An Assessment Framework for Research Data Reusability

Baharak Bakhtiari

Universiteit Utrecht

Master of Business Informatics 2018

Thesis date: March, 2019

Year: 2018
Title: Master thesis at the Utrecht University
Author: Baharak Bakhtiari

Acknowledgements

This is a place to give thanks to my supervisors Dr. Marco Spruit, Dr. A.L. Lamprecht, my daily supervisor Armel Lefebvre and the team of Tijmen Altena and Paul Tuinenburg in IDfuse B.V. whose incredible support and guidance helped me through this master thesis.

I would also like to give special thanks to my Mother and Father who have supported me during all my life decisions. All members of Mehrabi family and my supportive friends in Utrecht University.

Abstract

There are a growing amount of unstructured data produced by research communities. The high volume unstructured data, raise concerns about research data management. In the past few years funding agencies enforced open data policies and requirements for funded research, aimed to improve research data management.

The context of this study is Research Data Management Planning with the goal of improving research data reusability. Based on requirements of funding applications, researchers must follow certain procedures and guidelines when thinking about their data management and plan in advance. When researchers apply for a grant, there is a section in the proposal with a set of questions asking them how they are planning to manage their research data. In response, researchers need to be aware of the available solutions and ICT infrastructure that they need for their own research data management. The data section in the proposal is a set of questions based on FAIR principles in the Netherlands and the European Commission. FAIR principles are a set of guidelines used to produce machine-actionable/reusable data.

In this thesis, the data section in a proposal is selected as a case study to assess and improve the research data management planning of a research project. This thesis follows a Design Science Research approach and adapts the three cycles of relevance, rigor and design cycle [24]. These three cycles are implemented with the following actions: conducting a systematic literature review, building an initial Assessment Framework, conducting semi-structured interviews with experts, building and validating the final Assessment Framework, designing a user feedback tool based on the final framework and testing the accuracy of the design. The Assessment Framework improves research data reusability by using data quality indicators of reusable research data. The Assessment Framework is operationalised to the keywords which represent data reusability indicators in the RDMP deliverables. A set of 13 queries is designed, to measure 13 elements in the data management paragraph. Therefore, an experiment is conducted to run and test, and improve queries.

Data Management deliverables are human written text therefore in this study, an automated approach is proposed and implemented using an online text mining tool (the Impacter)[1]. In the automated approach, Natural Language Processing (NLP) techniques are also used in queries. The Impacter tool already produces user feedback on proposals. In this research, as well as keywords, NLP algorithms for the data management paragraph is designed, in order to measure reusability of the future data, and is explained in the data management paragraph section in the proposal.

[1] $https://$impacter.eu/

Contents

List of Figures

Chapter 1

Introduction

In data driven industries such as the financial industry, banking, logistics, manufacturing and health care the volume of data produced is growing rapidly. Scientific research is not an exception. Data produced in scientific research is high quality data and has immense potential to be used and reproduced in new research projects [60]. Research data, especially in funded research has to be reused and therefore data needs to be produced in a reusable manner. Research Data Management (RDM) is defined as organising data from entry into the research cycle to publication and archiving results [53]. RDM goes beyond publishing the results of the research, focusing on publishing the data set and raw data produced in the research cycle. The research data life cycle consists of creation, storage, security, preservation retrieval, and sharing and reuse, and these activities are associated with a need for technical, legal, ethical and governance solutions [35]. RDM brings solutions needed before, during and after the research data life cycle [35].

In the past few years, funding agencies have enforced an open data policy to promote research data sharing and raise awareness amongst researchers about the need for Research Data Management. Therefore researchers are required to follow the procedure of writing deliverables regarding the Research Data Management planning (RDMP) they have in mind when applying for funding. RDMP deliverables are for example the Data Management Plan and the Data Management paragraph submitted at different stages of their research project. In this study, planning, writing and submitting a RDMP deliverable is called RDMP. The document which is written and delivered in the context of RDMP is called a RDMP deliverable. RDMP deliverables contains a set of questions based on principles and guidelines used to produce reusable research data. FAIR principles are an example of these guidelines which help to produce machine-actionable or reusable data.

In the funding application researchers are more focused on their research topic and less attention is given to RDM and its planning. Many scientific studies on Research Data Management have focused on existing data sets and work flow [12, 17], or services [49] and policy implementation [56], with little attention given to the actual RDMP procedure from a researcher's point of view. Since writing a RDMP deliverable is the first step in starting RDMP in the research project it is important to think about it seriously rather than considering it as a formality in the paperwork. Serious thought about RDMP and planning based on potential solutions helps to circumvent challenges that may arise in the future of the project. Solving RDMP challenges in time, when the amount of data is already growing, means going back to manage data that was produced at the beginning of the project and it is indeed a tedious task. An automated approach to RDMP for measuring the quality of the RDMP

deliverable is needed in advance for a suitable RDMP. An automated approach can help researchers identify problems and their solutions regarding the RDM during the process of RDMP. This thesis proposes an assessment framework to provide help in writing the RDMP deliverables, which includes the design of an automated user feedback tool to provide support for researchers in their RDMP.

In this framework, Natural Language processing techniques (NLP) are applied to the RDMP deliverables to provide feedback on the RDMP and helps improve the RDMP where needed. The goal of this study is to make improvements on RDMP and eventually improve the reusability of research data in the project.

1.1 Structure of the thesis

The structure of this thesis is as follows:

- Chapter 1 provides an introduction, problem statement and the research method. In the problem statement the research questions are explained.
- Chapter 2 is an overview of actors and drivers involved in the RDMP.
- Chapter 3 contains the literature review and the proposed first design of the assessment framework for RDMP deliverables.
- Chapter 4 provides an empirical review based on the data from semi-structured interviews with reference to the roles of funding agencies and university libraries in the Netherlands. Thereafter the final design of the assessment framework is presented based on empirical findings combined with the results of Chapter 3.
- Chapter 5 explains the design of the user feedback tool with a report regarding the simulation conducted to validate the tool and measurement.
- Chapter 6 is the conclusion and discussion along with the limitations of this study.

1.2 Problem Statement

This thesis is a qualitative research project and the main research question is: *"How to improve reusability of research data by providing feedback on RDMP deliverables? "*

Funding agencies in the Netherlands have established a set of principles focused on extending the life cycle of research data and producing reusable data. Some of these guidelines are based on FAIR principles which is a set of guidelines used to produce *Findable*, *Accessible*, *Interoperable* and *Reusable* research data [56, 31].

FAIR guidelines are based on community standards and raises the question, are FAIR principles alone enough to assess research data reusability in the RMDP? This study aims to discover data quality indicators for research data reusability through a theoretical and empirical review, and represent all possible indicators in the proposed assessment framework. FAIR principles guidelines are also included and compared with real practices in the funding application.

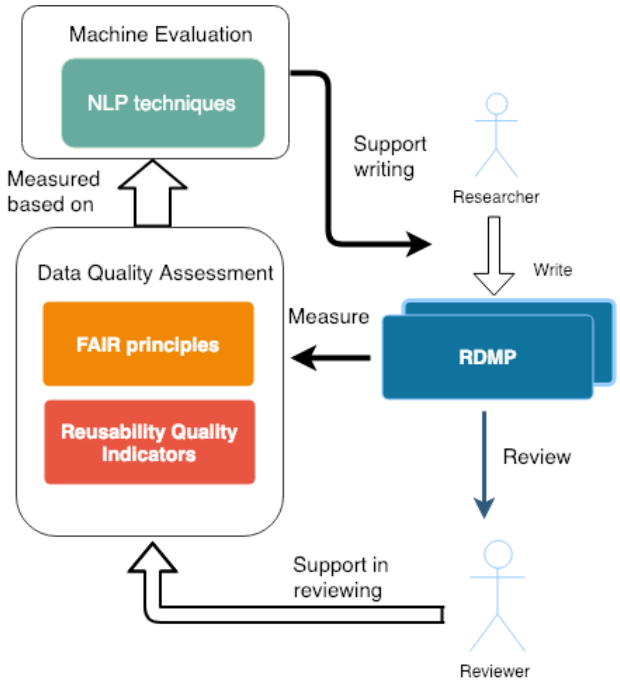


Figure 1.1: Conceptual Framework of this thesis

There is a lack of scalability in human capability to manage research data, since research data can be in high volume and with no structure and subsequently difficult to implement a suitable strategy. In this assessment, applying relevant NLP approaches is proposed due to the nature of RDMP.

The conceptual framework of this thesis is shown in Figure 1.3. In the conceptual framework concepts are associated with other concepts and actors. The Data Quality Assessment in this study is the data quality indicators of reusable data. Data quality indicators and FAIR principles are part of the ingredients of the assessment framework proposed in this thesis. To measure reusable data these concepts need to be clearly identified. Machine evaluation is using the quality assessment in an automated way to support a researcher in writing their RDMP deliverable.

The Data Quality assessments are the reusability metrics identified through the literature review in Chapter 3 and also in the empirical review in Chapter 4. The researcher and the reviewer are actors in this framework. Research in the university provides input for their RDMP and reviewers in the funding agency evaluate the RDMP deliverable based on the Data Quality Assessment.

1.2.1 Definition of Data in the RDMP

Data in the RDMP is considered ”a digital object” and it can be accessed in open sources [9]. Meta data is the information about the origin of data and the collection procedure, and the computation and tools associated with the data life cycle. In other words, meta data defines data in a machine readable format [40]. Research Data in the context of this project means elements of meta data and data itself.

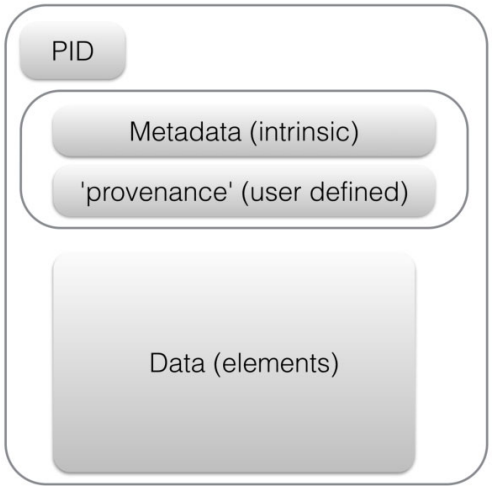


Figure 1.2: FAIR digital Object
[2]

1.3 Data Quality Assessment

1.3.1 FAIR Principles

The FAIR principles are a newly developed set of community standards established in a workshop called "Designing a Data Fairport", which took place in the Netherlands in 2014 [58]. The standards are based on private and academic stakeholders' interests 2.2.1. The main goal of FAIR principles is to improve RDMP by enhancing the quality of research data [58]. These measures were established to make research data both reusable and machine-readable [58] meaning publications and data sets that researchers collect during their research will be in a homogeneous format, findable and accessible to other researchers.

FAIR principles stands for Findable, Accessible, Interoperable and Reusable data (FAIR) [1]. The element of machine readable data, according to FAIR principles, are shown in Figure 1.2. Elements of data in the Figure 1.2 are explained with their related principles below:

To be Findable Findable data means data and metadata should be discover-able to humans and should be machine-readable. Metadata makes data Findable in repositories and the public domain [1]. A data identifier is a sequence of characters for an entity and a Persistent Identifier is used to reach data as "a digital object" in the data source through the Internet [21]. The unique and persistent identifiers called PID can be seen in Figure 1.2. A very common example of PID is the unique web link (for example $https://$) as an online metadata that directs users to the source of data via a browser [21]. In the data it is also called the Digital Object Identifier or the doi code.

To be Accessible Accessible means research data needs to be accessed through specified authorization [3] A standard protocol defines how data can be accessed and to whom they are accessible in a machine readable format [1]. The access protocol may affect the choice of data archiving as well [1]. Examples of protocols can be found here : W3.org [4].

To be Interoperable Interoperable is related to scientific work flow and data archiving in order to make data integrated to other data sets. As has been mentioned in Findable data,

[1] $https://$www.go-fair.org-accessed Sept 2018
[3] $https$://www.force11.org-Accessed Jan 2019
[4] $https://$www.w3.org/Protocols/-Accessed Sept 2018

a unique and persistent identifier is the element to make data Findable among other data sets and Interoperable is about the integration of Findable metadata with other Findable metadata [1].

To be Reusable Reusable data needs to have clear work flows of authors and should provide information and access protocols; other data sets that are used to produce the current data and conditions for reuse such as software that supports the data format[1]. Reusable data already contains elements of Findable, Accessible and Interoperable.

1.3.2 RDMP deliverables

Researchers need to fill in a Data Management section in their proposal. The European Commission calls (for example in Horizon 2020), this the Data Approach. [5]. In H2020 the Data Approach section is under the Dissemination and Exploitation section. In the NWO calls, the RDMP deliverable is a section called Data Management which includes 4 questions about reusable data. Data Management Paragraphs in the proposal needs to be filled in when applying for a grant.

The Data Management Plan is a set of questions about the RDMP which needs to be filled in by researchers when they receive their grant. H2020 and NWO both require RDMP deliverables and the questions are based on the FAIR principles.

In RDMP deliverables researchers explain their data sets and its elements. They clarify how their data will meet reusability criteria and what challenges they may have. Therefore RDMP deliverables is a suitable tool to capture their RDMP and acknowledge the reusability of their research data.

1.3.3 Reusability Quality Indicators

FAIR principles are new guidelines. Even though RDMP deliverables are based on FAIR principles, the RDMP can vary from the FAIR guidelines. The assessment framework proposed in this thesis identifies the reusability quality indicators in the context of RDMP and finds the gap between reusable data in reality and what is proposed by FAIR principles. In the proposed framework in this study, quality indicators related to reusable data are not considered only based on the requirements of FAIR but also how data can be reusable in the literature regardless of FAIR principles. Quality indicators are validated with the reviewers during interviews.

In Chapter 3, a systematic literature review is conducted to identify data reusability metrics and their relation to FAIR principles. There is an overlap between these two notions, reusable data according to the literature regardless of FAIR principles and the FAIR principles. Also there is a gap between data reusability measures in theory and in practice. Therefore there is a need for an assessment framework which focuses on the FAIR principles, reusable data quality indicators and real practice to measure the reusability of data.

[5] *http*://ec.europa.eu/research/participants/docs/h2020-funding-guide/grants/grant-management/dissemination-of-results -Accessed Dec 2019

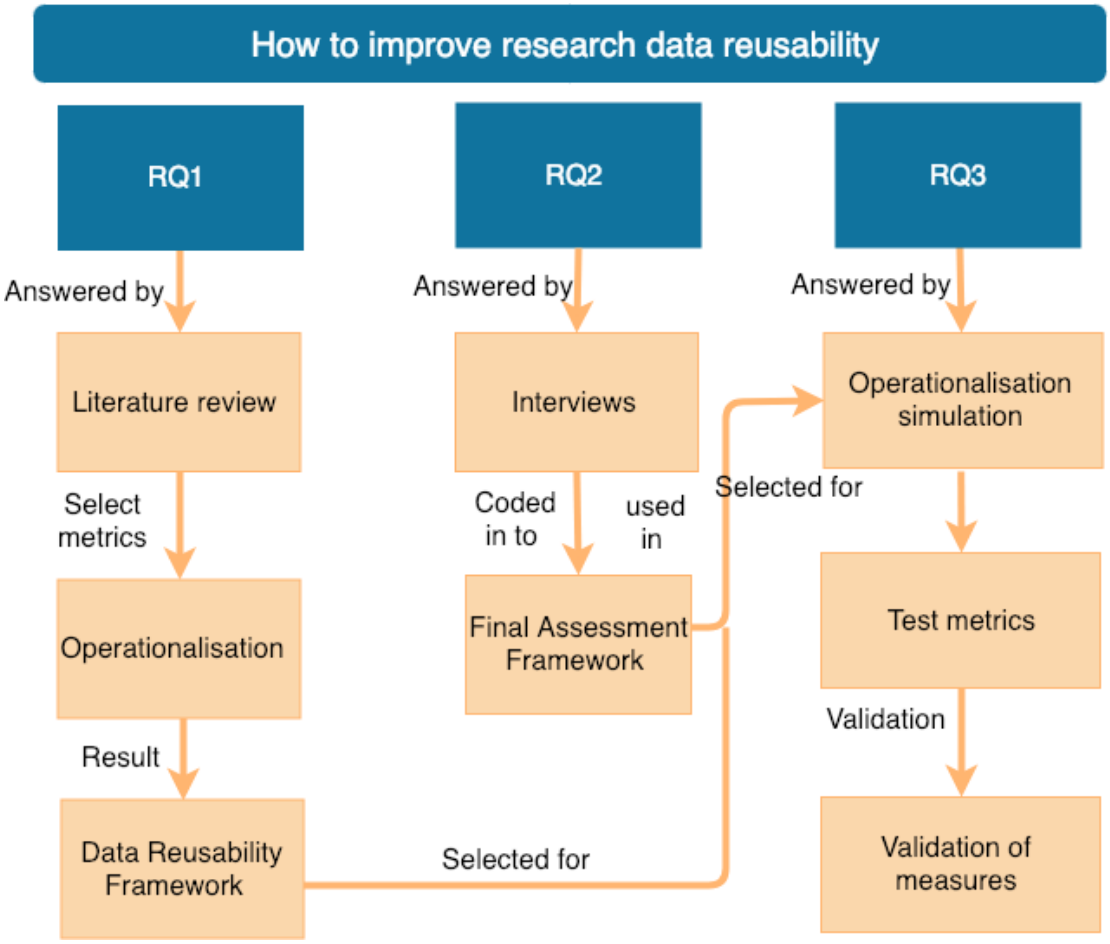


Figure 1.3: Structure of Answering RQs in the Thesis

1.4 Research Questions

The main research question of this thesis is *How to improve reusability of research data by providing feedback on RDMP deliverables?* and this question is divided into three sub-questions:

***RQ1**:What are the measures for data reusability?*

Rationale: There is a need for a homogeneous structure for data to improve data reusability. This question is answered based on current studies about research data reusability. Measuring research data reusability depends on the context of the data. In this question, quality indicators for reusable data are identified through a literature review. The literature about the quality indicators of reusable data and FAIR principles are investigated to answer this question. After finding reusability quality indicators, techniques to measure them are identified in the literature.

