments within the government should share records of their use of open data sources, including open source components within composite solutions. Second, 'Open standards' means that the government will specify requirements by reference to open standards and that they will require compliance with open standards. The UK government supports the use of HTML (ISO/IEC 15445:2000), Open Document Format (ISO/IEC 26300:2006), ISO 32000-1:2008("PDF"), and ISO/IEC 29500 ("Office Open XML formats"). In addition, Sheridan & Tennison (2010) state that dereferencing each URI has important benefits over interchange formats such as CSV or XML concerning reusability, because with these formats data can be changed or context lost as it is passed from hand to hand or from system to system. With URI, the publisher always controls what is returned when each URI is dereferenced. Finally, regarding Open Source techniques and reuse the UK government will use a standard OGC (Open Geospatial Consortium)-approved OJEU (Official Journal of the European Communities) clause to make clear that solutions are purchased on the basis that they may be reused elsewhere in the public sector. Solutions and licenses will have transferability across the public sector and into cloud based service environments.

Reusability as resultant

The activities described above are not only in the UK, it is happening in different countries around the world. Open government data projects can be found in the United States, Australia, New Zealand, but also in The Netherlands, Sweden, Spain, Austria and Denmark. And although the standards seem according to Sheridan & Tennison (2010) quite mature, capable and powerful, still much work needs to be done to translate those standards into simple and repeatable publishing patterns. Thereby, we observe that from this perspective reusability is more a goal in the policy than it is a quality indicator. However, at the same time Sheridan & Tennison (2010) conclude that the policy of the UK government makes it possible that the government data can be reused flexible and easily (for example through APIs) by data consumers. Thus, the basic concepts of Linked Data, like publishing and connecting, do have a positive influence on the extent data can be retrieved by the users without restrictions. And based on other literature we conclude this is the goal for reusability. From this perspective reusability, like interoperability, can be classified as a 'system functional feature', and ensures that information can be retrieved, downloaded, indexed, searched and visualized easily by the (end) user.

And then the question is: what is needed to realize this goal? According to Yu (2016) therefore data should be in an open format that is machine readable, platform independent, and made available without restrictions. This reminds us to the definition of accessibility; a connection between heterogeneous environments is desirable. And in the section on interoperability we saw that when distinct groups are interoperable sharing and reuse of data among these groups in their processes is realized (Malaverri et al., 2013). Which implies that storing contextual information is extremely important (definition findability). Therefore, we conclude that all other different quality indicators (FAI) 'realize' the underlying definition of reusability of Vries (2012): reusability of public data is about "putting the public data to use in new contexts and by other people than the original public-sector employees".

3.5.1 Conclusion

We conclude that reusability is often mentioned in literature as a result of other quality indicators and not as a quality indicator itself. The main goal is to edit data in terms of retrieving, downloading, indexing, searching, and visualizing. This is not in accordance with Wilkinson et al. (2016), since they mention provenance (which is about findability) as crucial to make reusability of data possible. Second, we conclude that agreement on (Open), standard formats is very important regarding reusability. Thereby again the importance of dereferencing URIs is mentioned. Finally, duplicated systems, designs, architectures but also data should be prevented, therefore 'sharing' and 'licenses' are essential.

Based on these findings we provide the following definition of reusability:

REUSABILITY is the ability to make easily use of data in a new context, in terms of retrieving, downloading, indexing, searching and visualizing the data without restrictions, as a result of findable, interoperable and accessible data.

3.5.2 Benefits new definition versus definition Wilkinson et al. (2016)

The new definition on reusability differs quite a lot from the definition of Wilkinson et al. (2016). The main reason for this is that based on literature reusability is described as the result of other improvements on data quality, like making data findable, accessible, and interoperable. Therefore, we decide to take this as foundation for our new definition, thereby:

- The definition of Wilkinson was too specific for a general definition of the concept, and the goal of reusability is missing.
 - Therefore, we include 'in terms of retrieving, downloading, indexing, searching, and visualizing the data without restrictions'.

3.5.3 Concept Table Reusability

We summarize the concepts regarding reusability as follows:

Reusability				
	Result of other quality indicators	Need for reusability	Dereference URIs	Goal of reusability
Debattista, Auer, Lange (2016)	X			
Fürber and Hepp, (2011)	X	X		
Malaverri, Mota, and Medeiros (2013)	X			
Matiasco, Zabovska, Zabovsky (2004)			X	
Action Plan of the UK government				X
Sheridan and Tennison (2010)			X	
Sherman (2014)		X		
Vries (2012) (cited in Lassinantti & Bergvall-Kareborn (2014))				X
Wang, Truptil & Benaben (2015)		X		
Yu (2016)				X

Table 7: Detailed Concept Table on Reusability

3.5.4 Results Evaluation

According to the survey 45% (see Figure 13) from the respondents chooses for our definition when we asked which definition describes reusability best. The respondents could choose for the definition of Wilkinson, and one other random formulated definition, namely: 'Reusability is putting the data to use in new contexts and by other people than the original sector employees'. This definition is from Vries (2012), and is chosen because it is in our opinion the simplest definition of reusability. The results from the survey are difficult regarding reusability. The respondents seem to be divided. Thereby, according to the evaluation results, based on the interview (provided in chapter 8), data providers argue that reusability is a quality indicator itself, and thereby an important part of their open data policy.

However, we decide not to reformulate the definition. In our opinion, the results are too unclear to make such a radical change in the definition. Therefore, we decide to let reusability be part of the four principles, so it is a guideline but to make clear in the definition that according to literature reusability has a different nature than the other three and must be seen as a resultant.

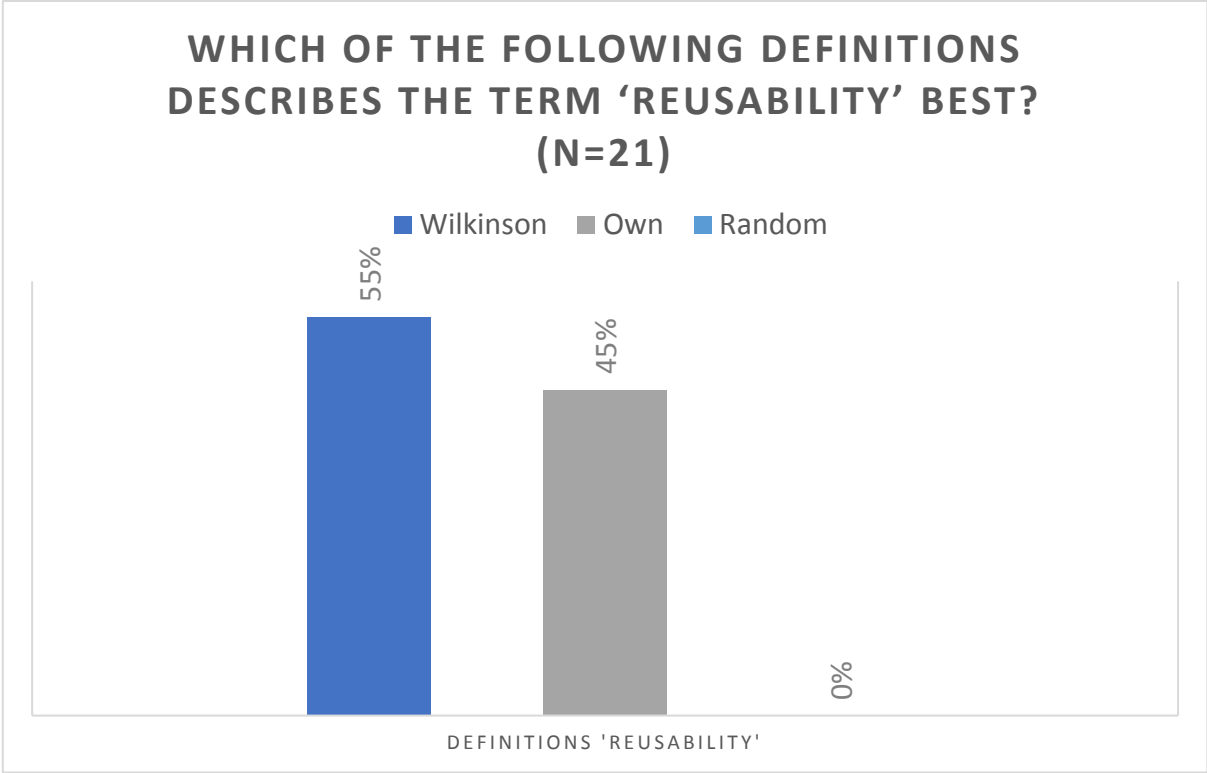


Figure 13: Results Survey regarding the concept Reusability

3.6 Statistical conclusions and summary Literature Research

3.6.1 Statistical conclusions

To answer the first research question ‘What are the definitions and interpretations of the FAIR data concepts in literature?’ we formulated per FAIR concept a definition based on the literature research, elaborated above. In addition to the detailed concept table per concept, we created two other types of concept tables (see Appendix A); one provides an overview from which concepts can be found in which papers, and the second gives the citation of these (indirect) definitions). We provide the ‘statistical’ results of these two tables in Table 8. Whereas ‘Definition’ stands for a literal definition from the concept, and ‘Indirect’ indicates a description of the concept.

	Findability	Accessibility	Interoperability	Reusability	Total:
Definition	1	9	6	3	19
Indirect	8	9	5	8	30
Total:	9	18	11	11	

Table 8: Statistics Literature Review

Explanation Table

- The total number of unique papers used is 26, wherein 19 definitions were found and 30 indirect definitions.
- Accessibility is the most frequent found and defined term. This is likely because ‘Accessibility’ is often used as the name of a quality dimension.
- Reusability and Findability are most unknown terms in literature. For reusability, it is likely this is because reusability is often mentioned as result of other quality indicators instead of a quality measure/indicator itself. For Findability, there is no clear reason why it is mentioned less, however it is quite a broad, generic term and a new term in the data quality domain.
- Regarding interoperability it is interesting to mention that there were the biggest differences between these (indirect) definitions.

3.6.4 Summary Literature Research

We can summarize the main results per concept based on the keywords in the detailed concept tables per concept (table 4-7). These are as follows:

- **Findability:** Versioning – Provenance – Unique Identifier – Uniqueness – Searchability
- **Accessibility:** Quality Dimension – Semantic perspective – Syntactic perspective – Social perspective – Importance (End) User
- **Interoperability:** Types of Interoperability – Need for interoperability – Goal of interoperability – Identity/ Minimal effort user
- **Reusability:** Result of other quality indicators – Need for reusability – Dereference URIs – Goal of reusability

Thereby, in short some additional explanation on the main results per concept in relation to the original definitions of Wilkinson et al. (2016):

- **Findability:** The definition of Findability is in line with the guidelines of Wilkinson et al. (2016). For both applies that ‘uniqueness’ is the keyword.
- **Accessibility:** The guidelines for Accessibility mentioned by Wilkinson et al. (2016) are not found in other literature. Especially a communication protocol is not mentioned.
- **Interoperability:** The concepts regarding Interoperability found in literature are quite in accordance to the guidelines of Wilkinson et al. (2016), especially that data must be interoperable with minimal effort (for (end) users).
- **Reusability:** Regarding reusability we conclude that reusability is more a consequence of the other three quality indicators, so the focus must be on those (Findability, Accessibility and Interoperability). This is in contrast with the definition of Wilkinson et al. (2016).

Based on these results we set up the following definitions:

- **Findability** (of data) is about the uniqueness and provenance of data, which means data described with rich meta data with focus on contextual information. This is achieved by versioning, associating URIs or embedding other kind of identifiers
- **Accessibility** is a semantic quality dimension, which comprises availability, security, performance, interactivity, flexibility of data, to ensure data access to (end) users regardless their different context and background.
- **Interoperability** is a system feature to connect and conform heterogeneous environments, to ensure sharing, reusing and exchanging of data between these environments without special effort from the (end) user.
- **Reusability** is the ability to make easily use of data in a new context, in terms of retrieving, downloading, indexing, searching and visualizing the data without restrictions, as a result of findable, interoperable and accessible data.

Finally, we summarize the most interesting conclusion of this literature research with the following metric: $(F + A + I)/3 = R$, which means Reusability is a resultant from Findability, Accessibility and Interoperability. In section 6.6 we will elaborate on this metric more in detail.

4. Research Question 2: Benefit of FAIR in Open Data Projects

WHAT IS THE BENEFIT OF FAIR IN THE CONTEXT OF OPEN DATA PROJECTS?

WHY describing FAIR in the context of Open Data Projects?

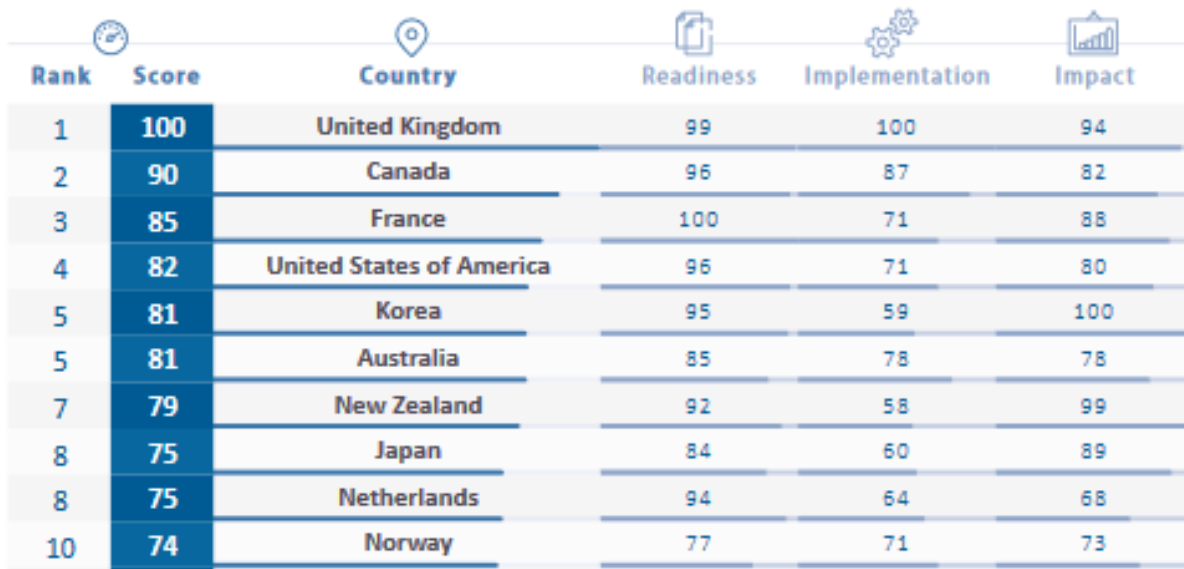
In research question 1 we defined each FAIR concept separately, and described the theoretical perspective on FAIR. However, due to our problem statement we should bridge the gap between theory and practice in how to deal with meta data management in open data repositories. Therefore, in this research question we will start with describing the open data landscape. And second, we will produce our first artifact: a roadmap for the data scouting process. This roadmap is created in collaboration with Berenschot Intellerts. The main goal of this artifact is to determine the place of FAIR in the context of open data projects. This is the first step to bridge the gap between theory (the FAIR guidelines for meta data management) and practice (the steps in open data projects). 'Data Scouting' is a term within Berenschot Intellerts to describe the process of gathering and storing the right data for a project. Therefore, this is also how we called this process.

4.1 Open Data Landscape

In research question one we pointed out the different definitions of the FAIR principles, and elaborated on other terms, concepts and frameworks which are invented regarding measuring quality of data and for 'structuring' (open)data. But to answer the overall research question, namely 'Which requirements should a data repository meet to satisfy the FAIR principles?', we also must know how the open data landscape looks like. What are the open data sources we are talking about? And especially what is the state of the open data supply in the Netherlands. In this section, we will give insights on the statistics about the data the Dutch government supplies. After this we will elaborate on the process of 'data scouting'; a roadmap for a so called 'data scout' is made. And finally, a taxonomy will be provided with the main open data sources divided per themes. In the whole chapter, the focus will be on mainly open government data.

Data landscape the Netherlands

In 2016 the Netherlands was in the top 10 of countries in the Open Data Barometer Project; on the 8th place. The exact numbers can be found in Figure 14 below.



Rank	Score	Country	Readiness	Implementation	Impact
1	100	United Kingdom	99	100	94
2	90	Canada	96	87	82
3	85	France	100	71	88
4	82	United States of America	96	71	80
5	81	Korea	95	59	100
5	81	Australia	85	78	78
7	79	New Zealand	92	58	99
8	75	Japan	84	60	89
8	75	Netherlands	94	64	68
10	74	Norway	77	71	73

Figure 14: Top-10 Open Data Barometer Project 2016

This Open Data Barometer project aims to uncover the true prevalence and impact of open data initiatives around the world. It analyses global trends, and ranks countries via an in depth methodology that considers: readiness to secure the benefits of open data, the actual levels of implementation, and the impact of such initiatives. The score

is between 0 and 100, and is given based on ten questions. Per question a maximum of 10 points can be given. The questions are as follows:

1. Does the data exist?
2. Are the data available in the government?
3. Is the data in a machine-readable format?
4. Is the machine-readable data available in bulk form?
5. Is the dataset available for free?
6. Is the data provided with an open license?
7. Is the dataset up-to-date?
8. Is the publication from this dataset sustainable?
9. Is it easy to find information about the dataset?
10. Are data APIs supplied for major components of the dataset?

Based on these questions the open data quality is measured per country. On the map in Figure 15 becomes clear that the Netherlands does quite well in comparison with the rest of the world.

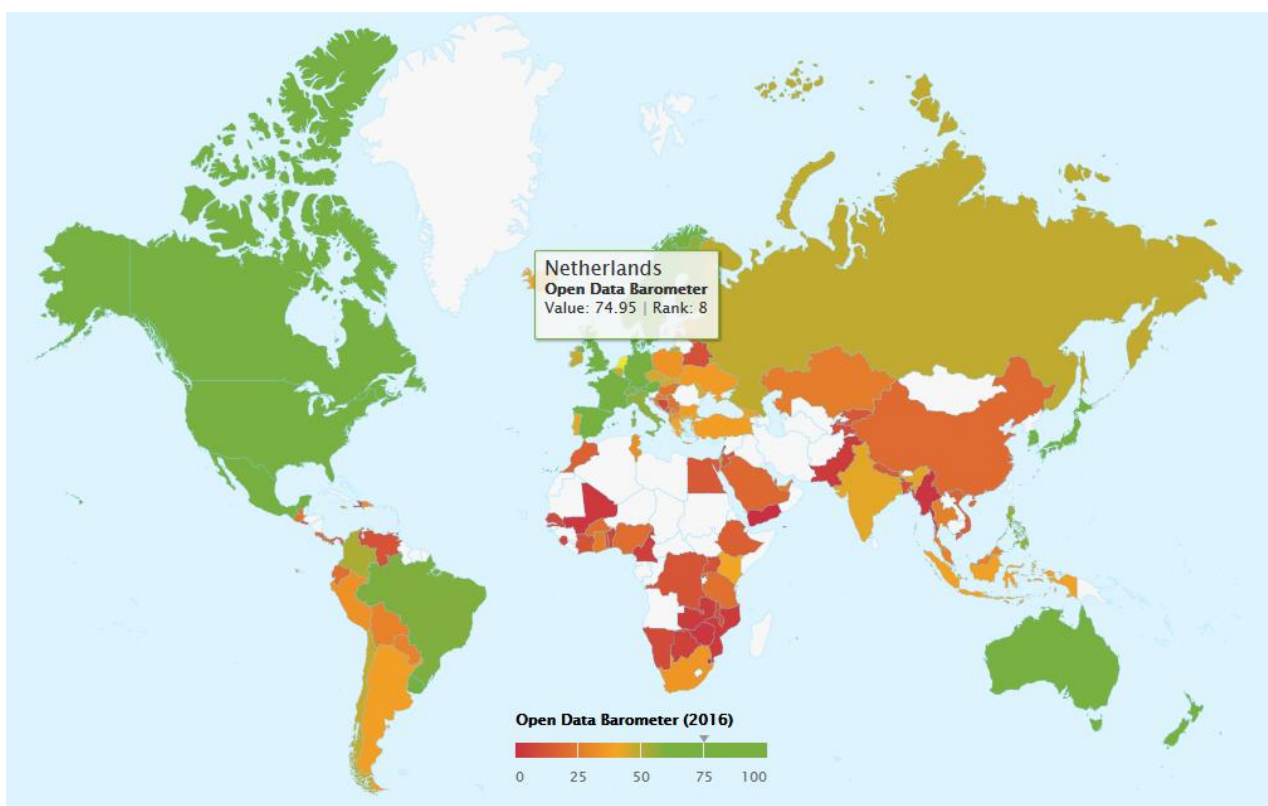


Figure 15: The Open Data status of the Netherlands compared to the rest of the world (Open Data Barometer, 2016).

However, there are always points of improvement. According to the report of the World Wide Web Foundation (2016) – initiators of the Open Data Barometer – the overall main findings are:

1. Nine out of 10 government datasets are not open.
2. Government data is typically incomplete and low quality.
3. Sustained political will is what makes or breaks the success of open data.
4. Governments are not publishing the data needed to restore citizens' trust.
5. Open government data risks reinforcing inequalities.

The focus of the paragraphs above is on governments, though also business and academia have their influence. In the end, it is beneficial for the Netherlands when open data is of good quality. Although it cost a lot of effort to

reach this, the results are visible: the Netherlands increased since 2013 already three places. According to the report of the World-Wide Foundation (2016) in the Netherlands open data policy is supported by a strong push from organized civil society groups, as well as support from those groups to stimulate the use of open data through hackathons and other activities. Researchers identified a much greater rate of open data publication in the Netherlands, where almost 50% of datasets surveyed qualified as open under the open definition. The UK is the number 1 for years. In Figure 16 becomes clear on which points they score better than the Netherlands (measured in 2015).

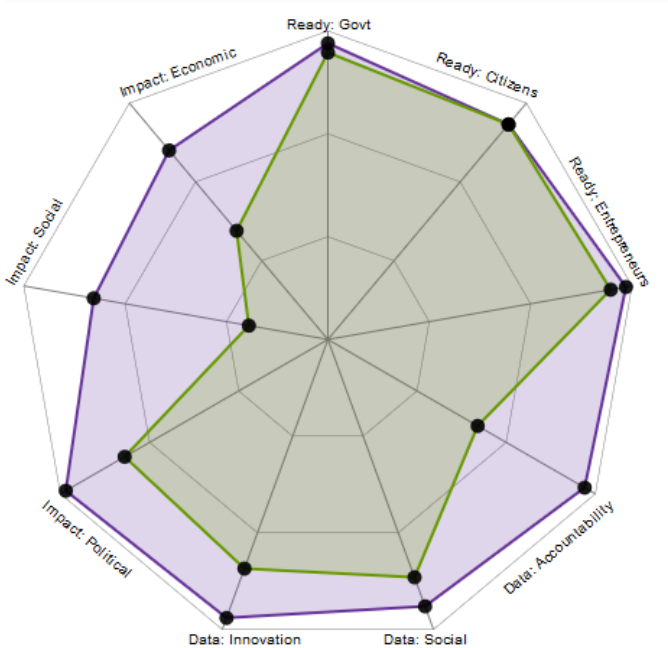


Figure 16: Netherlands (green) versus UK (purple) – 2015

The data portal data.gov.uk we already mentioned a few times in literature research. And, the Open Data Barometer project shows that the UK is an example for other countries. The lead on other countries perhaps decreases, however in 2013 they performed far better than Europe in total (see Figure 17). Because the UK is for years the number 1, we will shortly discuss what is so good over there. In 2009 the United Kingdom started an Open Government Data initiative. Since then open data has high policy priority to support innovation and economic growth. And this is paid out in the results. The United Kingdom leads. In 2012 the Open Data Users Group is set up, acting as a conduit for data requests and advising government on priority of data to release. Local authorities publish their own datasets and set up open data portals. Thereby, training on open data topics is increasingly available and open data hack-days, events and competitions are organized more and more. But the main reason for the success is that the government supports by policy and by supporting innovation funding to help new and existing businesses to engage with open data (report World Wide Foundation, 2016).

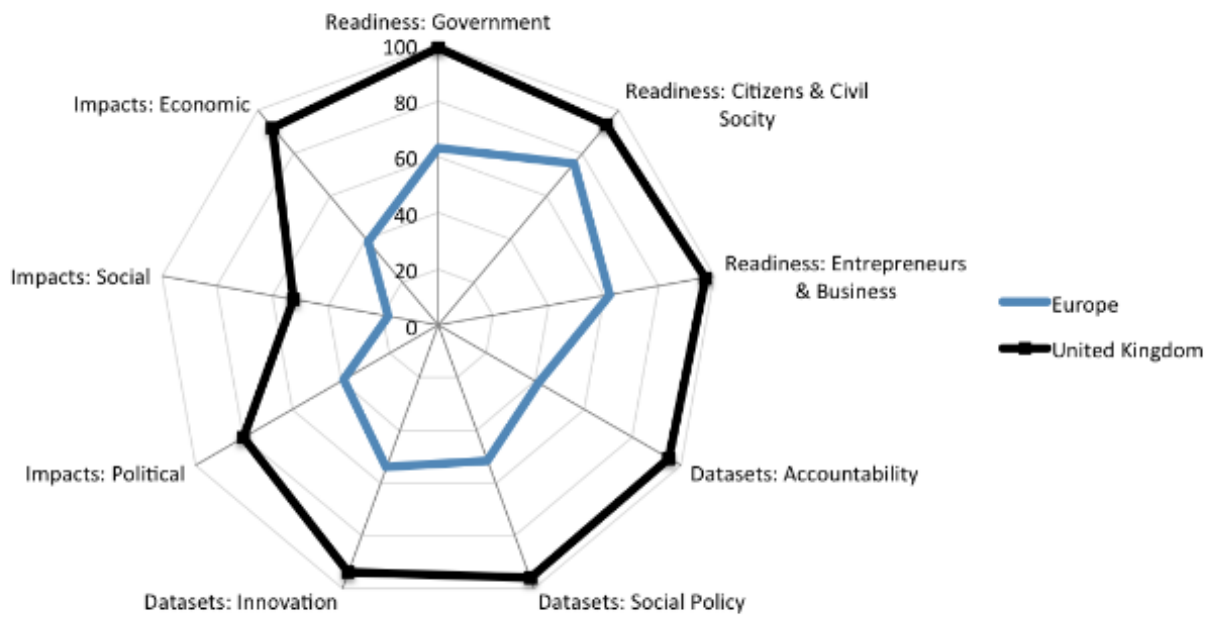


Figure 17: Radar chart of scaled sub-component scores. Comparison of UK and Europe average 2013.

We arise a more detailed view of the current situation regarding open data in the Netherlands when we zoom in on the different areas. Figure 18 shows how the Netherlands scored in the period from 2013 till 2015 according to the Open Barometer from the World Wide Web Foundation with the supply of open data within different policy areas (report Algemene Rekenkamer).

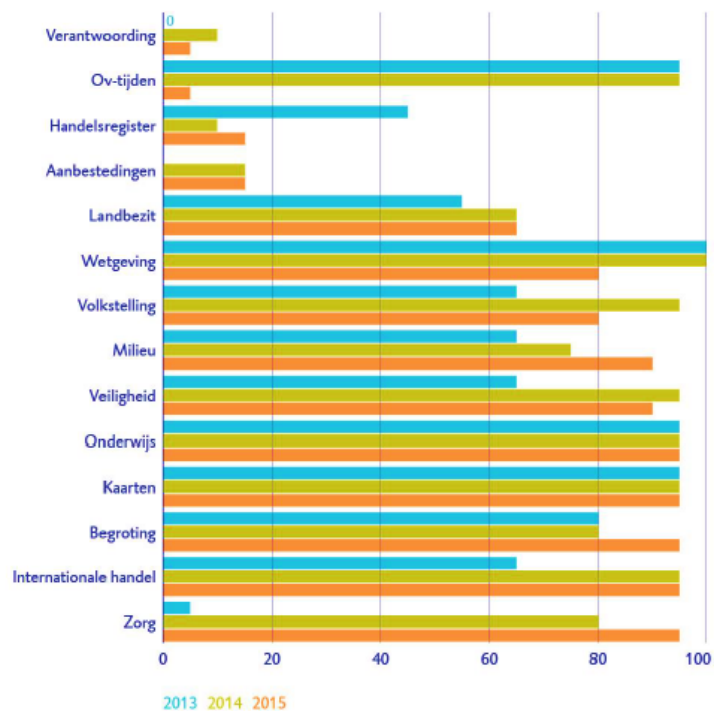


Figure 18: Open data supply policy areas the Netherlands 2013-2015 (Algemene Rekenkamer)

The cabinet put a lot of effort last few years in open data. In total per February 2016 about 7.400 datasets were made available. The data portal data.overheid provide next to open data of the 'rijksoverheid' also other data.

Although about 75% is from 'rijksoverheid'. The remaining datasets belong to provinces and municipalities. Thereby there are a few upcoming providers, like 'Gemeenschappelijke Regeling', 'Rechterlijke Macht', 'Europese Commissie', and 'Waterschap' (report Algemene Rekenkamer). The top 5 data providers of the Dutch government are: CBS, Rijkswaterstaat, KNMI, RIVM and Kadaster. They provided respectively about 3.877, 1.293, 68, 62, 53 data sources (measured in February 2015 for the government rapport Algemene Rekenkamer). That the Dutch government gives open data more and more priority appears also from the fact that they released in 2014 their spend data to the public to further increase the availability of open data. The datasets provide information on the suppliers from each ministry, divided into categories. In Figure 19 we provide an excerpt of the supplier network of all the eleven ministries of the Dutch government.