***RQ2**: How are RDMP deliverables reviewed by funding agencies?*

Rationale: This question is answered with a case study. The case study is the RDMP deliverables such as the Data Management paragraph (DM Paragraph) and Data Management Plan (DMP). In order to measure data reusability in RDMP deliverables, interviews are conducted with reviewers of RDMP deliverables. This case study is an observational case study. Therefore it is necessary to ask reviewers what they consider as reusability quality indicators.

This question is answered in Chapter 4.

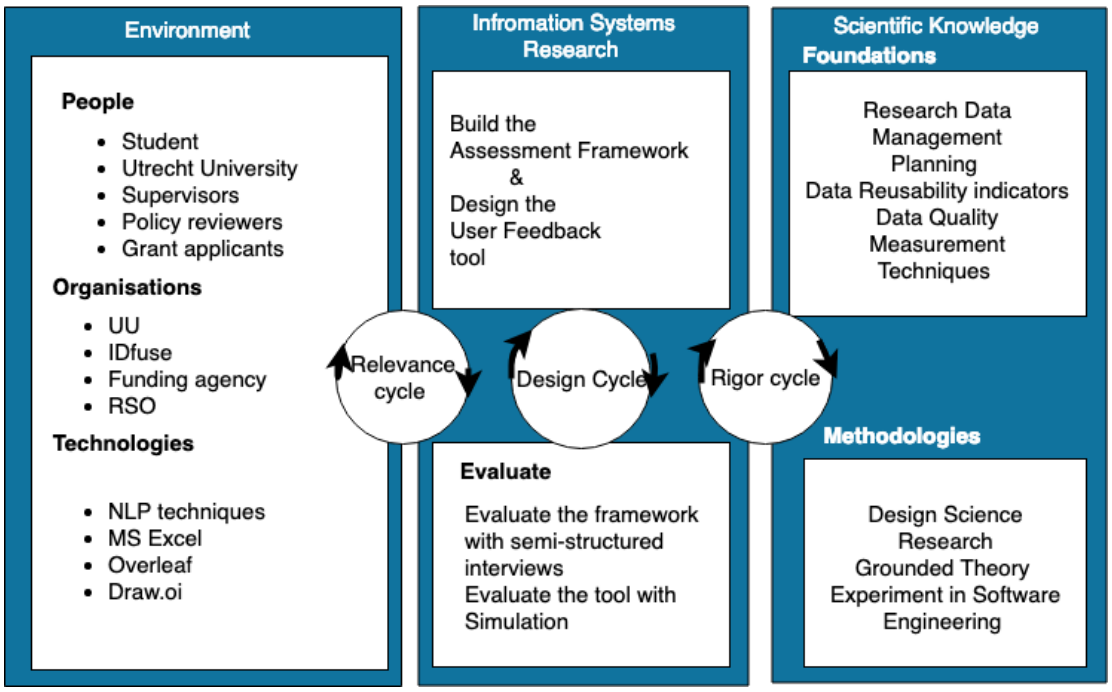


Figure 1.4: Information Systems Research Framework of own study based on Design science [24]

***RQ3**: How to design an artifact that generates feedback on RDMP deliverables to improve research data reusability?*

Rationale: The goal is to simulate the reviewers way of thinking in an automated approach to improve quality of RDMP deliverables. Quality indicators related to reusability of research data needs to be investigated in real practice.

The artifact is the tool Impacter which is based on various NLP techniques and the goal is to find out if data reusability in the data section paragraph can be measured. The Data Management Paragraph is a short text data and to measure the quality indicators, text mining tools are used. A large set of algorithms exist to extract data qualities from the text [2]. In the Figure 1.3 the structure of this thesis and how the research questions are answered are shown.

1.5 Research Design

1.5.1 Research Method

The research method in this thesis is Design Science by Hevner [24]. Design Science is a problem solving paradigm in Information Systems research suitable to designing, creating and evaluating IT artifacts to solve business and organisational issues [24]. In this thesis two IT artifacts are built and evaluated, one is the assessment framework validated through expert interviews and a tool which is evaluated by a simulation. The Design Science method has three integrated cycles which are described below 1.4.

1.5.2 Relevance Cycle

The relevance cycle is the iteration between the contextual environment and user requirements, business needs and challenges for the researcher as a user [24]. The iteration of the relevance cycle is between the designs in this thesis and environment. The roles involved in this cycle are student and supervisors, funding agencies, and university libraries. IDufse B.V. is where the tool Impacter is used and where an internship was completed, which is also involved in this cycle. In this cycle an observational case study is conducted by investigating RDMP deliverables.

The research data reusability indicators are defined from the reviewers point of view and was gathered during interviews. In conclusion, the result of the relevance cycle is creating the assessment framework and user feedback tool for the business environment and research context [24]. In this study, one iteration is by conducting interviews and applying real world context requirements of RDMP deliverables into the assessment framework [24].

1.5.3 Rigor Cycle

The rigor cycle concerns the knowledge based methodologies of the research in the state-of-art and the current artifact, and processes to have the research contribute to IS research [24]. A systematic literature review is conducted in the rigor cycle to investigate current solutions and the research gap. In the iteration of the rigor cycle an initial assessment framework based on the literature review is designed.

1.5.4 Design Cycle

The iteration of designing and refining IT artifacts in this research is conducted in the Design cycle [24]. In this thesis two IT artifacts are designed and for each artifact, iterations of the design cycle through building and evaluating these IT artifacts are completed [54]. The first iteration is done by designing the assessment framework and the second using the assessment framework to design the user feedback tool.

1.5.5 Design Proposition

There are two types of outcomes in this thesis, the first outcome is the assessment framework for research data reusability as the assessment method. The second outcome is the design of the user feedback tool to improve research data reusability on the RDMP deliverables based on the proposed assessment method.

1.5.6 The Validation of Artifacts in this thesis

Evaluation of the assessment framework is done during expert interviews. The evaluation of the user feedback tool is done by running a simulation (an experiment). In the simulation the proof of concept as a user feedback tool is evaluated and its measurements tested.

Chapter 2

Environment

2.1 Stakeholders of this Study

In this section the stakeholders of this thesis in the context of RDMP are explained. The different insights about RDMP from the stakeholders perspective is identified and compared. The external stakeholders are a broad range of funding agencies based all around the world and it is not feasible in the scope of this project to include them in the interview sample.

2.1.1 University and Supervisors

Stakeholders of this study are from the Applied Data Science Lab at Utrecht University. Supervisors of this thesis Dr. M.R. (Marco) Spruit, Dr. A.L. (Anna-Lena) Lamprecht and daily Supervisor, A.E.J. (Armel) Lefebvre and the author of this thesis as a student, are internal stakeholders of this thesis.

2.1.2 IDfuse and Impacter Tool

IDfuse B.V. is the other internal stakeholder of this thesis. An internship was completed at IDfuse B.V. in Europalaan, Utrecht, the Netherlands. This startup was founded in 2012 by Tijmen Altena and Paul Tuinenburg. Altena and Tuinenburg are graduates of Utrecht University. In addition, Martijn van Beers is the data scientist and is responsible for technical support. The core business in IDfuse B.V. is their tool called Impacter, which assists scientists in writing grant proposals before the final submission. Stakeholders of Impacter are researchers and research support service providers in universities.

Impacter is an online text mining tool and it is tailored to detect requirements of Knowledge Utilization paragraphs in grant proposals. Impacter provides automated feedback on grant proposals uploaded to its platform. Impacter uses NLP techniques in Python programming language. The functionality of Impacter is highly focused on the requirement of the Knowledge Utilization, which is required by NWO and the European Commission. The final outcome of this thesis is expected to add value to the tool as well as make a scientific contribution in the RDMP context.

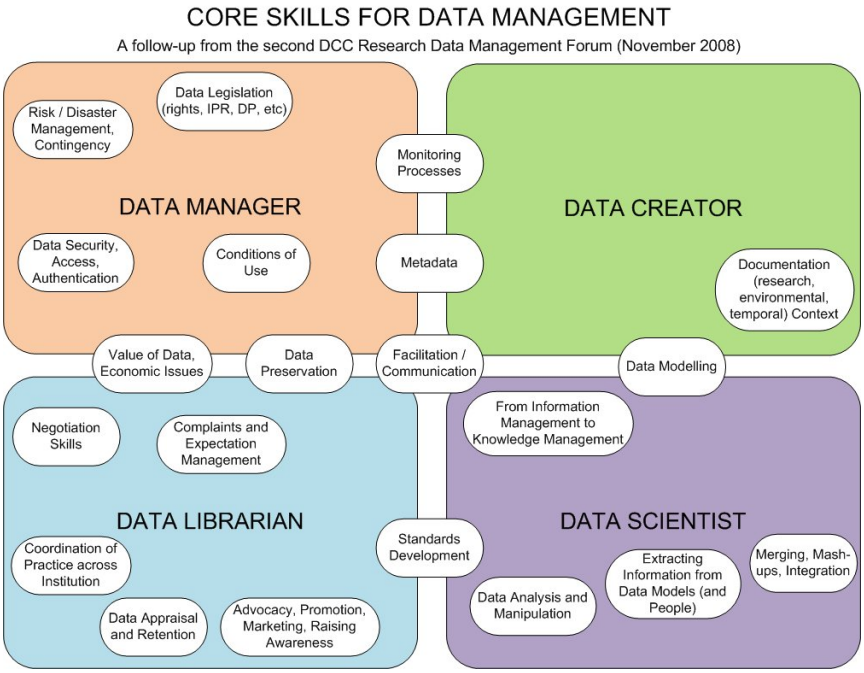


Figure 2.1: Core skills for Data Management
[37]

2.2 External Stakeholders And Environment

In this section roles involved in RDMP are explained. These roles are interlinked through the RDMP tasks.

2.2.1 Stakeholders that are involved in the Case Study

There are fours main roles involved in accomplishing effective Data Management, these are shown in the Figure 2.1 [37]. These roles are defined in the Data Management context in the DCC Research Data Management Forum as existing data related roles [37]. Therefore, they are adapted in the RDMP context [1] and the case study of RDMP deliverables.

Data creators are researchers who produce data in the context of RDMP, they are researchers in this thesis. Data scientists are the one who work in collaboration with data creators/authors and they share the role of managing ICT facilities for storage, access and, database operation and maintenance. In this research the data scientist role is handled by IDfuse B.V. with the tool Impacter and UU and Applied Data Science Lab as supervisors of this thesis and the student. 2.2.

Data librarians are the knowledge workers in the libraries who tend to manage storage based on data policy requirements. Data librarians in thesis are the Research Support Office and Data Management Department in universities.

2.2.2 Funding Agency

Funding agencies are responsible for developing data policies and standards, and thus, are the data managers in the RDMP context. The implementation and documentation of RDMP

[1] *https*://www.era.lib.ed.ac.uk/handle/1842/863-Accessed Jan 2019

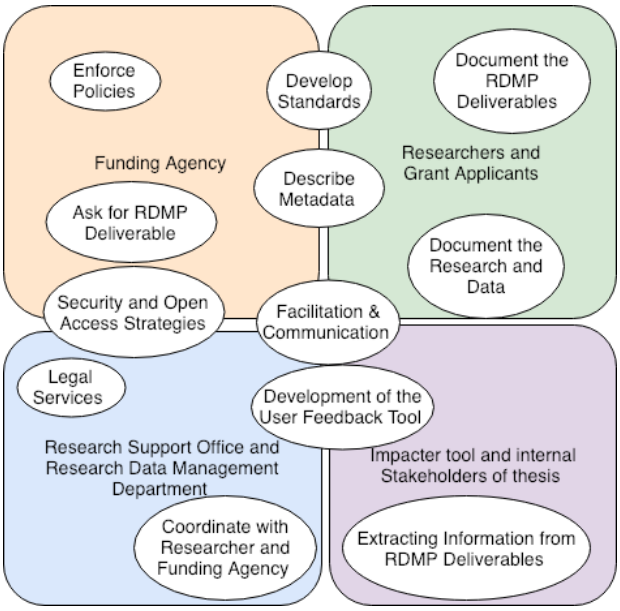


Figure 2.2: Core skills for Data Management adapted to RDMP deliverables and the thesis [37]

deliverables differ according to funding agencies and their locations. The goal of RDMP deliverables as a requirement is to improve research data reusability and promote data sharing in the research community. In table 2.1, RDMP requirements are compared in different funding bodies. The access time in the Table 2.1 is how long the funding agency requires researchers to make their data available after completion of the research project.

The European Commission launched the Open Research Data (ORD) pilot plan in 2014-2016. The EC dedicated 20 percent of its research funding budget to the pilot plan and evaluation of RDMP deliverables [15, 45]. RDMP deliverables in ORD is based on FAIR principles and researchers have the voluntary option to submit their RDMP deliverable for evaluation or not. [2]

UK funding agencies strictly consider funded researcher data as public goods which should be openly available with a few limitations. Researchers are able to use the public funding for their RDMP and Data sharing[3]. The Arts and Humanities Research Council (AHRC) requires a technical plan (will be called Data Management Plan from March 2019) [4] and summary of technical support, address preservation, sustainability from applicant and it should be accessible for at least three years after the end of the grant [45]. In the UK funding organizations like Cancer Research UK, MRC and BBSRC, deliverables are an integrated part of proposals however, RDMP requirements have different levels of details within different organizations.

Also, major funding agencies and the National Science Foundation (NSF) in the United States, follow a different approach and researchers are still obliged to submit RDMP deliverables in their proposal.[5]

[2] *http*://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management-en.htm-Accessed Jan 2019

[3] *https* : //www.ukri.org/funding/information-for-award-holders/data-policy/common-principles-on-data-policy-accessed Aug 2018

[4] *https* : //ahrc.ukri.org/documents/data/about-the-new-data-management-plan/-Accessed Jan 2019

[5] *https*://www.nsf.gov/eng/general/dmp.jsp-accessed Aug 2018

Agency	Area	Requires RDMP	Monitoring	Access time
NWO	Netherlands	Yes	Yes	10 years after the project
ESRC	UK	Yes	Yes	In Publication time
AHRC	UK	Yes	Yes	Min 3 years after
NSF	US	Yes	Yes	Within reasonable time
NIH	US	500K a year	Yes	In publication time
Horizon 2020	EU	Pilot	Yes	Not specified
ARC	Australia	No	No	Not specified

Table 2.1: A brief comparison in the use of RDMP[45]

In Australia, the Code for the Responsible Conduct of Research was launched in 2007 by the Australian Research Council, National Health Council and universities.[6]. Australia has large investments in RDM infrastructure, and as such the Australian National Data Service (ANDS) has become a leader in RDMP. [45]. RDMP polices in Australia have minor differences compared to other regions. They still place emphasis on the open access to research data by promoting the advancement of knowledge for peer review, decrease duplication and increase research benefits and citations, but do not require RDMP documented as a compulsory deliverable[45].

In conclusion, the difference between RDMP deliverables within different funding agencies are shown in Table 2.1. RDMP deliverables is requested by the UK, EU and the US. However, no unique RDMP deliverable template exists for them. The access time to the research data also differs, but not by more than three years access after completing the project.

In the RDMP context, stakeholders are involved in different stages of the research life cycle and based on their roles they have different responsibilities related to RDMP. The main responsibility of stakeholders is to follow data policy and guidelines to improve RDMP with a longer research data life cycle than only the research project. In Chapter 4 roles and responsibilities related to RDMP deliverables are presented using evidence gathered from interviewing stakeholders.

[6] $http://$www.ands.org.au/guides/code-awareness-accessed Aug 2018

Chapter 3

What are measures for data reusability?

The RQ1 is *"What are the measures for data reusability?"*. This question is answered by conducting a systematic literature review. The literature review is conducted based on three goals:

1. To identify quality indicators for reusable research data.
2. To find the research gap between FAIR principles and the quality indicators for reusable data.
3. To select the NLP approach suitable to measures selected quality indicators.

The systematic literature search, is conducted by using three search engines DLBP, Scopus and Google Scholar. All papers between the years 2010 to 2018 are included. A search based on queries related to the third goal within the time frame 2016-2018 are also included. The literature search and its details are mentioned in the literature review protocol in Appendix A. In the first search 180 papers were found based on the content of their abstracts and conclusions. In the second search, an in-depth reading of 77 papers was chosen as relevant to the goal. 4 papers related to Data Management and reusability of data were chosen. 25 papers were chosen in the search for FAIR principles keywords and 48 papers were chosen related to NLP techniques. All literature is stored and shared in the Mendeley repository. In the conclusion of findings of the systematic literature review, the initial assessment framework is designed based on the quality indicators found for reusable data.

3.1 Quality Indicators for Reusable Data

In the literature, the trend found for quality indicators of reusable data, identified two main indicators: metadata and data provenance.

3.1.1 Provenance Quality

Provenance is the process and work flow documented and associated with data, it can be captured by editing metadata [8]. Provenance is not only found in the final product of data but it is also in the process of creation of data [8]. A complete and clear provenance can produce high quality data and allow data users to trace back origins of data [18]. Provenance is found in eight references [8, 3, 47, 5, 18, 11, 3, 6] in the context of measuring data qualities

by using provenance and to measure data reusability by conducting the measurement on the model of provenance.

Also literature data quality measures are applied based on provenance. For example, Reproducibility characteristics is the special case of repeatability where the third party can replicate or reproduce the data and method [3]. Repeatable means that data contains services or processes and the execution order in other words details of its provenance quality [3]. Reproducibility is a metric related to the process of execution and integration of data sources with their metadata [6]. In other words, provenance is needed to simplify the data exploration and comparison of data in the digital ecosystem [11].

In the context of RDMP deliverables provenance quality will be measured with the same elements as metadata completeness. Since it is the planning phase of collecting the data, the details of provenance do not exist yet. Therefore mentioning potential ways of collecting data and clarifying owners and users is part of the provenance in the RDMP context.

3.1.2 Metadata Completeness

In the literature metadata is used as data quality indicators for reusable data [23, 21, 30, 44, 48, 25, 40, 34]. Metadata completeness is the main reusability factor for RDMP deliverables because metadata can still be available, even when the data set no longer exists [1]. Provenance quality and metadata completeness are interrelated information about data but with different levels of denials. Provenance completeness in this study has a close definition to reproduciblity of data [3].

RDMP deliverables are prepared prior to data collection, therefore planning to document and capture complete metadata and detailed provenance is an indicator for reusable data. In the initial framework 3.2 provenance quality and metadata completeness have a lot of overlap and for planning reusable data they are both acceptable based on the level of details they contain.

Provenance and metadata can be collected from various sources with different levels of details called granularity [46]. The level of details effect the usefulness of metadata and provenance in the abstract level of the data set [46]. In the initial assessment framework 3.2 the granularity of complete metadata based on the RDMP deliverables are named in the lowest level of the initial Assessment Framework.

Metadata granularity (Figure3.2) are authors and citation, data user, Persistent Identifier, and licenses name. The presence of elements in RDMP deliverables is considered to measure the completeness of metadata.

3.1.3 FAIR Principles and Data Quality Indicators for Reusable Data

There is overlap between FAIR metrics and, metadata completeness and provenance quality. [1] Provenance is also mentioned in the FAIR principles where ”the association of metadata with detailed provenance” is a reusability indicator [3]. Data availability indicators are the presence and absence of availability statements and, reasons and type of repository [22]. There are not many scientific studies that specifically apply FAIR principles as metrics for reusable data. This is not surprising since FAIR principles were established only in 2016. The recent trend has been that FAIRness of data is focused on repository attributes and infrastructure [57, 43].

[1] $https://$www.go-fair.org/fair-principles/-Accessed Oct 2018

The overlap of FAIR principles with provenance quality is emphasized in the existence of metadata and the association of metadata with its provenance [2]. FAIR principles are more focused on data archive and metadata, or machine-readable data.

3.2 Reusability Metrics in the Initial Assessment Framework

Data reusability indicators in RDMP deliverables are selected based on three indicators, namely FAIR principles, provenance quality and metadata completeness. In the Framework, metadata completeness represents elements from three quality indicators selected from the literature review 3.2. Reusablility indicators in detailed levels are measured based on the checklist of data elements required in RDMP deliverables. These elements are shaped based on indicators applicable to RDMP deliverables.

Metadata completeness in the RDMP deliverable is a Persistent Identifier (metadata), FAIR repositories (based on FAIR principles), citation, authors and data user (provenance) and licenses (FAIR principles) for reuse. According to Force 11 community data is equal to the data digital object + metadata + an Identifier. The reusability characteristic of data is based on the presence and level granularity of these elements in data as a digital object [31].

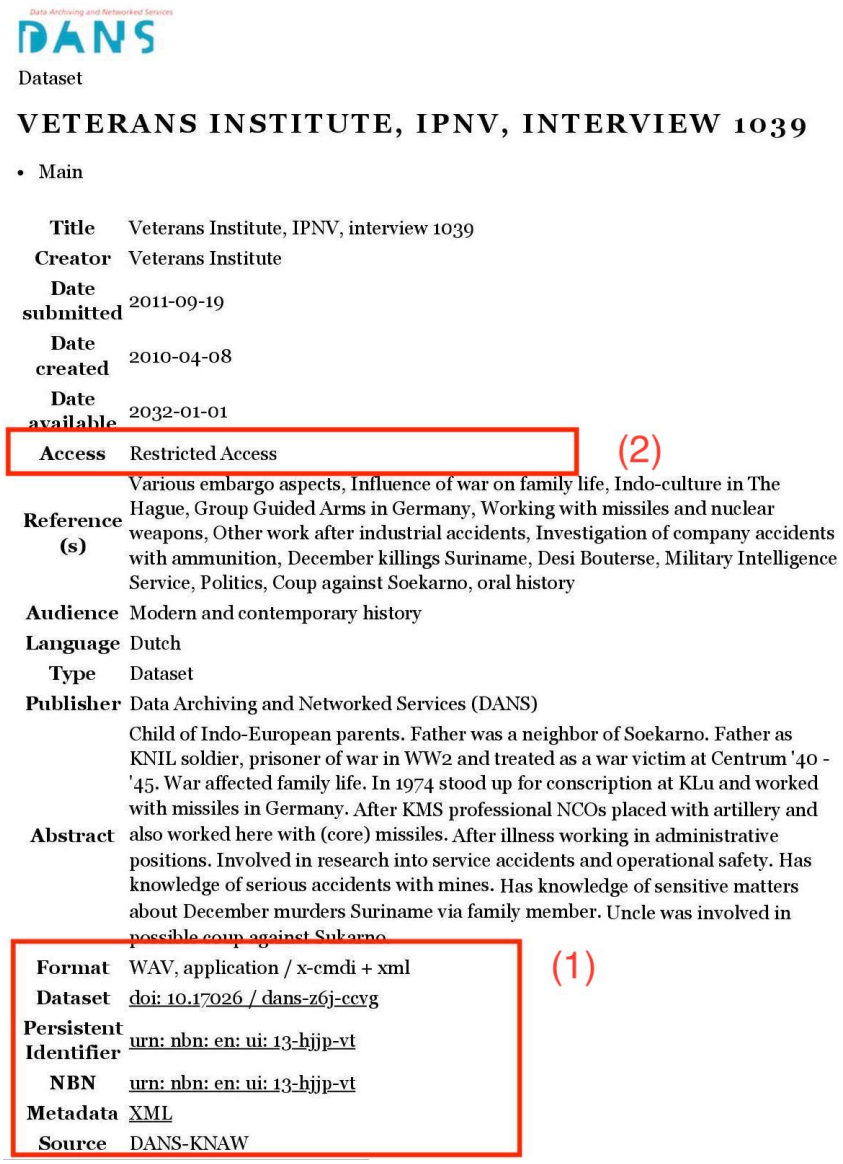
Metadata completeness values are considered complete at the presence of its elements according to FAIR principles. An example of these elements can be found in the Dublin Core Metadata Initiative (DCMI) where fifteen properties of metadata and their terms are described [2]. The DCMI originates from a workshop which took place in Dublin, Ohio and the properties are generic for different resources [3]. It is not expected that researchers provide all 15 properties in the RDMP, this depends on the time they have in preparing the RDMP deliverable. The number of elements that exist in the RDMP deliverables are based on questions in the template of deliverables as well. An example of complete metadata elements are shown in Figure 3.1. Information is directly related to the current data set stored in the repository.

3.2.1 Data Format and Persistent Identifier

The presence or absence of elements for reusability are considered as binary values. Keywords indicated in the example of metadata elements are shown in the Figure 3.1 in the number (1) marked section: Data format, Persistent Identifier is `urn:nbn:en:ui:13-bjjp-vt`, data set is doi, repository is the DANS-KNAW and NBN (The National Biodiversity Network metadata) are elements of data reusability and they are expected in RDMP deliverables. There are many standards and identifiers that can be used for the data set, in this example doi and data format both have a positive effect on data reusability and accessibility. To measure this effect score [1,0] each factor is considered and factors related to metadata have to be sufficient to make the metadata complete. Considering only metadata (1) in this example, metadata: XML[3] as a reusable format. In the example, reusable data format is present and PID are present. The granularity of metadata requires five elements to conclude data is

[2] http://dublincore.org/documents/dces/-Accessed October 2018

[3] XML(eXtensible Markup Language). The XML documents contain entities with parsed and unparsed data and the XML design supports many applications that can be used in different codes (like Python) [61]



Dataset

VETERANS INSTITUTE, IPNV, INTERVIEW 1039

- Main

Title	Veterans Institute, IPNV, interview 1039
Creator	Veterans Institute
Date submitted	2011-09-19
Date created	2010-04-08
Date available	2032-01-01
Access	Restricted Access
Reference (s)	Various embargo aspects, Influence of war on family life, Indo-culture in The Hague, Group Guided Arms in Germany, Working with missiles and nuclear weapons, Other work after industrial accidents, Investigation of company accidents with ammunition, December killings Suriname, Desi Bouterse, Military Intelligence Service, Politics, Coup against Soekarno, oral history
Audience	Modern and contemporary history
Language	Dutch
Type	Dataset
Publisher	Data Archiving and Networked Services (DANS)
Abstract	Child of Indo-European parents. Father was a neighbor of Soekarno. Father as KNIL soldier, prisoner of war in WW2 and treated as a war victim at Centrum '40 - '45. War affected family life. In 1974 stood up for conscription at KLu and worked with missiles in Germany. After KMS professional NCOs placed with artillery and also worked here with (core) missiles. After illness working in administrative positions. Involved in research into service accidents and operational safety. Has knowledge of serious accidents with mines. Has knowledge of sensitive matters about December murders Suriname via family member. Uncle was involved in possible coup against Sukarno
Format	WAV, application / x-cmdi + xml
Dataset	doi: 10.17026 / dans-z6j-ccvg
Persistent Identifier	urn: nbn: en: ui: 13-hjjp-vt
NBN	urn: nbn: en: ui: 13-hjjp-vt
Metadata	XML
Source	DANS-KNAW

Figure 3.1: Metadata Completeness Example

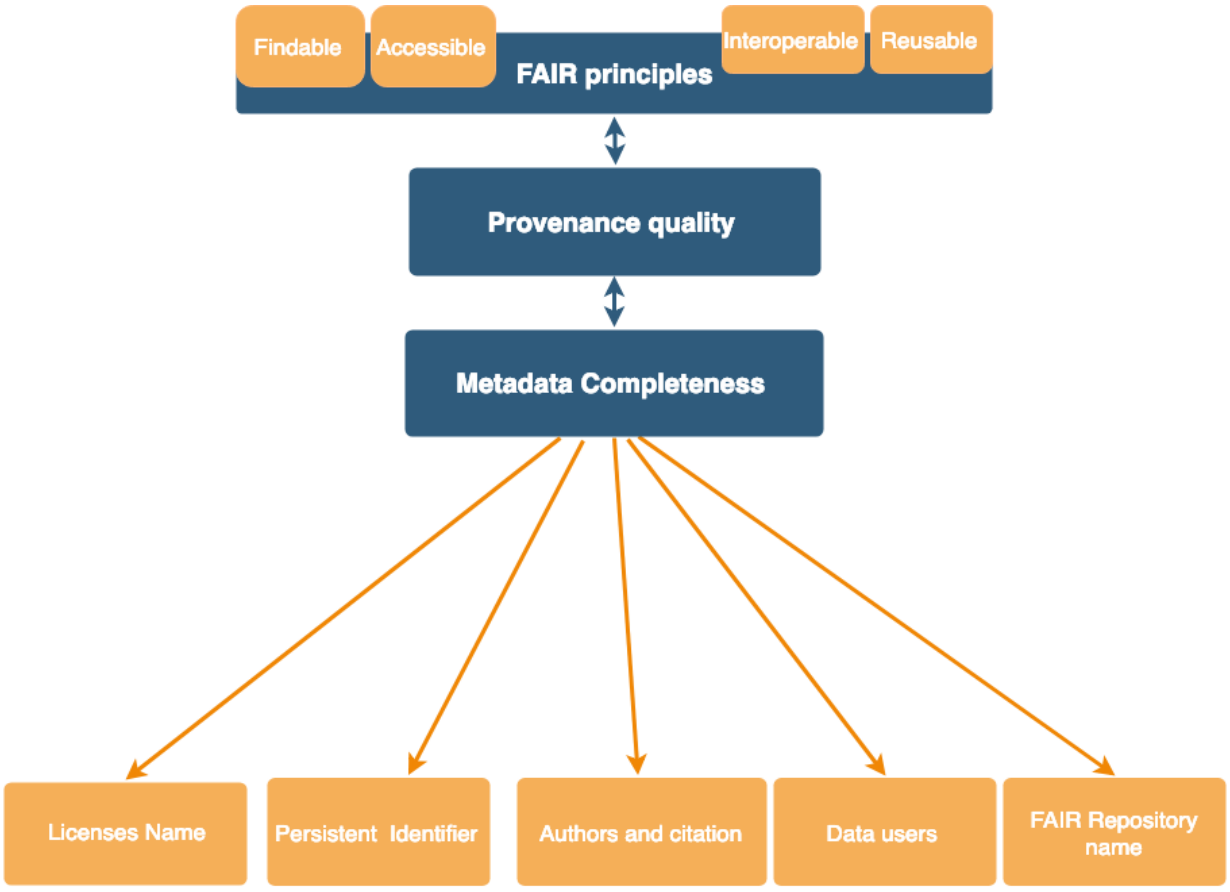


Figure 3.2: Operationalisation of Reusable Data in RDMP Deliverable

reusable in RDMP deliverables 3.2. The example of elements in a repository 3.1 and in a RDMP deliverable 3.1 are shown.

3.2.2 Authors and Citation

Name of authors and the citation is essential in the RDMP deliverable and if the writer is not who is in charge of RDMP in the project they will be mentioned. It goes without saying that the author's name and the type of citation are part of scientific work flows and the data generation process. The presence of data producers and the type of citation in the RDMP are also considered as a binary value.

3.2.3 Data User

The presence of data users is a factor in metadata completeness, shown in Figure 3.2 and measured with binary values [0 or 1] for presence or absence in the RDMP deliverable. Data will be reusable when the third party and data users are clearly identified. Authors mention that those who can use this data translates to the researchers' planning to generate reusable data.

DQ	Elements	Metrics	Example in a text
Metadata Completeness	Data format	Binary	Dataset is XML format.
	Third party and Data users	Binary	The dataset is usable by chemistry researchers
	Standard Identifier as description of metadata	Binary	The DOI code data has persistent identifier, URN.
	Vocabulary of FAIR principles and repository	Binary	Data will be accessible in DANS repository.
	Licenses and terms of use	Binary	Data license, CC -Y required for reuse.

Table 3.1: Definition of Data Reusability Indicators in RDMP deliverables

3.2.4 FAIR Repository

The [0,1] score for a FAIR repository is only one (1) when the repository is mentioned and allows third party access to data otherwise the presence of the repository in RDMP does not necessarily imply data reusability. In this example DANS-KNAW obtain the score 1 (DANS is a repository aligned with FAIR principles). When the selected repository does not affect the data reusability, the score will be zero [0]. The repositories certified with a Data Seal of Approval are considered as those repositories with open access to the public. [4].

3.2.5 Licenses to Reuse

There is a condition that affects metadata completeness and that is the license for reuse. The type of license imposed on data determines the level of accessibility and openness of the dataset. In the example 3.1 number (2) indicates restricted access to the data set and imposes an embargo on the data set . Licenses are chosen based on the level of security and privacy that researchers need for their data. For instance: restricted access in the example shown in Figure 3.1 number (2) means that data users need to ask for permission to access the data set and when this access is not granted then the data is not reusable. A common example of licenses are creative commons (CC0)[5]. The CC0 lets the data producer open their data set to the public domain under applicable copyright law[6]. When a researcher is using CC0 it means that data is available for reuse and access is not restricted, and this means a score of one (1) in the assessment framework 3.1. In RDMP, the presence of licenses are measured with a binary value. The overall scores is a sum of binary values found in the RDMP deliverable.

3.3 Use of NLP Techniques in RDMP deliverables

Detecting the right information about metadata in RDMP deliverables is executed using NLP techniques. NLP techniques are considered a solution since they enable identifying the vocabulary and translates keywords into readable values, vectors, etc.

NLP techniques enable the machine-driven evaluation of reusable data in the context of a RDMP deliverable. NLP generations aim to decrease ambiguity in the human writing for machine-driven evaluations and since the goal of reusable data is to become machine and human readable, it is a suitable approach for a reusability measurement in this study. NLP techniques are a sub-category of text mining. Text mining is an automated approach to create

[4] $https://$assessment.datasealofapproval.org/seals/-Accessed Oct 2018

[5] $https://$creativecommons.org/share-your-work/public-domain/cc0/-Accessed Oct 2018

[6] $https://$help.data.world/hc/en-us/articles/115006114287-Common-license-types-for-dataset-Accessed Oct 2018

reliable data based on requirements of an organization and it performs different tasks on the text [52]. The NLP algorithm prepares data, compresses and transforms it into identifiable information pieces for a machine [33].

Based on characteristics of theoretical linguistics, the natural language processing applies within different levels of analysis [26]. The phonetic level is human utterance and words from human voice [4]. The morphological stage of human language is related to the structure and meaning of a word. For example the word "think" is an atomic word, but thinking and rethink are from the same root in combination with other words and have a different meaning. A verb can be an noun or adjective like "study" which turns into "studying" by adding "ing" to it, an example is: "she likes studying" [4].

The lexical level is to match every word in the sentences as input to the sequence of digital characters suitable for computational programming, such as token, lexeme or a string as an output [26]. Lexical analysis is used to make words machine readable units and prepare them for further NLP approaches in the process [26]. 5. The syntax level is the first unit of analysis in a sentence and refers to the sequence of words [26]. Syntax parsing determines the grammatical and syntactical structure of sentences in the text to capture meaning of a sentence. It is necessary to remove errors and noise from the natural language and turn it into structured meaningful codes in context free grammar [26]. Lexical and syntax analysis is used in pre-processing steps in the user feedback tool, as presented in Chapter 5 The semantic level is the meaning of words and utterance [26]. The semantic analysis in the machine is to understand the result of parsing with common sense reasoning and circumstance [26]. Information retrieval, information extraction, text summarization, and machine translation are applications of semantic analysis [26]. The pragmatics and semantic level both refer to the meaning and use of language in context [4]. The highest level is at the discourse level which involves the context, the topic and finding the correlation between syntax, semantics and pragmatics [26]. The application of NLP techniques also refers to different linguistics levels, depending on the tasks needed to be done with the data.

3.4 NLP techniques in the literature review

The literature review is conducted on NLP techniques which were applied to measure provenance quality and metadata completeness. There is no literature found that applied NLP techniques based on the FAIR principles. There are two trends in the literature, one is based on the techniques used to check the quality of provenance and metadata completeness [18, 27, 19, 10]. The second is a trend which applies NLP techniques to measure data qualities [42, 19, 32, 41, 39, 1, 36]. In the table 3.2 the summary of approaches and algorithms used in the literature are shown.