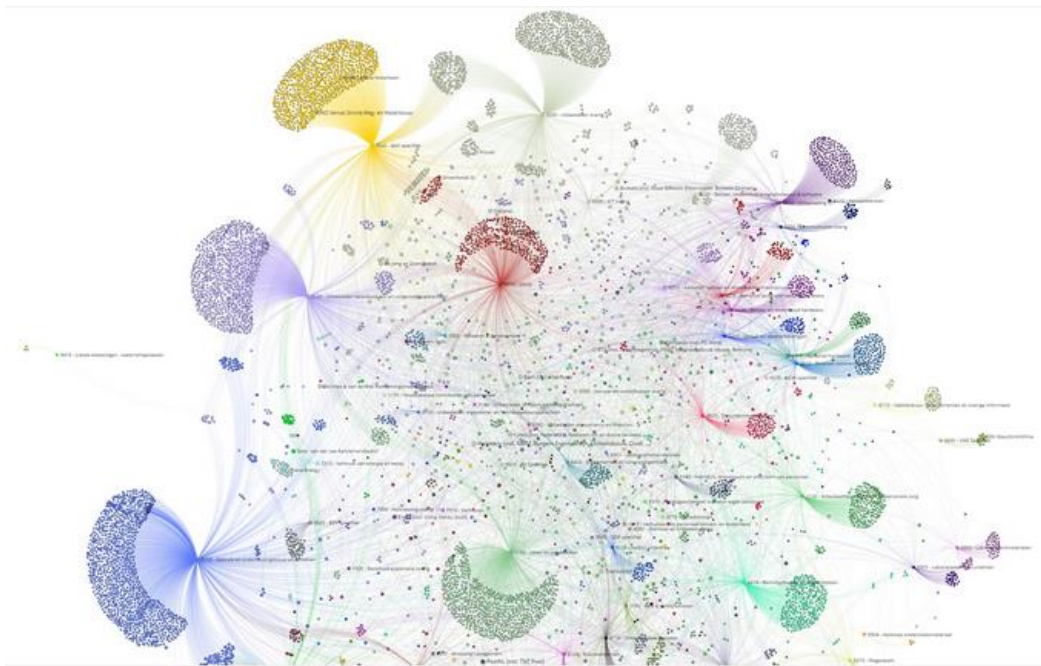


Figure 19: supplier network of all the eleven ministries of the Dutch government

In this section, we elaborate especially on open data within government context. This because open data associates strongly with open government. However, there are also a lot of open data initiatives from business side. Already in 2010 Bughin & Manyika (2010) distinguished 10 tech-enabled business trends with focus on clouds, big data, and smart assets. These are relevant, since most of them are still actual and senior executives still need to think strategically about how to prepare their organization for this challenging 'new' environment. We will just mention the trends here, for further elaboration we refer to the paper itself. The trends are:

1. Distributed cocreation moves into the mainstream.
2. Making the network the organization.
3. Collaboration at scale.
4. The growing 'Internet of Things'.
5. Experimentation and big data.
6. Wiring for a sustainable world.
7. Imaging anything as a service.
8. The age of the multisided business model.
9. Innovating from the bottom of the pyramid.
10. Producing public good on the grid.

Regarding trend 5 Bughin & Manyika (2010) conclude that Google, Amazon.com, eBay and financial institutions are the most experimenters, pioneers, when it is about approaches to make 'real time decisions'. But also, Ford Motor, PepsiCo, and Southwest Airlines are mining and analyzing consumer behavior on social-media. And we assume these initiatives are increased exponential since 2010, also a lot of start-ups dive in the niche of this

market. And with more and more open data, which means accessible and available data for everybody, even more is possible. And from academic perspective of course the FAIR principles are a worth mentioning initiative.

4.2 The place of FAIR in Data Projects: CRISP-DM and KDD Process

In the paragraphs before we described the context of the FAIR initiative with the focus on the Dutch open data landscape. However, in line with our research question we want to know: what is the actual benefit to apply fair on open data projects? And what is the place of FAIR in the existing workflows of data projects? Two famous initiatives to describe the process of data projects are the Cross Industry Standard Process for Data Mining (CRISP-DM) and the Knowledge Discovery in Databases Process (KDD). Therefore, to provide more insight of the place of FAIR in business context we describe in which steps in these processes FAIR should be applied.

First, the CRISP-DM Process. According to Wirth & Hipp (2000) the CRISP-DM project proposes a comprehensive process model for carrying out data mining projects. Thereby, the process model is independent of both the industry sector and the technology used. We provide this process in Figure 20 and marked the place of FAIR with red, namely the phase 'Data Preparation'.

According to Wirth & Hipp (2000) this phase covers all activities to construct the final dataset from the initial raw data. And according to them tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools. The benefit of our FAIR model regarding these tasks is that it provides guidelines in which information is most important to describe and which attributes are really indispensable regarding data storing. This is also the reason why FAIR is not placed in the Data Understanding phase, since this data understanding and description is more on content level of the data; which information from the dataset is interesting for analyses. While FAIR is not on content but about the quality of the meta data.

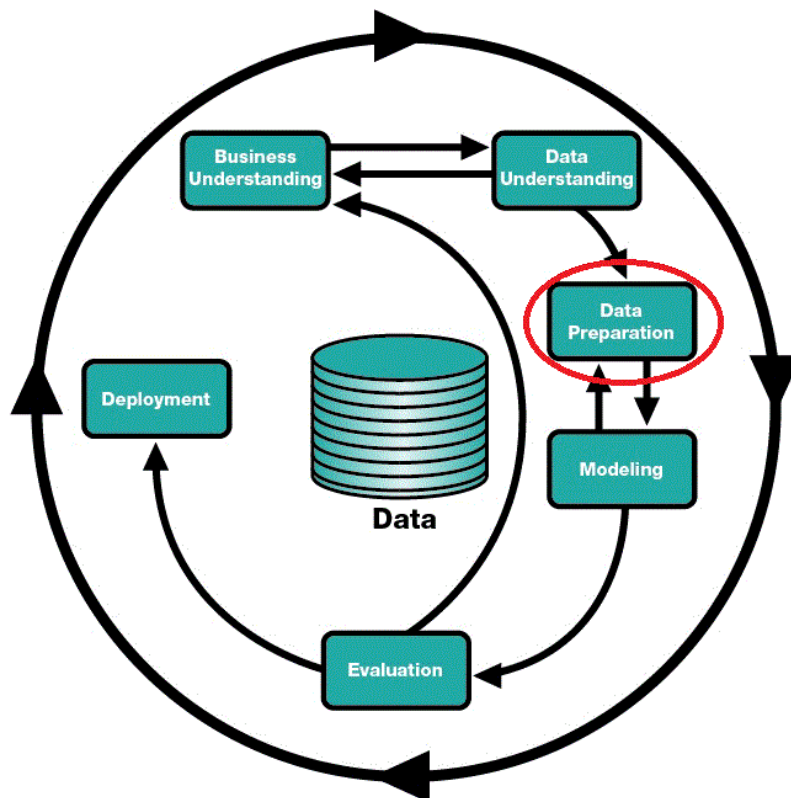


Figure 20: Place of FAIR in CRISP-DM

Second, the KDD process. In short, according to Fayyad, Piatetsky-Shapiro & Smyth (1996) the KDD process focuses on the overall process of knowledge discovery from data, including how the data is stored and accessed, how algorithms can be scaled to massive datasets and still run efficiently, how results can be interpreted and visualized, and how the overall human-machine interaction can be modeled and supported. We provide this step-by-step process in Figure 21 and marked the place of FAIR with red, namely the step ‘Preprocessing’.

According to Fayyad et al. (1996) preprocessing includes tasks like: removing noise or outliers if appropriate, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding DBMS issues, such as data types, schema, and mapping of missing and unknown values. These activities all focus on data cleaning and on the content of data. However, our FAIR model can really benefit this process since it provides guidelines for high quality meta data management, which results in data which is stored in such a way that it is easier to analyze on content and thereby easier to reuse. Therefore we advise to first store all ‘targeted datasets’ (according to Fayyad et al. (1996) part of phase ‘Selection’) in a FAIR repository and then focus on data cleaning.

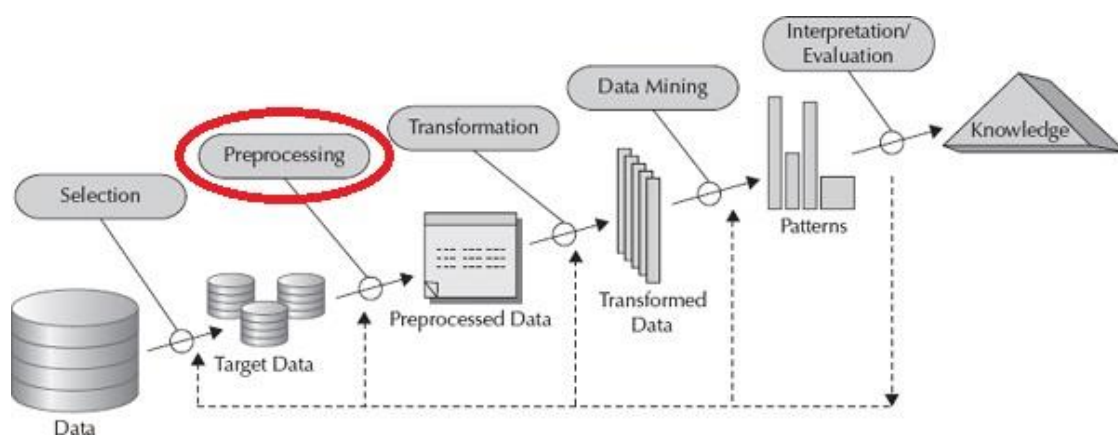


Figure 21: Place of FAIR in KDD Process

4.3 The place of FAIR in the Case Study Project: Roadmap Data Scouting Process

Regarding the second research question ‘What is the benefit of FAIR in the context of open data projects?’, we first described the context from theoretical perspective (open data landscape in the Netherlands), second we put FAIR in the context of two existing data processes (CRISP-DM and KDD), and finally we will put FAIR in the context of a Case Study project at Berenschot Intellerts. Since the existing models of CRISP-DM and KDD are not specific focused on open data projects, are still quite abstract, and most important comprise not exactly the steps the Berenschot Intellerts team follows during their projects, we decide to set up a Data Scouting Process Roadmap by ourselves as the starting point of our Case Study (with the main goal to apply our second artifact, the FAIR reference model, into practice).

We found within Berenschot Intellerts (based on conversations and written feedback) the set of steps provided in Figure 22. First, demarcate the domain of your project. This can be very global. Take for example as a starting point one of the themes from the next section: finance, energy, health, education, towns & cities, or infrastructure. Next step is start searching and identify the objectives. It is preferable to do first search without strictly delineated the objectives, just look which open data you find by googling on the theme. Do not think too fast the data is not valuable enough. Thirdly, it is time to structure and make clearer which questions you/the client want to be answered. Therefore, set up the key performance indicators of the project. Look at the data you already found in step two, and search targeted for what is still missing. And before starting the visualization and linking the data from the different (open) data sources, store the data by setting up a FAIR Data Repository. This is where our reference model comes in place (see grey box in Figure 22).

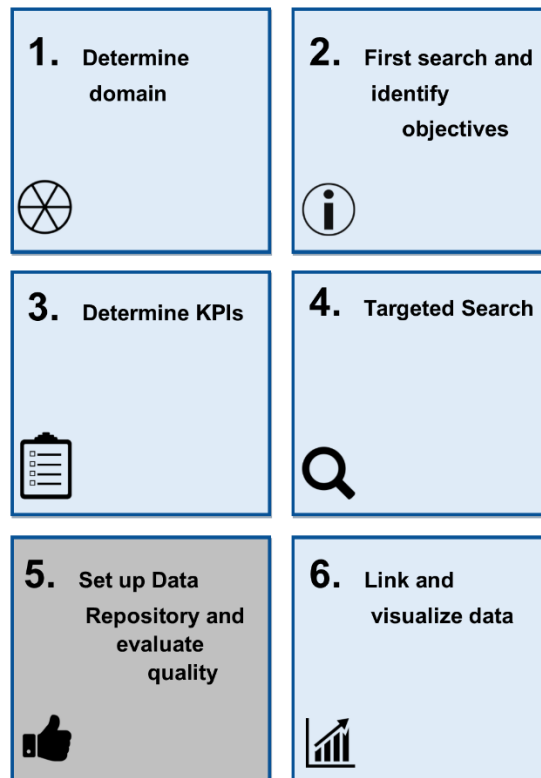


Figure 22: Data Scouting Process

4.4 Process and iterative evaluation Data Scouting Process Roadmap

The basis of the Data Scouting Process roadmap is quite simple. We asked at Berenschot Intellerts what the main steps are during their projects. Thereby, they had a clear wish to document this process and to understand how the FAIR principles could benefit their projects.

The main points of feedback during the setup of this roadmap was:

- Formulation of steps must be easy to understand
- Place of FAIR must be clear. For example, in an older set up for this roadmap (see Figure 23) the fifth step was described as 'Evaluation Data Quality', however this implies that the content of data is evaluated. Therefore, we changed this, because with our reference model not the content of data is evaluated but it ensures the quality of meta data.
- The design must be more attractive, and easy to understand. Therefore, we added also icons in the final design.

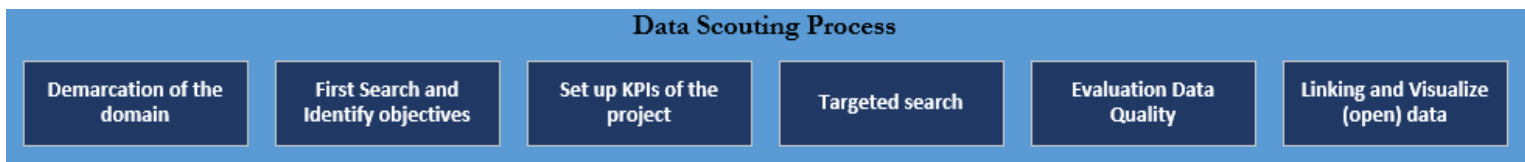


Figure 23: Older set up Data Scouting Process

4.5 Benefit of FAIR within Berenschot Intellerts

The grey box in Figure 22 shows us the place of FAIR in the Data Scouting Process; it is covered in the fifth step. This because FAIR, and especially the FAIR reference model we provide in this research, provides clear guidelines in how to ensure high quality in storing (open) data. The process in Figure 22 is designed based on the workflow of the projects within Berenschot Intellerts however this will not much differ from the workflows in other businesses doing the same kind of projects. Therefore, we assume that the following need for FAIR within Berenschot can be generalized to the need for FAIR in business in general.

Basically, we identified two main needs at Berenschot Intellerts:

1. A need for guidelines regarding meta data management in the process of storing open data in the 'data lake' (see Figure 24).
2. A need for an overview of the open data landscape in the Netherlands, and a standardized scouting process to find these sources.

The second need is fulfilled by providing the overall Data Scouting Process, and since FAIR provides clear guidelines for meta data management in the process of storing data it is interesting regarding the first need to apply these guidelines in their organization. We will discuss more detailed per FAIR concept why it is relevant in a business context.

Reusability in practice

We recall our definition of reusability, namely: "the ability to make easily use of data in a new context, in terms of retrieving, downloading, indexing, searching and visualizing the data without restrictions, as a result of findable, interoperable and accessible data". To illustrate this definition within practice we do not have to search deeply. Because Berenschot Intellerts is especially interested in the reusability of data and processes. Within their organization they use a lot of different (open) data sources. However, these sources are not stored in a structured way. Because of that for each new project, the different sources and eventually new sources are combined again. This should be easier, and the first step to realize this is to make sure that 'old' data sources, used sources, are 'ready' to use and combine again. Therefore, their aim is to realize the so called 'data lake'; a collection of data sources.

Interoperability in practice

To make the different data sources interoperable is a crucial phase in the working process of Berenschot Intellerts. This is also a very time-consuming part; therefore, it is interesting to research for possibilities how to facilitate the interoperability of new data sources, and to make different combinations of these data sources. Based on our definition interoperability is "a system feature to connect and conform heterogeneous environments, to ensure sharing, reusing and exchanging data between these environments without special effort from the (end) user". As we described, within Berenschot Intellerts interoperability is also about connecting data with the goal to share, exchange and reuse data. But especially also the last part of the definition "without special effort from the (end) user" is important for them too. Because the users of their final product (a dashboard), their clients, should not be charged with the difficulties of combining the different data sources. They just want to carry out their analyses and display the results.

Findability in practice

Our definition of findability is as follows: 'Findability (of data) is about the uniqueness and provenance of data, which means data described with rich meta data with focus on contextual information. This is achieved by versioning, associating URIs or embedding other kind of identifiers'. Reusability and Interoperability are the most important for Berenschot Intellerts. However, all principles are related, therefore also findability is relevant. For Berenschot Intellerts it is not so much about the external findability of data sources (this is out of their scope), but it is more about the internal findability; how can we make sure that all ever used data sources are findable for new projects?! Thus, internal searchability within the 'data lake'.

Accessibility in practice

Finally, Berenschot Intellerts has also to deal with accessibility. For example, which formats are the most accessible to use in tableau (the analytic tool used by Berenschot Intellerts to analyze the data and create the dashboards). Regarding our definition “accessibility is a semantic quality dimension, which comprises availability, security, performance, interactivity, flexibility of data, to ensure data access to (end) users regardless their different context and background”, we conclude that regarding accessibility Berenschot Intellerts is more internal and syntactical oriented while the definition is more focused on the semantical side.

5. Case Study Project: Data Scouting Process in practice

WHY applying the Data Scouting Process in practice?

In chapter 4 we set up the Data Scouting Process. In this chapter, we will follow these steps into practice during a Case Study Project at Berenschot Intellerts. The main goal to do this is that in this way we can apply our reference model, artifact 2, into practice. And we provide a clear context how FAIR can be applied in business. First, we will describe the processes and types of data within Intellerts to provide some more background information about this organization. Second, we will describe the context of the project. And thirdly, we walk through all the steps of the data scouting process to present the results from the Case Study Project.

5.1 Processes and Types of data Berenschot Intellerts

Within Berenschot they distinguish four types of data sources: open data (like CBS; often this data is published by governments), Client data (data conducted by their, e.g. temperature measures, sales etc.), Partner data (data sold by a company to another company), and data from the company itself (in this research: Intellerts data); the company already added value to this data (e.g. combining two sources, cleaned data etc.). In this research, we focus on the open data. Open data is data in an open machine-readable format, with no restrictions on re-use and available on the Internet.

As described we distinguish four main data sources, also called: oData (Open Data), iData (Intellerts Data), cData (Client Data), and pData (Partner Data). According to Intellerts, when the data 'comes in', it should be stored in the so called 'data lake'. Thus, within the data lake we can find the original, unedited sources. For these sources, the metadata should be described to make the data lake insightful and searchable. There is also a need to set up relevant themes, where the data sources can be categorized in, to create some structure. The next step in the process is the ETL-process, in this phase the data sources are edited. After this the combined, cleaned data files are stored as data marts in the data warehouse (called: 'Intellerts Data Garden'). A data mart is the access layer of the data warehouse environment that is used to get data out to the users. The data mart is a subset of the data warehouse that is usually oriented to a specific business question or team. An example of a data mart is the data mart 'KNMI/ NDW/Client Data'. So, in this data mart three separate sources from the data lake, namely 'KNMI', 'NDW', and the data of the client are combined. A visualization of this process is provided in Figure 24 below.

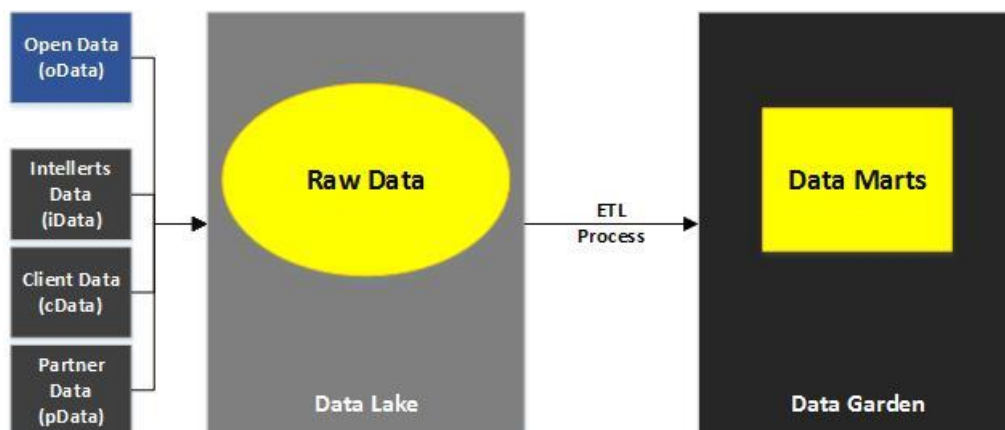
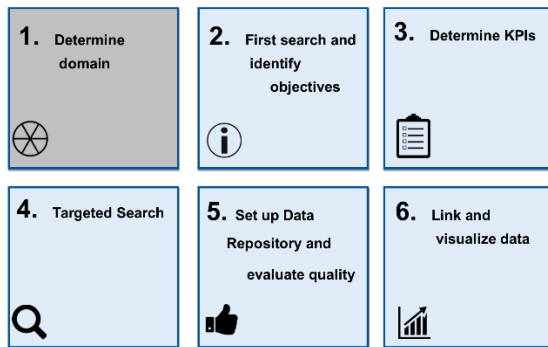


Figure 24: The Data Warehousing Process @ Berenschot Intellerts.

5.2 Step 1: Determine domain



In chapter 4 we tried to get some more insight in the open data landscape. What are frequent used open data sources? Which data sources are frequent used, and which topics do they include. The goal is to translate this into practice by setting up a taxonomy where different data sources are sorted per theme, so that the structure of the open data landscape becomes clearer. And so that when a project is started and the domain determined, data sources which might be useful are already structured for reuse. Both in business and academia this is from added value, because in both areas quantitative research on (open) data has a prominent place. The aim is not to provide a complete taxonomy in this report, but after each project useful, new sources should be add to the taxonomy to optimize the usefulness in the future.

The website data.overheid.nl consists a registry with information about and references to datasets from Dutch authorities. Thereby support on how to release (open)data is provided. In this data portal about ten thousand datasets in total can be found, however just thirty-three datasets are classified as ‘high value datasets’. Whether a dataset is a so called ‘high value dataset’ depends on the extent to which the dataset contributes to: transparency, statutory duty, cost reduction, target audience, and potential reuse. Although it is good to see there is a distinction between high quality and the rest, these terms are still quite vague terms to determine data quality. The datasets are divided into the following categories: “Bestuur”, “Cultuur en Recreatie”, “Economie”, “Financiën”, “Huisvesting”, “Internationaal”, “Landbouw”, “Migratie en integratie”, “Natuur en milieu”, “Onderwijs en wetenschap”, “Openbare orde en veiligheid”, “Zorg en gezondheid”, “Recht”, “Ruimte en infrastructuur”, “Sociale zekerheid”, “Verkeer”, and “Werk”. The Data owners of these datasets are the different ministries, but also the RDW (Dienst Wegverkeer), Kadaster, NDW (Nederlandse Databank Wegverkeergegevens), and the “Nationaal Archief”.

The UK government also has a data portal with government data, namely: data.gov.uk. This website uses more less the same themes: Education, Environment, Business and Economy, Crime and Justice, Defense, Government, Government spending, Health, Mapping, Society, Towns and Cities, and Transport.

From the themes above we choose – in consultation with the Intellerts people - the most relevant themes, namely: Finance (which comprises also accountancy), Energy & Environment, Transport & Infrastructure, Towns & Cities, Education, and Health. This choice of categories has similarities with the set of categories mentioned in the Open Data Guidebook of Bloomberg Philanthropies. According to them there is no consistent set of categories between open data portal, but the following are quite common and might serve as a starting point, namely: Business, Education, Environment, Finance, Health, Human (or Social) Services, Property, Public Safety, Recreation and Transportation.

Based on these six themes we set up a focus group to determine the most relevant open data sources per theme. Two data scientists and a consultant from Berenschot and Berenschot Intellerts provided their input, which resulted into the overview in Figure 25. The domain of the Case Study Project is Finance, since this comprises also accountancy.



Figure 25: Taxonomy of Open Data Sources

5.3 Step 2: First search and identify objectives



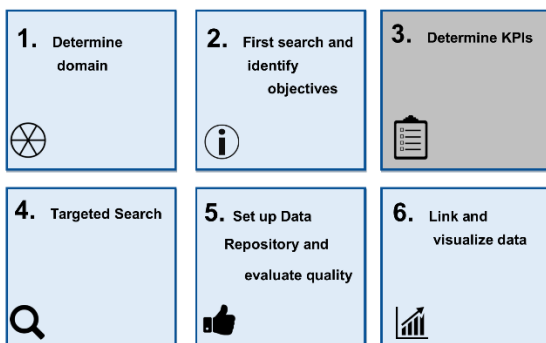
The project which counts as ‘case project’ is the accountancy project: the newest project at Berenschot Intellerts in the time we were there. In short, the project is a project from Berenschot Intellerts at an accountancy and consulting firm in the Netherlands. The project has two main goals: 1. Deliver new services for customers, and 2. Increase the efficiency. This are quite abstract goals, and therefore these are divided into several sub-goals.

To achieve new services for customers Berenschot Intellerts produces a data platform with applications on artificial intelligence and business intelligence. The platform contains four boxes: 1. Data universe (import and link sources, inclusive measuring quality of data), 2. Data factory (Integration, transformation and storing sources in database and add meta data), 3. Model factory (Data exploration), and 4. Model factory (Apply advanced analytics).

The second goal, to increase the efficiency in the work process of the client company, will be achieved by better data management, and by linking external data to the internal data of the accountancy firm. Therefore, also open data is needed, and this is where we came in. Our role in the project was to search for relevant open data regarding an accountancy firm; the role of a ‘data scout’. And thereby describe the steps we walked through to create a generalized road map for a data scout.

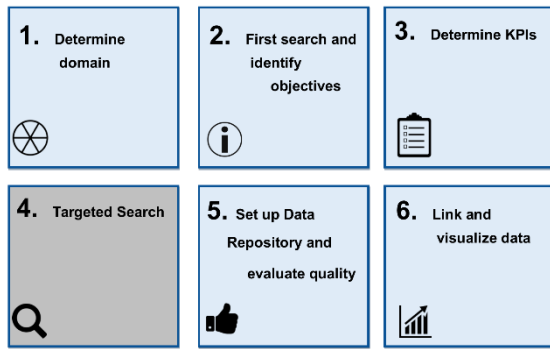
This project is a long-term project for Berenschot Intellerts, and therefore we have turned in the exploratory phase and startup phase. However, this research will be also relevant for the long-term of this project. Especially, regarding box 1 and 2 in the data platform we provide guidelines in terms of measuring and ensuring quality of data, and in terms of meta data management.

5.4 Step 3: Determine KPIs



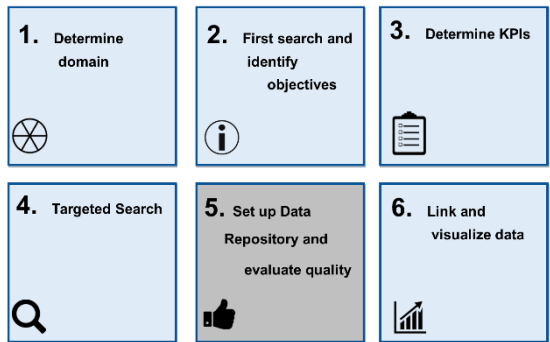
The third step is to translate the objectives to practice. Berenschot Intellerts defined a list of KPIs in consultation with the client. Since this is sensitive information, we cannot publish the KPIs in this report. An example of a general KPI is: the inflation in Europe measured over at least five years (see also visualization – step 6, top right).

5.5 Step 4: Targeted Search



During this step, we searched for open data we still missed to meet the KPIs, for example OECD we add later to the taxonomy. This step is very important during a project, because you really should make sure that the data you have provides enough information to satisfy the defined KPIs.

5.7 Step 5: Evaluate and ensure data quality



The fifth step is where our model comes in, because after searching for data we need to store the data. To ensure reuse and data quality a database needs to be FAIR. Therefore, we set, based on our model, up the data repository provided in Figure 26. After this we evaluated the quality based on the FAIR criteria, and determined which factors need some optimization. However, we provide the results of this step in chapter 7, after we first explain our model.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	Identifi	Title	Description	Category	Keywords	Indification Dat	Openl Clos	Source	License	Frequency updt	Temporal coverage	patial coverag	contact Informati	Language
1	Fin_S1	Conjunctuurklok	De stand en het	Finance	Nederlandse conjunctuur	14-2-2017	Open	https://www.cbs.nl/nl-nl/vs	unknown	maandelijks	2000 januari - 2017 februari	Landelijk NLD	x	NL
2	Fin_S2	Conjunctuurdashboard	Conjunctuurklok	Finance	Conjunctuurklok, Conju	12-8-2016	Open	https://www.cbs.nl/nl-nl/vs	unknown	jaarlijks	2001-2007	Landelijk NLD	x	NL
3	Fin_S3	Conjunctuurbeeld per bedrijfstak	Actueel samen	Finance	Conjunctuur	17-3-2017	Open	https://statline.cbs.nl/Statist	unknown	per kwartaal	2008 kwartaal 1 - 2017 kwarta	Landelijk NLD	x	NL
4	Fin_S4	Ziekteverzuimpercentage per bedrijfstak	Ziekteverzuim va	Finance	Ziekteverzuim	25-9-2015	Open	https://www.cbs.nl/nl-nl/vs	unknown	per kwartaal	1996 kwartaal 1 - 2016 kwarta	Landelijk NLD	x	NL
5	Fin_S5	ANBIs (belastingdienst)	Overzicht van Al	Finance	ANBI	27-8-2015	Open	https://www.belastingdiena	CC-O	Na aanpassing	2008-2017	Landelijk NLD	x	NL
6	Fin_S6	Financiën ministeries		Finance			Open	https://data.ovevheid.nl/dat		jaarlijks		Landelijk NLD	x	NL
7	Fin_S7	Uitgesproken faillissementen	Aantallen failliss	Finance	faillissementen	11-5-2017	Open	https://statline.cbs.nl/Statist	unknown	maandelijks	2009 januari - 2017 april	Landelijk NLD	x	NL
8	Fin_S8	Vestigingen van bedrijven	Aantal vestiging	Finance	Vestiging bedrijven	28-4-2017	Open	https://statline.cbs.nl/Statist	unknown	jaarlijks	2010-2017	Landelijk NLD	x	NL
9	Fin_S9	Ondernemingsklimaat	Internationaal ou	Finance	Ondernemingsklimaat, inn	23-2-2017	Open	http://statline.cbs.nl/Statist	unknown	jaarlijks	1990-2008	Landelijk NLD	x	NL
10	Fin_S10	Companyinfo	Jaarrekeningen	Finance	Jaarrekeningen	x	Closed	x	x	x	x	x	x	x
11	Fin_S11	Mijn gemeente		Finance, Gov			Open	www.waartatiegemeente				Landelijk NLD	x	NL
12	Fin_S12	Informatie voor Derden	Financiële gegee	Finance	Begrotingen, jaarrekening	10-5-2017	Open	https://www.cbs.nl/nl-nl/vs	unknown	per kwartaal	2010 kwartaal 1 - 2017 kwarta	Landelijk NLD	x	NL
13	Fin_S13	Rechterlijke uitspraken	Een deel van de	Finance	Rechterlijke uitspraken		Open	https://uitspraken.rechtspe				Landelijk NLD	x	NL
14	Fin_S14	OECD	Bank profitability	Finance	OECD, Bank, Profitability	unknown	Open	http://stats.oecd.org/Index	unknown	jaarlijks	1979-2009	Wereldwijd	x	EN
15	Fin_S15	AssetMacro	Economic Indica	Finance			Open	https://www.assetmacro.co				Wereldwijd	x	EN
16	Fin_S16	Business Confidence Index (BCI)	Business confid	Finance	BCI, Business, confidence	unknown	Open	https://data.oecd.org/leac	unknown	maandelijks	1974-2017	Landelijk NLD	x	EN
17	Fin_S17	Consumer Confidence Index (CIC)	Consumer confid	Finance	CIC, Consumer, confidenc	unknown	Open	https://data.oecd.org/leac	unknown	maandelijks	1973-2017	Landelijk NLD	x	EN
18	Fin_S18	CGI, Europe	Consumer confid	Finance	CGI, Consumer, confidenc	unknown	Open	https://data.oecd.org/leac	unknown	maandelijks	1973-2017	Europe	x	EN
19	Fin_S19	Gross Domestic Product (GDP)	Gross Domestic	Finance	GDP, BBP, National accou	unknown	Open	https://data.oecd.org/gdp/	unknown	jaarlijks	1969-2016	Landelijk NLD	x	EN
20	Fin_S20	GDP, Europe average	Gross Domestic	Finance	GDP, BBP, Europe, OECD	unknown	Open	https://data.oecd.org/gdp/	unknown	jaarlijks	1995-2016	Europe	x	EN
21	Fin_S21	UvV		Finance		x	Closed	x	x	x	x	x	x	x
22	Fin_S22	Branchemonitor	Benchmark bran	Finance	Bedrijven, Branches		Open	https://www.cbs.nl/nl-nl/vs				Landelijk NLD	x	NL
23	Fin_S23	Benchmarks	Rapporten Alger	Finance			Open	https://data.ovevheid.nl/vs				Landelijk NLD	x	NL
24	Fin_S26	Caos	Nieuwsite met all	Finance	cao	unknown	Open	https://www.loonwijzer.nl/	unknown	After modification		Landelijk NLD	x	NL

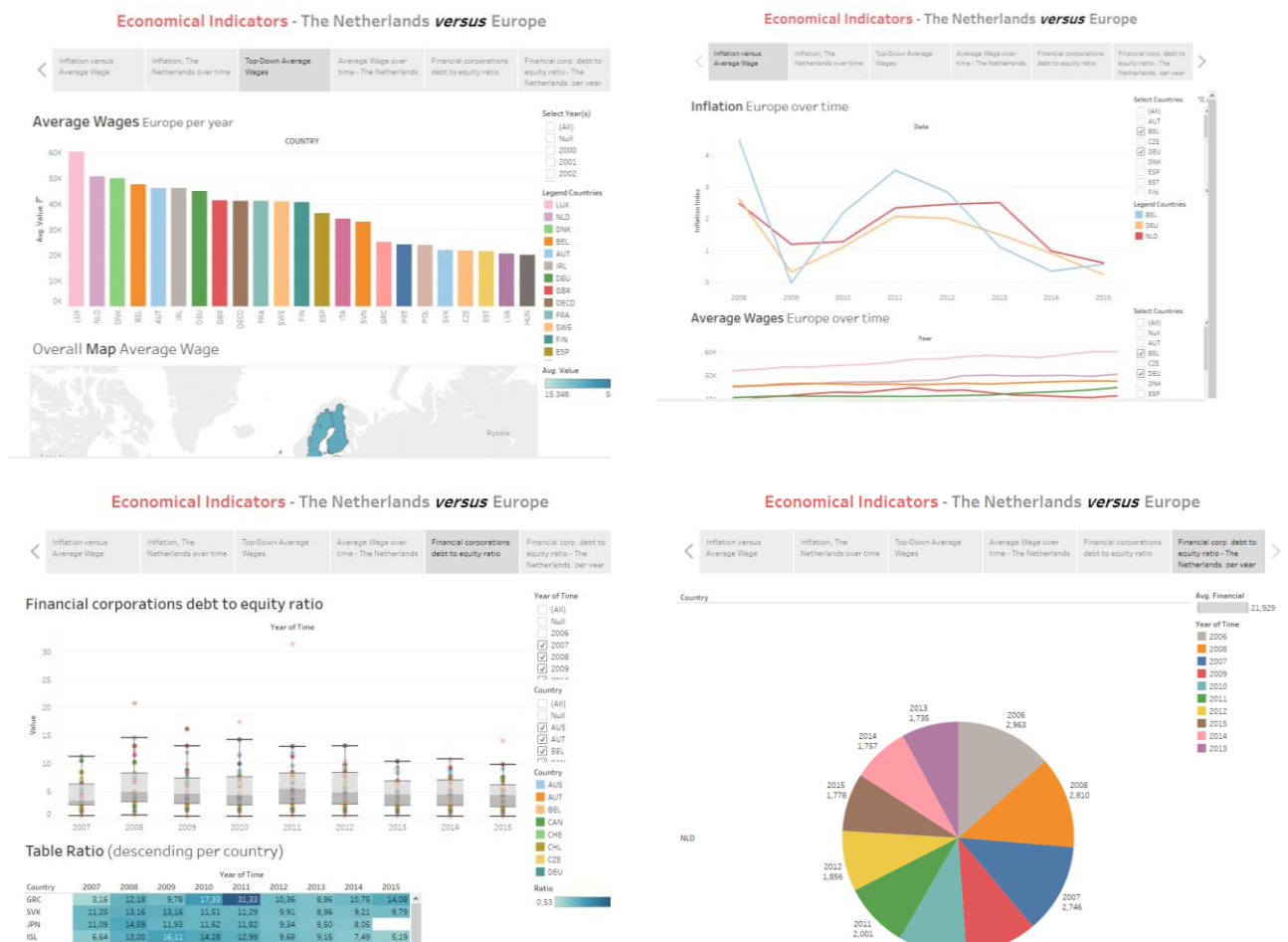
Figure 26: Data Repository Finance

5.8 Step 6: Linking and Visualization



After the Data, KPIs, and the repository is set up and evaluated, it is time for the 'real work'; linking and visualization. This step is where the final product for the customer is created, the dashboards. It can be easy to first create a data warehouse structure, like we did (see Figure 27), to know how the data can be linked (based on which key). In the Figure 27 we provide an example of such a data warehouse structure from the theme Finance (from the above taxonomy). But of course, this is not only applicable to these themes. In general, it is very important to document the keys per dataset, so that it is easier to reuse (to link it to other data).

Finally, we provide some examples of parts of the visualizations we created in Tableau (this can be any other visualization tool) below:



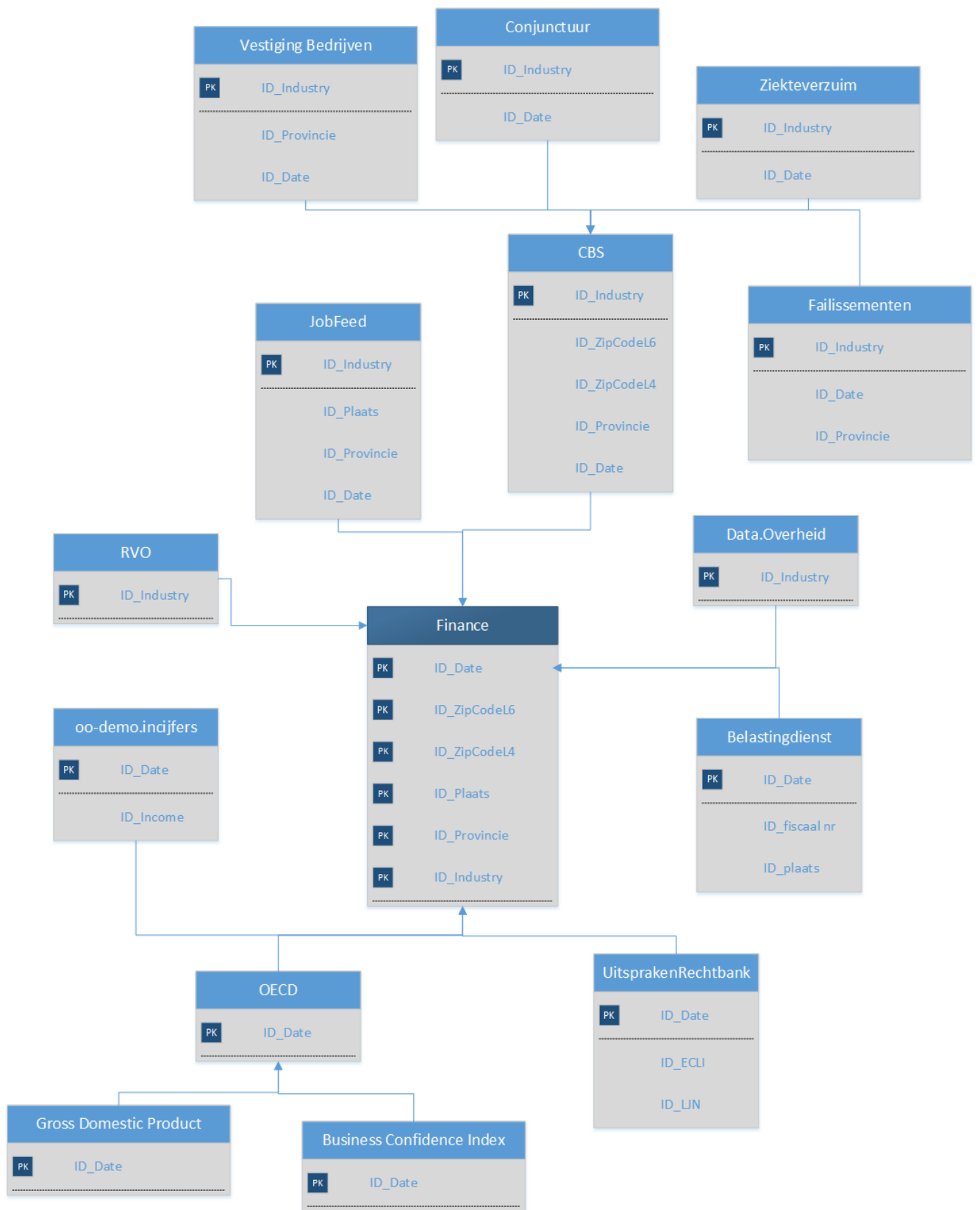


Figure 27: Data warehouse Structure Finance

Because the structure in Figure 27 is quite general, and a lot of datasets are behind the portals mentioned, we also provide a structure from another project. This is the Energy project and was actual performed at Berenschot Intellerts. We provide them the visualization in Figure 28. We see that when it is about one project the syntax is lot simpler; in this case even all datasets are linked based on the same key ('Date': Time dimension).

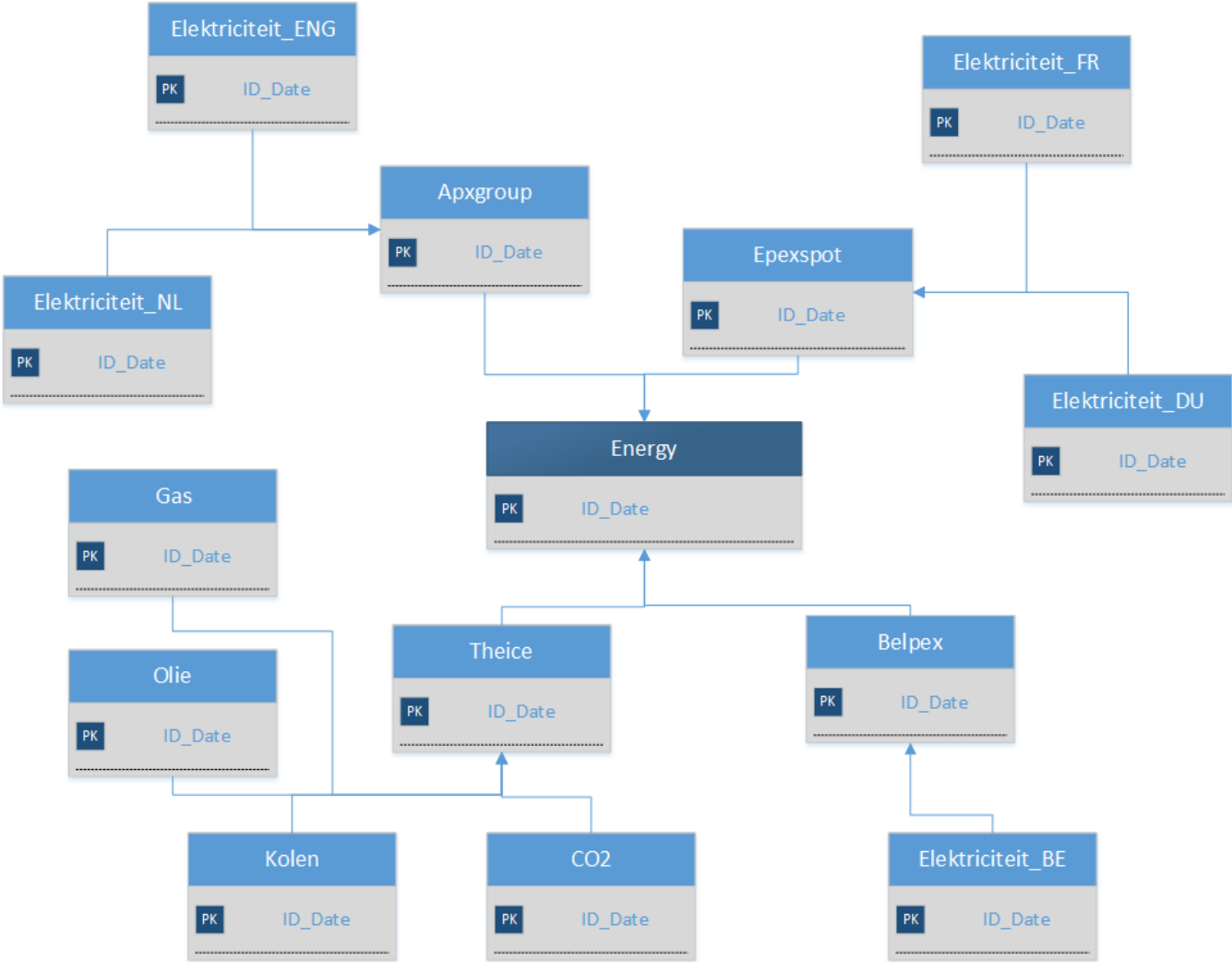


Figure 28: Data Warehouse Structure Energy project

6. Research Question 3: Meta Data Management

WHAT ARE THE META DATA REQUIREMENTS THAT SATISFY THE FAIR PRINCIPLES?

WHY translating FAIR definitions into meta data requirements?

In research question 1 the definitions of the FAIR concepts are provided, in answering the second research question the context is sketched. The final step in answering the main research question is to translate this theory into 'practical' meta-data requirements. The third research question is therefore: What are the meta-data requirements that satisfy the FAIR principles? This to determine what is the best way to describe and maintain (open) data sources. To answer this research question, we first describe the relevance of metadata and different types of metadata, then we will determine the basic elements for metadata of open data sources. Thirdly, we elaborate on the requirements for the 'FAIR Software'. And finally, we provide a list of requirements to ensure a FAIR data repository. The main goal is to make the FAIR concept tangible, in the sense that these concepts become a designation for data quality. This is the second step to, due to the problem statement, bridge theory with practice by applying FAIR into practice.

6.1 Metadata management in general

Already a few times the term 'metadata' is used without having been specified further. In this section, two perspectives on metadata are given. One straightforward definition of metadata is: data about data. However, this definition appears to not reflect the full scope of metadata. Hence, a more elaborated version used in this thesis is: "Metadata is the description of the data as it created, transformed, stored, access, and consumed in the enterprise" (p. 79, Sherman, 2014). It implies that metadata is about the context of data, and this corresponds to the elaboration on provenance in the first chapter, section 3.2: it provides you information on the what, when, where, how, by whom, and why of data.

Metadata is from added value from business and IT perspective. Business people need to know what data they are using, what the information they are analyzing represents; where did it come from, how was it transformed? Also for IT metadata is essential. Business requirements and coding specifications indicate what we intend to build, but it is the metadata that describes what has been built, tested, deployed, and is currently being used (Sherman, 2014). Sherman (2014) calls it 'business perspective', but this perspective can also be applied on academia. Academics also want to know what data they are using, and what the information they are analyzing represents. We observe an overlap in the need in business and academia here.

Resulting from those two perspectives Sherman (2014) also distinguished two types of metadata, namely: business and technical metadata. The Business metadata is according to him the description of information from the business perspective, like the weekly sales or budget variance reports. Most of the data the business people care about is not used by the software tools. Therefore, the Technical metadata is used. Technical metadata is the description of data as it is processed by software tools. Databases, for example, need to define columns (format, size, etc.), tables, and indexes. ETL tools need to define fields, mappings between source and targets, transformations, and workflows. And BI tools need to describe fields and reports. All this metadata is used to enable the software tools (not people) to understand and process data.

According to Sherman (2014) metadata management is not at the top of most BI teams or business people's priority lists. This was also confirmed in the interviews we did. However, handling metadata in a good way is essential to implement and manage technologies and products. Business people need to know how the data looks like they want to use for their business analytics, and IT people need to know what happened to the data to provide consistent, comprehensive and clean data for (deeper) business analytics. Metadata management is the way to determine the quality of data in a structured way.

6.2 Types and relevance of Metadata requirements

The distinction from Sherman (2014) between business metadata and technical metadata is still quite general. Although the Open Data guidebook of Bloomberg Philanthropies uses a comparable distinction, namely: metadata that provides an overview of data versus metadata that provides details about specific parts of your data. Riley (2017) makes in the primer publication of the national information standards organization a more detailed distinction: descriptive metadata, administrative metadata, structural metadata, and markup languages. Whereby descriptive metadata is for finding or understanding a resource. Administrative metadata can be divided into three subcategories: technical metadata (for decoding and rendering files), preservation metadata (long-term management of files) and rights metadata (intellectual property right attached to content). Structural metadata is about relationships of parts of resources to one another. And markup languages integrate metadata and flags for other structural or semantic features within content.

Metadata management is especially relevant in terms of knowledge management. When we analyze where corporate knowledge in organizations is stored. According to Marco (2000) 42% of the knowledge is stored in the brains of Employees, whereas the rest of the knowledge is in Electronic Documents, Paper Documents or other options (see Figure 29). The key for organizations is how to gather, retain and disseminate knowledge in a proper way. Meta data repositories are part of the solution to realize this.

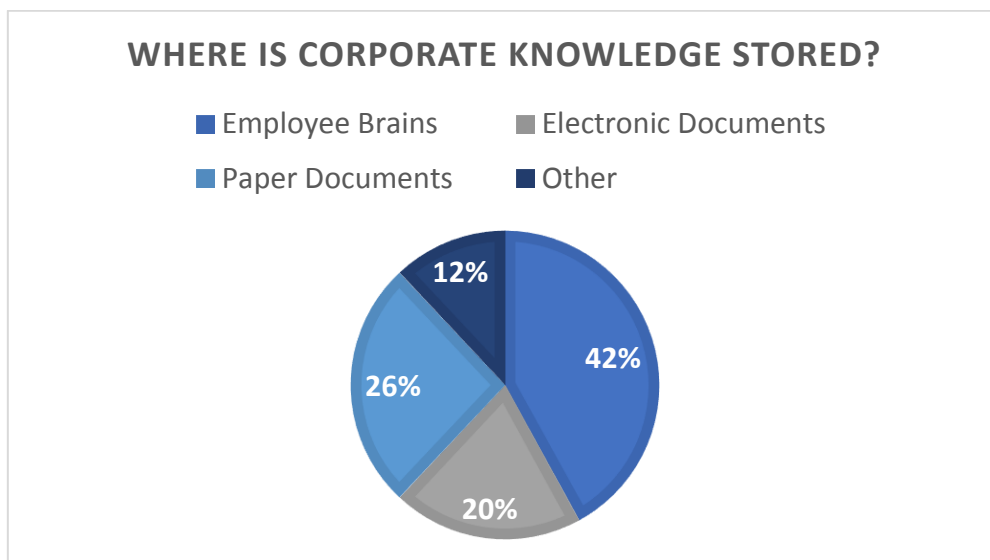


Figure 29: Where corporate knowledge within organizations is stored (Marco, 2000)

6.2 Realizations of metadata management

Metadata is only useful if it is understandable to software applications and people that use it. To realize this understanding a main aim of authorities and organization is to standardize metadata by predefining metadata sets. This can be done by standardizing syntax, e.g. make use of XML metadata vocabularies, known as schemas, element sets, or formats. In addition to standardize syntax, standardization of metadata can be also realized through control of the actual values used. One way to do this is to make use of controlled vocabularies. A controlled vocabulary is a predetermined list of terms on a certain topic or of a certain type. These lists typically identify one preferred word or phrase for a given concept, and sometimes provide mappings from other terms for the concept to the preferred one. A second option for standardizing the values that appear in metadata is the use of content standards. This are sets of guidelines that dictate how textual values in metadata should be structured. They can be also known as style guide (Riley, 2017).

The wish to standardize syntax for metadata causes a large set of all different kind of open metadata standards. We mentioned XML already before. XML is a common example of a metadata standards, and is a commonly used encoding, transfer, and occasional internal system storage mechanism for metadata (Riley, 2017). Another example, already mentioned in this thesis, is Linked Data and RDF. The concept of Linked Data was introduced by Tim Berners-Lee. Implementation of this idea involves organizations publishing their structured data on the Web,

explicitly naming entities in this data so they can be referenced by others, and linking to others' data to build a worldwide information network. Finally, also the well-known ISO-standards provide enough guidelines for metadata management.

However, these standards are in most of the cases specific per sector or not used within organizations since employees do not fill in all the needed information since lists of requirements are too long (Reference interviews). Therefore, we decide to use the relative short list of 'basic' metadata elements from center for government excellence² as starting point for this research. The elements are provided in the next section.

6.3 Basic Elements Metadata

Almost each dataset which will be published will include many of the following meta data elements:

- Title (or Name): Human-readable name for the data. It should be in plain English and include sufficient detail to facilitate search and discovery. Acronyms should be avoided.
- Description: Human-readable description (e.g. an abstract) with sufficient detail to enable a user to quickly understand whether the asset is of interest.
- Category (or Theme): Main thematic category of the dataset, usually chosen from a predefined list.
- Keywords (or Tags): Tags are generally single words which help visitors discover the data. These are terms that would be used by technical and non-technical users.
- Modification Date: The most recent date on which the dataset was changed, updated, or modified.
- Contact Information: The name and email address of the publisher of a dataset.
- License: Often datasets on open data portals are available in the public domain with no restrictions on reuse, however there may be circumstances where a specific dataset is offered using a different license.

These elements can be supplemented with elements like frequency with which the dataset is updated, temporal coverage (the range of time included in the dataset), and the spatial coverage (the geographic area for which this dataset is relevant).

6.4 FAIR Data Point Software Specification

Second part regarding the third research question is how to relate this meta data requirements to the FAIR principles. Therefore, we first elaborate on some FAIR initiatives which already exist. First of all, we discuss in short a (draft) document called 'FAIR Data Point Software Specification and the relating requirements from Gavai, Kuzniar, and Kaliyaperumal (2016). And second, we elaborate on the initiative (also in set up phase) from Doorn & Dillo (2016), which combines the 'DSA requirements' (The European Framework for Audit and Certification of Digital Repositories) with FAIR.

In October 2016 Gavai, Kuzniar, Kaliyaperumal, and Burger published a (draft) document called 'FAIR Data Point Software Specification'. In this document, they published requirements for so called FDP Software. According to them FDP is the main goal of this software that is allows data owners to expose datasets in a FAIR manner, and allows data users to discover metadata about these datasets and, if license conditions allow, to access this data. Thereby the developers want to make it also possible for existing data repositories to implement FDPs and API and metadata content, behaving in this way also as a FDP. Especially this 'metadata content' is interesting, because this is also where this research is about. However, a pilot of this project is upcoming. Currently [February-September 2017] the first versions of FDPs are being designed and developed in the context of ODEX4ALL, BBMRI 2.0, RD-Connect and the Elixir Rare Disease pilot (see also: <http://www.dtls.nl/fair-data-points-as-eudat-services/>).

For now, we will look at the requirements for this software. These are still quite generic; however, it is good to keep it in mind. Gavai et al. (2016) defined different usage scenarios to derive the requirements for data storage and accessibility infrastructure. They distinguish: Data discovery, Data access, Data publication, and Data metrics gathering. Data discovery and Data access speaks for itself. The FDP software is focused on academia, especially on life sciences (biological datasets), therefore Data publication is less important in business context. And Data metrics gathering is about the number of users accessing the metadata; who they are, where they are coming

² <https://centerforgov.gitbooks.io/open-data-metadata-guide/content/dataset-metadata.html>

from and such. These scenarios result in the following requirements/ goals for the software to ensure FAIR datasets:

- Allow data owners to expose their datasets in a way that complies with the FAIR Data Principles.
- Allow data consumers to discover information about the FAIR Data Point, its offered datasets and the actual data items from each of the datasets
- Allow data consumers to access the data. Whenever the license of a dataset imposes further restrictions, the FDP should enforce these restrictions.
- Allow the data owner to gather access metrics about the offered (meta)data.
- Allow interaction for both humans (GUI) and software agents (API).

These goals are modelled in an architecture in the ArchiMate modelling language and can be found at: <https://dtl-fair.atlassian.net/wiki/display/FDP/FAIR+Data+Point+Software+Specification>.

6.5 DSA versus FAIR

The requirements mentioned in the paragraph before are still quite generic. Therefore, we will discuss in this section the so called DSA requirements. The *European Framework for Audit and Certification of Digital Repositories* was defined in a memorandum of understanding signed in July 2010 between Consultative Committee for Space Data Systems (CCSDS), Data Seal of Approval (DSA) Board and German Institute for Standardization (DIN) working group.

The framework is intended to help organizations in obtaining appropriate certification as a trusted digital repository and establishes three increasingly demanding levels of assessment:

- Basic certification: self-assessment using 16 criteria of DSA
- Extended Certification: Basic certification and additional externally reviewed self-audit against ISO 16363 or DIN 31644 requirements
- Formal certification: validation of the self-certification with a third-party official audit based on ISO 16363 or DN 31644.

For now, we will just look at the 16 DSA criteria/requirements. These are as follows:

1. *The repository has an explicit mission to provide access to and preserve data in its domain.*
2. *The repository maintains all applicable licenses covering data access and use and monitors compliance.*
3. *The repository has a continuity plan to ensure ongoing access to and preservation of its holdings*
4. *The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms*
5. *The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission*
6. *The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse, or external, including scientific guidance, if relevant)*
7. *The repository guarantees the integrity and authenticity of the data*
8. *The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users*
9. *The repository applies documented processes and procedures in managing archival storage of the data*
10. *The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way*
11. *The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations*
12. *Archiving takes place according to defined workflows from ingest to dissemination*
13. *The repository enables users to discover the data and refer to them in a persistent way through proper citation*
14. *The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data*

15. *The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community*
16. *The technical infrastructure of the repository provides for protection of the facility and its data, products, services and users*

We classified the DSA requirements per FAIR principle as follows to give an indication of how strongly related these requirements are with the FAIR concepts. Although it becomes clear that the DSA requirements are even more detailed.