In the literature, according to Mathew Gamble, the trustworthiness of provenance is measured on web data by using a probabilistic approach [17]. In the literature metadata completeness is measured by using keywords and assigning values to them [30]. Gamble's study used a provenance graph as a data model and applied a multi-entity Bayesian Network algorithm to measure the completeness of provenance on web data [18]. This graph based approach is adaptable to the RDMP text but it requires various lexical and syntax parsing to construct the graph from text data. It requires a complicated algorithm and calculation for very short text. This approach is suitable when data is a web page or an existing dataset.

Article	Approach	Algorithms
[7]	Statistical	Graph analysis, Clustering
[62]	Dictionary based	CRF- bootstraping
[13]	Rule-based	POS,CRF
[29]	Probabilistic	Defacto (Deep Fact validation)
[28]	Unsupervised method	(LSA), (SVD), TF-IDF, Clustering
[27]	Rule-based	AIR Web rule language
[14]	Statistical	Bayesian network
[16]	Rule based, supervised	Classifier, n-gram
[5]	Linguistic	Name Entity Recognition
[17]	Probabilistic graph based	Multi-Entity Bayesian
[38]	Unsupervised method	Clustering

Table 3.2: Algorithms used to measure data quality based on literature

The two approaches in the literature are relevant to the RDMP context. The most suitable approach based on the data in a RDMP deliverable, is keywords extraction algorithms. Since data is either a paragraph or a template with questions and metadata elements are named entities; extracting keywords of completeness elements, as shown in Table 3.1 is the most suitable approach.

3.5 Adaptation of NLP techniques to RDMP deliverables

3.5.1 Text pre-processing

In the RDMP deliverable it is necessary to apply NLP techniques at the syntax and lexical level of data. In syntax parsing the words need to be clear data differentiated from the text that is not needed for RDMP (eg. questions, template, answers to other questions, ect.). Lexical analysis is needed to find the right words as elements of metadata.

In the RDMP deliverable, text segmentation is used to convert the text string into component words and sentences [26]. In computational linguistics these components are called tokens [26]. Therefore, tokenization is the first step of processing text documents, and it is a sub-task of parsing [26]. Tokenization algorithms convert a document into encoded characters and bytes, and finally compares patterns in the document to find the code that fits to bytes [26].

The SpaCy is an open source library for advanced NLP with various linguistic features such as: Non-destructive tokenization, named entity recognition, part-of-speech tagging, labelled dependency parsing, syntax-driven sentence segmentation, etc.[7]. SpaCy can extract dependency of words in a sentence. To design the text mining artifact, SpaCy[8] is used for

[7] $https://$spacy.io/usage/linguistic-features-Accessed October2018
[8] $https://$spacy.io/-Accessed Oct 2018

syntax and lexical parsing.

3.5.2 Dictionary Based Approach and Key words Extraction

In the second step identical keywords of metadata completeness need to be extracted from the RDMP deliverable. To conduct this step a dictionary is needed with metadata keywords. A dictionary that contains elements of metadata includes: PID, license, data users, types of data format and FAIR repositories. A dictionary tags the concept from corpora and identifies it based on a database and it requires the building of a domain relevant dictionary [51]. It is a time consuming task to build a dictionary from scratch and pre-defined tools like WordNet [9] can be used to build the dictionary in English. WordNet is the lexical database for the English language that creates syn-sets (set of synonyms) from nouns, verbs, adjectives and adverbs, this connects associated words with the keywords and their part of speech tag (verb, noun, adjective) [10]. WordNet is used for RDMP deliverables as the text is generated by humans with the high chance of using similar semantics about names, for example: archive, storage, repository. The challenge is that there is no dictionary for certain elements such as FAIR repositories and licenses. As specific names and vocabulary need to be found in the RDMP deliverable, a relevant dictionary is expected to yield more precise words than a pre-defined dictionary.

In order to examine the right approach for RDMP deliverables and achieve the relevant outcome semi-structured interviews are conducted with reviewers of RDMP deliverables and RSO in universities in the Netherlands.

[9] $https://$wordnet.princeton.edu/-Accessed October2018
[10] $https://$wordnet.princeton.edu/-Accessed Oct 2018

Chapter 4

Assessment Framework for RDM

In this chapter, the data collection and empirical review is presented. The first version of the assessment framework is based on findings in the literature review. The second version of the framework is developed by including interview findings in the initial framework.

4.1 Data Collection

Data was collected during semi-structured interviews with funding agencies and Research Support Offices (RSO) in universities in the Netherlands. Reviewers are those who review and assess final RDMP deliverables on behalf of funding agencies. The Research Support officers help researchers conform to RDMP deliverable requirements.

Interviews are transcribed and coded anonymously with no trace back to the interviewees or their organization. The interview data represents, criteria applied in reviewing RDMP deliverables and challenges from people who help researchers make an acceptable RDMP deliverable.

4.1.1 Sampling

The sample is a cluster sample including ten organizations in the Netherlands. Five out of ten interviewees in the sample, review or reviewed RDMP deliverables. The funding agencies in the sample are the two main funding bodies in the Netherlands, shown as Dutch Funding Agency (as DFA in the table). One data archive expert (shown as DA in the table) and two other experts are from the European Commission, mentioned as EFA in Table 4.1. There are five data managers from the RDM office (RDM in the table) based in university libraries. Related information of participants are in the table 4.1. The field of the interviewees are not mentioned in order to prevent tracing back.

This case study considers a specific RDMP deliverable, the Data Management Paragraph in NWO proposals, therefore the sample is focused on the Netherlands funding agencies. Since funding agencies in the Netherlands are partners with the European Commission their policy reviewer and national contact point in the EU are also included in this sample.

ID	Organization	Experience	Opinion about FAIR
F1	EF	4 years	Heared it
F2	DF	7-8 years	Supportive
F3	DF	8 months	Critical
F4	Repo	7-8 years	Supportive
F5	EF	4 years	Supportive
R1	RDM	2 years	Supportive
R2	RDM	3 months	Supportive
R3	RDM	10 years	Supportive
R4	RDM	5 years	Supportive
R5	RDM	4 years	Supportive

Table 4.1: Sample of interview participants

4.1.2 Interview Process

Some interviews were conducted remotely and three interviews were conducted face to face. To comply with the ethical code, each interviewee was given an informed consent form (see appendix B) and the interview protocol (see appendix D) three days before the scheduled interview. One interviewee dropped out after receiving the protocol. However, the rest of the sample agreed to participate after conducting a pre-interview call. Sending the informed consent and protocol and information participant letter(see appendix C) gave the interviewees the opportunity to share the questions with their colleagues and collect more information about the protocol.

The interviews were semi-structured and had open-ended questions. The interviewee had enough freedom to explain challenges and major issues regarding RDMP based on their organisation, including his/her insights and experiences. Pilot interviews were conducted three times with two different people, in order to adapt the questions. These test runs ensured the actual questions aligned with the interview goals.

4.1.3 Data Reduction and Interpretation

A popular method for the interpretation of symbolic data like interview transcriptions, is the Grounded Theory Method (GTM) [54]. In the GTM, themes are extracted from interview transcriptions [54]. Interview transcriptions were coded using NVivo software. The first step involved open coding; concepts and rich descriptions were extracted with line-by-line coding, resulting in the theory being built [55]. The themes were extracted from codes and rules derived from those themes. Constant comparison is the process of constantly com-

paring coded data and nodes or labels of each unit of data with other units, in the same category[50]. To check the validity of interpretations, interviews were coded by doing a constant comparison between coded nodes and new nodes [50, 54]. Interview themes were based on producing facts about the first framework and challenges in the RDMP for all the roles in the RDMP process [54]. Interviews were also coded by the daily supervisor, with the resulting interpretations serving as the validation method of triangulation [54]. The code book is in the appendix E.

4.2 RQ2. How are the RDMP deliverables reviewed?

The model in Figure 4.1 describes the process of reviewing RDMP deliverables. The model is a schematic view of the reviewing process identified in the interviews. There are three main roles in the review process. Reviewers are external experts who are hired by the funding agency to review RDMP deliverables. Researchers or applicants need to prepare their RDMP deliverables in several versions for the funder. The first version is the data section in their application, with more complete plans required when they receive the grant. Thereafter there are also midterm and final plans required by some funding agencies. They have to deliver each version of their deliverable to the funding agencies in the particular time frame of their project. The university library or RDM offices are in close contact with researchers and the funding agencies. Their main role is to help researchers write their RDMP deliverables. This help means providing transparency for researchers and general solutions to raise the awareness of researchers to their RDMP. RDM offices work in collaboration with funding agencies, relaying updates to researchers, trouble-shooting challenges and flagging ambiguities with funding agencies as well. In the review model in Figure 4.1, arrows are the indication of the ongoing process of writing deliverables and reviewing them.

Funding agencies have two main objectives for the requirement of RDMP deliverables:

- Raise awareness about research data management
- Enhance quality of research data

The funding agency invests in promoting research data quality, to make researchers more aware of the requirements of reusable data. By writing the RDMP deliverables, the researcher will look for solutions such as where he/she can deposit the data, what metadata to use, etc.

4.2.1 RDMP Challenges Identified For Researchers

As mentioned earlier, the review of the RDMP deliverables is a learning process that involves iterations between the three roles. The iteration and learning process is needed to cover the ongoing challenges in this movement. The challenges identified in interviews are below:

1. Lack of awareness about current infrastructure (repository)
2. Not knowing the concept of metadata
3. Lack of transparency about FAIR principles (mainly Interoperability)
4. Lack of some community effort to improve RDM

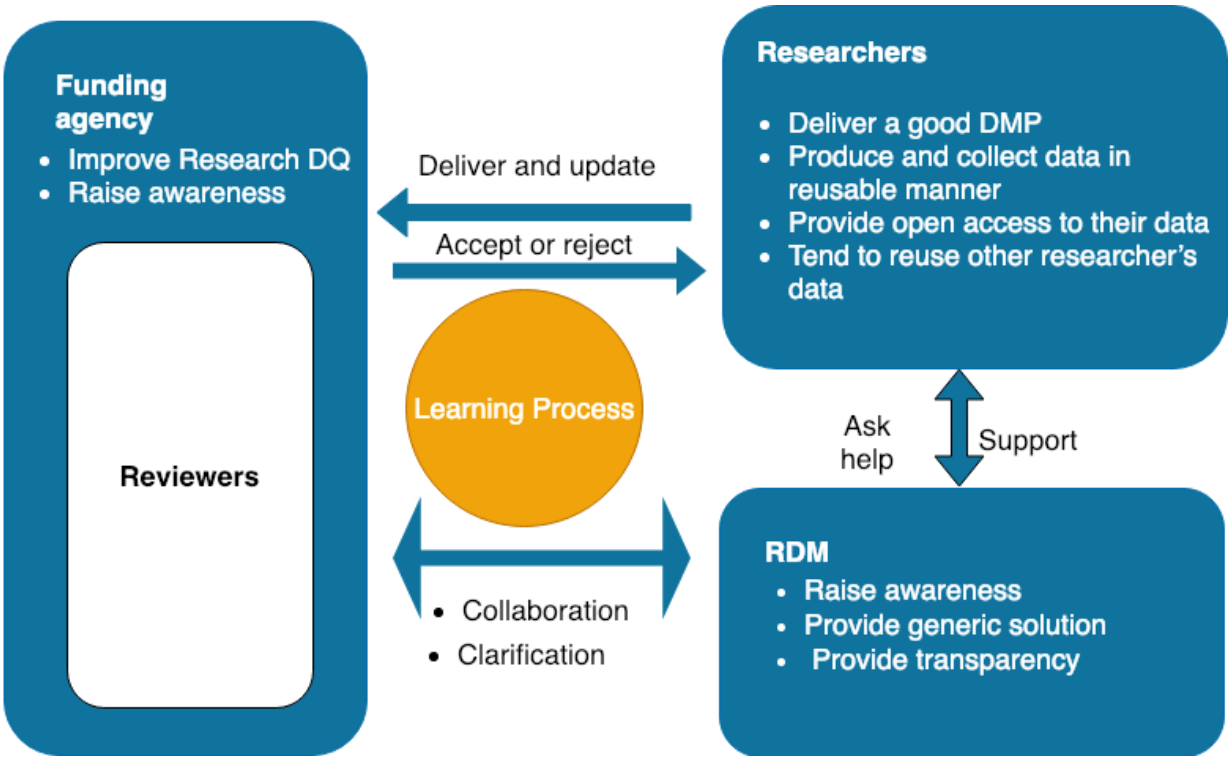


Figure 4.1: The Workflow of delivering and reviewing DMPs

5. Lack of RDM budget to make data available for the long term (after the project)

6. Deal with legal issues relating to open access of their data

7. Lack of a data management expert to review the RDMP deliverable

8. No guidelines to provide feedback on RDMP deliverables

4.3 Assessment Rules

There are elements identified from interviews and there are rules associated with them. The elements that are checked are shown in Table 4.2. In Table 4.2, the first column contains the most mentioned issues by interviewees. For example, the interoperability of data has been discussed in eight of the ten interviews and the main category for interoperability of data is a sub-category of metadata node in interviews. The metadata node contains mentions of challenges related to finding the right metadata. The interoperability of data as an assessment criterion means the metadata completeness has to be checked for interoperability of the metadata information provided in the RDMP deliverable.

The first column contains the elements and terms that interviewees mentioned. In the second column, Reusability metrics are matched with the terms used in the framework based on the literature review. In the third column, the relationship between criteria and FAIR principles is displayed.

4.3.1 The Overlap With The Initial Framework

Elements of the framework 3.2 in chapter 3 has common criteria with what was mentioned in interviews and these elements describe, metadata completeness, license name for legal issues and open access to data. In the final framework 4.2 metadata completeness is coined an acceptable metadata. There are eight rules derived from the interview data and these are used for the measurement, solution, and keywords references to build the Assessment Framework.

Acceptable Metadata

Acceptable metadata in a RDMP deliverable should have the following elements listed below:

Rule1. Metadata elements are complete based on FAIR principles or 5 key items required. ZonMW requires the following main elements: doi code, president identifier, a link to a repository during the project and after the project, and metadata standards that allow data to be linked to other data collections [1]. The 5 key items are similar to the F, A and I in the FAIR principles, which is required by Dutch organisations.

Rule 2. Information about types of data (personal data, type of data, data set, data sets, existing data and the reuse of existing data) are elements of Acceptable metadata.

Rule 3. The Dublin Core Metadata Initiative(DCMI) and DataCite are the standards used to generate reusable metadata.

Rule 4. Data deposited in the certified repository can be protected with usage licences. For example licenses of use or restricted access.

Rule 5. Data deposited in a certified repository are required to have reusable metadata and the repository generates unique identifiers for the data set.

Rule 6. Data deposited in a certified repository (Data Seal Approval and Core Trust Seal) are reusable during or after the project.

Opening Their Data

The willingness of researchers to open their data to the public; the data should be findable and accessible.

Rule 7. The openness of data may depend on the type of data and sometimes it is a challenge to balance that with legal issues and the use of licences.

Applying Domain Specific Guidance

RDM offices claim that they have to read parts of the proposal to understand the topic and to be able to help researchers find the right solution.

Rule 8. To generate reusable metadata or deposit data, domain relevant repositories and common practice in the same discipline provide solutions for the RDMP.

[1] *https* : //www.zonmw.nl/en/research-and-results/fair-data-data-management/format-data-management-plan- Accessed Jan 2019

The most mentioned criteria in Interviews	Overlap with Initial Framework	F	A	I	R
Interoperability of data	metadata completeness	x	x	x	
Data archiving requirements	FAIR repository name		x		
Standard key items-FAIR	metadata completeness		x	x	x
Legal issues	Licences name				x
Good practices by discipline	-				x
Vocabulary by discipline	-				x
Living document	-				
Revision	-				
Unexpected changes	-				

Table 4.2: Mapping the findings from interview and literature review and their overlap with FAIR principles

Interview	Literature	Final Criteria
Interoperability of data	Metadata completeness	Acceptable Metadata
Data archiving requirements	FAIR repository name	
Standard key items	Metadata completeness	
Legal issues	License name	Opening their data
Good practices by discipline		Applying domain specific guidance
Vocabulary by discipline	-	
Living document	-	Promising future updates
Revision	-	
Unexpected changes	-	

Table 4.3: The Final elements used in the assessment framework are based on the elements from interviews and the initial Framework developed from the literature review

Promise Future Updates

RDMP deliverables are living documents and delivering more complete versions of RDMP deliverables during the project life cycle is required by funding agencies.

Rule 9. The reviewers would like to see the applicant acknowledge that a RDMP deliverable is a planing tool for their RDMP and it has to be updated during the project based on changes in the project.