Findability

- 5 (clear governance)
- 8 (metadata for understanding)
- 11 (sufficient info available)
- 13 (proper citation)

Accessibility

- 1 (access)
- 2 (licenses)
- 3 (ongoing access)
- 7 (integrity and authenticity)

Interoperability

- 9 (documented processes and procedures)
- 10 (long-term preservation)
- 12 (archiving based on defined workflows)
- 15 (well-supported operating systems)
- 16 (infrastructure for protection)

Reusability

- 4 (ethical norms)
- 6 (help of experts)
- 14 (reuse of the data over time)

We found these requirements and the strong relation with FAIR by ourselves, however, during this research we also found out that Peter Doorn and Ingrid Dillo hold in December 2016 a webinar about a possible combination between FAIR and DSA (<http://aims.fao.org/activity/blog/put-fair-principles-practice-and-enjoy-your-data>).

We did not find a scientific paper (it is a quite new idea), but we will elaborate this idea and combine it further with our ideas. So, from now we will refer to this idea as 'Doorn & Dillo (2016)'. For both of us the starting point is to set up quality criteria for datasets. Doorn & Dillo (2016) consequently speak about 'research datasets', we pull it wider and assume that there is also a demand for this from business perspective.

Doorn & Dillo (2016) conclude correctly that both DSA and FAIR do not make value judgements about the content of datasets but rather qualify the fitness for data reuse in an impartial and measurable way. Thereby, we even concluded that at least the FAIR principles are not measurable (in terms of tangible requirements) at all. The difference is according to Doorn & Dillo (2016) that the FAIR principles present quality criteria for target individual datasets, and the DSA presents quality criteria for digital repositories.

Finally, what we did with the classification above, Doorn & Dillo (2016) summarized in Figure 30. Roughly the only difference is that we decided to classify 'CITABLE' under Findability (in accordance to our definition that data should be identified in a unique way), and they mention it as separate concept.

DSA Principles - for data repositories

FAIR (Findable, Accessible, Interoperable, Reusable) data - both for machines and for people

The data can be found on internet	FINDABLE
The data are accessible (clear rights and licenses)	ACCESSIBLE
The data are in a usable format	INTEROPERABLE
The data are reliable	REUSABLE
The data are identified in a unique and persistent way to be referred to	(CITABLE)

Figure 30: Overlap between DSA and FAIR requirements

6.6 Metric and Scoring Mechanism Approach

In this section we will provide, in accordance to Doorn and Dillo (2016), a possible scoring mechanism to determine the FAIRness of a data repository. One of the main conclusions in our literature research is that reusability is the result of the other three concepts. Doorn & Dillo (2016) also confirm that reusability is the resultant of the other three. Therefore, they approach the same metric as we want to present, namely: $(F + A + I)/3 = R$.

An example of how FAIR then can be determined is as follows:

Dataset X has FAIR profile F4-A3-I2, which results in R=3 (on a scale from 1 to 5)

According to Doorn and Dillo (reference) we should interpret this as:

ID with limited metadata (F4)

Accessible with some restrictions (A3)

Fairly low interoperability (I2)

Thereby they suggest to additionally indicate the number of assessments, reviews and downloads as well (See Figure 31).



Figure 31: Example FAIR Profile Dataset X

We conclude that the disadvantage of determine the FAIRness of a whole database in this way, is that first it should be measured per single dataset, and only then the average FAIR score can be calculated. Therefore, this process should be automated; a program needs to be written. And thereby, we discover another need, namely: how to interpret this FAIRness into 'normal language' and which steps of preparation are needed to make sure FAIR can be measured. The need for a program which automates this process is in production under guidance of Doorn &

Dillo (2016) Therefore, especially the second need to provide step-by-step the activities to ensure the 'FAIR measures' can be carried out on a data repository, is the starting point for building our models. We elaborate how we build and provide these models in the next chapter.

7. Model Design Reference Models

Model Construct Reference Models

In this chapter, we present our reference models, how to interpret this model and the steps which were required to design this. All three research questions come together in this chapter. The literature research on the definitions (RQ1) and the inventory of metadata requirements (RQ3) provide the input for the models, and the Case Study (RQ2) provides the context and practical implementation of the models in an actual open data project. In Figure 32 we provide in the steps of our final model construct, to give more insight in the different phases in the construction of the model.

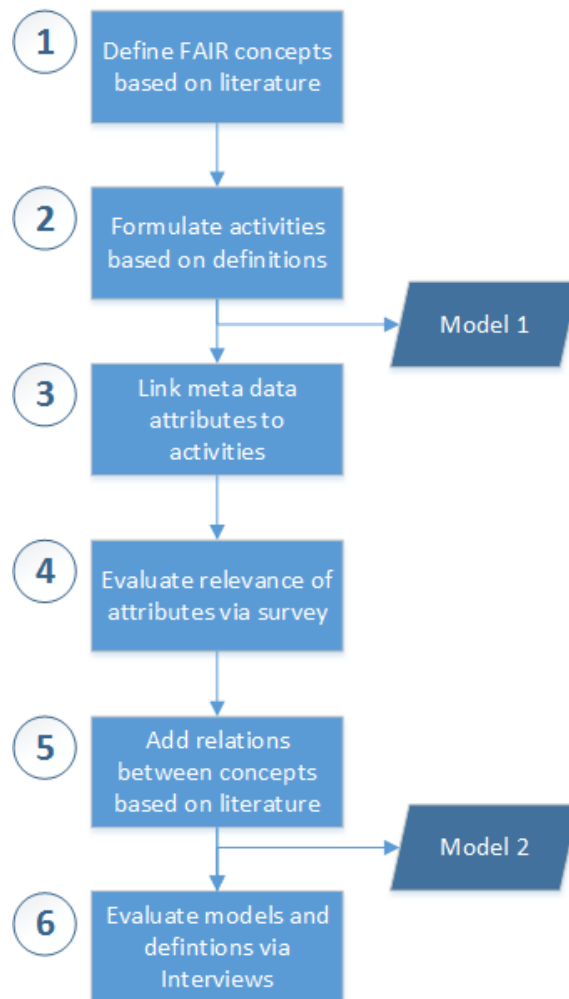


Figure 32: Visualization Model Construct

7.1 Key parts within FAIR definitions

In the first research question a literature research is provided, and based on these four new definitions for the FAIR concepts are defined. The definitions are listed below again, and the key parts are highlighted in red. These key parts form the basis for the activities within the model.

Definition 1

FINDABILITY (of data) is about **the uniqueness** and **provenance** of data, which means data described with rich meta data with focus on contextual information. This is achieved by versioning, associating URLs or embedding other kind of identifiers.

Definition 2

ACCESSIBILITY is a semantic quality dimension, which comprises **availability**, **security**, performance, **interactivity**, flexibility of data, to ensure data access to (end) users regardless their different context and background.

Definition 3

INTEROPERABILITY is a system feature to **connect** and **conform heterogeneous environments**, to ensure sharing, reusing and exchanging of data between these environments without special effort from the (end) user.

Definition 4

REUSABILITY is the ability to **make easily use of data** in a new context, in terms of retrieving, downloading, indexing, searching and visualizing the data without restrictions, **as a result of findable, interoperable and accessible data**.

7.2 Model 1: FAIR activities model

Based on the key parts highlighted in the section before activities per concepts are abstracted. This results in the 'FAIR activities model'. Whereby the concepts consist of the following activities:

Id	Activity
1	Determine a unique identifier per dataset
2	Collect the provenance of data
3	Ensure the data is available
4	Ensure the data is secure and legal to use
5	Ensure that interactivity between data is possible
6	Conform the different data sources
7	Connect the different data sources
8	Determine the plausibility of reuse
9	Ensure that editing the data is possible

Table 9: Id Activities Reference Model



Figure 33: Model 1 FAIR Activities

7.3 Relations between the concepts

In chapter 3 the concepts are discussed separate. In this section, we provide a model for the mutual relations between the four concepts. The model is shown in Figure 34.

Based on the definitions provided in chapter 3, we roughly conclude that ‘Findability’ and ‘Accessibility’ form the fundament for ‘Interoperability’, which results in ‘Reusability’. In short, when a set of data sources – a data repository – contains findable and accessible data, it is easier to connect sources (interoperability between sources), and this makes data suitable for reuse.

However, we have two make two comments on this model. First, the relations between these concepts are not rigid, which means they are not indeterminate. We can imagine there could be a data repository which is interoperable, but not totally accessible. Therefore, we should keep in mind this model is just a guideline to get the concepts clearer in the basics, and to give attention to the mutual cohesion. But the concepts function separate from each other.

Second, the meta data attributes which are added in the model are not sufficient to get a totally FAIR data repository. However, on the other hand they are an indication, based on the results from literature, to describe the practical interpretation of each concept. In the next section, we provide an explanation per attribute why it is linked to a specific concept, and how it is rated by the respondents (experts) in a survey.

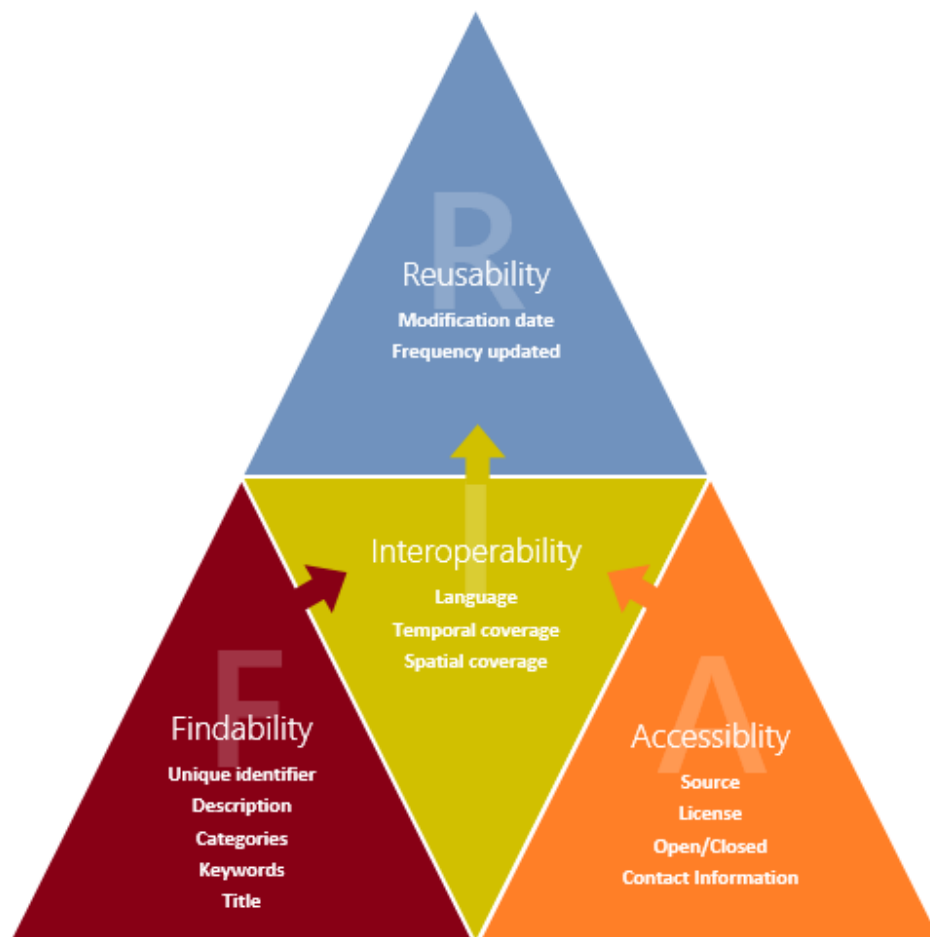


Figure 34: Model mutual relations FAIR concepts

Based on the founded set of basic meta data elements in section 6.3, we make a classification of these attributes per FAIR concept (See Table 4). Thereby, we discussed with some people at Intellerts (based on their experience) which attributes seem to be optional and which ones are required to have, and compared this with the (overlapping) set of metadata attributes in the FAIR Data Point Software Specification (<https://dtl-fair.atlassian.net/wiki/display/FDP/FAIR+Data+Point+Software+Specification>). This set of attributes form the basis for the design of our reference model. During the evaluation (chapter 8) we check whether this classification is right.

Attribute	FAIR concept	Optional/Required
Unique Identifier	Interoperability, Findability	Required
Title	Findability	Required
Description	Findability	Optional
Category	Findability/ Reusability	Required
Keywords	Findability/ Reusability	Required
Modification Date	Reusability	Optional
Source	Accessibility	Required
License	Accessibility	Optional
Frequency updated	Reusability	Optional
Temporal coverage	Interoperability	Required
Spatial coverage	Interoperability	Required
Open/Closed Data	Accessibility	Required
Contact Information	Findability	Optional
Language	Interoperability	Optional

Table 9: Classification Meta Data Attributes

7.4 Final Model

Thirdly, we link the activities – based on the definitions – to the basic meta data attributes. And these attributes are rated by experts in a survey how relevant each attribute really is. This results in the table below; the first column contains the four concepts, the second refers to the activity id in table 9 section 7.2, the fourth column contains the indication for relevance (1-5) based on the survey, and the fifth column provides an explanation why every concept consists of these activities. Finally, table 10 provides the distribution of the answers from the respondents per attribute, and for more details on the survey we refer to chapter 8.

Concept	Activity Id	Attribute	Value (based on survey: rated 1-5)	Explanation
Findability	1	Unique Identifier	4,67	Based on literature research and the survey the identifier is the most important meta data attribute. Provenance is a wide term. It starts with the basics: title, description, category, and keywords. These attributes are standard for creating good findability of information/data. Whereby 'Description' is ranked as most important and 'Keywords' are the least important.
Findability	2	Description	4,13	
Findability	2	Category	3,57	
Findability	2	Keywords	3,01	
Findability	2	Title	3,93	

Accessibility	3	Source	3,64	If open data is available it means it is accessible via a proper URL. If data is open interactivity is possible. Always check whether the total needed data is open.
Accessibility	3/5	Open/Closed	3,36	
Accessibility	4/5	License	3,21	License is important for open and closed data, to check whether it is legal to use the (whole) dataset. If Data is closed it can be useful to get in contact with the data provider. There is a chance interactivity is possible after this.
Accessibility	4	Contact Information	3,42	
Interoperability	6/7	Language	3,29	To connect different sources, it is important datasets are on the same level. These need to be preferably 'conform' in language, Temporal coverage (period data is available for), and Spatial coverage (on which level; zip code, worldwide etc.).
Interoperability	6/7	Temporal coverage	3,08	
Interoperability	6/7	Spatial coverage	2,92	
Reusability	8/9	Modification date	4,07	Check before reusing data always the modification data and last updates to know how to interpret the data. After this you can edit the data for what you want.
Reusability	8/9	Frequency updated	2,93	

	1 (unnecessary)	2	3	4	5 (indispensable)	N.A.	Totaal	Gewogen gemiddelde
Title	0,00% 0	11,76% 2	11,76% 2	29,41% 5	29,41% 5	17,65% 3	17	3,93
Unique Identifier	0,00% 0	5,56% 1	0,00% 0	11,11% 2	66,67% 12	16,67% 3	18	4,67
Description	0,00% 0	0,00% 0	11,11% 2	50,00% 9	22,22% 4	16,67% 3	18	4,13
Category	0,00% 0	5,88% 1	29,41% 5	41,18% 7	5,88% 1	17,65% 3	17	3,57
Keywords	0,00% 0	22,22% 4	38,89% 7	5,56% 1	11,11% 2	22,22% 4	18	3,07
Modification Date	0,00% 0	0,00% 0	16,67% 3	44,44% 8	22,22% 4	16,67% 3	18	4,07
Source (URL)	0,00% 0	16,67% 3	5,56% 1	44,44% 8	11,11% 2	22,22% 4	18	3,64
License	16,67% 3	5,56% 1	16,67% 3	22,22% 4	16,67% 3	22,22% 4	18	3,21
Frequency updated	5,56% 1	27,78% 5	16,67% 3	33,33% 6	0,00% 0	16,67% 3	18	2,93
Temporal Coverage	5,56% 1	5,56% 1	44,44% 8	11,11% 2	5,56% 1	27,78% 5	18	3,08
Spatial Coverage	5,56% 1	5,56% 1	50,00% 9	11,11% 2	0,00% 0	27,78% 5	18	2,92
Open/ Closed Data	5,88% 1	11,76% 2	17,65% 3	11,76% 2	17,65% 3	35,29% 6	17	3,36
Contact Information	0,00% 0	23,53% 4	11,76% 2	17,65% 3	17,65% 3	29,41% 5	17	3,42
Language	5,56% 1	5,56% 1	38,89% 7	16,67% 3	11,11% 2	22,22% 4	18	3,29

Table 10: Distribution of the answers from the respondents in the survey

Based on all information above we created first the model provided in Figure 36. But after the evaluation we decided to change this model into the model in Figure 35. This will be further motivated in chapter 8. This model is a combination of the activities, the relations between the concepts and the corresponding (see colors) set of meta data attributes.

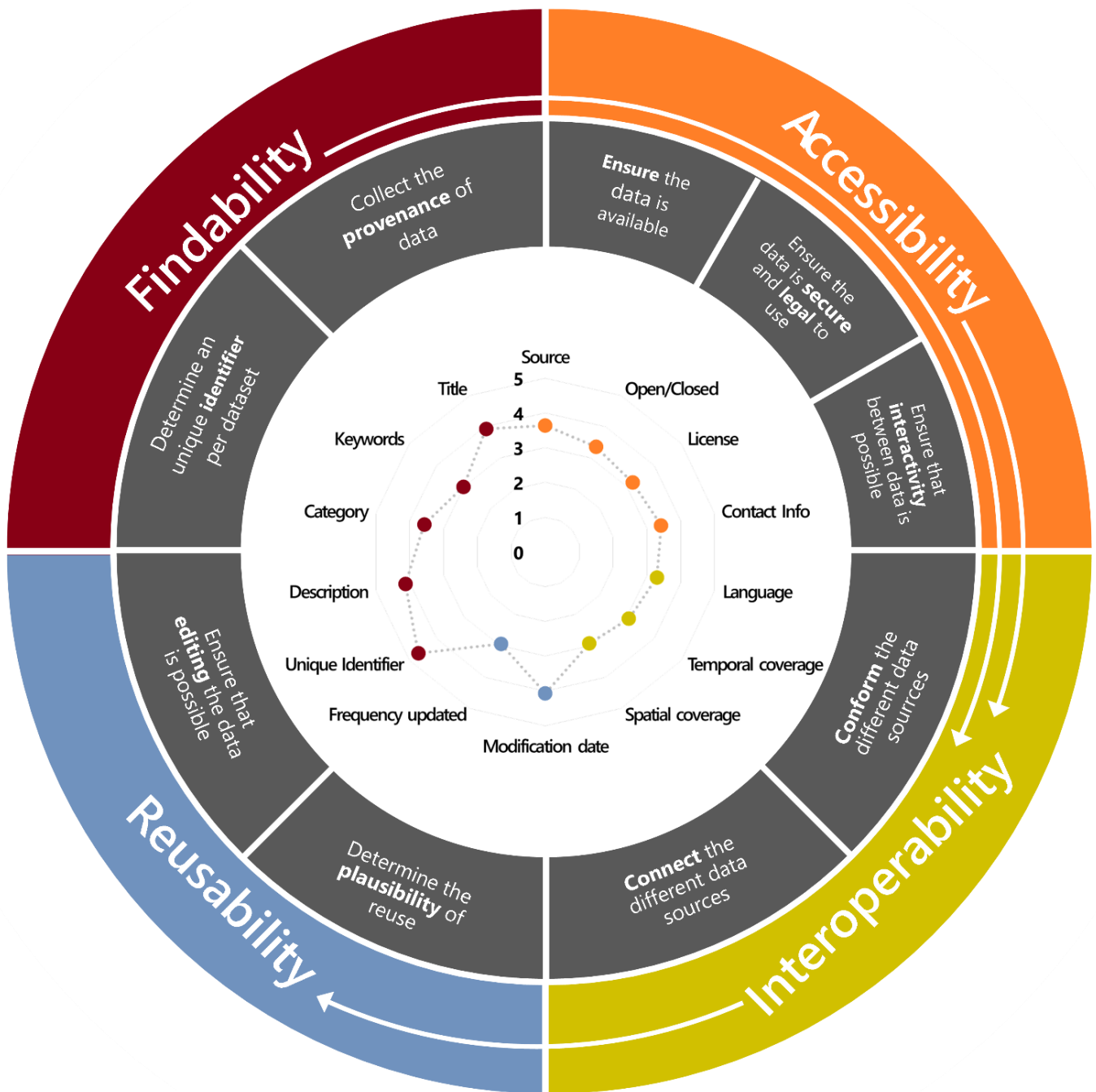


Figure 35: Final Reference Model 2

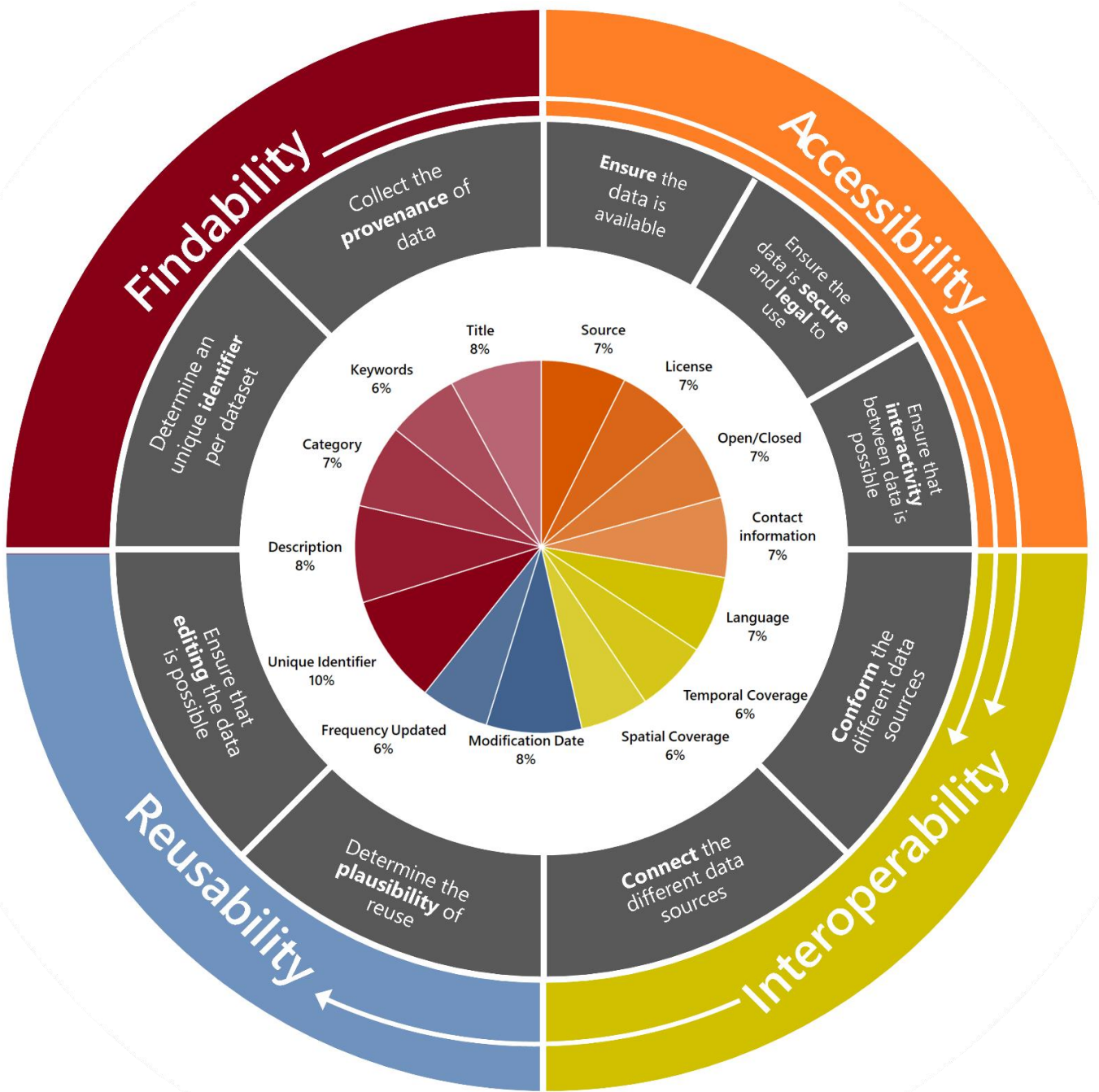


Figure 36: Draft Reference Model 2

7.5 Interpretation Final Model

7.5.1 Interpretation outer circle

Creation

The outer circle (see Figure 37) of the final reference model basically is a combination of the first reference model (Figure 33) and the relations between the concepts (Figure 33).

Main goal

The main goal is to provide a step-by-step set of activities to ensure a data repository contains all aspects to measure FAIR, in normal language, and which is applicable in the public and private sector.

The reference model is especially useful for setting up new repositories, however also existing ones can be edit.

Meaning

Every colored block (red, orange, green, blue) contains one of the four concepts, and their corresponding activities (grey). The white arrows show the relationship between the activities, as explained in section 7.3, and at the same time the order of the activities.



Figure 37: Outer Circle Reference Model

7.5.2 Interpretation middle part

Creation

The middle part (see Figure 38) consists of the set of meta data attributes, and the ranking based on the survey results.

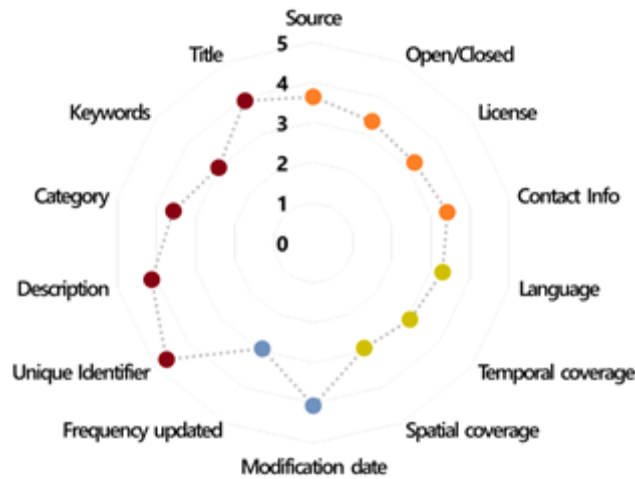


Figure 38: Middle Part Final Reference Model

Main goal

The main goal is to provide a tangible* and practical achievable** set of meta data attributes to measure FAIR, so that a standard set of requirements is created for FAIR data repositories in the public and private sector.

*i.e. Easy to interpret, and appealing to use

**i.e. A list which is not too long

Meaning

Every meta data attribute is ranked on scale 0-5. We should interpret this as follows:

Relevance per attribute	
Score per attribute	Interpretation
1-2	This attribute is optional, and evaluated as 'unnecessary'
2-4	This attribute is recommended, and evaluated as 'useful' attributes.
4-5	This attribute is required, and evaluated as 'indispensable'.

Based on the ranking we know which attribute has the highest priority to optimize. But how to improve this? Therefore, we provide a table with advised syntax and semantics per attribute. This table is set up in cooperation with Berenschot Intellerts.

Syntax versus Semantics			
Attribute	Relevance (1-5)	Syntax (grammar)	Semantics (meaning)
Unique Identifier	4,67	1. Consistent whole repository	1. Preferably letters/numbers corresponding with 'Category'
Title	3,93	2. English 3. No acronyms	1. Includes sufficient data to facilitate search and discovery 2. Human-readable
Description	4,13	1. English 2. Maximum 15	1. Sufficient detail to enable understanding 2. Human-readable
Category	3,57	1. English 2. Preferable one word/term – maximum 3	1. preferably chosen from a predefined list
Keywords	3,07	1. English 2. Single words	1. Understandable for technical and non-technical users
Modification Date	4,07	1. Consistent whole repository, preferably: 00-00-00.	1. Most recent data
Source	3,64	1. URL	
License	3,21	1. Use abbreviations/ acronyms	
Frequency updated	2,93	1. English 2. Use terms like: Daily, Weekly, Monthly, Yearly.	
Temporal coverage	3,08	1. Period in years, like: '2010-2014'	1. If extra information available add, e.g. 'February 2010- December 2014' (English)
Spatial coverage	2,92	1. English 2. Use of terms like: 'Europe', 'NLD (abbreviations countries), 'worldwide'.	
Open/ Closed Data	3,36	1. Consistent use of 'C' and 'O', or 'Open', 'Closed'.	
Contact Information	3,42	1. Name: Initials + last name 2. E-mail address	
Language	3,29	1. Use abbreviations, like 'NL', 'EN'.	

Example

We provide an example (see Figure 39) to make the reference model more concrete. In this case we set up the repository by our self. We followed the steps in the reference model (see Case Study), therefore it contains all relevant meta data attributes (see checklist fourth column). But the key question is: Is the below data repository FAIR? Therefore, we determine per dataset whether the meta data attributes are available. This gives us an indication how well the data repository scores among the different attributes and on the different concepts. Thereby, the table above about syntax and semantics can be used to optimize FAIR in general per attribute, and from the middle part can be derived which factors are most relevant to give priority to. In this case License, Contact Information, and Modification Date score the worst, whereas the ranking shows Modification Date (4,07) is most relevant to upgrade, then Contact Information (3,42), and then License (3,21).