4.4 Final Assessment Framework for RDM

The assessment framework 4.2 contains the high level criteria needed to review the RDMP deliverables. The main criterion breaks down into elements which are required to be in the RDMP deliverable. The list of requirements are based on the deliverable and funding scheme. In the next chapter, the framework is used to design the artifact that provides user feedback

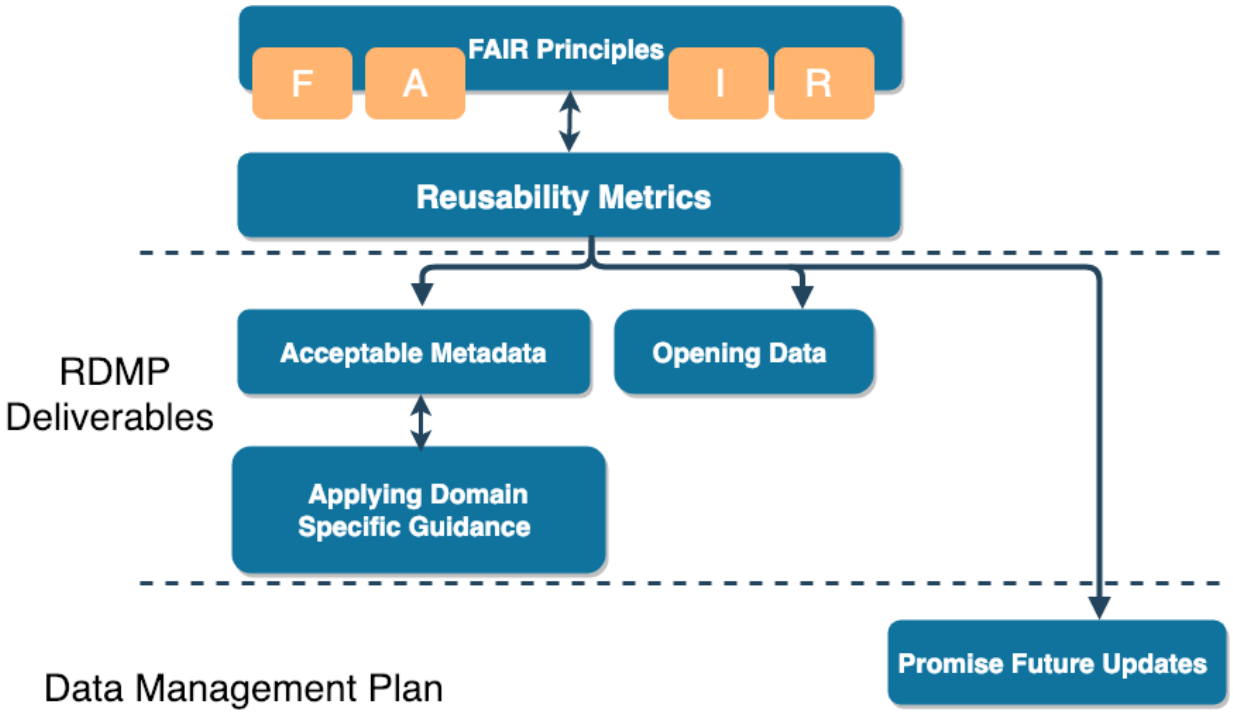


Figure 4.2: The assessment framework elements for RDM planning

on the Data management paragraph(DMP) in NWO proposals.

As mentioned in the first framework, the reusability metrics and FAIR principles have a lot in common. Therefore, the assessment framework 4.2 has a two-sided arrow to describe the relationship between them. Basically to follow FAIR principles in real RDMP deliverables we need to check the reusability metrics. The FAIR principles could appear in more detail in the RDMP deliverable, depending on the stage of the project. For example, the data section in a proposal is very generic and the researcher may provide information about what type of metadata standards they are going to have and in which repository they will deposit their data but not provide the actual link to the repository. In the midterm RDMP deliverable, more specific plans such as having the doi code and the metadata link to the repository where the data exists, is required.

4.4.1 Acceptable Metadata

The reusability metrics in the assessment framework 4.2 have three main criteria, namely acceptable metadata, opening their data and a promise of future updates. The acceptable metadata includes the presence of metadata elements and links to the RDMP deliverable. Rules 1-6 mentions all the metadata elements.

4.4.2 Applying Domain Specific Guidance

Domain relevant metadata and repositories can be used based on the research topic and discipline 4.3.1. Using domain specific elements for acceptable metadata increases findability of data and indicates precise knowledge about the RDMP of the project.

4.4.3 Opening Data

In order to achieve reusable data, the researcher needs to open their data to the public. The openness level of the data may vary depending on the type of data and challenges faced in opening their data. As mentioned in rule 7 (in 4.3.1) for example, data may be findable but not accessible due to privacy issues. This data is findable for other researchers and they can request access from the owner. While another data set can be fully accessed in the same repository.

4.4.4 Promise Future Updates

As explained in rule 9 4.3.1 promising future updates on the plan indicates that the researcher has the knowledge that they need to follow their plan taking into account future changes and delivering new versions based on the changes in the project. This is only applicable when the deliverable is the Data Management Plan and there are previous or future versions of it.

In the next chapter the framework, which has been used with an automated approach for RDMP based on the DMP, is presented.

Chapter 5

Design of the artifact to generate feedback on the Data Paragraph

5.1 Design of the Artifact

RQ3 is "*How to design an artifact that provides user feedback on the DM paragraph?* As it has been mentioned in interviews, the goal of funding agencies by requiring RDMP deliverables is to raise awareness of researchers. Therefore, an automated user feedback approach can improve the awareness of users (in this case researchers) about the missing requirements of RDMP deliverable.

In this chapter, the assessment framework 4.2 is used to design the quality measurements for an automated tool to provide user feedback on DM paragraphs and to answer RQ3. In order to do so, the final framework is operationalised with specific concepts sourced from Table 5.1. From these specific concepts, keywords are generated and a query is produced that measures the presence of keywords in a RDMP deliverable.

The RDMP deliverable used to design this artifact is the Data Management Section in a proposal. Impacter already provides user feedback on the proposal of the Data Management Paragraphs. The artifact is designed to be an extension for Impacter and used to provide feedback on the Data Management Paragraph. The paragraph in this chapter is referred to as the Data management section in the sample.

5.2 Use of NLP techniques

In the chapter 3 the NLP approaches proposed are based on the literature review. The proposed approach is to build a dictionary and extract keywords from RDMP deliverables.

In this study the sample is RDMP deliverables, which means the RDMP is present and in paragraphs data is not collected yet. The dictionary-based approach is suitable for Data Management Paragraphs too. In the literature [20] a semi-automated keywords extractor proposed to capture the quality requirements of the software. In the semi-automated keyword extractor, first, the step is to remove stop words from the sentences. The keywords are extracted from standard definitions of quality attributes and their stem words and synonyms. A set of keywords matched to a relevant quality attribute. The keywords and paired keywords are extracted. A set of weight assigned to the keywords. In the semi-automated approach, rules are

used to classify the weight and to select the relevant quality attributes for the classification found in a sentence.

5.3 Key words Extraction

Design of the automated user feedback is adopted from the semi-automated [20] above, to capture quality requirements of RDMP deliverables. The data model used in this design is based on the assessment framework proposed in chapter 4. The source of keywords is from the interviews and the result of the operationalisation of the assessment framework into keywords. The keywords are encoded to the machine-readable data (queries). Queries ran on each sentence in the Data management paragraphs. The binary values are set as a weight for the keywords found in a paragraph. The classification of the keywords is matched to the user feedback. The user feedback tool is destined in 5 steps. The step 1 to 5 are adapted from the semi-automated approach in the literature [20]. Steps 6 and 7 are the evaluation of measurements and design of the tool. The evaluation steps are 6 and 7 which are designed and implemented based on the book Experiment design in software engineering [59] (see appendix F). The evaluation steps are an experiment/simulation of the script ran on the database to evaluate the measurement design in the tool. Since the goal of this design is to provide feedback for users, the step 8 is implemented based on the requirements and finding from interviews.

In this design of the user feedback tool is implemented with the following steps:

1. Finding keywords based on the assessment framework in Chapter 4
2. Design queries based on references mentioned by reviewers
3. Remove false positives from queries
4. Cleaning data and removing noise from the sample
5. Run queries for the first time
6. Validation of queries based on the results of step 5
7. Design new queries with NLP techniques and improve the old ones
8. Design feedback by using all sets of improved queries

5.3.1 Finding Keywords and Designing Queries

As mentioned above, keywords are derived from the measurement criteria in the assessment framework 4.2. For each criterion in the Framework, keywords are determined based on interviews and finally the actual keywords are designed based on references mentioned in the interviews with reviewers.

Acceptable metadata is measured based on the Metadata items and Metadata standards. Metadata items are required by Zenodo[1]. Key items are doi code, online metadata and catalogue where data is indexed and registered, link to repository, terms and conditions of use

[1] *https*://www.zonmw.nl/en/research-and-results/fair-data-data-management/data-management-at-zonmw/zonmws-key-items-for-monitoring/-Accessed Feb 2019

Criteria in the Framework	Keywords concepts	Specific Keywords
Acceptable Metadata	Meta data (key items)	Persistent Identifiers/data format
	Metadata Standards	Dublin core/DataCite
	Certified repository	Data Seal Approval/Core Trust
Opening their data	Usage License	Creative Common licenses
Applying domain specific guidance	Domain specific repository	Data Seal Approval

Table 5.1: Operationalisation of the Assessment Framework into Keywords and Queries

for third party. The other keywords mentioned about metadata included persistent identifiers, doi code, and data format. The keywords also include synonyms. Metadata standards are also mentioned in the discussion of Metadata, with the DCMI and DataCite standards being mentioned as references. Certified repositories such as Data Seal Approval and Core Trust Certified are derived from the list of general repositories.

Licenses imposed on the data has implications for the opening of researchers' data. This is circumvented by the use of Creative Commons. The Creative Commons family is the corresponding reference to derive the keywords. The application of domain-specific guidance is measured with a list of domain-specific repositories certified with DSA. The full list of queries are in the appendix G. In Table 5.1, the first column is criteria in the framework and the second column is concepts mentioned by reviewers. The last column shows the specific keywords derived from the references for each of the concepts.

5.3.2 Remove False Positives from Queries

Lists of words for the six variables, persistent identifier, data format, Dublin Core and DataCite, Data Seal Approval repository, Creative Common licenses and domain specific repository, are prepared in the form of queries in Python. Queries are tested by running a Python script on the database in two runs. The queries are named entities and abbreviations. The abbreviations are wrapped in spaces for example "ARCHE"[2] is the repository for Humanities in Austria but the "ARCHE" is part of the word "rese**arche**r". Therefore, the word "ARCHE" in the repositories classification is " ARCHE ".

5.3.3 Cleaning Data and Removing Noise from the Sample

For the simulation of running Python script, sample (subjects) are selected from proposals in the (Impacter) database. The proposal template has section 2.e Data Management paragraph. Only when this section is included in the template is the proposal included in the sample. The section 2.e in the NWO proposals is the Data Management section, contains four questions about RDMP of the project [3]. Not all proposals contain a paragraph 2.e Data management (some only include paragraph 2b. Knowledge Utilisation). Additionally, users on average submit their proposal 3.5 times. Only the last version uploaded is selected for this sample, resulting in a final sample of 98 proposals. Before conducting the simulation, pre-processing steps are done to prepare the sample and remove noise from data:

[2] *https*://arche.acdh.oeaw.ac.at/browser/-Accessed Feb 2019

[3] *https*://www.nwo.nl/en/policies/open+science/data+management+chapter-Accessed Feb 2019

1. Splitting paragraphs, custom front end extension splitting on the basis of Table of Contents, Headings, Regular Expressions and user feedback (in that order). This step is already part of the Impacter software.

2. Removing the template text, using SpaCy that matches sentences to the template document.This is a customized extension of the NLP pipeline that matches sentences to the template document and labels them as such, so that they can be excluded from the analysis.

3. Split the questions in the 2.e Data Management Paragraph. Sub-string splitting is done on the basis of exact phrases used in the questions of the template.

4. Run queries on the sentences

    - The first run: Finding the sub-string (case-insensitive)
    - The second run: SpaCy TargetMatcher

5.3.4 Run Queries for the First Time

After removing possible false positives, the queries in `.json` file are transferred and run in the staging environment on the production data of Impacter. There are six `.json` files corresponding to the six variables in Table 5.1, the variables are measured by the absence and presence of them in paragraphs. Since questions in the paragraph don't have a corresponding one-word answer and is predicted to involve some reasoning, it is unknown in which questions the user is going to talk about their repository. To avoid false positives the queries are executed per question and the result for each variable is converted to a binary value corresponding to either the presence (1) or absence (0) of a keyword from the list in the proposal. As the specific terms are not recorded, but only binary levels are registered for anonymised proposal numbers, the results cannot be traced back to the confidential proposals themselves.

5.3.5 Validation of the Queries based on the Result of the Previous Step

After the first run, validation is sought between real data and the result of queries done by a human reader. Since the data is confidential, validation is conducted by IDfuse and the result is shown in Table 5.2. The validation criteria converts to outcome measures in binary values based on four patterns in the paragraphs shown in Table 5.2. The outcome of the first round of running queries yields the results shown in Table 5.3 and the performance of queries is calculated by a F1 score based on the precision and recall calculated. Accuracy of the queries based on the True Positives (TP) and True Negatives (TN) is 0.602. The goal is to improve this measure as well as the F1 score by improving queries. In Table 5.3, 40 TP were found and a number of false negatives, which is relatively high. False negatives are mostly the paragraph with no specific words and they are not found, since queries contain only specific words. So, in the case where a researcher explains how their data will be open but does not have a license for them yet, it is counted as a false negative. In the design of the next treatment, the goal is to reduce this number and not at the cost of increasing false positives.

Validation criteria	Binary value	Definition
Generic answers	1	User explained about reuse of their data in general terms or part of data is reusable
Reusable data	1	User have reusable data and specifically explained their RDMP
Empty paragraphs	0	There is no answers in the 2.e section
Not reusable data	0	They have text but they explained data can not be reused with open access

Table 5.2: Validation criteria and conversion to binary values

Treatment	True Positive	True Negative	False Positive	False Negative
Treatment 1	40	19	1	38
Treatment 2	45	18	2	33

Table 5.3: Outcome of running the first and the second treatment

With reference to Table 5.4, in order to measure performance of the first treatment the precision and recall are calculated. A F1 score is calculated based on the harmonic average of precision and recall. The F1 score is between 0 the lowest, to 1 for high performance.

As the result of the first simulation shows, the F1 score is 0.66 which means that there is a gain in more than half of the performance value in the queries but there is room to increase the performance. Returning to the validation step, there are many generic answers found in paragraphs, where the user has explained about their data reusability but did not use any of the specific entities in the queries. In conclusion for the second treatment, more generic words for queries are considered.

5.3.6 Design New Queries and using NLP techniques

In order to reduce false negatives, a new set of queries is designed at a higher level of the assessment framework. The higher level measures in the assessment framework is Acceptable Metadata and Opening data. In this level there is lower granularity of metadata which is lower than in the first treatment.

Acceptable Metadata

To conduct the second run, a new set of queries with generic terms of metadata, data format and metadata standards is designed to measure Acceptable Metadata. For example, in the query for metadata standards included terms such as Dublin Core and Data Cite are in the generic query, only variations of the words ”Dublin core metadata initiative”

	T1	T2
Accuracy	0,60	0,64
Precision	0,97	0,96
Recall	0,51	0,57
F1	0,67	0,72

Table 5.4: The result of performance between two treatments

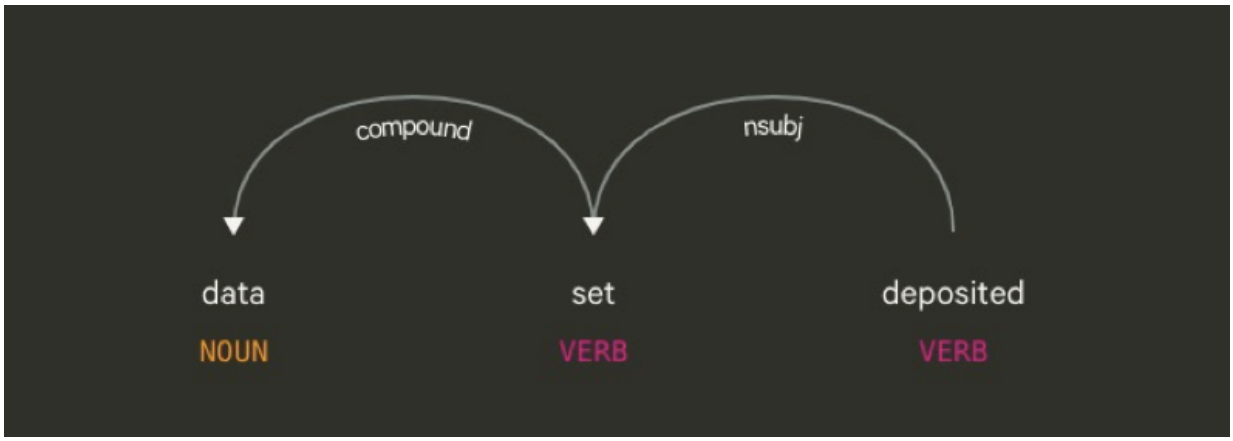


Figure 5.1: Dependency Parsing example for Generic Queries

and ”DataCite” are used. In the data format query instead of `".txt", "text","image" and "video"` are used. In the query related to Metadata, instead of the specific `"URN:NBN identifier project"` there are generic words used such as :`"Interoperable meta data ", "Persistent Identifier`. Generic words in the query are general words defining what is used in the first run, their possible variations, plus related words mentioned in the discussion with interviewees.

The generic terms of repository made by use of part-of-speech and dependency parsing as key phases that explains data as stored in an open access repository. In the second set of queries, instead of the repository name like `" ARCHE "`, the way the user may explain about depositing their data is considered for example:

```
[ " LOWER ": "Publicly " , "LOWER ": " available " , " LOWER ": "archive ", "POS ": " NOUN " ]
 [ "LOWER " : "dataset ", "DEP ": "nsubj ",  "LOWER ": "deposited "]
```

In these queries NLP techniques are used to capture more generic sentences about generic terms for example : `["LOWER": "Publicly ", " LOWER ": " available " , " LOWER" : "archive ", "POS" : "NOUN "]`

”Publicly available archive” is a part-of-speech (POS) tag in a sentence and when it is used as a noun for a repository query, it means the user is discussing a repository with open access. Also, dependencies parsing is used

```
[ "LOWER": "dataset","DEP": "nsubj", "LOWER": "deposited "]
```

to find a ”data set deposited”. Figure 5.1 shows how the dependencies are selected for the query.

Opening Data

In order to use a generic version of query for opening data and to consider possible answers about how the user wants to share their data, NLP techniques are used instead of Creative Commons Family licenses. An example of this query is as follow:

```
["LOWER":"dataset","DEP":"nsubj", "LOWER":"deposited"]
["LOWER":"partly","DEP":"advmod","LOWER":"available"]
```

By adding the generic version of answers and NLP queries, the performance of queries are expected to improve and the outcome to run a new set of queries are presented in Table 5.3. In addition, the risk of increasing false positives will be higher because of using generic

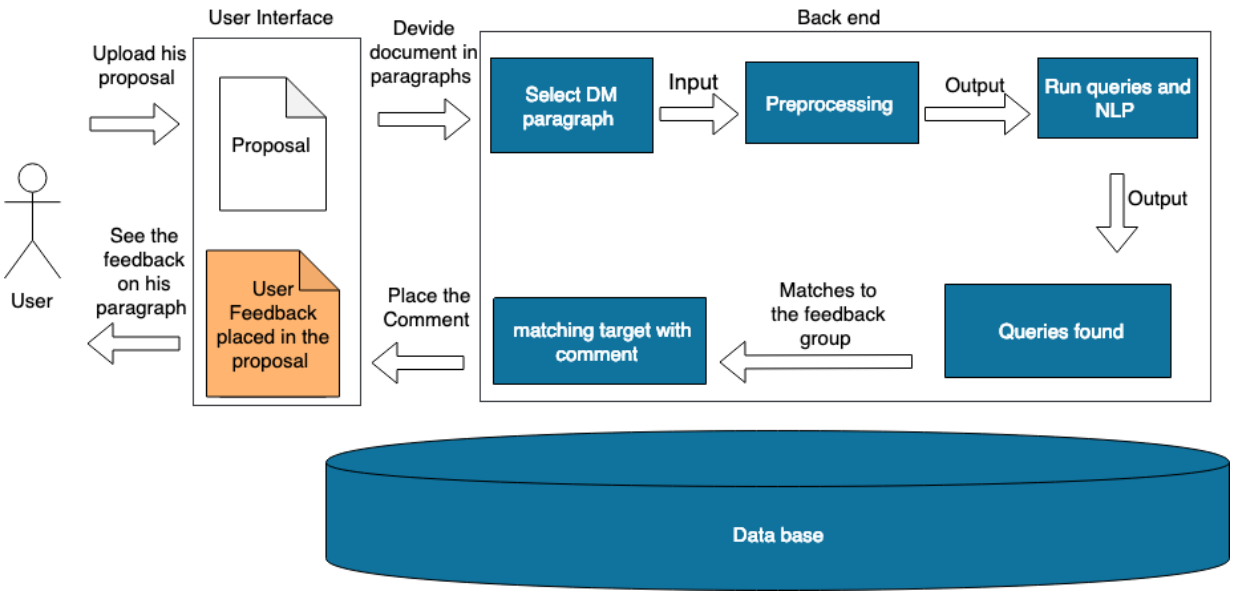


Figure 5.2: The Design Of The User Feedback Tool

terms, therefore the first set of specific queries used are added beside the new set. In the Table 5.3 the number of false negatives reduced from 38 to 35 and true positives added up to 45 while it was 40 in the first run. There are two false positives in both runs but the false positive is from two different files.

On the first run, the paragraph that expresses their data is not reusable and the other one is an empty text. In the second run both paragraphs are not reusable data.

The performance measure is calculated for the second set of queries in the Table 5.4. Accuracy has improved to 0.64 and the F1 score is slightly higher than the F1 score in the first run.

5.3.7 Design User Feedback based on Queries

As explained earlier, Impacter is a user feedback online tool, where queries and code are designed to run in the same tool. The architecture of the current tool can not be explained in fine detail due to confidentiality. The process of providing user feedback designed in this thesis is shown in the Figure 5.2. In the front end, the user uploads the .docx file including the grant application template. Paragraphs are divided based on the template of the proposal. The section Data Management is selected from the paragraphs and the pre-processing questions are removed to avoid false positives. Sentences are selected and queries are run on each sentence in the answers of the DM section. The output from queries shows which category of query got score 1 (exists) in the whole paragraphs. Feedback corresponding to the category found in the document will be selected and placed in the .docx file. In the end the user can see the comments on the DM section as well as the rest of the proposal on his/her screen.

Feedback Corresponding to Category of Queries

Queries are used to measure the reusability of paragraphs and the feedback matches to query found in a paragraph. In Table 5.5 feedback is shown corresponding to their target query in a

category. The feed backs are referenced to interviewees who has mentioned them. Solutions provided by reviewers or RDM officers in the interview sample are the source of feed backs.

The feedback design based on the level of details that the user specified in a paragraph, inform the user to be able to provide more detailed information. For example, if the user specifies the data format, feedback asks questions to specify online metadata. When generic queries are found in the paragraph, the feedback asks to provide more specific details about their data format and metadata.

5.4 Evaluation of the User Feedback Tool

The Assessment Framework contains RDMP reusability metrics and through the python script, metrics and queries has been validated. The process of user feedback tool in Figure 5.2 adapted from steps conducted during the experiment and the current capabilities of the tool Impacter. The evaluation of the user feedback tool requires implementation of the all the steps in the front end and back end of Impacter. Therefore it is not feasible to be done at the scope of this thesis. In the future, the queries will be used in Impacter. The Assessment Framework is validated through the experiment and it can be generalised for RDMP deliverables.

Category of Query	Corresponding Feed backs
Data Format 1 and 2	What documentation do you need to make your data format more authentic and transparent [R1]. If your data do not have a common format, please provide a relevant convert tool [EF1].
Persistent Identifier 1	The DOI code and catalogue and repository and data format are the critical requirements of Data Management for your project [F2]. ”Thinking of good license is indeed not really a challenge but you need to do it!”[F4]
Metadata 2	Meta data is important for for find-ability of data [R4], to produce reusable data you need to be more specific about metadata. You can follow Dublin Core standards [R4, R2, F3].
Metadata Standards2 and Dublin Core	”Most repositories do confirm at least Dublin Core standards”[R4]. You can find list of trusted repositories on the NWO website [R4, F3, R3].
Certified Repository 1	Open access repositories are able to keep your data even after the project! It is better to calculate the cost of your data maintenance during and after the project [F3, EF5, R1, R3].
Repository 2	Is the repository that you are going to use, certified repositories with open access? You can find list of trusted repositories on the NWO website [F3, R1, R3, F2].
Licenses 1	Thinking of good license is indeed not really a challenge but you need to do it to protect your data [F4].
Licences 2	Data set needs to be findable and have a good description and licence or how to access it and who to access it and what is possible to do with it, who to contact if you can’t get direct access [F4]
Domain specific Repository	Data management in different research field varies a lot and domain expert will be reviewing your data management plan later [EF5] therefore you have to comply with the demand of your faculty [R5] and get domain specific guidance [F4] .
Empty Paragraph	At this point of time you need to talk about the basics of your data management [EF1]. The DOI code and catalogue and repository and data format are the critical requirements of Data Management for your project [F2]. The main aim is to create awareness of good research data management and after the grant is awarded only the project leader are requested to write Data Management Plan according to the template [F3]

Table 5.5: User Feedback designed corresponding to category of queries

Chapter 6

Conclusion and Discussion

6.1 Findings

The main research question in this study is: *"How to improve reusability of research data by providing feedback on RDMP deliverables? "* In this thesis the main research question is answered by exploring the answers of the following three sub-questions:

6.1.1 What are the measures for data reusability?

Data reusability is about a transparent structure for data which makes data readable for humans and machines [58]. Research Data reusability quality indicators emphasize the quality of metadata and provenance. Metadata completeness indicates data reusability with elements such as persistent identifier, data user, data author and citation, license to access data, the archive where the data will be stored and data provenance. Data as digital object carries the elements to clarify it origins, that is, how it can be accessed and who is the owner of the data. These elements are in detail, specific and unique for each data set. Unique elements of data will make a homogeneous structure for data where everyone can access and produce reusable data. FAIR principles are guidelines used to achieve reusable research data and are adopted for each of the data elements. There is overlap between quality indicators of reusable data and FAIR principles, which can create reusable data.

6.1.2 How RDMP deliverables are reviewed by funding agencies?

This question is answered based on semi-interviews conducted with funding agencies in the Netherlands and the European commission. In real practice the review of RDMP deliverables is conducted with experts in the same domain. According to interviewees, the main challenge in preparing RDMP deliverables is to find the right metadata and to understand the concept of metadata. There are no specific guidelines to review RDMP deliverables; the domain expert reviews the paragraph based on the domain standards and/or common practices.

Metadata becomes the common criterion in both the initial and final frameworks. The domain specific guidance and promises for future updates are added to the final Framework. Promise for changes and updates is a solution for the issues in the project that the researcher has not yet found a solution for. In future versions there is an opportunity to submit RDMP deliverables, even at the end of the project. If the researcher acknowledges that they will in

future improve for certain RDMP matters, this means that they are aware of their RDMP. The reviewers consider raising the awareness of researchers is more important than delivering a perfect answer based on university templates. Reviewers need to perceive from the RDMP deliverable where the RDMP stands in the research project. These criteria collected in the Assessment Framework, is based on the literature review, and adjusted to the findings from interviews, by adding domain specific guidance and promise of updates in the plan, this contributes to reusability measures for the RDMP deliverables.

6.1.3 How to design an artifact that generates feedback on RDMP deliverables?

The artifact is the user feedback tool which is designed based on the assessment framework for research data reusability with regards to the DM paragraph. The user feedback tool simulates the reviewer's thinking about the RDMP deliverables. Review criteria that was mentioned by reviewers is operationalised in a set of queries based on two categories, Opening data and Acceptable metadata. The queries were tested on the data and designed through pilots run to avoid false positive errors. The tool is designed as a prototype and a simulation was run to test this prototype within the implementation capabilities of this thesis, with queries which are going to be included in a dictionary in the final tool. Feedback is designed for each category of queries to provide information on the answers written. The main issues in the RDMP found during this design, based on the result of the simulation are:

1. There are twenty-five out of ninety-eight generic answers provided which could not be found with name entities and required a generic solution.

2. There is one case where the researcher who wrote the grant may not be in charge of the RDMP in the project and not aware of the RDMP.

3. Researcher mentioned access to publication but not to data

4. A case in which a researcher needed to calculate a budget for their RDMP

In the simulation generic answers were addressed with a set of NLP techniques, this disregarded paragraphs with generic answers about their reusable data and thus avoided false negatives.

6.2 Validation

This study was associated with some threat and certain action was taken to remove influence of validity threats on the design of the assessment framework and the user feedback tool. The simulation was conducted to validate the metrics. Metrics were target keywords based on the assessment framework criteria 4.2. The assessment framework criteria are based on the most common patterns mentioned by reviewers and the challenges mentioned by RDM officers.

6.2.1 Construct Validity

The threats associated with the design of this study is the explication of constructs [59]. The constructs in the assessment framework defined explicitly based on the interviews. By using grounded theory in coding the interview data, the threat of vague definition of constructs in the assessment framework criteria are mitigated.

In the user feedback tool, threats in the design are reduced by deriving the queries from interview data. That is, the interviews were conducted to validate the initial assessment framework design which was based on the literature review. The measurement in the feedback tool is considered in binary values. Binary value corresponds to the presence or absence of query as a solid measurement to consider for the checklist of Acceptable metadata and Opening data to measure research data reusability.

In order to test the measurement design, before running the simulation, several pilots were run to test the queries and reduce unnecessary false positives caused by similarity of words in abbreviations.

6.2.2 External Validity

External threats in the design of the assessment framework are related to the interviews. In the interviews, since people review the RDMP deliverables, they may choose not to share complete information about the review procedure. Ethical issues are considered for interviews and they require interviewees to sign informed consent forms before the interview. The interview protocol was shared with interviewees to avoid short answers and give them sufficient time to think about the questions and their answers. To mitigate this threats the informed consent, information letter and interview protocol are prepared and shared with interviewees in advanced.

The metrics were designed based on the context of RDMP in the past three years. The time frame has an impact on the quality of paragraphs provided by researchers nevertheless the metrics are still applicable to RDMP deliverables. User feedback can change over time with more detailed guidance, for example, in case the majority of paragraphs are found to be reusable.

The assessment framework can be used for RDMP deliverables therefore it is possible to adapt the assessment framework to other platforms that provide RDMP services to measure data reusability through preparation of the RDMP deliverables.

6.2.3 Internal Validity

The internal validity threats are in the simulation of the user feedback tool. The threats are the time and social effects of the RDMP deliverables requirements at the time of the simulation.

There are certain times of the year, near deadlines for the call, that influence the quality of proposals. To avoid the influence of time this study considers only the last upload of each proposal in its sample, the quality of paragraphs are better as there is insufficient time to change the last submission.

A social threat is that researchers still do not have the funding and they do not need to fill in this paragraph yet [59]. In addition, there is no obligation to fill the paragraph, and this accounts for the number of empty paragraphs in the sample. However, the number of empty

paragraphs will not affect the measurement in the simulation but could provide a bigger sample.

6.2.4 Reliability

The reliability of the assessment framework is validated by experts and to have reliable measures in the tool, queries are references taken from the interviews.

The threats to conclusion validity are in the sample selection during implementation of the relevance cycle [59]. The action taken to mitigate this threat is that subject selected for interviews are from the Netherlands in order to avoid heterogeneity of criteria in the sample the EU policy advisor are added to the sample [59]. In order to cover the challenges and issues from researchers the Research Support Offices in the universities are interviewed as well.

The sample selection in the interview reflects metrics and finally keywords used in the simulation. The threats involved in the simulation of user feedback tool was including the data in the proposal that were no relevant to the RDMP of the project to the measurements. In the user feedback tool, the pre-processing step is conducted on the sample to remove the template text and exclude other information in the rest of the proposal in the measurement. Therefore the conclusion can be generalized for all the RDMP deliverables required by the European Union and funding agencies in the Netherlands.

6.3 Limitations

Limitations in this thesis were time and access to a bigger sample size for interviews and the simulation of the user feedback tool. If researchers were aware that they will receive feedback on the DM paragraph in a period of 8 months, a much bigger sample could be made available. It was not predicted that from 400 uploads there would be 98 DM templates included. In future research some possibilities that require a big time frame need planning prior to the project.

6.4 Research Contribution

There are two types of practical and scientific research contributions in this study. The scientific contribution of this research is to provide transparency on the quality indicators of reusable data and implement them in the real case study of RDMP deliverables. This Framework contributes to raising awareness on the research community about data management and measuring the quality of the data set they are planning to produce. Research data management has already been practiced however, not all the research domains have the same level of awareness about it. In conclusion, the main factor for research data reusability is to provide accessible and findable data sets during and after a research project. The other scientific contribution is in finding to provide transparency on FAIR principles at the RDMP stage. The assessment framework built in this research is applicable to assess research data reusability in the planing phases of the project. This study contributes FAIR metrics for RDMP. FAIR principles are the guidelines to share data during and after the project, while in the assessment framework the FAIR metrics are applied in to RDMP before the research project. The

assessment framework developed in this thesis can be used as a checklist of elements that are needed to produce reusable data and achieve different levels of reusability. The framework has a high level of abstraction which can be generalised in different contexts in RDMP. In this research the assessment framework is used for RDMP deliverables but it is possible to adapt the assessment for research data during the research as well. The assessment framework contributes to planning for RDMP in a research project. Criteria added with interviews and a literature review validated what is needed to achieve reusable data in the research project based on the current situation.

The practical contribution of this study, is the design of the artifact supporting researchers with the informative user feedback tool to improve their RDMP deliverable. Current and future users of the Impacter tool will receive tailored feedback on the Data Management section of their proposals based on the tool designed in this study. This tool will raise the awareness of the researcher to RDMP, by flagging quick feedback which may initiate suitable solutions to prevent a lot of RDMP challenges later on during the project.

6.5 Future Work

The RDMP is growing and there are different levels of awareness among different research domains. Some research disciplines like astronomy have practiced RDMP over the years and in some disciplines it has not been practiced strongly. Therefore, RDMP assessment can be tailored for research domains to provide precise solutions to common problems in that domain. Due to a lack of time and numerous research domains, it was not feasible to conduct research in all or many specific disciplines. Further research is required in domain specific assessments of reusable data in specific research domains. In addition, assessment criteria which are determined by domain expert leaders to create reusable data based on specified needs in the domain and solution sharing among researchers, is required. RDMP deliverables are changing every year while funding agencies push to enforce the requirement of more detailed plans in the RDMP context. Rapid changes in the environment raise the need for future work on RDMP based on the new requirements and policies. Finally, in future RDMP will become an obligation in research project applications and thus, a need for more research and new automated approaches based on new guidelines and requirements will be inevitable.

Appendix A

Literature Review Protocol

A.1 Search

Papers were found through three search engines: DBLP, SCOPUS and Google Scholar. For all queries the time frame 2010-2018 is used. Since the number of papers were not enough to achieve a conclusion based on DBLP and SCOPUS, papers from Google scholar were added to the search. Two queries related to NLP approaches in Google Scholar are considered from 2016-2018 to narrow down the number of papers within the scope of related data reusability quality indicators found in the previous search and NLP approaches. Backward and forward searches were conducted on the selected papers where needed and 77 papers that were read in-depth were chosen based on this strategy.

A.2 Exclusion

- Patent
- Papers related to health care data

A.3 Inclusion

- Papers which used web page data
- Papers included in 10 first pages of search results

A.4 Key Words

Key words used in the search are shown in Table A.1.

F	Query	Date	Result	Abstract	B
Scopus	TITLE-ABS-KEY ("NLP" AND "Data Quality")	All	16	5	4
	ALL ("NLP techniques" AND "provenance" AND "data quality")	0	0		
	TITLE-ABS-KEY ("NLP" AND "Information Quality")	All	9	5	
	TITLE-ABS-KEY ("Text summeraziation" AND "Data Quality")	All	0	0	
	TITLE-ABS-KEY ("text summarisation" AND "Information Quality")	All	2	0	
	TITLE-ABS-KEY (fairness AND principles AND "data management") AND DOCTYPE (ar)	All	0	0	
	ALL ("Information Extraction" + "Name Entity Recognition" + "dictionary based")	All	7	1	
	ALL ("Keyword extraction" AND dictionary AND "information extraction")	All	58	5	
	ALL ("Keyword extraction" AND "NLP techniques" AND data AND quality)	All	5	0	
DLBP	"NLP" AND "Data Quality"	All	2	1	
	"NLP"AND "Information Quality"	All	3	2	
	"Text summeraziation" AND "Data Quality"	All	0	0	
	"Text summeraziation" AND "Informaiton Quality"	All	0	0	
	"FAIR principles"	All	16	0	
	ALL ("Information Extraction" + "Name Entity Recognition" + "dictionary based")	All	0	0	
	"Keywords extraction "AND Dictionary based AND "data quality"	All	0	0	
	Provenance AND trust AND data quality	All	2	2	
Scholar	NLP techniques AND "Information Quality "	2010-2018	529		48
	ALL ("Keyword extraction" AND dictionary AND "information extraction")	2010-2018	350	13	
	"NLP"AND "Information Quality"	2010-2018	146	38	
	allintitle: AND "text summarisation" OR "Information quality" "Data quality"	2010-2018	68	8	
	"Reusability "OR "Data reusability"AND"NLP"OR"Natural language processing"	2016-2018	18600	13	
	"Keywords extraction "AND "word embedding AND "data quality"	2016-2019	5	2	
	TITLE-ABS-KEY (fairness AND principles AND "data management")	2010-2018	16	1	
	"FAIR principles"	2010-2018	619	13	25
	"NLP techniques" AND "provenance"AND "data quality"	2010-2018	85	13	
	"scientific data" management	2010-2018	41,700	36	
Total				158	77

Table A.1: List of Queries Used to Conduct Systematic Literature Review

Appendix B

Interview Consent Form

Declaration of consent for participating in:

Data management and Stewardship Study

I confirm that:

- I am well informed about the study after reading the accompanying information letter;
- I have had the opportunity to ask questions about this study and my questions were answered to my satisfaction;
- I have had the opportunity to think carefully about participating in this study;
- I participate voluntarily.