A	B	C	D	E	F	G	H	I	J	K	L	M	N
Identif	Title	Description	Category	Keywords	Modification Date	Open/Closed	Source	License	Frequency updated	Temporal coverage	Spatial coverage	Contact Information	Language
Fin_S1	Conjunctuurlokl	De stand en het	Finance	Nederlandse conjunctuur	14-2-2017	Open	https://www.cbs.nl/nl/nl/vs	unknown	maandelijks	2000 januari - 2017 februari			NL
Fin_S2	Conjunctuairdashboard	Conjunctuurlokl	Finance	Conjunctuurlokl, Conjunction	12-8-2016	Open	https://www.cbs.nl/nl/nl/vs	unknown	jaarlijks	2001-2007			NL
Fin_S3	Conjunctuurbeeld per bedrijfstak	Actueel samen	Finance	Conjunctuur	17-3-2017	Open	http://statline.cbs.nl/Statw/	unknown	per kwartaal	2008 kwartaal 1- 2017 kwartaal 1			NL
Fin_S4	Ziekteverzuimpercentage per bedrijfstak	Ziekteverzuim va	Finance	Ziekteverzuim	25-9-2015	Open	https://www.cbs.nl/nl/nl/vs	unknown	per kwartaal	1996 kwartaal 1- 2016 kwartaal 1			NL
Fin_S5	ANBIs (belastingdienst)	Overzicht van Al	Finance	ANBI	27-8-2015	Open	https://www.belastingdienst.nl/	CC-0	Na aanpassing	2008-2017			NL
Fin_S6	Financiën ministeries		Finance			Open	https://data.oversheid.nl/data		jaarlijks				NL
Fin_S7	Uitgesproken faillissementen	Aantallen faillisse	Finance	Faillissementen	11-5-2017	Open	http://statline.cbs.nl/Statw/	unknown	maandelijks	2009 januari - 2017 april			NL
Fin_S8	Vestigingen van bedrijven	Aantal vestiging	Finance	Vestiging bedrijven	28-4-2017	Open	http://statline.cbs.nl/Statw/	unknown	jaarlijks	2010-2017			NL
Fin_S9	Ondernemingsklimaat	Internationaal ov	Finance	Ondernemingsklimaat, inn	23-2-2017	Open	http://statline.cbs.nl/Statw/	unknown	jaarlijks	1990-2008			NL
Fin_S10	Companyinfo	Jaarrekeningen	Finance	Jaarrekeningen		Closed							x
Fin_S11	Mijn gemeente	Finance_Gov	Finance			Open	www.waartsaaliegemeente.nl/						x
Fin_S12	Informatie voor Denden	Financiële gege	Finance	Begrotingen, jaarrekening	10-5-2017	Open	https://www.cbs.nl/nl/nl/vs	unknown	per kwartaal	2010 kwartaal 1- 2017 kwartaal 1			NL
Fin_S13	Rechterlijke uitspraken	Een deel van de	Finance	Rechterlijke uitspraken		Open	https://uitspraken.rechtspraak.nl/						NL
Fin_S14	OECD	Bank profitability	Finance	OECD, Bank, Profitability	unknown	Open	http://stats.oecd.org/index	unknown	jaarlijks	1979-2009			EN
Fin_S15	AssetMacro	Economic Indica	Finance			Open	https://www.assetmacro.com						EN
Fin_S16	Business Confidence Index (BCI)	Business confid	Finance	BCI, Business, confidence	unknown	Open	https://data.oecd.org/leac/	unknown	maandelijks	1974-2017			EN
Fin_S17	Consumer Confidence Index (CIC)	Consumer confid	Finance	CIC, Consumer, confidenc	unknown	Open	https://data.oecd.org/leac/	unknown	maandelijks	1973-2017			EN
Fin_S18	CGI, Europe	Consumer confid	Finance	CGI, Consumer, confidenc	unknown	Open	https://data.oecd.org/leac/	unknown	maandelijks	1973-2017			EN
Fin_S19	Gross Domestic Product (GDP)	Gross Domestic	Finance	GDP, BPP, National accou	unknown	Open	https://data.oecd.org/leac/	unknown	jaarlijks	1963-2016			EN
Fin_S20	GDP, Europe average	Gross Domestic	Finance	GDP, BPP, Europe, OECD	unknown	Open	https://data.oecd.org/leac/	unknown	jaarlijks	1995-2016			EN
Fin_S21	UwV	Finance	Finance			Closed							x
Fin_S22	Branchemonitor	Benchmark bran	Finance	Bedrijven, Branches		Open	https://www.cbs.nl/nl/nl/vs						NL
Fin_S23	Benchmarks	Rapporten Alger	Finance			Open	https://data.oversheid.nl/data						NL
Fin_S26	Caos	Nieuws site met all	Finance	cao	unknown	Open	https://www.loonwijzer.nl/	unknown	After modification				NL

Figure 39: Example Data Repository (Based on the Case Study Project)

Concept	Activity	Attributes	Relevance (1-5)	Attribute available?	Nr datasets/ Total
Findability	Determine a unique identifier per dataset	Unique Identifier*	4,67	✓	24/24
	Collect the provenance of data	Description	4,13	✓	21/24
		Category	3,57	✓	24/24
		Keywords	3,07	✓	19/24
		Title	3,93	✓	24/24
Accessibility	Ensure the data is available	Source	3,64	✓	22/24
	Ensure the data is secure and legal to use	Open/Closed	3,36	✓	24/24
		Ensure that interactivity between data is possible	License	3,21	✓
	Contact Info	3,42	✓	0/24	
Interoperability	Conform the different data sources	Language	3,29	✓	22/24
		Connect the different data sources	Temporal Coverage	3,08	✓
	Spatial Coverage	2,92	✓	22/24	
Reusability	Determine the plausibility of reuse	Modification Date	4,07	✓	9/24
	Ensure that editing the data is possible	Frequency Updated	2,93	✓	17/24

*Fin_S1 (Finance Source 1), Fin_S2, Fin_S3, etc.

8. Results Survey and Interviews

We performed three types of evaluation: iterative expert evaluation, survey, and interviews. The iterative expert evaluation is carried out within Berenschot Intellerts. Almost every two weeks we discussed the progress and results.

With the survey, we evaluated the definitions and validated the input for the model. The survey was among experts from business and academic side, thereby also some exclusive FAIR experts.

Third type of evaluation are the interviews. These are carried out to evaluate the definitions (also how these concepts are interpreted in the organizations) and to evaluate the model (design and content). These interviews were among data providers.

We will not discuss the results of the iterative evaluation separately; the whole document is the result. Therefore, we first discuss the survey results and finally the interview results.

8.1 Survey

We choose to do a survey because:

- With a survey, it is possible to reach a relatively large group of people. Thereby, it is low-threshold for FAIR experts (who may be difficult to contact).
- The answers cannot be influenced by the interviewer, and the questions are standardized. This is especially interesting regarding the ranking of meta data attributes and the choice for the best definition per concept.

The total number of respondents is 21, and we provide the expert list below:

No	Function	Years of Experience	Organization
1	Software Developer	25	ING
2	Chapter Lead	2 (in management)	ING
3	Managing Consultant	25	Berenschot Intellerts
4	Data Scientist	10 (software innovation & development)	Berenschot Intellerts
5	Development Engineer	2	ING
6	International manager FAIR Data		
7	Managing Partner		Ockham Group
8	Senior Scientist	15	
9	Account manager	17	AFAS
10	Senior BI Consultant	7	
11	Product manager	4	AFAS
12	Director	16	Sogeti
13	CTO FAIR Data	11 (in semantic interoperability)	
14	Program manager	11	Vanderlande
15	Manager		ELIXIR NL
16	Deep Learning Engineer	1	CentERdata
17	Director		Berenschot Intellerts
18	Data Scientist	18 (in BI)	Berenschot Intellerts
19	Software Engineer	2	Dutch Techcentre for Life Sciences
20	Software Engineer Associate		Infor
21	Professor Enterprise Engineering	25	The Standish Group

8.2 Main Findings Survey

We discussed already in chapter 3 (definitions) and chapter 7 (input model) the main results of the survey. Thereby, we add all the questions and answers in Appendix C.

A few extra interesting findings are:

- 52% of the respondents knows what FAIR means on beforehand, and 48% does not. However, people still can be expert on the meta data management side without knowing the specific FAIR term (survey question 1)
- Most of the people would describe the relation between meta data management and data quality as a correlation; both topics influence each other. Which agrees with our assumption that with good meta data management also the quality of the content of data can be influenced in a positive way (survey question 6)
- The respondents do make use of open data, however not really that much (together they named 16 different sources. We think the taxonomy of sources would be useful when it is publicly accessible, so that the awareness regarding open data grows and people can make use of the knowledge of others (survey question 7).

8.3 Interviews

We choose to do interviews because:

- To learn more about the practical interpretation of the FAIR concepts within organizations; the view of the data providers (where the data comes from, and by whom it is stored).
- With the survey, we got a ranking (from 0-5) regarding the meta data attributes, with the interviews we thereby got an explanation.
- To evaluate the design and content of the model. Especially the design is quite subjective, so reactions are can be determined best by an interview.

We provide the list of interviewees in the table below. **Interviewee 1** is a relevant interviewee because he is the initiator of the Open Data Program at the Gemeente Amsterdam, thereby he is intimately involved with the DataLab in Amsterdam. The knowledge of **Interviewee 2** is relevant because he is the initiator of open data projects at CBS, and **Interviewee 3** because he is the current manager of Open Data at CBS.

No	Interviewee	Function	Organization
1	Interviewee 1	CTO innovatiemanagement	Gemeente Amsterdam
2	Interviewee 2	Project Manager	CBS
3	Interviewee 3	Manager Open Data	CBS

8.4 Main Findings Interviews

We summarize the main findings of the interviews by a collection of quotes and conclusions. We distinguish three categories: main findings regarding definitions, main findings regarding the models, and thirdly the main findings regarding the meta data attributes.

Main findings Definitions

1. Findability

- “Make a distinction between ‘searchability’ and findability on meta data level” (**Interviewee 1**).
- “Regarding findability it is advisable to set up editorial rules (guidelines in writing descriptions)” (**Interviewee 2**)
- Provenance is an important part of findability.
- “Categories/themes are important regarding findability” (**Interviewee 3, Interviewee 1**).
- Findability regarding numbers becomes more actual

2. Accessibility

- “Make distinction between content accessibility versus technical accessibility” (**Interviewee 1**)
- Regarding technical accessibility, security is most important (tested by hackers) → suggestion for definition: ‘regardless the technologies they use’ (**Interviewee 1**).

3. Interoperability

- Interoperability can be mainly focused for internal usage (internal region code at CBS, which is not available for external parties) or the focus is more on interoperable with repositories from other organizations (depends on size own organization)
- Interoperability is most time-consuming part when data sources are linked

4. Reusability

- Within organizations reusability is not seen as ‘a consequence’ of the other concepts (as we concluded from literature). It is, however, a significant part of the policy, and much attention is paid to it (**Interviewee 2**, **Interviewee 1**)
- “Bigger organizations especially focus on internal reusability, so reusability for the data provider itself, which leads to efficiency in the work/production processes.” (**Interviewee 3**)

Main Findings Model 1

- Everybody was enthusiastic about model 1; the FAIR activities model. Main comments were that the design was clear and attractive. **Interviewee 2** says he was “positively surprised” about the model, and “happy that someone from university comes to CBS” since he knew someone who graduated after four years on the topic open data “but never visited the biggest open data provider in our country”. Additionally, **Interviewee 3** says he “really likes the appealing design” and “thinks it would be useful to use instead of the 5-star model” they are using right now. This ‘5-star model’ of Tim Berners-Lee is provided in Figure 40. **Interviewee 2** also confirms this.
- **Interviewee 2** expresses the benefit of our reference model over the ‘5-star model’ as follows: “it looks further than the fifth phase of the 5-star model [Link your data to other data to provide context], and provides us more details and guidelines in how to manage open data repositories”.

We conclude that the current model has enough potency, and therefore we do not change it after evaluation.

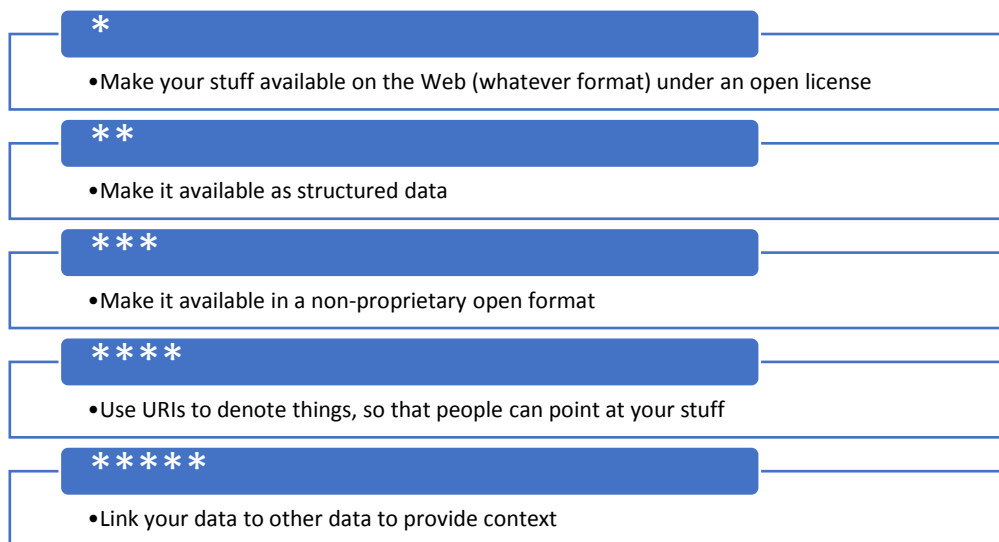


Figure 40: 5 Star Model Open Data, Tim Berners-Lee

Main Findings Model 2

- **Interviewee 1** says he “especially likes the set of meta data attributes”. He also has experience with long lists of attributes, and then “everybody fills in what he likes”, this is a “doable set of elements”.
- **Interviewee 3** again states that “the model could be really useful, because it is more detailed than guidelines which are available in this area until now, because the open data discipline is quite new”. He also thinks other companies would have benefit from such kind of model. When he looks around to colleagues he concludes that “especially in big companies they are all still a bit searching for guidelines to set up, determine and demarcate their open/big data policy”.
- More critical is **Interviewee 2** about this model. He states that “this set of attributes is insufficient to satisfy the FAIR concepts”, and thereby he does not understand the percentages in the middle part. This was also confirmed by **Interviewee 1** and **3**.
- Another point of critic is that this model “still does not tell me anything about the quality of the data itself”, **Interviewee 2** expresses it as follows: “when an organization sets up all kind of repositories with these attributes, it does not tell anything about the actual impact on the organization”.

We conclude that the middle part of the model must be changed. Therefore, we removed the relative percentages pie chart, and add a scale 0-5 to express the relevance per attribute. Second, we conclude that the set of meta data attributes does not completely includes the FAIR concepts. We will elaborate more on this in the discussion chapter.

Main Findings Meta data attributes

Most important findings regarding the rating of the meta data attributes are:

- The Unique Identifier is by all people (survey and interviews) marked as most important.
- Mentioned twice is that Category and Keywords could be comprised by a Description
- Mentioned once is that Language could be unnecessary, because it's something you see at glance
- Although license is mentioned as important by all parties, no one has a clear overview on which licenses are available. 'CC-BY' is mentioned as frequent occurring, but especially for open data the wish is 'CC-0' (public domain). When it is about a repository with data from different sources, the policy is often that the responsibility regarding the license is placed to the original data owner.
- The main point of critical feedback was that the attributes used are not complete to comprise FAIR. However, which elements are missing concrete was also hard to say. Although two aspects were mentioned as addition to the current attributes:
 - Notification of available services from the data owner/provider (like helpdesk at CBS)
 - Tuning of the repository to the target group/users (e.g. students, app developers, business)
- Finally, we received from the municipality of Amsterdam a document (created in cooperation with municipalities Amsterdam, Eindhoven, Den Haag and Utrecht) with a set of useful meta data attributes. These attributes are categorized in: 'must have', 'should have', and 'could have' and are provided in the table below.
- Although this set of attributes differs not that much of our set of attributes, and therefore can be seen as a confirmation, two things are notable:
 - The Unique Identifier is missing
 - They mention a lot of attributes regarding Source

Must	Should	Could
Naam/Titel	Meer informatie	Verstrekker (Organisatie)
Beschrijving inhoud	Geografische eenheid (gemeente, wijk, buurt etc.)	Uitgiftedatum
Eigenaar (Organisatie)	Trefwoorden	Taal
Contactpersoon (Contact)	Rechten	Versie
Dekking in tijd (van/tot)	Toelichting versie	Grondslag
Geografisch gebied	Source status (actueel, historisch, concept etc.)	Source grootte bestand (bijv. 10.1MB)
Wijzigingsfrequentie	Source wijzigingsdatum	Source uitgiftedatum
Wijzigingsdatum		
Thema		
Licentie		
Toegepaste standaard(en)		
Source titel		
Source Omschrijving		
Source Formaat		
Link/URL/Bestand		

8.5 Evaluation Reference model

In Chapter 7 we already provide an example of how the model must be applied on a data repository. However, this model was set up by our own, and what do the results mean in comparison to other data repositories? Therefore, we asked during the interviews for example data repositories, where we applied our model on as well. We present the results in this paragraph. The first one is the repository of the 'Energy Project' within Berenschot Intellerts we mentioned in section 5.3. The second one is the repository of the municipality of Amsterdam.

Energiedrager	Land	Hub	Periode	Site	Opmerkingen
Elektriciteit	NL	APX	DA	https://www.apxgroup.com/market-results/apx-power-nl/dashboard/	
Elektriciteit		APX	Intra-day	?	
Elektriciteit	DU	EPEX	DA	https://www.epexspot.com/en/market-data/dayaheadauction/auction-table/2016-12-21/DE	
Elektriciteit		EPEX	Intra-day	https://www.epexspot.com/en/market-data/intradaycontinuous/intraday-table/-/DE	
Elektriciteit	BE	Belpex	DA	http://www.belpex.be/market-results/the-market-today/dashboard/	
Elektriciteit		Belpex	Intra-day	?	
Elektriciteit	FR	EPEX	DA	https://www.epexspot.com/en/market-data/dayaheadauction/auction-table/2016-12-21/FR	
Elektriciteit		EPEX	Intra-day	https://www.epexspot.com/en/market-data/intradaycontinuous/intraday-table/-/FR	
Elektriciteit	ENG	APX	DA	https://www.apxgroup.com/market-results/apx-power-uk/dashboard/	LET OP: pond ipv euro
Elektriciteit		APX	Intra-day	?	LET OP: pond ipv euro
Gas		TTF	NATURAL GAS WITHIN-DAY INDEX	https://www.theice.com/marketdata/reports/168	
Gas		TTF	NATURAL GAS DAY-AHEAD INDEX	https://www.theice.com/marketdata/reports/168	
Gas		TTF	Futures	https://www.theice.com/products/27996665/Dutch-TTF-Gas-Futures/data	
Gas		NBP	Futures	https://www.theice.com/products/910/UK-Natural-Gas-Futures/data	
Olie		Brent	Futures	https://www.theice.com/products/219/Brent-Crude-Futures/data	
Olie		WTI	Futures	https://www.theice.com/products/213/WTI-Crude-Futures/data	
Kolen		ARA	Futures	https://www.theice.com/products/243/Rotterdam-Coal-Futures/data	
CO2		EUA	Futures	https://www.theice.com/marketdata/reports/82	

Figure 41: Repository Energy Project Berenschot Intellerts

The Energy Repository is a slightly disappointing repository in terms of FAIR. As the results below show, the repository misses crucial attributes. It contains only: spatial coverage, temporal coverage, source, theme, and keywords. Therefore, the repository is not qualified as a FAIR repository (at least it must contain all basic attributes).

Concept	Activity	Attributes	Relevance (1-5)	Attribute available?	Nr datasets/ Total
Findability	Determine a unique identifier per dataset	Unique Identifier	4,67	✗	0/18
	Collect the provenance of data	Description	4,13	✗	0/18
		Category	3,57	✓	18/18
		Keywords	3,07	✓	18/18
		Title	3,93	✗	0/18
Accessibility	Ensure the data is available	Source	3,64	✓	15/18
	Ensure the data is secure and legal to use	Open/Closed	3,36	✗	0/18
	Ensure that interactivity between data is possible	License	3,21	✗	0/18
		Contact Info	3,42	✗	0/18
Interoperability	Conform the different data sources	Language	3,29	✗	0/18
	Connect the different data sources	Temporal Coverage	3,08	✓	18/18
		Spatial Coverage	2,92	✓	5/18
Reusability	Determine the plausibility of reuse	Modification Date	4,07	✗	0/18
	Ensure that editing the data is possible	Frequency Updated	2,93	✗	0/18



Figure 42: Data Repository Gemeente Amsterdam

The Repository of Gemeente Amsterdam (<https://data.amsterdam.nl/>) can be qualified already some higher when we look from FAIR perspective. Although “this repository is an example of kind of ‘dumping place’ for open data without taking too much care on data quality (quality of meta data, but also quality of content)” (Interview Interviewee 1). Notable is that the most relevant attribute, Unique ID, is missing. Thereby, Language, Temporal Coverage, Spatial Coverage, and Frequency Updated are not available. In this case we did not count the number of datasets, since it is a bit too much work for this kind of illustration. As said, most relevant is to add a unique id, then to add Language, then Temporal Coverage, and finally Frequency Updated and Spatial Coverage.

Concept	Activity	Attributes	Relevance (1-5)	Attribute available?	Nr datasets/ Total
Findability	Determine a unique identifier per dataset	Unique Identifier	4,67	✗	
	Collect the provenance of data	Description	4,13	✓	
		Category	3,57	✓	
		Keywords	3,07	✓	
		Title	3,93	✓	
Accessibility	Ensure the data is available	Source	3,64	✓	
	Ensure the data is secure and legal to use	Open/Closed	3,36	✓	
	Ensure that interactivity between data is possible	License	3,21	✓	
		Contact Info	3,42	✓	
Interoperability	Conform the different data sources	Language	3,29	✗	
	Connect the different data sources	Temporal Coverage	3,08	✗	
		Spatial Coverage	2,92	✗	
Reusability	Determine the plausibility of reuse	Modification Date	4,07	✓	
	Ensure that editing the data is possible	Frequency Updated	2,93	✗	

8.6 Main improvements after evaluation

Summarized the main improvements after evaluation are:

- Change of design of the final reference model (see Figure 35 and 34).
 - BECAUSE of misinterpretation of the original model.
- The set of meta data attributes remains the same.
 - BECAUSE all attributes were ranked around 3 and higher on relevance in the survey.
 - BECAUSE there were no wider supported, concrete suggestions to add attributes (a few were only mentioned once).
- The definition on Findability changed.
 - BECAUSE the survey results were disappointing (see section 3.2.4 for detailed explanation).
- The Data Scouting Process changed.
 - BECAUSE of feedback within Berenschot Intellerts from data scientists (see section 4.3).

9. Overall Conclusion

In this research, we formulated three sub-questions. In section 9.1 we elaborate on the main findings and conclusions per research question. In the next section, we explain how this altogether answers the main research question.

9.1 Answering sub-questions

1. What are the definitions and interpretations of the FAIR concepts in literature?

First, we conclude the FAIR concepts are still quite new in literature. This is not strange when we realize that the paper of Wilkinson et al. is from 2016. However, the concepts appear in literature separately. Based on a systematic literature review, we defined each concept and argued why this definition is an improvement or addition to the definition of Wilkinson et al. (2016).

Regarding the first concept, findability, we conclude that three aspects are important: versioning, provenance, and a unique identifier. Therefore, findability has all to do with uniqueness. These findings are in line with the definition of Wilkinson et al. (2016), however the new definition contains (in contrast to the definition of Wilkinson et al., 2016) a description of the concept itself, how this can be realized, and the goal of findability.

Regarding accessibility the main conclusion is that it is important to realize whether we approach it from semantic perspective versus syntactic perspective. We conclude the semantic side is more dominant in literature; accessibility is seen and described most as semantic quality dimension. Thereby, it is important that the focus is on the ease of user for the (end) user, so that data can be easily perceived, interpreted, and applied. From this perspective, the main shortcoming of the definition of Wilkinson et al. (2016) is that the definition is too 'small'; it is too detailed and specific.

Thirdly, we conclude that interoperability in the literature is about the interaction between heterogeneous environments, and again the (end) user is the main stakeholder. Thereby, we note that cooperation between systems should require minimal effort from their users to be qualified as interoperable. Thus, interoperability is needed to connect sources with minimal or no special effort from the user to realize data can be shared. The concepts regarding Interoperability found in literature are quite in accordance to the guidelines of Wilkinson, especially that data must be interoperable with minimal effort (for (end) users).

Finally, for reusability we conclude based on literature that reusable data is 'realized' by all other quality indicators (FAI), so that the data can be used in a new context by other people. This contrasts with the definition of Wilkinson et al. (2016). Although, Doorn & Dillo (2016) confirm this, we found a little support for this statement in the evaluation (interviews and survey). Therefore, we conclude this new definition is not indisputable.

The findings above and the evaluation via survey and interviews result in following new definitions:

- Findability (of data) is about the uniqueness and provenance of data, which means data described with rich meta data with focus on contextual information. This is achieved by versioning, associating URIs or embedding other kind of identifiers
- Accessibility is a semantic quality dimension, which comprises availability, security, performance, interactivity, flexibility of data, to ensure data access to (end) users regardless their different context and background.
- Interoperability is a system feature to connect and conform heterogeneous environments, to ensure sharing, reusing and exchanging of data between these environments without special effort from the (end) user.
- Reusability is the ability to make easily use of data in a new context, in terms of retrieving, downloading, indexing, searching and visualizing the data without restrictions, as a result of findable, interoperable and accessible data.

2. What is the benefit of FAIR in the context of open data projects?

The main result of this research question is the roadmap of the Data Scouting Process. But before we designed this artifact we described the open data landscape in the Netherlands, and we put FAIR in the context of existing models: CRISP-DM and KDD Process.

We conclude the Netherlands performs good regarding open data comparing to the rest of the world. The UK does even better, and is for years the number 1 according to the Open Data Barometer. Especially on the areas 'social impact', 'economic impact', and 'Data Accountability' we can learn from them.

Based on the input provided by Berenschot Intellerts we conclude that the 'Data Scouting Process' contains the following steps: 1. Determine domain, 2. First search and identify objectives, 3. Determine KPIs, 4. Targeted Search, 5. Set up Data Repository and evaluate quality, 6. Link and Visualize data.

The fifth step of the Data Scouting Process is where FAIR comes in place. We conclude the FAIR principles are relevant within Berenschot Intellerts because: there is a need for better internal searchability within used data sources and material from projects in the past (Findability), there is a need for knowledge about relevant formats (syntactical side of Accessibility), there is a wish to make the interoperability process between different data sources less time consuming (still the most intensive phase) and more efficient (Interoperability), and finally the goal is to use and store sources, processes and other project materials in such a way it can be reused for new projects in order to save time and money, and to share knowledge (Reusability). We conclude, based on the interviews, these needs are within business in general, and in the public sector the same (although the focus will be more on knowledge sharing then).

Based on our reference model (as results from RQ3) we created during the Case Study Project a FAIR repository for Berenschot Intellerts, and evaluated what eventually points of improvement are. At the same time, this repository is a template to ensure FAIR meta data management in a data repository. And finally, we provided during the Case Study Project a 'open data taxonomy' with mainly used open data sources and portals during open data projects at Intellerts until now.

3. What are the meta data requirements that satisfy the FAIR principles?

In the third part of this research we made the translation from 'theory' into 'practice' by determine relevant meta data requirements for a FAIR repository, and by based on that, in combination with the definitions (RQ1), create the final reference model.

First, we described the relevance and types of meta data in general. We conclude that meta data management is from added value for all layers within an organization/company, whereas the distinction between business and technical meta data is important. Thereby we conclude that meta data management is especially relevant in terms of knowledge management; in this way, the knowledge in the 'heads of people' can be stored and therefore it is easier to reuse the knowledge, and specific employees are no longer indispensable.

Second, we conclude that the existing DSA requirements have a lot of overlap with the FAIR concepts. This is also confirmed by Doorn & Dillo (2016), and they are working on a program to automate the process of measuring FAIRness in a dataset. We conclude that they also confirm that reusability must be seen as 'a resultant of the other three concepts', therefore they approach, in accordance to us, the following metric: $(F+A+I)/3 = R$. Since the initiative of Doorn & Dillo (2016) is still in process we decided to not write a method for measuring FAIR, but to set up a reference model, so that this research is a relevant complement on their initiative.

Finally, our search for meta data elements, in accordance to the FAIR concepts, led to the following set of basic elements: Title, Description, Category, Keywords, Modification Date, Contact Information, License, Frequency updated, Temporal Coverage, and Spatial Coverage. These attributes are ranked by experts based on a survey, and led together with the definitions of FAIR, and the activities to our final reference model.

9.2 Answering main research question

WHICH REQUIREMENTS SHOULD A DATA REPOSITORY MEET TO SATISFY THE FAIR PRINCIPLES?

In this research, we provided two main artifacts to answer this main research question: the Data Scouting Process roadmap, and a Reference Model which can be applied in step 5 of the Data Scouting Process. The main goal of the Data Scouting Process is to determine the place of FAIR in the context of open data projects. And the main goal of the Reference Model is to provide a step-by-step set of activities to ensure a data repository contains all aspects to measure FAIR, and which is applicable in the public and private sector. Thereby, an important starting point was that the set of meta data attributes, provided in the reference model, must be 'doable'. Which means that the list of attributes must not be too long since otherwise no one will fill in all the information per attribute.

To design the two artifacts, we split the research into three main parts, corresponding to the three main research questions. Regarding research question 1 we looked from a semantic perspective on the FAIR principles. Based on this we derived four new definitions. In the second research question, we looked at the business context of FAIR; we described the open data landscape, and set up the data scouting process (artifact 1). Finally, in research question 3 we looked more from a syntactic perspective; how to translate the theory into practice by linking meta data attributes to the FAIR requirements? The output of research question 1 and 3 together is the input for the reference model (artifact 2).

The first artifact, the Data Scouting Process roadmap, is evaluated by iterative expert evaluation from Berenschot Intellerts. This means they had a big impact on the design of the roadmap and the content of the steps. The first time we applied this roadmap to an open data project in their organization was during the Case Study Project. We conclude based on this experience that the roadmap provides clear guidance during a project, and makes it easier to understand where FAIR can benefit the process, namely during the setup of a data repository.

The main input for the second artifact, the reference model, were the FAIR definitions (RQ1), the relations between the concepts (RQ1) and the set of meta data attributes (RQ3). We derived based on the FAIR definitions from research question 1 the following activities: Determine a unique identifier per dataset, Collect the provenance of data, Ensure the data is available, Ensure the data is secure and legal to use, Ensure that interactivity between data is possible, Conform the different data sources, Connect the different data sources, Determine the plausibility of reuse, and Ensure that editing the data is possible. Second, the relations between the different concepts can be described by the following metric: $(F+A+I)/3 = R$. We conclude reusability is a resultant of the other three concepts: Findability, Accessibility and Interoperability. And thirdly, the set of meta data attributes consists of the following attributes: Title, Description, Category, Keywords, Modification Date, Contact Information, License, Frequency updated, Temporal Coverage, and Spatial Coverage.

The model is evaluated via interviews at Gemeente van Amsterdam and CBS, and by iterative expert evaluation from Berenschot Intellerts. Based on these results we conclude that the model is positively received; all parties want to use the model and see it as relevant addition to their current open data policy. Thereby, via contacts within CBS also Ministerie van Binnenlandse Zaken wants to use the model. We conclude this success is due to our practical, comprehensible and accessible approach in combination with attention for a good design. And finally, we conclude that therefore these artifacts satisfy the problem statement by bridging between theory to practice regarding meta data management.

10. Discussion and Future Work

In this discussion section, we first elaborate on four critical statements, which also will be disproved. Second, we describe some shortcomings regarding the research approach. And finally, we mention the most relevant opportunities for future work.

Statement 1: This research is more from business perspective and too little from academic perspective, although the aim of the model is that it is useful in both contexts.

Defense Statement 1: First, the definitions are set up from academic perspective, and therefore also indirectly the activities and elements in the model. Second, from academic perspective it was already clear that there is a wish for FAIR data repositories. Finally, in general business is less accessible for this kind of initiatives, so when it works there it will be also fine for academia. However, for future research a specific implementation of the model in academic context would be useful.

Statement 2: Meta data attributes are not suitable to comprise FAIR concepts.

Defense Statement 2: We agree that meta data attributes are not completely adequate. However, in our opinion to make something 'tangible' and practically accessible we should do some compromise. One of the main goals of this research is to 'translate' the FAIR concepts into practice, and from that perspective we choose for meta data attributes.

Statement 3: The model is insufficient regarding determining the data quality on content.

Defense Statement 3: True, it would be a great improvement on the model when the actual content of the meta data would be included. Since this is more on the data governance side we did not include it in this research.

Statement 4: The connection between the meta data attributes and concepts should be more thorough

Defense Statement 4: we partly agree. We evaluated that the set of attributes is not totally complete to comprise the FAIR concepts, however, at the same time no other attributes were specifically mentioned. Therefore, for now we keep the original set. Thereby, maybe even more important is that it must be a 'doable' set of attributes, otherwise no data user will fill in all the information. Therefore, this is a constant dilemma between theoretical quality and practical applicability.

Second, we discuss some shortcomings in general. It would be useful for future research to let people link the attributes to the concepts, and to give them the opportunity for their own suggestions. Maybe this leads to great insights in setting up a list of most relevant meta data attributes to cover FAIR. Thereby, the balance between business versus academic people could be better. And finally, it would be great to implement the model in an organization and to observe over a long time how people use it and what the added value really is. To answer questions like: does it change their perception on the usefulness of meta data management? Do they understand the different steps within the model? Do they experience benefits in searching for data and storing data after using the model?

Finally, summarized some concrete suggestions for future work regarding this research are as follows:

- Implement the model at CBS, Berenschot Intellerts and Ministerie van Binnenlandse Zaken, and interview them after a few months how it influenced their process, and what can be better. Thereby it would be useful to also search for an academic context.
- Second, there is enough space for improvement on the definitions. These are now only based on one original paper (Wilkinson et al., 2016), and this research.
- Thirdly, Carry out an additional, more detailed evaluation on the set of meta data attributes to cover FAIR.
- In the fourth place, make the taxonomy publicly accessible, so that it can be supplemented and occurs more awareness of the benefits of open data.
- Finally, it would be great to look for collaboration with the initiative from Doorn & Dillo (2016). Since this research (the reference models) hopefully can prepare the way by making the FAIR concepts appealing, useful, tangible and understandable for their automated program to measure FAIR.

References

- Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3), 16.
- Beebe, N. L., & Walz, D. (2005). An Empirical Investigation of the Impact of Data Quality and its Antecedents on Data Warehousing. *AMCIS 2005 Proceedings*, 28.
- Bloomberg Philanthropies. Getting meta with metadata. Open data guide. Retrieved at: <https://www.gitbook.com/book/centerforgov/open-data-metadata-guide/details>
- Bughin, J., Chui, M., & Manyika, J. (2010). Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. *McKinsey Quarterly*, 56(1), 75-86.
- Cappiello, C., Francalanci, C., & Pernici, B. (2003). Preserving Web sites: A data quality approach. *International Conference on Information Quality (ICIQ)*, 331-344.
- Charalabidis, Y., Alexopoulos, C., & Loukis, E. (2016). A taxonomy of open government data research areas and topics. *Journal of Organizational Computing and Electronic Commerce*, 26(1-2), 41-63.
- Chatfield, A., Reddick, C., & Al-Zubaidi, W. (2015). Capability challenges in transforming government through open and big data: Tales of two cities.
- Chong, S., Skalka, C., & Vaughan, J. A. (2015). Self-identifying data for fair use. *Journal of Data and Information Quality (JDIQ)*, 5(3), 11.
- Cui, T., Ye, H. J., Teo, H. H., & Li, J. (2015). Information technology and open innovation: A strategic alignment perspective. *Information & Management*, 52(3), 348-358.
- Cupola, P., Earley, S., & Henderson, D. (2014). DAMA-DMBOK2 Framework. Retrieved at: <https://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf>.
- Data Plan; What's Your Data Plan? Retrieved at: http://researchdata.wisc.edu/wp-content/uploads/2010/04/data_plan_guide.pdf
- Debattista, J., Auer, S., & Lange, C. (2016). Luzzu—A Methodology and Framework for Linked Data Quality Assessment. *Journal of Data and Information Quality (JDIQ)*, 8(1), 4.
- Doorn, P. & Dillo, I. (2016). Put FAIR principles into practice and enjoy your data! Retrieved at: <http://aims.fao.org/activity/blog/put-fair-principles-practice-and-enjoy-your-data>.
- Doorn, P. (2017). Do you want to play fair? Retrieved at: https://gallery.mailchimp.com/7e239f27770173adf851141f2/files/3a6f5b10-d1bd-4e16-95b9-8ab83e97587a/Edata_Research_februari_2017_DEFweb.pdf
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), 27-34.
- Fürber, C., & Hepp, M. (2011). Swiqa-a semantic web information quality assessment framework. In *ECIS (Vol. 15, p. 19)*.
- Gavai, Kuzniar, Kaliyaperumal, and Burger (2016). FAIR Data Point Software specification. Retrieved at: <https://dtl-fair.atlassian.net/wiki/display/FDP/FAIR+Data+Point+Software+Specification>

- Geisler, S., Quix, C., Weber, S., & Jarke, M. (2016). Ontology-Based Data Quality Management for Data Streams. *Journal of Data and Information Quality (JDIQ)*, 7(4), 18.
- Harn, J., Lee, J. N., Kim, D., & Choi, B. (2015). Open Innovation Maturity Model for the Government: An Open System Perspective.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 75-105.
- Jayewardene, V., Sadiq, S. W., & Indulska, M. (2012). Practical Significance of Key Data Quality Research Areas. In *PACIS* (p. 179).
- Kallinikos, J., Aaltonen, A., & Marton, A. (2013). The Ambivalent Ontology of Digital Artifacts. *Mis Quarterly*, 37(2), 357-370.
- Lassinantti, J., & Bergvall-Kåreborn, B. (2014). Open data in Europe: mapping user groups to future innovation impacts. In *Information Systems Research Seminar in Scandinavia: IRIS 37: Designing Human Technologies 10/08/2014-13/08/2014* (pp. 160-176). AIS Electronic Library (AISeL).
- Levin (2015). Open Access, Open Data, Open Science... What does "openness" mean in the first place? Retrieved at: <http://somatosphere.net/2015/02/open-science.html>.
- Malaverri, J. E. G., Mota, M. S., & Medeiros, C. B. (2013). Estimating the quality of data using provenance: a case study in eScience. In *19th Americas Conference on Information Systems, AMCIS 2013-Hyperconnected World: Anything, Anywhere, Anytime*.
- Marco, D. (2000). Building and managing the meta data repository. A full lifecycle guide.
- Matiaško, K., Záborská, K., & Záborský, M. (2004). Building the Unified Data Access Framework. *Journal of Information, Management and Control Systems*, 2(2).
- McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., ... & Spies, J. R. (2016). How open science helps researchers succeed. *Elife*, 5, e16800.
- Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., ... & Plale, B. (2011). The open provenance model core specification (v1. 1). *Future generation computer systems*, 27(6), 743-756.
- Mosley, M. (2008). DAMA-DMBOK Functional Framework, v 3.02. Dama International. Retrieved at: https://www.dama.org/sites/default/files/download/DAMA-DMBOK_Functional_Framework_v3_02_20080910.pdf
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science*, 348(6242), 1422-1425.
- Olbrich, S. (2010). Warehousing and Analyzing Streaming Data Quality Information. In *AMCIS* (p. 159).
- Open standards and Re-use Government Action Plan of the UK government (2010). Retrieved at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/61962/open_source.pdf
- Phillips-Wren, G., Iyer, L. S., Kulkarni, U., & Ariyachandra, T. (2015). Business analytics in the context of big data: a roadmap for research. *Communications of the Association for Information Systems (CAIS)*, 37(1), 448-472.
- Rao, R. (2003). From unstructured data to actionable intelligence. *IT professional*, 5(6), 29-35.

- Riley, J. (2017). NISO Primer Understanding Metadata. What is metadata and what is it for? Retrieved at: http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf
- Scannapieco, M., & Catarci, T. (2002). Data quality under a computer science perspective. *Archivi & Computer*, 2, 1-15.
- Sherman, R. (2014). *Business Intelligence Guidebook: From Data Integration to Analytics*. Newnes.
- Sheridan, J., & Tennison, J. (2010). Linking UK Government Data. In Ldow.
- Shakir, G., Loutas, N., Peristeras, V., & Sklarß, S. (2013). Towards semantically interoperable metadata repositories: The asset description metadata schema. *Computers in Industry*, 64(1), 10-18.
- Tilly, R., Posegga, O., Fischbach, K., & Schoder, D. (2015). What is Quality of Data and Information in Social Information Systems? Towards a Definition and Ontology.
- Vaishnavi, V. & Kuechler, W. (2004). Design Science Research in Information Systems. Retrieved at: <http://www.desrist.org/design-research-in-information-systems/>.
- Vries, M. (2012). Charging for PSI re-use: a snap shot of the state of affairs in Europe. Retrieved at: <http://epsiplat-form.eu/sites/default/files/Topic%20Report%20pricing%20final%201.3.pdf>.
- 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.
- Wang, T., Truptil, S., & Benaben, F. (2015). A general model transformation methodology to serve enterprise interoperability data sharing problem. In *International IFIP Working Conference on Enterprise Interoperability* (pp. 16-29). Springer Berlin Heidelberg.
- Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, 13-23.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 160018.
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39).
- Yu, C. C. (2016). A value-centric business model framework for managing open data applications. *Journal of Organizational Computing and Electronic Commerce*, 26(1-2), 80-115.
- Zhang, R., Jayawardene, V., Indulska, M., Sadiq, S., & Zhou, X. (2014). A Data Driven Approach for Discovering Data Quality Requirements.

Appendix A: Literature Research Set up

Overview: which paper contains which concepts (alphabetical order)

<i>The concept is literally defined</i>	X			
<i>The concept is described indirectly</i>	X			
Authors/ Concept definitions	Findability	Accessibility	Interoperability	Reusability
Batini, Cappiello, Francalanci, Maurino (2009)	X	X	X	
Charalabidis, Alexopoulos, Loukis (2016)		X	X	
Chong, Skalka and Vaughan (2015)	X			
Debattista, Auer, Lange (2016)		X	X	X
Fürber and Hepp, (2011)	X	X		X
Geisler, Quix, Weber, Jarke (2016)		X		
Harn, Kim, Lee, Choi (2015)		X		
Kallinikos, Aaltonen, Marton (2013)	X	X	X	
Beebe & Walz (2005)		X		
Malaverri, Mota, and Medeiros (2013)	X		X	X
Matiasko, Zabovska, Zabovsky (2004)		X	X	X
Moreau, Freire, Futrelle, Groth,..., Plale (2011)	X			
Olbrich (2010)		X		
Open standards and Re-use Government Action Plan of the UK government (2010)				X
Philips-Wren, Iyer, Kulkarni & Arivachandra (2015)		X		
Sheridan and Tennison (2010)	X			X
Sherman (2014)	X	X	X	X
Shukair et al (2013)			X	
Tilly, Posegga, Fischbach, Schoder (2015)		X		
Vries(2012) cited in Lassinantti (2014)				X
Wang and Strong (1996)		X		
Wang, Truptil & Benaben (2015)			X	X
Wilkinson et al. (2016)	X	X	X	X
Yu (2016)		X	X	X
Zhang, Indulska, Jayawardene, Sadiq, Zhou		X		

Overview: citations per paper (alphabetical order)

<i>The concept is literally defined</i>				
<i>The concept is described indirectly</i>				
Authors/ Concept definitions	Findability	Accessibility	Interoperability	Reusability
Batini, Cappiello, Francalanci, Maurino (2009)	To support the preservation process of the entire life cycle of information, <u>associate a new URL</u> when old data is still valid. (p.33)	Accessibility as quality dimension: "Accessibility measures <u>the ability of users to access data, given their culture, physical status and available technologies</u> , and is important in cooperative and network-based information systems." (p. 27)	Distributed information system: "Data can be stored in different databases, but interoperability is guaranteed by <u>the logical integration of their schemas.</u> " (p.11) CIS: "In CISs, data are not logically integrated, since they are stored in separate databases according to different schemas. (...) <u>Integration is realized at a process level.</u> " (p.11)	
Charalabidis, Alexopoulos, Loukis (2016)		Open Government Data Infrastructures "includes research topics concerning various important technological aspects of the ICT infrastructures developed by government agencies to make OGD accessible to different groups of actors, <u>such as their architectures, APIs provision, and personalization capabilities</u> " (p.40)	"Interoperability <u>is a highly important feature of all types of information systems</u> , and this gave rise to the development of a well-established research domain, which attracts considerable research interest, motivated by the increasing need of data exchange among organizations." (p.51)	
Chong, Skalka and Vaughan (2015)	They provide "a scheme for embedding <u>a provenance identifier</u> in environmental datasets, that <u>associates metadata with datasets</u> in a tightly coupled manner that does not rely on external structure such as XML formats or database schema. We say that such datasets are ' <u>self-identifying</u> '." (p.2)			
Debattista, Auer, Lange (2016)		"A Category is a group of quality dimensions in which a common type of information is used as <u>quality indicator</u> (e.g., Accessibility, <u>which comprises not only availability but also dimensions such as security or performance</u>)." (p.14)	"Regarding interoperability, Luzzu [the framework] is <u>accompanied by a set of ontologies for capturing quality-related information for re-use, including quality measures, issues, and</u>	<u>Reusability as consequence:</u> "Having high quality datasets and even more importantly being aware of the quality indicators ensures reusability and thus helps to decrease

		<p>"Dimension is a characteristic of a dataset relevant to the consumer (e.g., Availability of a dataset). (...) Metric is a concrete quality measure for a concrete quality indicator usually associated with a measuring procedure." (p.14)</p>	<p><u>reports</u> that can be re-used in other semantic frameworks and tools." (p. 2)</p>	<p>the number of duplicate and redundant resources on the Web." (p. 30)</p>
Fürber and Hepp, (2011)	<p><u>Uniqueness</u> is "the degree to which data is free of redundancies in breadth, depth, and scope." (p.)</p>	<p><u>Completeness</u> is "the extent to which data are of sufficient breadth, depth, and scope for the task at hand". <u>Timeliness</u> "reflects how up-to-date the data is with respect to the task it's used for." (p.)</p>		<p>The internet is currently evolving from the "<u>Web of Documents</u>" into the "<u>Web of Data</u>" where data is available on web-scale in the so called Semantic Web to retrieve information <u>or for data reuse</u>, e.g. within applications for more automation. (p.2)</p>
Geisler, Quix, Weber, Jarke		<p>Accessibility is a <u>quality dimension</u>.</p>		
Harn, Kim, Lee, Choi (2015)		<p>They mention a few challenges regarding open data. And according to them the <u>most identified barriers</u> include lack of comprehensive data policies, lack of validity, completeness of datasets, lack of motivation within public sector, lack of technical and semantic interoperability, lack of technical ability within public and private sectors, <u>and inaccessible datasets</u></p>		
Jayewardene, Sadiq, Indulska		<p>Accessibility is a <u>quality dimension</u>: "various dimensions of information quality, such as accessibility, believability, completeness, and so on." (p.3)</p>		

<p>Kallinikos, Aaltonen, Marton</p>	<p>Relation 'findability' and 'accessibility': "search-driven information accessibility and retrieval should be understood not only in terms of the <u>immediate findability of digital artifacts</u> but also in terms of the second order effects that emerge from reactions to the contingent findability of webpages through <u>search engines</u>." (p. 363)</p>	<p>Accessibility defines together with leverage, adaptability, ease of mastery, and transferability, <u>the functional identity and innovativeness of generative technologies</u>. (p. 258)</p>	<p>Interoperability is "an important <u>condition of the digital ecosystem</u>." (p. 360) "Digital objects certainly admit investigation in terms of the technical and organizational requirements that ensure their interoperability and growth." (p. 367)</p>	
<p>Beebe & Walz (2005)</p>		<p>The model measures data quality along three <u>dimensions</u>: accuracy, relevancy, and accessibility.</p>		
<p>Malaverri, Mota, and Medeiros (2013)</p>	<p><u>Provenance</u> is "a key piece to evaluate the quality of data". They state that Data Quality is a <u>subjective concept</u>. Data which is good for one organization can be bad for another. Therefore, according to them <u>the context</u> is important to consider. And that makes provenance so indispensable. (p.)</p>		<p>"by making use of ontologies to represent provenance we allow <u>interoperability among groups</u>, enabling them to share and compare the information produced in their work." (p.2) "Interoperability across <u>distinct groups that want to share and reuse data sets in their processes</u>." (p.7) "Data quality assessment is a key factor in data-intensive domains. The data deluge is aggravated by an increasing need for interoperability and cooperation across groups and organizations." (p.1)</p>	<p><u>Reusability as consequence</u>: "This enhances interoperability across distinct groups that want to share and reuse data sets in their processes." (p.7)</p>
<p>Matiasko, Zabovska, Zabovsky (2004)</p>		<p><u>Necessity of accessibility</u>: "Data stored inside these database systems is under strong pressure to be accessible directly in the semantic form." (p. 227)</p>	<p><u>Requested interoperability</u> consists of "Data-type interoperability, Specification-level interoperability, and Semantic interoperability." (p.223)</p>	<p>Importance of <u>dereferencing of URIs</u> regarding reusability of data.</p>

				<p><u>Semantic interoperability</u> is “the ability of user to access, consistently and coherently, similar (though autonomously defined and managed) classes of digital objects and service distributed across heterogeneous repositories, with federating or mediating software compensating for site-by-site variations.” (p.224)</p>	
Moreau et al. (2011)	<p>Provenance → the tracking of historical information concerning the creation of a dataset. It is a kind of <u>metadata that gives information about what, when, where, how, by whom, and why a dataset was created.</u></p>				
Olbrich (2010)		<p>“Data quality describes the suitability or utility of data for one respective data processing application. To evaluate this suitability, Wang et al. (1996) empirically analyzed <u>semantic categories of data quality</u>, such as intrinsic, representational and contextual quality as well as accessibility.” (p.4)</p>			
Open standards and Re-use Government Action Plan UK					Nr of different ‘actions’ to make reusability of data better.
Philips-Wren, Iyer, Kulkarni & Arivachandra (2015)		<p>“We need techniques that increase accessibility of data analytics to a larger number of users. <u>New dashboard designs</u> are being developed with new requirements such as <u>interactivity and data flexibility.</u>” (p. 461)</p>			
Sheridan and Tennison (2010)	Paragraph 3.3 about ‘provenance’ (p.3)				In this model “each fact or data point is <u>associated with a URI</u> and that URI can be resolved. The publisher determines what information is returned when a request is made and can serve

				whatever additional context or provenance information they deem necessary. (...) The data can be copied, adapted and re-used, but <u>the publisher always controls what is returned when each URI is dereferenced</u> . This is an important benefit over interchange formats such as CSV or XML where data can be changed or context lost as it is passed from hand to hand or system to system." (p.2)
Sherman (2014)	' <u>Consistent</u> ': to avoid confusion about whose version of the data is the correct one.	' <u>Current</u> ': the business needs to base decisions on whatever currency is necessary for that type of decision. This means that the data in some cases needs to be up to the minute. ' <u>Comprehensive</u> ': business people should have all the data they need to do their jobs, <u>regardless of where the data comes from and its level of granularity</u> .	' <u>Conformed</u> ': the business needs to analyze the data <u>across common, shareable dimensions of business people across the enterprises</u> , so that the same information for decision-making is used.	' <u>Clean</u> ': because dirty data has missing items, invalid entries, and other problems that wreak havoc with automated data integration and data analysis. They state that most source data are dirty to some degree, which is why data profiling and cleansing are critical steps in data warehousing. <u>Dirty data is not useful and not reusable</u> .
Shukair et al (2013)			Based on this variety of topics we can conclude that <u>interoperability has many aspects, mainly technical, semantic, and organizational</u> . And it <u>becomes more and more important in government</u> , because the <u>different interpretations of data, the lack of common metadata, and the absence of universal reference data</u> .	
Tilly, Posegga, Fischbach, Schoder		This distinction between data and information is also reflected in definitions of traditional DQ and IQ. A common definition of traditional DQ		

		<p>is the degree of integrity and correspondence of data to external phenomena, which comprises, for example, completeness, unambiguity, meaningfulness, and correctness (for example, in Orr (1998); Price and Shanks (2005); Wand and Wang (1996)). Conversely, a common definition of traditional IQ is “fitness for use,” that is, the extent to which information can easily be perceived, interpreted, and applied to a task by the consumer of that information, based on data s/he receives (see, for example, Ballou et al. (2003); Madnick et al. (2009); Strong et al. (1997); Wang and Strong (1996)). This may include dimensions such as accessibility, suitability of presentation, understandability, security, and flexibility.</p> <p>(p. 4)</p>		
Vries (2012) cited in Lassinantti (2014)				<p>reusability of public data is about “putting the public data to use in new contexts and by other people than the original public-sector employees”.</p>
Wang, Strong (1996)		<p>Accessibility is the “ability to identify errors” (p.11)</p> <p>“Representational DQ and accessibility DQ emphasize the importance of the role of systems.” (p.6)</p> <p>“Accessibility Data Quality Information systems professionals understand accessibility DQ well. Our research findings show that <u>data consumers also recognize its importance</u>. Our findings appear to differ from the literature that treats accessibility as distinct from information quality” (p. 21)</p> <p>“there is little <u>difference between treating accessibility DQ as a category of overall data quality, or separating it from other categories of data quality</u>. In either case, accessibility needs to be considered.” (p.21)</p>		
Wang, Truptil & Benaben (2015)			<p>Interoperability is “one of the key competition factors for modern enterprises” and “describes the ability to establish partnership activities in an environment of unstable market” (p.16)</p>	<p>Traditional model transformation practices have their weaknesses: “low reusability, repetitive</p>

			<p>"Interoperability is <u>the ability of a system or a product to work with other systems or products without special effort from the user</u>" (p.16)</p> <p>"Interoperability is <u>a measure of the degree to which diverse systems, organizations, and/or individuals are able to work together to achieve a common goal.</u>" (p.16-17)</p>	tasks, huge manual effort." (p.17)
Wilkinson et al. (2016)	See Problem Statement	See Problem Statement	See Problem Statement	See Problem Statement
Yu (2016)		Accessibility is "an <u>(End) User Acceptance Factor</u> " (p.97)	Interoperability is "a <u>System Functional Feature</u> " (p.97)	Reusability is "a <u>System Functional Feature</u> " (p.97) "(...), and reusability emphasize that <u>government published information can be retrieved, downloaded, indexed, searched, and visualized easily, and should be stored in an open format that is machine readable, platform independent, and made available without restrictions</u> to impede the re-use of information." (p.96)
Zhang, Indulska, Jayawardene, Sadiq, Zhou		Accessibility is (in combination with availability) about " <u>ease of use, maintainability and control of the data from end users' perspective</u> " (p.4)		

Appendix B: Process Systematic Literature Review

Overall Inclusion and exclusion criteria:

- Language: English
- Recommended journals
- Set of predefined search terms
- Published since 1990
- Terms like 'data quality', 'linked data', 'ontology', 'open innovation' in title

Legend: Column 'Relevant'

X	Not relevant for this research
!	Doubtful
V	Useful for this research

	Access methods
	Data integration
	Linked Data
	Ontology
	Missing Data
	Provenance
	Data warehousing
	Data quality dimensions
	Open Government Data
	Reusability
	Open Data Marketplace
	Industry/ open innovation
	Open Data Science
	Remaining

Total amount of useful papers: 47 papers

Useful papers per topic			
	Authors, year	Title	Topic
Access methods			
1	Gaede, Günther (1998)	Multidimensional access methods	data structures, multidimensional access methods
2	Batini, Cappiello, Francalanci, Maurino (2009)	Methodologies for data quality assessment and improvement	Systematic and comparative description of methodologies for data quality assessment and improvement
3	Fürber, Hepp (2011)	SWIQA – A semantic web information quality assessment framework	provides a framework for information quality assessment of Semantic Web data called SWIQA
4	Matiasko, Zabovska, Zabovsky (2004)	Building the unified data access framework	The main aim of our work is to allow the unified data access on the international level for educational, commercial and security purposes.
Data integration			
5	Sheth, Larson (1990)	Federated database systems for managing distributed,	Database design and integration, heterogeneous DBMS, schema integration

		heterogeneous, and autonomous databases	
6	Martin, Poulouvasilis, Wang (2014)	A methodology and Architecture Embedding Quality Assessment in Data Integration	Data integration methodology, iterative quality assessment and improvement of the integrated resource
Linked Data			
7	Yu, Dietze, Pedrinaci (2011)	A linked data compliant framework for dynamic and web-scale consumption of web services	propose to apply RDF to expose Web services and Web APIs and introduce a framework in which service registries as well as services contribute to the automation of service discovery, and hence, workload is distributed more efficiently.
8	Sheridan, Tennison (2010)	Linking UK Government Data	Guidelines Linked Data, UK Government's data website as an example
9	Debattista, Auer, Lange (2016)	Luzzu – A Methodology and Framework for Linked Data Quality Assessment	describes a conceptual methodology for assessing Linked Datasets, and Luzzu; a framework for Linked Data Quality Assessment.
Ontology-based Data Quality Management			
10	Kallinikos, Aaltonen, Marton (2013)	The ambivalent ontology of Digital Artifacts	(1) provenance and authenticity of digital documents within the overall context of archiving and social memory and (2) the content dynamics occasioned by the findability of content mediated by Internet search engines.
11	Tilly, Posegga, Fischbach, Schoder (2015)	What is Quality of Data and Information in Social Information Systems? Towards a Definition and Ontology	new definition of DIQ in social IS based on the notion of "matching" between dynamic, voluntary, and heterogeneous supply and demand of data/information. We illustrate our definition with an ontological framework and discuss its implications.
12	Geisler, Quix, Weber, Jarke (2016)	Ontology- based Data Quality Management for Data Streams	an ontology-based data quality framework for relational DSMS that includes DQ measurement and monitoring
Missing Data			
13	Li (2009)	A Bayesian Approach for Estimating and Replacing Missing Categorical Data	Two alternative methods for replacing the missing value are proposed
14	Tremblay, Dutta, Vandermeer (2010)	Using Data Mining Techniques to Discover Bias Patterns in Missing Data	Data quality problem in data repositories: missing data
Provenance			
15	Joana, Gonzales Malaverri, Mota, Bauzer Medeiros (2013)	Estimating the quality of data using provenance: a case study in eScience	presents a strategy to provide information to support the evaluation of the quality of data sets. This strategy is based on combining metadata on the provenance of a data set and quality dimensions
16	Chong, Skalka, Vaughan (2015)	Self-Identifying Data for Fair Use	Introduces a technique to directly associate provenance information with sensor datasets
Data Warehousing			
17	Beebe & Walz (2005)	An Empirical Investigation of the Impact of Data Quality and its Antecedents on Data Warehousing	reviews system success and data quality literature and proposes a new model for data warehousing success

18	LeRouge, Gjestland (2002)	A typology of data warehouse quality	Data warehouse quality is divided into Information quality and system quality and attributes
19	Olbrich (2010)	Warehousing and Analyzing Streaming Data Quality Information	How to efficiently provide applications with information about data quality > a novel concept to stream and warehouse data together with its describing data quality info
20	Neely (2002)	Data Quality Knowledge management: a tool for the collection and organization of metadata in a data warehouse	This paper describes a relational database tool, the DQKM, which captures and organizes the metadata associated with a data warehouse project
Data Quality Dimensions			
21	Yeoh, Want, Verbitskiy (2012)	Describing Data Quality Problem through a Metadata Framework (Download at UU)	A set of data quality dimensions by examining the data quality management principles and current BI environment. Thereby a high-level metadata framework is proposed.
22	Jayewardene, Sadiq, Indulska (2012)	Practical Significance of Key Data Quality Research Areas	Key data quality research themes + practitioner views on these seven data quality factors
Open Government Data			
23	Charalabidis, Alexopoulos, Loukis (2016)	A taxonomy of open government data research areas and topics	a detailed taxonomy of research areas and corresponding research topics of the Open Government Data (OGD) domain is presented
24	Chatfield, Reddick, Al-Zubaidi (2015)	Capability Challenges in Transforming Government through Open and Big data: Tales of Two cities	explores organizational capability challenges in transforming government through big data use (systematic research)
25	Harn, Lee, Kim, Choi (2015)	Open Innovation Maturity Model for the Government: An open system perspective	this study aims to understand data-driven open innovation practices in government by developing a government-level open innovation maturity model, evaluating the status of open innovation of the government, and suggesting appropriate future directions and guidelines for the government.
Reusability			
26	Lassinantti, Bergvall-Kareborn (2014)	Open Data in Europe – Mapping User Groups to Future Innovation Impacts	Opening of Data in Europe, reusability
Industry; Open innovation			
27	Cui, Ye, Teo, Li (2015)	InformationTechnology And open innovation: A strategic alignment perspective	proposes a model to explain the performance of organizational open innovation
28	Blohm, Leimeister, Krcmar (2013)	Crowdsourcing: How to Benefit from (Too) Many Great Ideas	focuses on how companies can cope with the enormous volume and variety of data (big data) that is acquired on crowdsourcing platforms from the worldwide community of Internet users.
29	Kaasenbrood, Zuiderwijk, Janssen, de Jong, Bharosa, (2015)	Exploring the Factors Influencing the Adoption of Open Government Data by Private Organizations	A framework for identifying factors influencing the adoption of Open Government Data by private organizations
30	Balazinska (2015)	Big Data Research: Will Industry Solve all the Problems?	whether industry will solve all the problems or whether there is a place for academic research in big data and what is that place
Open Data Science			
31	Allen, Burk, Davis (2006)	Academic Data Collection in Electronic Environments: Defining Acceptable Use of Internet Resources	Two major legal challenges to the use of automated data collection agents for academic research use are based on the legal doctrines of trespass and copyright.