I give permission:

- that my data will be used for scientific reasons and will be saved as is formulated in the information letter;
- that if applicable, audio recordings are made for scientific purposes

I understand that:

- I have the right to withdraw my permission to use my data after participation, without the obligation to give reasons why;
- my data will be used according to the stipulations formulated in de Code of Conduct of the VSNU (www.vsnu.nl/gedragscodes).

Name participant: _____Maarten Goldberg_____ Date of birth: __28/04/1959

Consent for possible reuse of data

(1) Researchers share quite frequently their research data with other researchers, in this way data is used most optimally. Naturally, these data are completely anonymous. Do you agree that your anonymous data could be shared in the future with other researchers?

(Please, mark the intended answer with a cross, and sign if agreed)

[X] Yes, I agree. Signature: _____ **[ ]** No, not agreed.

(2) Sometimes result of interview is presented during a scientific conference or course (which again are sometimes placed on the internet). Do you agree that such recordings are used for these abovementioned purposes?

[X] Yes, I agree. Signature: _____ **[ ]** No, not agreed.

To fill out by the researcher:

I declare that I have explained to the participant what participation involves and I will ensure that the data will be treated anonymously.

Name: _____

Date: ___ / ___ / _____ (dd/mm/yyyy)

Signature:

Appendix C

Interview Information Participant

Universiteit Utrecht

Information for participants

Data Management and Stewardship Study

1. Introduction

You have indicated your willingness to participate in a scientific study conducted by researchers from the Information and Computing Science Department from Utrecht University. The overall goal of the study is the design of information systems which are capable of automatically generating feedback about research data paragraphs and research data management plans submitted by researchers.

2. What is expected of you?

This is a semi structured interview and takes about 45-60 minutes, in one session.

3. What are the possible advantages and disadvantages of taking part in this research?

Participating in this study does not offer you any direct advantage, however in the future the study may lead to useful knowledge about improving research data reusability.

4. Voluntary participation

Your participation is voluntary. If you decide not to take part in the research, you do not need to do anything, nor sign any document. You do not have to explain why you decide not to participate in the research. If you do decide to participate, you can always reconsider this decision and stop at any given moment – also during the interview.

5. What happens with the data that we collect?

Data that are collected in this research will be stored in complete anonymity on protected servers of Utrecht University One drive private folder. Your personal data will not be stored in the system.
Your personal data are taken care of by Baharak Bakhtiari and Armel Lefebvre. In case you would like to update your details, you can contact her at the email address: b.bakhtiari@students.uu.nl or A.E.J.Lefebvre@uu.nl.

We are obliged to keep the research data – anonymized – for 10 years. By participating in this research, you are giving us permission to do that. If you do not like us to keep these anonymized details, you may not take part in this interview.

6. More information on this research?

Would you like to have more information on this research? Please feel free to contact:
Baharak Bakhtiari
Email: b.bakhtiari@students.uu.nl
Tel: 0031619247650

Daily supervisor: Armel Lefebvre, MSc
Email: A.E.J.Lefebvre@uu.nl

First Supervisor: dr. Marco Spruit
Email: m.r.spruit@uu.nl

Appendix D

Interview Protocol of grant reviewers

Interview protocol for grant reviewers

- Introduction:

  Dear sir/madam,

I am Baharak from Utrecht University and I am working on my thesis focused on Data Management. I would like to thank you for your time and support in my project and conduct this semi-structured interview. The goal of this interview is to assess the challenges and issues that reviewers are facing during reviewing the Data Management Plan (paragraph).

- Ethical code:

I would like to inform you that this interview will be recorded and the information and content of the interview will be used for analysis of my master thesis. You have the right to stop or refuse to answer the questions at any time you want.

- Opening: Please Introduce yourself?
- tell me about your position and background?
- How many years you have been in this position?

1. Would you please describe the grant application procedure within the NWO/your agency and how you proceed with the applications?
First, ask about the plan and paragraph and what is he/she is reviewing? (make a definition about the DM instrument)
   1.a. What do you mean by DMP? (paragraph or plan or any other name?)
2. Why it is important to have the DMP/paragraph in the proposal?
3. Do you believe that scientists pay enough attention to the DMP?
4. What are the common mistakes that a researcher have in DMP? Or in other words, What type of feedback do you provide for them?
5. Would you please explain in details how do you assess and review this part of the proposal (DMP)?
6. What are the challenges for you in reviewing the DMP?
7. What are the benefits of DMP for reviewers and for the applicant in terms of facilitating Data Management? What do you personally think about DMP challenges and benefit? Future of DMP and changes.
8. What in this challenge is hard to solve, is the time constraints etc?
9. What could help you more to overcome this challenge(maybe a tool or software)? (user requirements) criteria to measure.
10. What's your opinion on FAIR principles?

- Closing: If you would like to check the transcription of this interview, I can send it by email and please feel free to let me know your comments.
- Thank you for your time and effort. If you know that there will be some colleagues in your organization would be willing to share their opinion can you please put me through their contact.

Appendix E

Interview Code Book

Name	Description	Files	References
Deliverables and requirements	Funder expectation in the DMP and what has to be delivered	11	509
completeness checklist	Reviewers checks the completeness of DMP	10	261
convert tool		2	2
Cost calculation		1	4
generic answers	General answered are in the paragraph	5	9
openness of data.		8	77
a mix access regime		1	3
Access to data		4	9
anonymisation		2	7
data availability		2	3
Legal issues	Environmental factors reflect reusability	6	36
GDPR		5	12
IPR		3	8
privacy issues		4	8
Security		2	4
Opening data		4	8
personal data		2	3
standard key items-FAIR		10	128
catalogue		1	2
Data archiving requirements	Reusability measures	8	31
backup routine and storage		2	4
certified repository		4	8
DANS		2	3
data volume		3	5
refuse to deposit		1	2
the core trust seal		1	1
Usage License		1	3
DIO		2	3
interoperability		8	14
machine actionable		4	4
metadata		10	72
a generic meta data standards		1	2
data reusability		10	34
dataset.		1	2
existing data		4	4
Findability		4	5
Form of data		2	6
workflow		1	1
dataCite		1	1
Dublin core		1	3
persistent identifier		2	3
The use of template	Template is an instrument provided by Fs to facilitate writing the DMP	6	16
Consistency check	Reviewers look for consistency on the DMP	6	20
Consistency of proposal with DMP	The relationship between the grant proposal and data section	6	20
data management paragraph		7	36
DM approach		1	7
data management plans		9	55
Domain specific subjectivity	DMP and RD requirements are domain specific	9	75
Challenge of Subjectivity		2	2
Good practices by discipline	The examples of good DMP and awareness to the specific domain	8	46
Astronomers data sharing protocol		2	3
data science		2	4
domain specific guidance	Funder expect more specific information in the DMP	4	16

life science		4	11
responsible medicine		1	2
social sciences-good practices		4	5
the medical center		1	1
Subject related repository		1	2
subjective meta data		4	5
vocabulary by discipline		7	18
FAIR data		5	12
fulfill the funder requirements	RDM offices and researcher try to fulfil what has been required depend c	9	34
Evaluation criteria		1	4
Expectation of the DMP	Reviewer exception when looking at the DMPs	6	17
generic guidance		1	5
Negative points on the DMP		1	2
Timeliness	Adapt the DMP versions	8	28
consistancy of plan with changes		3	5
living document		3	5
revision		3	3
unexpected changes		2	2
Main Drivers	Goals of funders by requiring the DMP from researcher	11	306
awarenss of researcher		11	61
future incentives		3	4
Challenges to improve data reusability	The environmental factors affects RDM	8	86
Deal with consortium agreement	Consortium agreement about data	3	17
future uncertainty		6	15
data maintenance after project		4	7
infrastructure	Lack of awareness about infrastructure	4	8
lack of Reproducibility	Due to project characteristics	1	1
Lack of resources to provide feedback	Funding agencies challenges to deliver feedback about DMP	4	8
RDM lack of budget		6	27
cost of data center		5	18
Shortage of DMs		3	10
Enhance quality of research data		7	12
Budget RDM in proposal		3	6
DM initiatives		1	4
DMP main goal		7	12
funding agencies		9	36
Dutch organization		7	25
NWO		1	1
EU pilot		1	5
Learning process	Mutual collaboration between funders and RDMs to find solutions	9	111
Collaboration and discussion	Ongoing conversation between RDM and Funding agencies	5	10
Clarification of concepts	Clarity of Concepts of FAIR, DMP and etc for all the stakeholders	8	13
confusion of publication with data		4	7
Lack of transparency on FAIR	RDMs claim that FAIR principles are not clear to them	8	23
Critical view about FAIR		6	15
reusability perception		2	4
The conflict of open data and GDPR		1	1
End user effort	Supports provided by university and researcher responsibility to improve	5	22
adaptability to change		4	6
Responsibilities of researchers	ZonW perspective	4	16
Institutional support	The services that RDMs provide for researchers.	8	57
Current tools and requirement of new one		7	15
data stewardship wizards.		1	3
DMPOnline		5	13

DTL		1	1
FAIR metrics		3	9
the Collectica		1	3
Institution repository		3	5
iteration to deliver the plan	The process between RDM and researcher to deliver the DMP	2	5
iterative review process		7	16

Appendix F

Simulation protocol

F.1 Goal

Analyze queries designed for the user feedback tool for the purpose of evaluation of RDMP deliverables with respect to RDMP assessment framework from the point of view of the Data management Paragraphs in the NWO proposals in the context of RDMP, providing user feedback for researchers.

F.2 Context Selection

The context is RDMP deliverables. The simulation is for an extension code for the Impacter tool designed to detect reusability metrics on the grant proposal in the Data Management section of proposals. This is an offline experiment to check reusability metrics on the document in the database of Impacter in order to design a user feedback automated tool.

F.3 Hypothesis formulation

The use of keywords extraction is a proposed approach, here we validate the two sets of treatments based on interview keywords and NLP techniques.

F.4 Dependent variables

- Acceptable metadata
- Domain-specific guidance
- Opening data

F.5 Independent variables

- Data format
- Dublin Core

- License
- Persistent Identifier
- Certified Repository
- Domain-specific

F.6 Subjects

A sample of 98 DMP were selected from 400 proposals, it is a convenient sampling because only proposals which included the DMP section in their uploaded template were selected. Not all proposals contain the paragraph 2.e Data Management (some online include paragraph 2b. Knowledge Utilisation). Additionally, users on average conduct 3.5 runs on a proposal. Only the last versions are selected, resulting in 98 DM paragraphs.

F.7 The Experiment Design

The experiment is one factor two treatments and treatments are compared against each other. The factor is the design of queries and the treatments are the first and second set of queries.

F.8 Instrumentation

Measurement instruments are queries based on the extracted keywords based on the interview and the references mentioned in the interviews. The instrument is designed based on the Assessment framework **??**

F.9 Validity Evaluation

Validation of queries are done based on a check for false positives and empty documents, which were a result of the first treatment. The second treatment used improved queries and NLP techniques. User feedback was validated based on semi-structured interviews and solutions provided by reviewers.

F.10 Population

The population is the RDMP deliverables, Data Management Plan and the data section in the proposals.

F.11 Objects of Study

The Data Management paragraph is a RDMP deliverable in the early stages of a research project and it is similar to other RDMP deliverables as there are questions about reusability of research data during and after the project.

F.12 Construction of a Sample

The sampling method is convenience sampling, since proposals are already in the database. Data required is confidential and researchers do not want to share part of their proposal when they are not using Impacter.

The sample is 98 proposals which included the 2.e Data Management section in the template. The 2.e section did not exist in the proposals before 2016. Therefore only the uploads from 2016 are included in the sample.

F.13 Measurement Design

The values are binary based on the presence or absence of keywords in the sample. Measurement tools are the queries in .json file in personal laptop and USB drive. The data is confidential therefore only results derived from Data Management Paragraphs is used either as 0 (absence) or 1 (presence). Data is interpreted and stored in excel sheets in Google drive.

F.14 Treatments

There are two treatments used in this experiment:

- First treatments are the first set of specific queries.
- The second treatment is the set of improved queries and added NLP codes.

Appendix G

Queries Designed in the First Round of Simulation

G.1 Acceptable Metadata

G.1.1 Data Format 1

```
[ " BWF ",".bwf "," MXF ",".mxf ","Matroska",
".mka"," FLAC ",".flac "," OPUS"," MXF ",".mxf "," Matroska",".mkv ",
" AutoCAD DXF v. R12",".dxf "," CAD "," GIS
","GML",".gml"," MIF/MID",".mif ",".mid "," mid ",
" mif"," gml ",".gmlGeoTIFF",".tif",".tiff "," tif"," ASCII GRID",
".asc"," Raster"," WaveFront Object","(.obj) ","  X3D ",
".x3d ", " CAQDAS ",".pdf ",".txt ",".xlsx"]
```

G.1.2 Persistent Identifier 1

```
["ORCID HTTP"," FTP"," SMTP"," SMTP"," FTP DLT",
"DLT Identifier","Identifier","ISSN"," ISSN",
" URN:NBN "," NBN identifier project","URN:NBN",
"identifier project","identifier project",
"NBNidentifier",
"NBN a Pubmed article","a PubChem chemical",
"Pubchem","Pub chem NCBI taxon"," NCBI",
"NCBItaxon ",
" CGNC gene","CGNC",
"CGNCgene ","a BOLD taxon"," BOLDtaxon",
"a GRIN taxon"," GRIN","GRINtaxon",
"Local IDs"," Local ID","LocalID","CMDI",
" Component MetaData Infrastructure"]
```

G.1.3 Dublin Core Standards 1

```
["Data cite","DATAcite","Dublin core standards",
```

```
"Dubline core meta data initiative","DCMI",
"DublinCoremetadata ",
"Accural method","Dublincore metadata ",
" accrual periodicity"," accrual periodicity ",
" Dublin core meta data ","bibilographic",
"bibilographic citation",
" citation ", " conforms To",
"date Accepted","date Copyrighted","has Format
","has Version","instructional Method ",
" Screen reader support enabled","Dubline core standards",
" Resource Type ",
"  Alternate Identifier"," Related Identifier" ,
" GeoLocation"," Funding Reference"," organisation Identifier",
"personal Identifier"]
```

G.1.4 Certified Repository 1

```
[ "4TU.Datacentrum","4TU.Datacentrum"," 4TU.Datacentrum",
"3TU.datacentrum",
" datacentrum","4tu"," https://data.4tu.nl/",
" Australian Data Archive",
" ADA "," ADA-"," Long Term Preservation at
the Bavarian State Library ",
" BABS  "," BABS-"," LZA"," BABS-LZA ",
"Banco de Información para la Investigación Aplicada
en Ciencias Sociales",
"(BIIACS)"," BIIACS "," BIIA ",
" CINES : Long-term Preservation Platform (PAC)",
" CINES","DANS: Electronic Archiving SYstem (EASY)",
" DANS "," DANS.nl ", " DANS. "," Electronic archiving system",
"DataFirst Data Portal","De Digitale Koepel"," Koepel",
"DIGITAL.CSIC" ,".CSIC "," digital.CSIC",
" DKRZ-","-LTA "," DKRZ-LTA ",
" DRUM (The Data Repository for University of Minnesota)",
" DRUM "," Edinburgh DataShare",
" Datashare "," datashare",
" edinbrugh-datashare",
" Edition Topoi Collections"," Edition.topoi.collections",
" topoi-edition.org "," Edition Topoi Collections "," EROS ",
" EROS center "," EROS.center ",
" German National Library/
Deutsche Nationalbibliothek (DNB) ",
" DNB "," Deutsche Nationalbibliothek ",
" German national library archive ",
" German national library  "," Huygens ING: eLaborate ",
" Huygens "," ING "," e Laborate ","  Huygens ",
```

```
" HZSK Repository "," HZSk. "," HZSK repository ",
" HZSK "," IDS Repository "," IDS ", " IMS Repository ",
" IMS ",
" Inter-university Consortium for
Political and Social Research ",
" LISS panel data "," LISS "," LISS-panel-data "," LISS. ",
" Meertens Institute "," Meertens "," Meertens.institute ",
" Mendeley Data ", " Menedeley- "," Menedeley ",
" Oxford Research Archive for Data (ORA-Data) ",
" ORA- "," ORA-data "," PROFILES Registry "," PROFILES ",
" PROFILES-registry ",
" PUB-Publications at Bielefeld University "," PUB- ",
" PUB-publications ",
" Permanent Service for Mean Sea Level "," psmsl ",
" RU-IIEc "," RU-IIEc "," RU-IIEc. "," https://snd.gu.se/en ",
"Scholars'Mine "," Scholarsmine "," scholar'smine ","SNDS ",
" Swedish National Data Service "," https://snd.gu.se/en ",
" Talkbank. "," talk bank "," talk/bank ",
" Tilburg University Dataverse "," DataverseNL ",
" dataverse. "," data archive UK "," UK data archive ",
" UK.data.archive ","Data-Hub ","  datahub ","  data_hub ",
" data.hub "," zenodo "," UC3 Merritt "," UC3merrit "," UC3-merrit "
," mmPort ","QualitativeDataRepository ",
"Qualitative Data Repository "
," QDR ","  CSIRO Data Access Portal "," CSIR ",
" CSIROData Access ",
" CSIRO Data Portal "," CSIRO "," CSIRO Data Access Portal ",
" CSIRODataAccessPortal"]
```