32	Nosek, Alter, Bank, Borsboom, Bowman, Breckler, Contestabile, (2015).	Promoting an open research culture	Author guidelines for journals could help to promote transparency, openness, and reproducibility
33	Levin (2015)	Open Access, Open Data, Open Science... What does "openness" mean in the first place?	Critical paper on open science challenges
34	McKiernan, Erin, et al. (2016)	How open science helps researchers succeed	The benefits of open science for the researcher
35	Wilkinson, et al. (2016)	The FAIR Guiding Principles for scientific data management and stewardship	FAIR Data Principles
Remaining			
36	Wang, Strong (1996)	Beyond Accuracy: What Data Quality Means to Data Consumers	a framework that captures the aspects of data quality that are important to data consumers
37	Batini, Scannapieco (2016)	Data and Information Quality: Dimensions, Principles and Techniques	Data Quality Dimensions, Information Quality Dimensions, models, Data integration, open IQ problems
38	Benenson (2016)	The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences	Key concepts
39	Josuttis (2007)	SOA in practice: the art of distributed system design	Service Oriented Architecture
40	Chien-Chih (2016)	A value-centric business model framework for managing open data applications	introduces an Open Data Application (ODA) applicable value-centric Business Model (ODA-vBM) framework for guiding the development of operational business models
41	Chaturvedi, Dolk, Drnevich (2011)	Design Principles for Virtual Worlds	IS Design theory, virtual world systems, emergent knowledge processes, agent-based simulation, deep structure, platform as a methodology (PaaM), user-developed content (UDC)
42	Wang, Truptil, Benaben (2015)	A General Model Transformation Methodology to Serve Enterprise Interoperability Data Sharing Problem	a general model transformation methodology
43	Klein, Lehner (2009)	Representing Data Quality in Sensor Data Streaming Environments (Not whole text available?)	restricted quality of sensor data due to limited sensor precision and sensor failures
44	Philips-Wren, Iyer, Kulkarni & Arivachandra (2015)	Business Analytics in the Context of Big Data: A Roadmap for Research	Big data access, big data governance
45	Kokemueller (2011)	An empirical investigation of factors influencing data quality improvement success	Empirically analyzing the factors influencing the success of data quality improvements. Organizational implementation success is positively associated with perceived data quality.
46	Zhang, Jayawardene, Indulska, Sadiq, Zhou (2014)	A Data Driven Approach for Discovering Data Quality Requirements	an approach for discovering data quality issues using generic exploratory methods
47	Smith, Ofe, Sandberg (2016)	Digital Service Innovation from Open Data: Exploring the Value Proposition of an Open Data Marketplace	open data marketplaces can lower the threshold of using open data by providing better access to open data and associated support services, and increasing knowledge transfer within the ecosystem.

ACM (Association for Computing Machinery)

<http://dl.acm.org/results.cfm?query=data%20quality&filtered=&within=acmPubGroups%2EacmPubGroup%3DJournal&dte=1990&bfr=&srt=score>

1. Search term: 'Data Quality' → 9,039 results
2. Published since 1990 → 7,827 results
3. >500 citations → 12 results (see table)
4. Reading the abstracts, determine topic, based on that; is the paper relevant for this research? → 2 results (see table)

Authors, year	citations	Title	Topic	Relevant	Link
Jain, Murty, Flynn (1999)	1,986	Data clustering: a review	overview of pattern clustering methods from a statistical pattern recognition perspective	X	http://eprints.iisc.ernet.in/273/1/p264-jain.pdf
Herlocker, Konstan, Terveen, Riedl (2003)	1,037	Evaluating collaborative filtering recommender systems	key decisions in evaluating collaborative filtering recommender systems	X	http://scholar.google.nl/scholar?q=Evaluating+collaborative+filtering+recommender+systems&btnG=&hl=nl&as_sdt=0%2C5
Järvelin, Kekäläinen (2002)	893	Cumulated gain-based evaluation of IR techniques	in large database environments; IR methods based on their ability to retrieve highly relevant documents	!	http://scholar.google.nl/scholar?q=Cumulated+gain-based+evaluation+of+IR+techniques&btnG=&hl=nl&as_sdt=0%2C5
Herlihy, Wing (1990)	800	Linearizability: a correctness condition for concurrent objects	concurrency, correctness, linearizability, multiprocessing, serializability	X	http://www.doc.ic.ac.uk/~gbd10/aw590/Linearizability%20-%20A%20Correctness%20Condition%20for%20Concurrent%20Objects.pdf
Cytron, Ferrante, Rosen, Wegman, Zadeck (1991)	687	Efficiently computing static single assignment form and the control dependence graph	Control dependence, Control flow graph, clef-use chain, dominator, optimizing compilers	X	https://www.researchgate.net/profile/Jeanne_Ferrante/publication/213879567_Efficiently_Computing_Static_Single_Assignment_Form_and_the_Control_Dependence_Graph/links/549459020cf25e15dda2420.pdf
Rother, Kolmogorov, Blake	676	"GrabCut": interactive foreground extraction using iterated graph cuts	Interactive Image Segmentation, Graph Cuts, Image Editing, Foreground extraction, Alpha Matting	X	http://pages.cs.wisc.edu/~dyer/cs534-fall11/papers/grabcut-rother.pdf

Herlihy (1991)	593	Wait-free synchronization	Linearizability, wait-free synchronization	X	https://www4.informatik.uni-erlangen.de/DE/Lehre/WS14/PS_KVBK/papers/waitfree.pdf
Van Gelder, Ross, Schlipf (1991)	546	The well-founded semantics for general logic programs	Unfounded sets And well-founded Partial models	X	https://pdfs.semanticscholar.org/ad69/24abcce554dc66819fe05de9c88bd3fd43d8.pdf
Gaede, Günther (1998)	533	Multidimensional access methods	data structures, multidimensional access methods	V	http://cs.unibo.it/~montesi/CBD/Articoli/MultidimensionalAccessMethods.pdf
Sheth, Larson (1990)	528	Federated database systems for managing distributed, heterogeneous, and autonomous databases	Database design and integration, heterogeneous DBMS, schema integration	V	http://csis.pace.edu/~marchese/CS865/Papers/p183-sheth.pdf
Aurenhammer (1991)	523	Voronoi diagrams – a survey of a fundamental geometric data structure	Voronoi diagrams	X	http://www.dsg.nutn.edu.tw/msrg/home%20page/member/96patrick/pdf/Voronoi%20Diagrams%20A%20Survey%20of%20a%20Fundamental%20Geometric%20Data%20Structure.pdf
Yilmaz, Javed, Shah (2006)	520	Object tracking: a survey	Object tracking, feature selection, object detection	X	http://dl.acm.org/citation.cfm?id=1177355

1. Search term: 'Data Quality' → 9,039 results
2. Published since 1990 → 7,827 results
3. Top-10 most relevant (due to the 'relevance' filter dl.acm.org) → 10 results (see table)
4. Reading the abstracts, determine topic, based on that; is the paper relevant for this research? → 6 results (see table)

Authors, year	Citations	Title	Topic	Relevant	
Fan, Geerts, Wijsen (2012)	4	Determining the Currency of Data	Data currency	!	http://www2.cs.siu.edu/~dche2/files/datacurrency.pdf
Martin, Poulouvasilis, Wang (2014)	2	A methodology and Architecture Embedding Quality Assessment in Data Integration	Data integration methodology, iterative quality assessment and improvement of the integrated resource	V	http://dl.acm.org/citation.cfm?id=2567663
Batini, Cappiello, Francalanci, Maurino (2009)	58	Methodologies for data quality assessment and improvement	Systematic and comparative description of methodologies for data quality assessment and improvement	V	http://dimacs-algorithmic-mdm.wdfiles.com/local--files/start/Method

					ologies%20for%20Data%20Quality%20Assessment%20and%20Improvement.pdf
Heinrich, Klier, Kaiser (2009)	8	A Procedure to Develop Metrics for Currency and its Application in CRM	Procedure which can be adjusted to specific characteristics of data attribute values	X	http://epub.uni-regensburg.de/23166/1/heinrich.pdf
Debattista, Auer, Lange (2016)	0	Luzzu – A Methodology and Framework for Linked Data Quality Assessment	describes a conceptual methodology for assessing Linked Datasets, and Luzzu; a framework for Linked Data Quality Assessment.	V	(only UBU access)
Geisler, Quix, Weber, Jarke (2016)	0	Ontology- based Data Quality Management for Data Streams	an ontology-based data quality framework for relational DSMS that includes DQ measurement and monitoring	V	(only UBU access)
Collins, Janssens (2012)	0	Creating a General (Family) Practice Epidemiological Database in Ireland – Data Quality Issue Management	outlines the process of data quality issue management undertaken	X	http://dl.acm.org/citation.cfm?id=2378018
Tremblay, Dutta, Vandermeer (2010)	0	Using Data Mining Techniques to Discover Bias Patterns in Missing Data	Data quality problem in data repositories: missing data	V	https://datapro.fiu.edu/campusedge/files/articles/charnitrm2956.pdf
Klein, Lehner (2009)	9	Representing Data Quality in Sensor Data Streaming Environments (Not whole text available?)	restricted quality of sensor data due to limited sensor precision and sensor failures	V	
Li (2009)	4	A Bayesian Approach for Estimating and Replacing Missing Categorical Data	Two alternative methods for replacing the missing value are proposed	X	http://dl.acm.org/citation.cfm?id=1515695

1. Search term: 'fair data'+ published since 1990 → 6,840 results
2. Scanning by title + abstract → 1 result (see table)

Authors, year	citations	Title	Topic	Relevant	Link
Chong, Skalka, Vaughan (2015)	0	Self-Identifying Data for Fair Use	Introduces a technique to directly associate provenance information with sensor datasets	V	https://dash.harvard.edu/bitstream/handle/1/22043260/14114565.pdf?sequence=1

AIS (Association for Information Systems)

<http://aisel.aisnet.org/do/search/?q=Data%20Quality&start=0&context=509156>

1. Search term: 'Data Quality' → 18,995 results
2. Published since 1990 → 18,629 results
3. Top-15 most relevant (due to the 'relevance' filter aisel.aisnet.org) → 15 results (see table)
4. Reading the abstracts, determine topic, based on that; is the paper relevant for this research? → 6 results (see table)

Authors, year	Title	Topic	Relevant	Link
Strong, Madnick, Hartman, Peace, Thompson (1994)	Data Quality: A Critical Research Issue for the 1990s and Beyond	The purpose of this panel, therefore, is to assess the state-of-the-art of research on data quality and to discuss emerging research issues.	X	http://aisel.aisnet.org/icis1994/18/
Beebe & Walz (2005)	An Empirical Investigation of the Impact of Data Quality and its Antecedents on Data Warehousing	reviews system success and data quality literature and proposes a new model for data warehousing success	V	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.837.3370&rep=rep1&type=pdf
LeRouge, Gjestland (2002)	A typology of data warehouse quality	Data warehouse quality is divided into Information quality and system quality and attributes	V	http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1365&context=amcis2002
Olbrich (2010)	Warehousing and Analyzing Streaming Data Quality Information	How to efficiently provide applications with information about data quality > a novel concept to stream and warehouse data together with its describing data quality info	V	http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1165&context=amcis2010
Aisbett, Gibbon, Lear (1997)	The Quality of Data and its Effect on Information Usage	Suggestions for describing the quality of data content	!	http://www.pacis-net.org/file/1997/66.pdf
Kokemueller (2011)	An empirical investigation of factors influencing data quality improvement success	Empirically analyzing the factors influencing the success of data quality improvements. Organizational implementation success is positively associated with perceived data quality.	V	https://pdfs.semanticscholar.org/b349/59fbb21e4e8a2f38289395c703519aa245ba.pdf
Robbert, Senne (2003)	Teaching GIGO: Data Quality in the Curriculum	reviews current texts for inclusion of quality and notes little change in the model curriculums inclusion of quality	X	http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1762&context=amcis2003
Baghi, Otto, Oesterle (2013)	Controlling Customer Master Data Quality: Findings from a Case study	a single case study describing the process of implementing a comprehensive data quality controlling system. The study focuses on controlling activities defined in the field of business management.	X	https://www.researchgate.net/profile/Boris_Otto/publication/259943236_Controlling_Customer_Master_Data_Quality_Findings_from_a_Case_Study/links/579f113a08ae5d5e1e172a85.pdf
Al-Abdullah, Weistroffer (2011)	A Framework to Enhance Decision Outcomes: Data Quality Perspective	Framework that relates data quality aspects to decision support based on the published literature	!	https://www.researchgate.net/profile/Heinz_Weistroffer/publication/228424355_A_Framework_to_Enhance_Decision_Outcomes_Data_Quality_Perspective/links/0046351912fe3964e0000000.pdf

Shanks (2001)	The Impact of Data Quality Tagging on Decision Outcomes	Draws together concepts from a semiotic-based theory on data quality and normative theories on decision-making to examine the impact of data quality tagging about data accuracy on decision outcomes	X	http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1068&context=acis2001
Yeoh, Want, Verbitskiy (2012)	Describing Data Quality Problem through a Metadata Framework (Download at UU)	A set of data quality dimensions by examining the data quality management principles and current BI environment. Thereby a high-level metadata framework is proposed.	V	Only UBU access
Jayewardene, Sadiq, Indulska (2012)	Practical Significance of Key Data Quality Research Areas	Key data quality research themes + practitioner views on these seven data quality factors	V	http://www.pacis-net.org/file/2012/PACIS2012-070.pdf
Becker, Poeppelbus, Gloerfeld, Bruhns (2009)	The Impact of Data Quality on Value Based Management of Financial Institutions	Data quality issues in the financial sector	X	http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1512&context=amcis2009
Kerr, Norris, Stockdale (2007)	Data Quality Information and Decision Making: A Healthcare Case Study	Development of a data quality evaluation framework for the NZ health sector	X	https://www.researchgate.net/profile/Karolyn_Kerr/publication/228354908_Data_quality_information_and_decision_making_A_healthcare_case_study/links/09e4150582c751ea67000000.pdf
Neely, Lin, Gao, Koronios (2006)	The Deficiencies of Current Data Quality Tools in the Realm of Engineering Asset Management	Discuss the actual data quality problems with the operation-level and middle-level managers in engineering asset management organizations	X	http://scholarworks.rit.edu/cgi/viewcontent.cgi?article=1576&context=article

1. Search term: 'Data Quality open data' + since 1990 → 10,392 results
2. Top-100 most relevant (filter ACM) → 100 results
3. Screening on title → 12 results (see table)
4. After reading abstract still relevant? → 9 results (see table)

Authors, year	Title	Topic	Relevant	Link
Neely (2002)	Data Quality Knowledge management: a tool for the collection and organization of metadata in a data warehouse	This paper describes a relational database tool, the DQKM, which captures and organizes the metadata associated with a data warehouse project	V	http://scholarworks.rit.edu/cgi/viewcontent.cgi?article=1444&context=other
Fürber, Hepp (2011)	SWIQA – A semantic web information quality assessment framework	provides a framework for information quality assessment of Semantic Web data called SWIQA	V	http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1075&context=ecis2011

Zhang, Jayawardene, Indulka, Sadiq, Zhou (2014)	A Data Driven Approach for Discovering Data Quality Requirements	an approach for discovering data quality issues using generic exploratory methods	✓	https://pdfs.semanticscholar.org/2213/f668e596647709c0f3d43e691ec57f07f22c.pdf
Chatfield, Reddick, Al-Zubaidi (2015)	Capability Challenges in Transforming Government through Open and Big data: Tales of Two cities	explores organizational capability challenges in transforming government through big data use (systematic research)	✓	https://pdfs.semanticscholar.org/4558/e6a25bb045dbe0f8f9c23bc75eb20e49f4e.pdf
Alanazi, Chatfield (2012)	Sharing Government-Owned Data with the Public: A cross-country analysis of Open Data Practice in the Middle East	Open government policies in the middle East	✗	http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1296&context=eispapers
Philips-Wren, Iyer, Kulkarni & Arivachandra (2015)	Business Analytics in the Context of Big Data: A Roadmap for Research	Big data access, big data governance	✓	http://aisel.aisnet.org/cgi/viewcontent.cgi?article=3888&context=cais
Joana, Gonzales Malaverri, Mota, Bauzer Medeiros (2013)	Estimating the quality of data using provenance: a case study in eScience	resents a strategy to provide information to support the evaluation of the quality of data sets. This strategy is based on combining metadata on the provenance of a data set and quality dimensions	✓	http://www.repositorio.unicamp.br/bitstream/REPOSIP/88829/1/2-s2.0-84893254834.pdf
Harn, Lee, Kim, Choi (2015)	Open Innovation Maturity Model for the Government: An open system perspective	this study aims to understand data-driven open innovation practices in government by developing a government-level open innovation maturity model, evaluating the status of open innovation of the government, and suggesting appropriate future directions and guidelines for the government.	✓	https://pdfs.semanticscholar.org/8633/35ff33f617d0de6d9203204e5c6d30b25ae6.pdf
Otto, Aier (2013)	Business Models in the Data Economy: A Case study from the Business Partner Data Domain	Business models, case study, data quality, data resource management, resource-based view	✗	https://www.alexandria.unisg.ch/220968/1/Otto.Aier.2013.DataProviderBusinessModels.pdf
Omar, Bass, Lowit (2014)	A Grounded Theory of Open Government Data: A Case Study in the UK	Importance and effects of Open Government Data	!	http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1016&context=ukais2014
Lassinantti, Bergvall-Kareborn (2014)	Open Data in Europe – Mapping User Groups to Future Innovation Impacts	Opening of Data in Europe, reusability	✓	https://www.researchgate.net/profile/Josefin_Lassinantti/publication/271835916_Open_Data_in_Europe_-_Mapping_User_Groups_to_Future_Innovation_Impacts/links/54d35a210cf250179181ff18.pdf

Tilly, Posegga, Fischbach, Schoder (2015)	What is Quality of Data and Information in Social Information Systems? Towards a Definition and Ontology	new definition of DIQ in social IS based on the notion of "matching" between dynamic, voluntary, and heterogeneous supply and demand of data/information. We illustrate our definition with an ontological framework and discuss its implications.	V	https://www.researchgate.net/profile/Oliver_Posegga/publication/285598152_What_is_Quality_of_Data_and_Information_in_Social_Information_Systems_Towards_a_Definition_and_Ontology/links/5679167e08aebcdda0ed443f.pdf
---	--	--	---	---

Web of Science

<https://apps.webofknowledge.com>

1. Search terms: "Open data sources", "Interoperability open data"
2. Search term: "Open innovation" → 9,391 results
3. Filter "Computer science information systems" → 723 results
4. Publication Data – newest to oldest
5. Scanning by title → 90 results
6. Filtering by reading abstract → 5 results (see table)

Journal	Author, year	Title	Topic	Relevant	Link
Journal of Organizational Computing and Electronic Commerce	Charalabidis, Alexopoulos, Loukis (2016)	A taxonomy of open government data research areas and topics	a detailed taxonomy of research areas and corresponding research topics of the Open Government Data (OGD) domain is presented	V	http://www.tandfonline.com/doi/pdf/10.1080/10919392.2015.1124720?needAccess=true (only UBU access)
Journal of Organizational Computing and Electronic Commerce	Chien-Chih (2016)	A value-centric business model framework for managing open data applications	introduces an Open Data Application (ODA) applicable value-centric Business Model (ODA-vBM) framework for guiding the development of operational business models	V	http://www.tandfonline.com/doi/pdf/10.1080/10919392.2015.1125175?needAccess=true (only UBU access)
49th Hawaii International Conference on System Sciences (HICSS)	Smith, Ofes, Sandberg (2016)	Digital Service Innovation from Open Data: Exploring the Value Proposition of an Open Data Marketplace	open data marketplaces can lower the threshold of using open data by providing better access to open data and associated support services, and	V	http://www.scdi.se/wp-content/uploads/2014/06/Digital-Service-Innovation-from-Open-Data.pdf

			increasing knowledge transfer within the ecosystem.		
49th Hawaii International Conference on System Sciences (HICSS)	Chen, Kazman, Haziye (2016)	Big Data as a service: A Neo-Metropolis Model Approach for Innovation	Neo-Metropolis model-a variant of the Metropolis model-that offers an organized, coherent set of open-world innovation opportunities for vendors as well as for the platform's edge customers	!	https://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=5&SID=X2n1svQ3pHNch86JNQG&page=6&doc=55 (Not Full text available?)
Proceeding of the VLDB Endowment	Balazinska (2015)	Big Data Research: Will Industry Solve all the Problems?	whether industry will solve all the problems or whether there is a place for academic research in big data and what is that place	✓	https://pdfs.semanticscholar.org/140e/eb927e77b759a94716303fcfb2456c507a.pdf
Information and Management, Elsevier Science	Cui, Ye, Teo, Li (2015)	Information technology and open innovation: A strategic alignment perspective	proposes a model to explain the performance of organizational open innovation	✓	http://www.sciencedirect.com/science/article/pii/S0378720614001566

1. Search term: "Reusable Data" → 1,616 results
2. Filter "Computer science information systems" → 249 results
1. Publication Data – newest to oldest
2. Scanning by title → 30 results
3. Filtering by reading abstract → 1 result (see table)

Journal	Author, year	Title	Topic	Relevant	Link
6th International IFIP Working Conference on Enterprise Interoperability (IWEI)	Wang, Truptil, Benaben (2015)	A General Model Transformation Methodology to Serve Enterprise Interoperability Data Sharing Problem	a general model transformation methodology	✓	http://link.springer.com/chapter/10.1007%2F978-3-662-47157-9_2

MIS Quarterly

1. Search terms: "Open data", "Linked data", "Interoperability", "unique identifier open data", "findable open data", "metadata open data"
2. Scanning by title → 9 results (see table)
3. After reading abstract still relevant? → 3 results (see table)

Authors, year	Title	Topic	Relevant	Link
---------------	-------	-------	----------	------

Blohm, Leimeister, Krcmar (2013)	Crowdsourcing: How to Benefit from (Too) Many Great Ideas	focuses on how companies can cope with the enormous volume and variety of data (big data) that is acquired on crowdsourcing platforms from the worldwide community of Internet users.	✓	https://www.alexandria.unisg.ch/229504/1/JML_464.pdf
Chaturvedi, Dolk, Drnevich (2011)	Design Principles for Virtual Worlds	IS Design theory, virtual world systems, emergent knowledge processes, agent-based simulation, deep structure, platform as a methodology (PaaM), user-developed content (UDC)	✓	https://www.researchgate.net/profile/Paul_Drnevich/publication/220259848_Design_Principles_for_Virtual_Worlds/links/55118d540cf20bfdad4ea7e3.pdf
Chen, Chiang, Storey (2012)	Business Intelligence and Analytics: From Big Data to Big Impact	BI&A research framework	!	http://s3.amazonaws.com/academica.edu/documents/32970305/FROM_BIG_DATA_TO_BIG_IMPACT.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1481716240&Signature=PpZUCBdjzx57F8BYGqXz%2FQaREsw%3D&response-content-disposition=inline%3B%20filename%3DSPECIAL_ISSUE_BUSINESS_INTELIGENCE_RESE.pdf
Kallinikos, Aaltonen, Marton (2013)	The ambivalent ontology of Digital Artifacts	(1) provenance and authenticity of digital documents within the overall context of archiving and social memory and (2) the content dynamics occasioned by the findability of content mediated by Internet search engines.	✓	https://www.researchgate.net/profile/Jannis_Kallinikos/publication/231521578_The_Ambivalent_Ontology_of_Digital_Artifacts/lin

				ks/555ade9b08ae6fd2d8283893.pdf
McGrath (2016)	Identity verification and societal challenges: explaining the gap between service provision and development outcomes	Social mechanism, trust, distrust, suspicion, ambivalence, national identity cards, comparative study, socioeconomic development, financial reform, generative mechanism	X	http://aisel.aisnet.org/cgi/viewcontent.cgi?article=3294&context=misq
Rui, Chen et al. (2013)	Data Model Development for fire related extreme events: an activity theory approach	This study contributes to the literature in interoperability and data modeling; it also informs practice in emergency response system design	!	http://www.som.buffalo.edu/isinterface/papers/rui-chen-et-al-MISQ-2013.pdf
Lyer, Henderson (2010)	Seven capabilities cloud computing	predict that cloud strategies will lead to more intense ecosystem-based competition	X	Only UBU access
Wixom, Watson (2001)	An empirical Investigation of the factors affecting data warehousing success	Implementation factors and the success of data warehousing	!	http://s3.amazonaws.com/academia.edu/documents/46812813/Wixom_and_Watson_The_Factors_that_Affect_DW_Success.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1481717099&Signature=521a2iYfT7%2BB3FOIX8ZMMRa5No0%3D&response-content-disposition=inlineline%3B%20filename%3DAn_Empirical_Investigation_of_the_Factor.pdf
Zhu, Kraemer, Gurbaxani, Xu (2006)	Migration to Open-Standard Interorganizational Systems: Network Effects, Switching Costs and Path Dependency	Open standards, standards diffusion, Interorganizational Systems	X	https://escholarship.org/uc/item/7ws3n2jw#page-2

Journal of Management Information Systems

1. Search terms: "Open data standard", "open data quality", "data repository" →
2. Scanning by title → 5 results (see table)
3. After reading abstract still relevant? → 4 results (see table)

Authors, year	Title	Topic	Relevant	Link
Allen, Burk, Davis (2006)	Academic Data Collection in Electronic Environments: Defining Acceptable Use of Internet Resources	Two major legal challenges to the use of automated data collection agents for academic research use are based on the legal doctrines of trespass and copyright.	V	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.4283&rep=rep1&type=pdf
Matiasko, Zabovska, Zabovsky (2004)	Building the unified data access framework	The main aim of our work is to allow the unified data access on the international level for educational, commercial and security purposes.	V	https://www.researchgate.net/profile/Michal_Zabovsky/publication/228951802_Building_the_Unified_Data_Access_Framework/links/02bfe51113c43da9a2000000.pdf
Wang, Strong (1996)	Beyond Accuracy: What Data Quality Means to Data Consumers	a framework that captures the aspects of data quality that are important to data consumers	V	http://courses.washington.edu/geog482/resource/14_Beyond_Accuracy.pdf
Boh, Yellin (2006)	Using Enterprise Architecture Standards in Managing Information Technology	four key governance mechanisms for EA standards management and how these affect the use of EA standards.	X	http://s3.amazonaws.com/academica.edu/documents/33532192/Using_Enterprise_Architecture_Standards_in_Managing_Information_Technology.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1481715552&Signature=QJXpO79cMmnsLICuQh7jIWuRCH0%3D&response-content-disposition=inline%3B%2

				Ofilename%3DUSING_ENTERPRISE_ARCHITECTURE_STANDARDS.pdf
Yu, Dietze, Pedrinaci (2011)	A linked data compliant framework for dynamic and web-scale consumption of web services	propose to apply RDF to expose Web services and Web APIs and introduce a framework in which service registries as well as services contribute to the automation of service discovery, and hence, workload is distributed more efficiently.	V	http://oro.open.ac.uk/28899/1/Paper90.pdf

Google Scholar

4. Search term: "Open data science" → 5,940,000 results
5. Filter "since 2015" → 208,000
6. Filter by title → 40 results
7. Filter by reading abstract → 5 results (see table)

Authors, year	Title	Topic	Relevant	Link	Journal
Nosek, Alter, Bank, Borsboom, Bowman, Breckler, Contestabile, (2015).	Promoting an open research culture	Author guidelines for journals could help to promote transparency, openness, and reproducibility	V	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4550299/pdf/nihms-714651.pdf	-

Benenson (2016)	The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences	Key concepts	V	ONLY UBU Access	Geography Research Forum
Levin (2015)	Open Access, Open Data, Open Science... What does "openness" mean in the first place?	Critical paper on open science challenges	V	http://somasphere.net/2015/02/open-science.html	-
Kaasenbrood, Zuiderwijk, Janssen, de Jong, Bharosa, (2015)	Exploring the Factors Influencing the Adoption of Open Government Data by Private Organizations	A framework for identifying factors influencing the adoption of Open Government Data by private organizations	V	http://s3.amazonaws.com/academia.edu/documents/45511090/Exploring_the_Factors_Influencing_the_Ad20160510-17826-sgo3lb.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1481725308&Signature=SRkWKzDdgzcwVM3dRO7942Xj0MQ%3D&response-content-disposition=inline%3B%20filename%3DExploring_the_Factors_Influencing_the_Ad.pdf	International Journal of Public Administration in the Digital Age (IJPADA)
McKiernan, Erin, et al. (2016)	How open science helps researchers succeed	The benefits of open science for the researcher	V	https://elifesciences.org/content/5/e16800?utm_campaign=BM40104U&utm_medium=BM40104U&utm_source=Teradata	-

Remaining

Authors, year	Title	Topic	Relevant	Link	Journal
Sheridan, Tennison (2010)	Linking UK Government Data	Guidelines Linked Data, UK Government's data website as an example	V	http://wtlab.um.ac.ir/images/e-library/linked_data/2010/ldow2010	LDOW

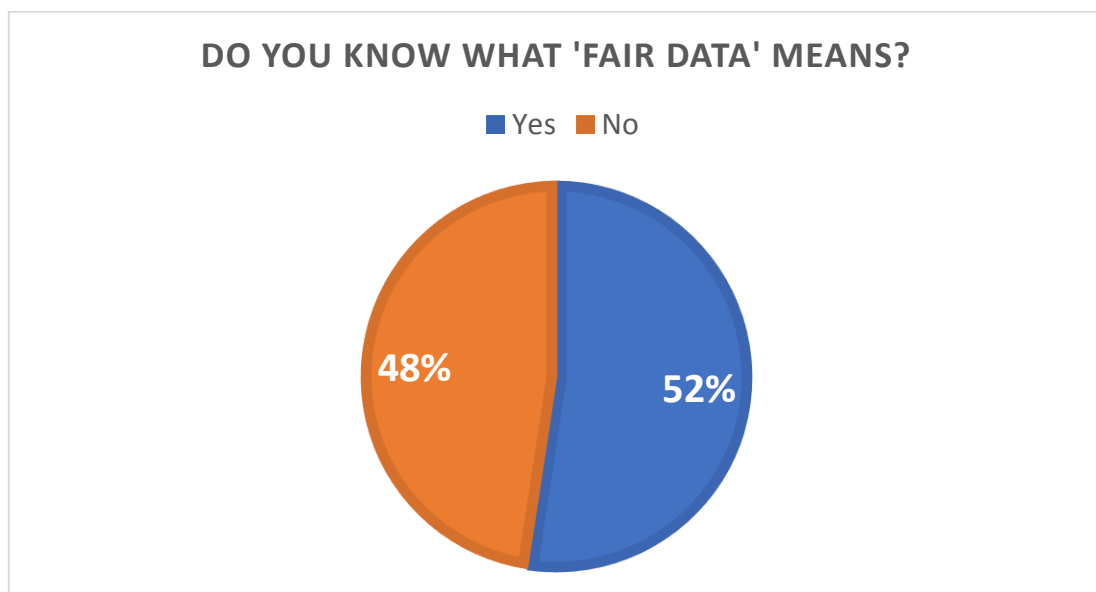
				_paper14.pdf	
Cram, Brohman, Gallupe (2016)	Information Systems Control: A Review and Framework for Emerging Information Systems Processes.	Integrate existing IS control constructs and relationships into a comprehensive IS control model	!	Only UBU Access	Journal of the Association for Information Systems
Batini, Scannapieco (2016)	Data and Information Quality: Dimensions, Principles and Techniques	Data Quality Dimensions, Information Quality Dimensions, models, Data integration, open IQ problems	✓	Only UBU Access	Book
Wilkinson, et al. (2016)	The FAIR Guiding Principles for scientific data management and stewardship	FAIR Data Principles	✓	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/	
Josuttis (2007)	SOA in practice: the art of distributed system design	Service Oriented Architecture	✓	Private property	Book

Appendix C: Survey Results

Question 1

Do you know what the term 'FAIR data' means?

ANTWOORDKEUZEN	REACTIES	
Yes	52,38%	11
No	47,62%	10
Totaal		21

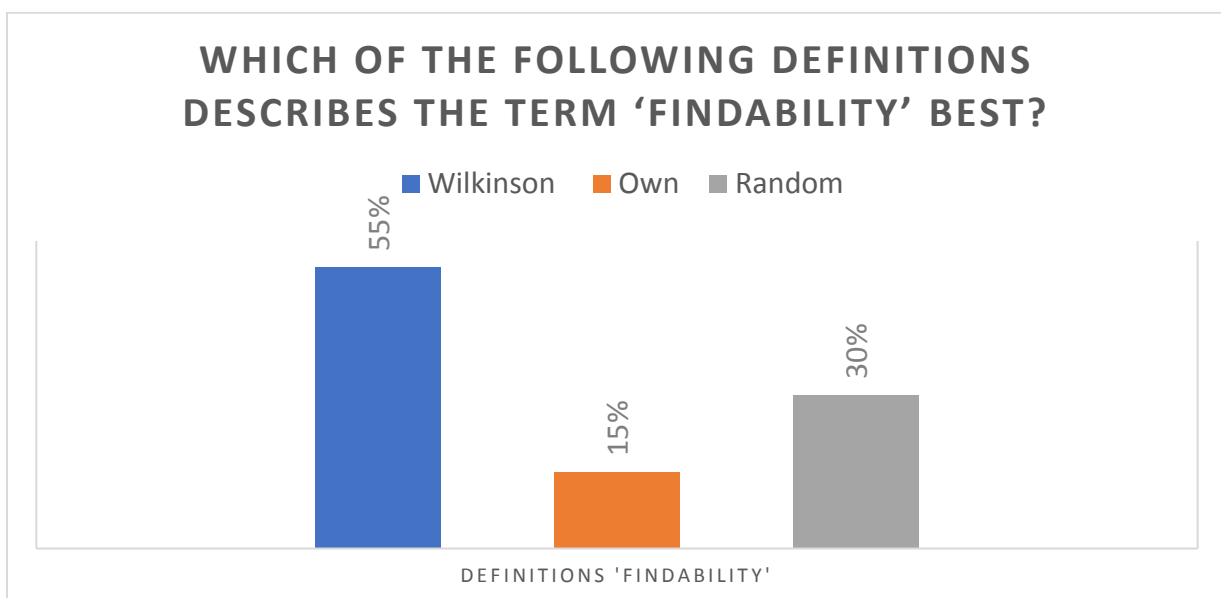


Conclusion: 50/50

Question 2

Which of the following definitions describes the term 'Findability' best?

ANTWOORDKEUZEN	REACTIES
▼ Data is Findable when the (meta) data 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 the searchable resource.	55,00% 11
▼ Findability is about the uniqueness and provenance of data to ensure data sources are searchable, free of redundancies and the context is clear, and this is achieved by versioning, associating (new) URLs or by embedding other kind of (provenance) identifiers.	15,00% 3
▼ Findability is about the ease with which information can be found using search engines.	30,00% 6
Totaal	20



Conclusion: Wilkinson definition!

Additional Response:

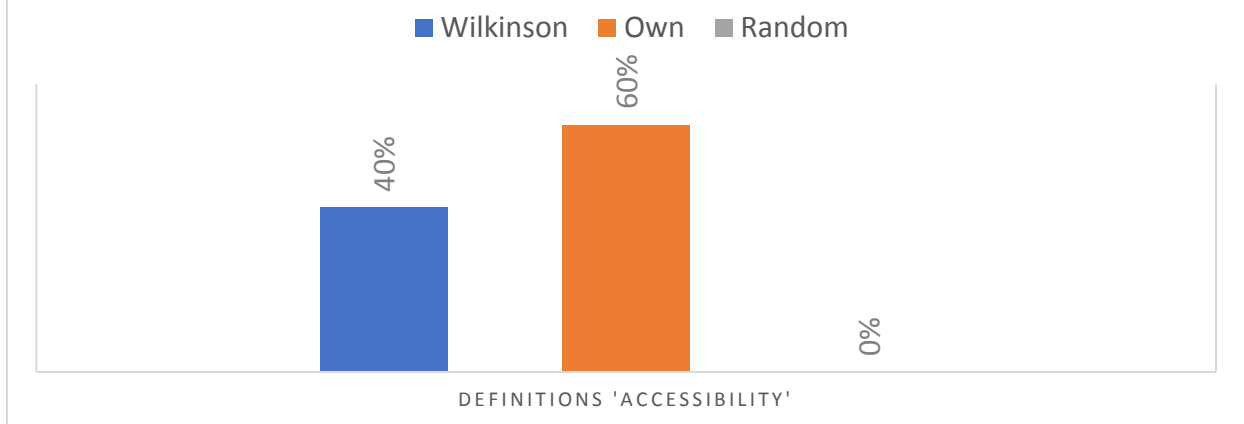
Findable – Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets;

Question 3

Which of the following definitions describes the term 'Accessibility' best?

ANTWOORDKEUZEN	REACTIES
▼ Accessibility is defined as a 'Quality dimension' to 'measure', describe and ensure data quality.	0,00% 0
▼ Data is Accessible when (meta)data are retrievably by their identifier using a standardized communication protocol, whereby the protocol is open, free, and universally implementable.	40,00% 8
▼ Accessibility is a semantic quality dimension, which comprises availability, security, performance, interactivity, flexibility of data, to ensure data access to (end) users regardless their different context and background.	60,00% 12
Totaal	20

WHICH OF THE FOLLOWING DEFINITIONS DESCRIBES THE TERM 'ACCESSIBILITY' BEST?



Conclusion: New definition!

Additional Response:

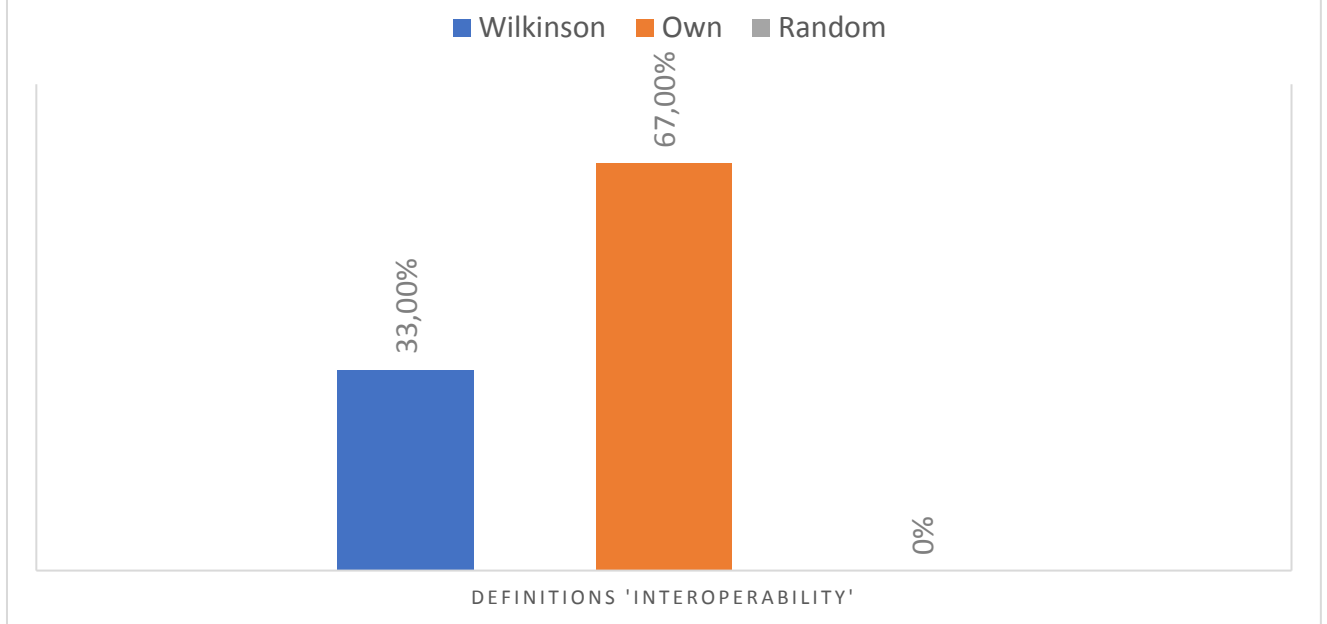
Accessible – Stored for long term such that they can be easily accessed and/or downloaded with well-defined license and access conditions (Open Access when possible), whether at the level of metadata, or at the level of the actual data content;

Question 4

Which of the following definitions describes the term 'Interoperability' best?

ANTWOORDKEUZEN	REACTIES
▼ Interoperability describes the ability to establish partnership activities in an environment of unstable market.	0,00% 0
▼ Interoperability is the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort.	33,33% 6
▼ Interoperability is a system feature to connect and conform heterogeneous environments, to ensure sharing, reusing and exchanging of data between these without special effort from the (end) user.	66,67% 12
Totaal	18

WHICH OF THE FOLLOWING DEFINITIONS DESCRIBES THE TERM 'INTEROPERABILITY' BEST?



Conclusion: New definition!

Additional Response:

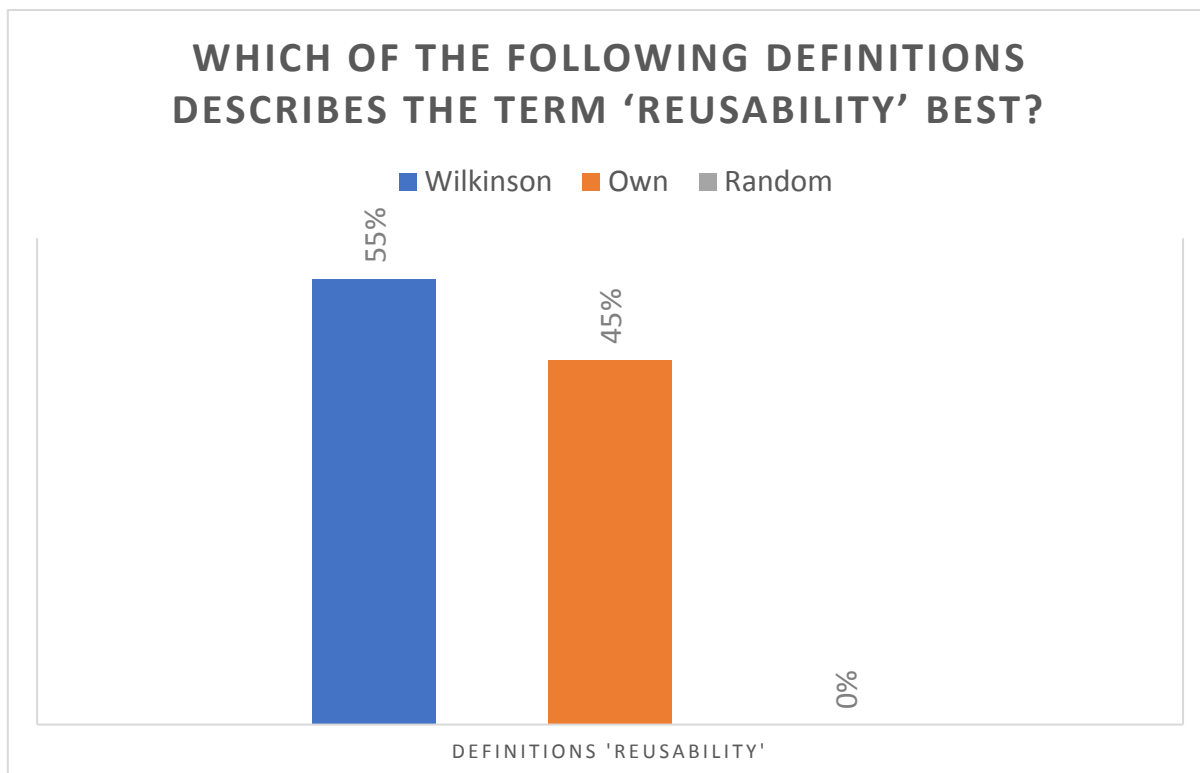
(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

Interoperable – Ready to be combined with other datasets by humans as well as computer systems;

Question 5

Which of the following definitions describes the term 'Reusability' best?

ANTWOORDKEUZEN	REACTIES
▼ Reusability is putting the data to use in new contexts and by other people than the original sector employees.	0,00% 0
▼ Reusability is the ability to make easily use of data in a new context, in terms of retrieving, downloading, indexing, searching and visualizing the data without restrictions, as a result of findable, interoperable and accessible data.	45,00% 9
▼ Data is Reusable when (meta) data are richly described with a plurality of accurate and relevant attributes, which means that (meta)data are released with a clear and accessible data usage license, (meta)data are associated with detailed provenance, and (meta)data meet domain-relevant community standards.	55,00% 11
Totaal	20



Conclusion: Wilkinson definition!

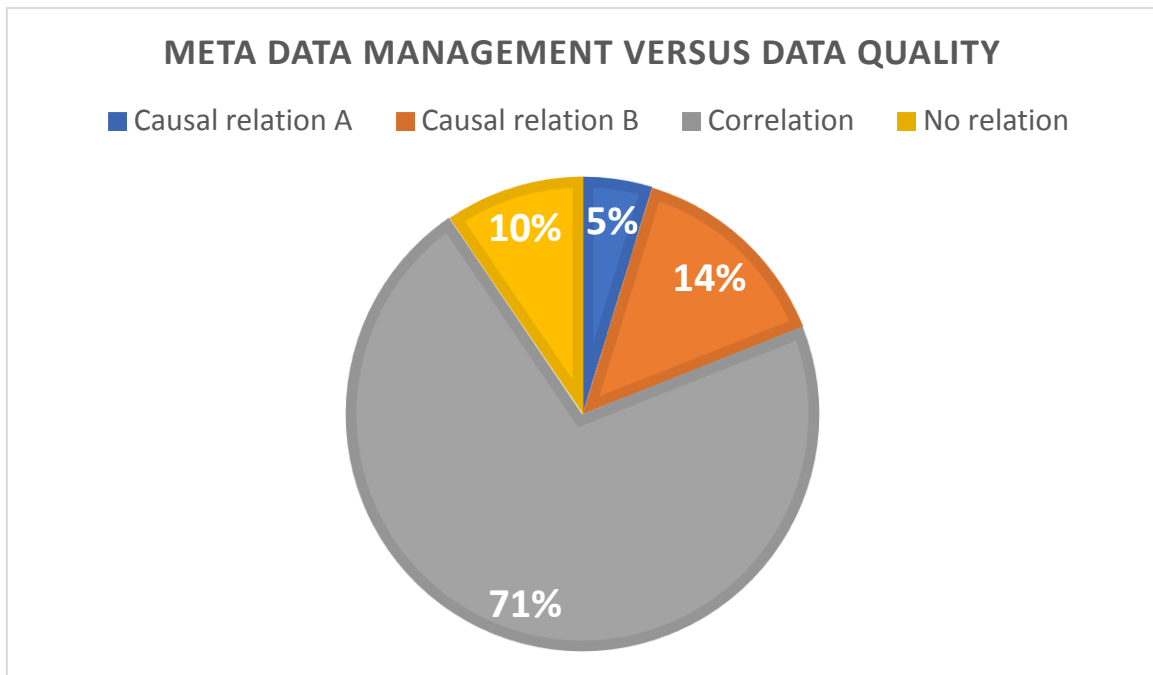
Additional Response:

Reusable – Ready to be used for future research and to be processed further using computational methods.

Question 6

How can the relation between meta data management and data quality be described (based on your personal experience in an organization/academia)?

ANTWOORDKEUZEN	REACTIES	
▼ Causal relation A: good meta data management ensures high data quality.	4,76%	1
▼ Causal relation B: the quality of data determines the status of meta data management.	14,29%	3
▼ Correlation: both influences each other.	57,14%	12
▼ All of the above answers.	14,29%	3
▼ No relation.	9,52%	2
Totaal		21



Additional Response:

Some aspects of DQ are affected by MDM, but not all.

Question 7

Do you make use of open data sources in your work/research? If yes, which sources?

Wikipedia, CBS, KNMI, Uniprot, ncbi, pdb, BAG, Weather data, Health data, UCI machine learning datasets: <http://archive.ics.uci.edu/ml/datasets.html?sort=nameUp&view=list> , Reference datasets like uniprot, Open vocabularies, ontologies, taxonomies, Company.info, KVK Webservices, Adress information LinkedIn, Phenotype database (dbnp.org), energy consumption.

Question 8

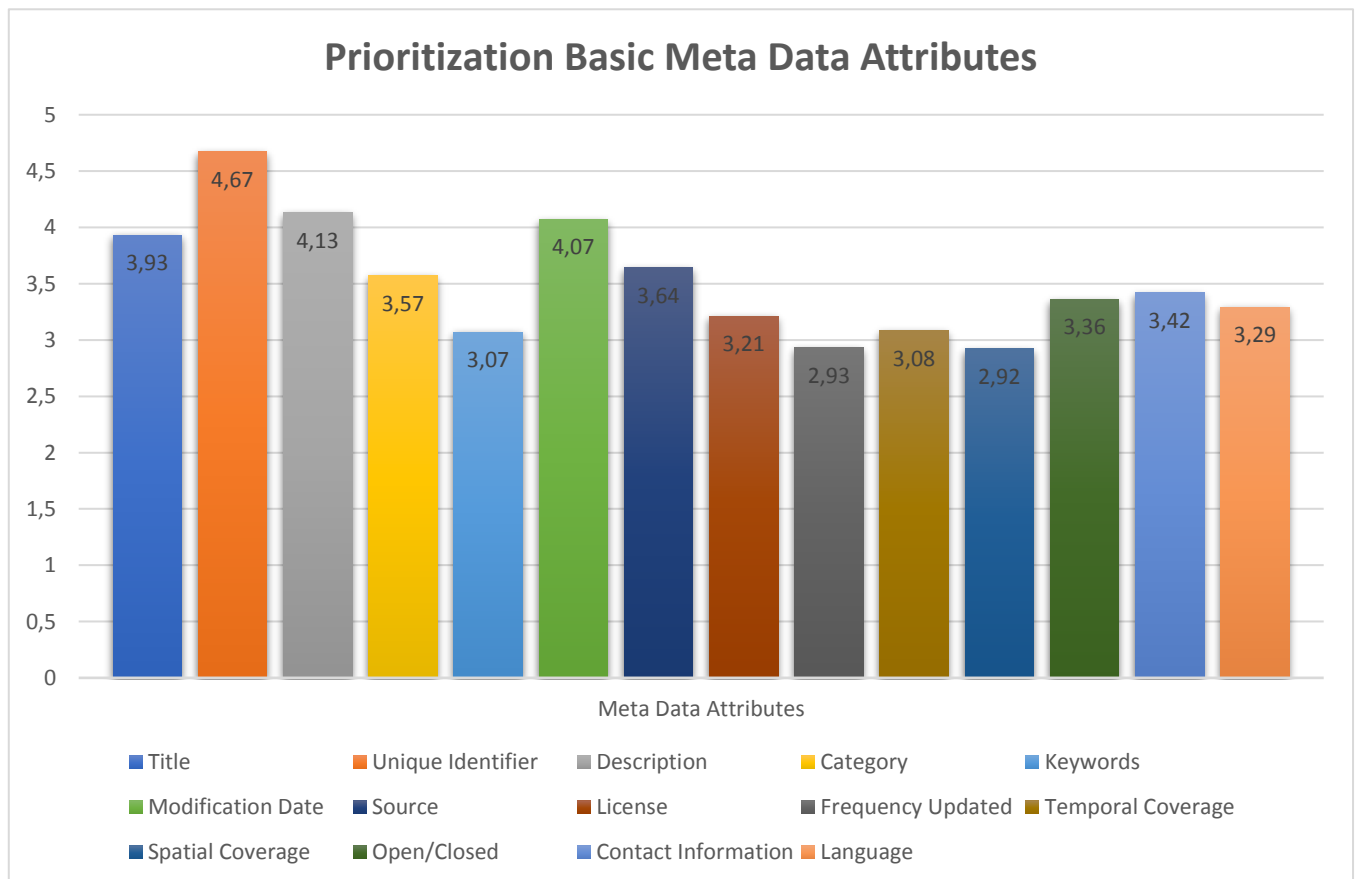
Are you in the possession of one or more data repositories? If yes, describe which attributes (column names) are in there.

- CRM database of AFAS Software BV 6pp Dutch postal code database (<https://www.pro6pp.nl>)
- CompanyId, ProjectId, Title, LedgerTransactionType, PostingType, CostType, AccountType, CostCenter, MainAccountId, MainAccount, Date, Description, Voucher, FundId, Fund, AmountAccountingCurrency
- Phenotype database (www.dbnp.org): too many to include here. Include a template structure to include all relevant meta data.

Question 9

What is the priority of the following (meta data) attributes in a data repository? From unnecessary (1) – indispensable (5).

▼ Title	0,00% 0	11,76% 2	11,76% 2	29,41% 5	29,41% 5	17,65% 3	17	3,93
▼ Unique Identifier	0,00% 0	5,56% 1	0,00% 0	11,11% 2	66,67% 12	16,67% 3	18	4,67
▼ Description	0,00% 0	0,00% 0	11,11% 2	50,00% 9	22,22% 4	16,67% 3	18	4,13
▼ Category	0,00% 0	5,88% 1	29,41% 5	41,18% 7	5,88% 1	17,65% 3	17	3,57
▼ Keywords	0,00% 0	22,22% 4	38,89% 7	5,56% 1	11,11% 2	22,22% 4	18	3,07
▼ Modification Date	0,00% 0	0,00% 0	16,67% 3	44,44% 8	22,22% 4	16,67% 3	18	4,07
▼ Source (URL)	0,00% 0	16,67% 3	5,56% 1	44,44% 8	11,11% 2	22,22% 4	18	3,64
▼ License	16,67% 3	5,56% 1	16,67% 3	22,22% 4	16,67% 3	22,22% 4	18	3,21
▼ Frequency updated	5,56% 1	27,78% 5	16,67% 3	33,33% 6	0,00% 0	16,67% 3	18	2,93
▼ Temporal Coverage	5,56% 1	5,56% 1	44,44% 8	11,11% 2	5,56% 1	27,78% 5	18	3,08
▼ Spatial Coverage	5,56% 1	5,56% 1	50,00% 9	11,11% 2	0,00% 0	27,78% 5	18	2,92
▼ Open/ Closed Data	5,88% 1	11,76% 2	17,65% 3	11,76% 2	17,65% 3	35,29% 6	17	3,36
▼ Contact Information	0,00% 0	23,53% 4	11,76% 2	17,65% 3	17,65% 3	29,41% 5	17	3,42
▼ Language	5,56% 1	5,56% 1	38,89% 7	16,67% 3	11,11% 2	22,22% 4	18	3,29



Appendix D: Interview Protocol

Interviewer: Jorien van Ginkel

Introduction

1. Can you tell something about the open data policy at your organization? What are the developments/innovations? What is the status in comparison with recent years?
2. Can you tell something about your role?

Theoretical Part

1. Do you know the term 'FAIR Data'?
2. Can you tell me something the status of data management within the organization? What are requirements? Goals?
3. What type of open data is in your repositories/ data portal? What is the information/meta data you collect about these sources?
4. What about the attributes in the repositories?
5. Is there any difference in data management for open data or for closed data? If yes, can you elaborate more on these differences.
6. Can you tell in your own words what you think is Findability/Accessibility/Interoperability/Reusability of data?

Definition 1: Findability (of data) is about the uniqueness and provenance of data to ensure data sources are searchable, free of redundancies and the context is clear, and this is achieved by versioning, associating (new) URLs or by embedding other kind of (provenance) identifiers.

Definition 2: Accessibility is a semantic quality dimension, which comprises availability, security, performance, interactivity, flexibility of data, to ensure data access to (end) users regardless their different context and background.

Definition 3: Interoperability is a system feature to connect and conform heterogeneous environments, to ensure sharing, reusing and exchanging of data between these without special effort from the (end) user.

Definition 4: Reusability is the ability to make easily use of data in a new context, in terms of retrieving, downloading, indexing, searching and visualizing the data without restrictions, as a result of findable, interoperable and accessible data.

7. Judge and discuss the above definitions.

8. Can you make a classification of the following attributes per activity? Are attributes missing/ unnecessary? Explain why.

Attribute	Value (0-5)
Unique Identifier	
Description	

Category	
Keywords	
Title	
Source	
Open/Closed	
License	
Contact Information	
Language	
Temporal coverage	
Spatial coverage	
Modification date	
Frequency updated	

9. Given the definitions and attributes... what would be your classification of attributes per concept?

Models

1. What about the visualizations/design/colors of the models?
2. What do you think it means?

<Explanation models>

3. Would it be relevant?
4. How can the model be improved?