G.2 Opening Data

G.2.1 Creative Common license 1

```
["CC BY","CC0 1.0","CC BY 4.0"," CC BY NC 3.0"," MIT ",
"Apache-2.0","BSD 3-clause","BSD 2-clause","GPLv3",
"CERN OHL","TAPR OHL","Creative common","CCBY-NC-SA",
" CCBY-NC-SA","CCBYNCSA", " CCBY-NC-SA ","CCBY NC SA",
" CCBY NC SA"," CCBYNCSA"," CCBY-NCSA ",
" CC BYNCND","CC BY-NC"," CC BY4.1"," BSD",
" CC BY-ND"," CC BY-SA"]
```

G.3 Domain Specific Guidance

G.3.1 Domain Specific Repository


```
["Banco de Información para la Investigación Aplicada
en Ciencias Sociales  repository",
"ADP - Social Science Data Archives",
"Archaeology Data Service",
" ARCHE ","CISER Data Archive",
"CLARIN Center BBAW","CLARIN Center INL",
"CLARIN Portal INT","CLARIN Center IvdNT",
"Cornell Institute for Social and
Economic Research repository",
"CLARIN-UDS","CLARIN Centre Vienna",
"CLARIN-D Resource Center Leipzig",
"Clarin-PL Repository","CLARIN.SI Repository",
"CLARIND-UDS","CLARINO Bergen Repository",
"Goportis Digital Archive - German National
Library of Economics (ZBW)",
"Goportis Digital Archive - German National
Library of Science and Technology (TIB)",
"LINDAT-Clarin - Centre for Language Research
Infrastructure in the Czech Republic",
"Netherlands Institute for Sound and Vision ",
"(NISV)","National Geoscience Data Centre (NGDC)",
"Pacific and Regional Archive for Digital Sources
in Endangered Cultures (PARADISEC)",
"Roper Center for Public Opinion Research",
"The Clarin centre at the University of Copenhagen",
"Språkbanken CLARIN Repository",
"The Finnish Social Science Data Archive (FSD)",
" The ILC4CLARIN
Centre at the
Institute for
Computational Linguistics",
"ILC4CLARIN",
"Max Planck Institute for
Psycholinguistics",
"The Language Bank
of Finland",
"University Information System RUSSIA archive",
"Australian Antarctic Data Centre","CELR META-SHARE",
" Chinese Astronomical Data Center","Norwegian Marine Data
Centre (NMD)"," Banco de Información para la
Investigación Aplicada en Ciencias Sociales
(BIIACS)","CINES"," Long-term Preservation Platform (PAC)",
```

```
" CLARIN Virtual Language Observatory",
"Digital Repository of Ireland",
"DHS Data Access",
"Czech Social Science Data Archive (CSDA)",
"DARIS","FDAT","(GAMS)",
"Geisteswissenschaftliches Asset Management System",
"GESIS Data Archive for the Social Sciences",
"Inter-university Consortium for Political
and Social Research (ICPSR)",
" Irish Social Science Data Archive (ISSDA)",
"LASA","LDC Catalog","NSD's Research Data Archive",
"Odum Institute Data Archive",
"Repository of Charles University in Prague",
"SLUBArchiv","Strasbourg Astronomical Data Center (CDS)",
"TRAILS","UCD Digital Library","ISRIC WDC - Soils",
"Banco de Información para la
Investigación Aplicada en Ciencias Sociales (BIIACS)",
"www.adp.fdv.uni-lj.si/",
"arche.acdh.oeaw.ac.at"," BBAW"]
```

Appendix H

Queries Designed in the Second Round of Simulation

H.1 Acceptable Metadata

H.1.1 Data Format 2

```
["Audio","Video","Computer Aided Design",
"Geographical Information",
"Geo referenced images","Raster GIS",
"3D"," RDF ","Computer Assisted Qualitative Data Analysis",
"text"," excel sheets","image"]
```

H.1.2 Metadata 2

```
["International Standard Name Identifier (ISNI) ",
" International Standard Name Identifier (ISNI)",
" ISNI "," International Standard Name Identifier",
" ORCID iD"," online metadataURIs CURIE",
"online meta data", "data discovery",
"Standard key items",
" Interopetrable metadata",
" meta data ","Catalogue",
" Catalogue"," DIO"," doi ",
" Complete Metadata",
" actionable data",
" Machine actionable data",
" Interoperable","Interoperable metadata",
"meta data","Persistent identifier",
"Universally unique identifier"," DCAT ",
" Universally unique identifier"," DLT ",
" DCAT"," ECAT7"," DLT FAIRifier tool",
" ECAT7 data format"," URI"," DOI "," doi ",
" Digital Object Identifier",
```

```
" DigitalObjectIdentifier",
" DigitalObjectIdentifier"," URI,URN ",
" URI "," URN ", "DataCite"," Data Cite",
" Data Cite"," Uniform Resource Identifier",
"UniformResourceIdentifier", "UniformIdentifier ",
" Digital Object Identifier"]
```

H.1.3 Meta Data Standards

```
["Dublin core","datacite","metadata standards","meta data standards"]
```

H.1.4 Repository 2

```
[
    [{"LOWER":"data set","DEP":"nsubj"},{"LOWER":"stored"}],
    [{"LOWER":"dataset","DEP":"nsubj"},{"LOWER":"deposited"}],
    [{"LOWER":"data","DEP":"nsubj"},{"LOWER":"deposited"}],
    [{"LOWER":"data","DEP":"nsubj"},{"LOWER":"deposited"}],
    [{"LOWER": "accessible" },
    {"LOWER": "storage", "POS":"NOUN"}],
    [ {"LOWER": "accessible" },
    {"LOWER": "reporistory", "POS":"NOUN"}],
    [{"LOWER": "accessible" },
    {"LOWER": "archive", "POS":"NOUN"}],
    [{"LOWER":"open"}, {"LOWER": "access" },
    {"LOWER": "archive", "POS":"NOUN"}],
    [{"LOWER":"open"}, {"LOWER": "access" },
    {"LOWER": "storage", "POS":"NOUN"}],
    [{"LOWER":"open"}, {"LOWER": "access" },
    {"LOWER": "repository", "POS":"NOUN"}],
    [{"LOWER":"Public"}, {"LOWER": "available" },
    {"LOWER": "repository", "POS":"NOUN"}],
    [{"LOWER":"Publicly"}, {"LOWER": "available" },
    {"LOWER": "archive", "POS":"NOUN"}],
    [{"LOWER":"Public"}, {"LOWER": "available" },
    {"LOWER": "storage", "POS":"NOUN"}],
    [{"LOWER":"certified"},
    {"LOWER": "repository" , "POS":"NOUN"}],
    [{"LOWER":"trusted"},
    {"LOWER": "repository" , "POS":"NOUN"}],
    [{"LOWER":"FAIR"},
    {"LOWER": "repository" , "POS":"NOUN"}],
    [{"LOWER":"digital"},
    {"LOWER": "repository" , "POS":"NOUN"}],
```

```
        [{"ORTH":"UK"},{"ORTH":"DCC"}],
        [{"ORTH":"GEISIS"}],
        [{"LOWER":"figshare"}],
        [{"LOWER":"deposit"},{"LOWER":"data"},{"LOWER":"in"},
        {"LOWER":"reporistory","POS":"NOUN"}]
    ]
```

H.2 Opening Data

```
[
   [{"LOWER":"will","DEP":"aux"},{"LOWER":"be","DEP":"aux"},
   {"LOWER":"available","DEP":"acomp"}],
   [{"LOWER":"will","DEP":"aux"},{"LOWER":"be","DEP":"aux"},
   {"LOWER":"accessible","DEP":"acomp"}],
   [{"LOWER":"partly","DEP":"advmod"},{"LOWER":"accessible"}],
   [{"LOWER":"partly","DEP":"advmod"},{"LOWER":"available"}],
   [{"LOWER":"legal","POS":"NOUN"}],
   [{"ORTH":"IPR"}],
   [{"ORTH":"GDPR"}],
   [{"LOWER":"personal","DEP":"amod"}],
   [{"LOWER":"anonymising","DEP":"advmod"},
   {"LOWER":"efficiently"}],
   [{"LOWER":"privacy","DEP":"compound"},{"LOWER":"issues"}]
]
```

Appendix I

Data Analysis and Simulation Result

Paragraph ID	Validated data	Binary conversi	Total score 1	Binary score 1	POS	NEG	FALSEPOS	FALSENEG	Total score 2	Binary score 2	POS	NEG	FALSEPOS	FALSENEG
1	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
2	Generic	1	0	0	0	0	0	1	1	1	1	0	0	0
3	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
4	Reusable	1	1	1	1	0	0	0	0	0	0	0	0	1
5	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
6	Reusable	1	0	0	0	0	0	1	0	0	0	0	0	1
7	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
8	Reusable	1	0	0	0	0	0	1	0	0	0	0	0	1
9	Reusable	1	2	1	1	0	0	0	3	1	1	0	0	0
10	Reusable	1	1	1	1	0	0	0	2	1	1	0	0	0
11	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
12	No text	1	0	0	0	0	0	1	0	0	0	0	0	1
13	Reusable	1	0	0	0	0	0	1	1	1	1	0	0	0
14	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
15	Reusable	1	3	1	1	0	0	0	4	1	1	0	0	0
16	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
17	Reusable	1	3	1	1	0	0	0	6	1	1	0	0	0
18	No text	1	0	0	0	0	0	1	0	0	0	0	0	1
19	No text	1	0	0	0	0	0	1	0	0	0	0	0	1
20	No text	1	0	0	0	0	0	1	0	0	0	0	0	1
21	No text	1	0	0	0	0	0	1	0	0	0	0	0	1
22	Reusable	1	1	1	1	0	0	0	1	1	1	0	0	0
23	No text	1	0	0	0	0	0	1	0	0	0	0	0	1
24	Not Reusable	0	0	0	0	1	0	0	0	0	0	1	0	0
25	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
26	No text	1	0	0	0	0	0	1	0	0	0	0	0	1
27	No text	1	0	0	0	0	0	1	0	0	0	0	0	1
28	Not Reusable	0	0	0	0	1	0	0	0	0	0	1	0	0
29	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
30	No text	1	0	0	0	0	0	1	0	0	0	0	0	1
31	No text	1	0	0	0	0	0	1	0	0	0	0	0	1
32	Reusable	1	1	1	1	0	0	0	2	1	1	0	0	0
33	Reusable	1	1	1	1	0	0	0	3	1	1	0	0	0
34	Reusable	1	3	1	1	0	0	0	5	1	1	0	0	0
35	Reusable	1	2	1	1	0	0	0	3	1	1	0	0	0
36	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
37	Reusable	1	2	1	1	0	0	0	1	1	1	0	0	0
38	Reusable	1	2	1	1	0	0	0	5	1	1	0	0	0
39	Reusable	1	3	1	1	0	0	0	5	1	1	0	0	0
40	No text	1	0	0	0	0	0	1	0	0	0	0	0	1
41	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
42	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
43	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
44	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
45	Reusable	1	1	1	1	0	0	0	2	1	1	0	0	0
46	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
47	Reusable	1	0	0	0	0	0	1	2	1	1	0	0	0
48	Reusable	1	1	1	1	0	0	0	3	1	1	0	0	0
49	Generic	1	0	0	0	0	0	1	1	1	1	0	0	0
50	Reusable	1	1	1	1	0	0	0	1	1	1	0	0	0
51	Reusable	1	1	1	1	0	0	0	4	1	1	0	0	0
52	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
53	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
54	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
55	Reusable	1	1	1	1	0	0	0	1	1	1	0	0	0
56	Reusable	1	1	1	1	0	0	0	1	1	1	0	0	0
57	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
58	Generic	1	1	1	1	0	0	0	2	1	1	0	0	0
59	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
60	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
61	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
62	Reusable	1	1	1	1	0	0	0	0	0	0	0	0	1
63	Reusable	1	2	1	1	0	0	0	4	1	1	0	0	0
64	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
65	Generic	1	2	1	1	0	0	0	2	1	1	0	0	0
66	Reusable	1	1	1	1	0	0	0	2	1	1	0	0	0
67	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
68	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
69	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
70	Reusable	1	1	1	1	0	0	0	1	1	1	0	0	0
71	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
72	Generic	1	0	0	0	0	0	1	1	1	1	0	0	0
73	Not Reusable	0	0	0	0	1	0	0	1	1	0	0	1	0
74	Generic	1	0	0	0	0	0	1	1	1	1	0	0	0
75	Generic	1	0	0	0	0	0	1	1	1	1	0	0	0
76	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
77	Reusable	1	1	1	1	0	0	0	1	1	1	0	0	0
78	Reusable	1	1	1	1	0	0	0	0	0	0	0	0	1
79	Reusable	1	2	1	1	0	0	0	4	1	1	0	0	0
80	Reusable	1	1	1	1	0	0	0	1	1	1	0	0	0
81	Reusable	1	1	1	1	0	0	0	1	1	1	0	0	0
82	Reusable	1	3	1	1	0	0	0	6	1	1	0	0	0
83	Reusable	1	3	1	1	0	0	0	6	1	1	0	0	0
84	Reusable	1	3	1	1	0	0	0	6	1	1	0	0	0
85	Reusable	1	2	1	1	0	0	0	3	1	1	0	0	0
86	Reusable	1	1	1	1	0	0	0	2	1	1	0	0	0
87	Reusable	1	1	1	1	0	0	0	1	1	1	0	0	0
88	Generic	1	0	0	0	0	0	1	1	1	1	0	0	0

89	Reusable	1	2	1	1	0	0	0	5	1	1	0	0	0
90	Reusable	1	2	1	1	0	0	0	3	1	1	0	0	0
91	Not Reusable	0	1	1	0	0	1	0	0	0	0	1	0	0
92	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
93	Generic	1	0	0	0	0	0	1	0	0	0	0	0	1
94	Reusable	1	2	1	1	0	0	0	3	1	1	0	0	0
95	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
96	No text	0	0	0	0	1	0	0	0	0	0	1	0	0
97	No text	0	0	0	0	1	0	0	1	1	0	0	1	0
98	Reusable	1	1	1	1	0	0	0	2	1	1	0	0	0
					40	19	1	38			45	18	2	33

Bibliography

[1] C Abi Chahine et al. *Context and Keyword Extraction in Plain Text using a Graph Representation*. Tech. rep. 2009. URL: http://www.w3.org/2004/OWL/,.

[2] Mehdi Allahyari et al. *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. Tech. rep. 2017. URL: http://en.wikipedia.org/wiki/Statistics.

[3] Sean Bechhofer et al. "Why linked data is not enough for scientists". In: *Future Generation Computer Systems* 29.2 (Feb. 2013), pp. 599–611. URL: https://www-sciencedirect-com.proxy.library.uu.nl/science/article/pii/S0167739X11001439?via=ihub%20http://linkinghub.elsevier.com/retrieve/pii/S0167739X11001439.

[4] Ted Briscoe. "Introduction to Linguistics for Natural Language Processing". In: *October* (2011), pp. 1–37. DOI: 10.1007/978-3-642-80960-6{_}3. URL: https://www.cl.cam.ac.uk/teaching/1314/L100/introling.pdf.

[5] Davide Ceolin et al. "Combining User Reputation and Provenance Analysis for Trust Assessment". In: *Journal of Data and Information Quality* 7.1-2 (Jan. 2016), pp. 1–28. ISSN: 19361955. DOI: 10.1145/2818382. URL: http://dl.acm.org/citation.cfm?doid=2888577.2818382.

[6] Paolo Ceravolo et al. "Big Data Semantics". In: *Journal on Data Semantics* 7.2 (June 2018), pp. 65–85. ISSN: 1861-2032. DOI: 10.1007/s13740-018-0086-2. URL: http://link.springer.com/10.1007/s13740-018-0086-2.

[7] Authors Cheah et al. "Provenance analysis: Towards quality provenance Publication Date Provenance Analysis: Towards Quality Provenance". In: *Lawrence Berkeley National Laboratory* (2017). URL: https://cloudfront.escholarship.org/dist/prd/content/qt1kj6w1qn/qt1kj6w1qn.pdf.

[8] Guillem Closa et al. "W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed environment". In: *Computers, Environment and Urban Systems* 64 (July 2017), pp. 103–117. ISSN: 0198-9715. DOI: 10.1016/J.COMPENVURBSYS.2017.01.008. URL: https://www.sciencedirect.com/science/article/pii/S0198971517300558.

[9] S Higgins curation. "The DCC curation lifecycle model". In: *International journal of digital curation* (2008). URL: http://www.ijdc.net/index.php/ijdc/article/view/69.

[10] Fariz Darari et al. "Enabling Fine-Grained RDF Data Completeness Assessment". In: Springer, Cham, 2016, pp. 170–187. DOI: 10.1007/978-3-319-38791-8{_}10. URL: http://link.springer.com/10.1007/978-3-319-38791-8_10.

[11] Susan B Davidson and Juliana Freire. “Provenance and Scientific Workflows: Challenges and Opportunities”. In: ACM, 2008, pp. 1345–135.

[12] Ewa Deelman and Ann Chervenak. “Data Management Challenges of Data-Intensive Scientific Workflows”. In: (2008). DOI: 10.1109/CCGRID.2008.24. URL: https://www.researchgate.net/publication/4340883.

[13] L. Deleger et al. “Large-scale evaluation of automated clinical note de-identification and its impact on information extraction”. In: *Journal of the American Medical Informatics Association* 20.1 (Jan. 2013), pp. 84–94. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2012-001012. URL: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2012-001012.

[14] M Edwards. “Data quality measures for identity resolution”. In: (2018). URL: http://eprints.lancs.ac.uk/124402/.

[15] European Commission. *OPEN RESEARCH DATA IN HORIZON 2020*. Tech. rep. 2016. URL: http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf#view=fit&pagemode=none.

[16] Marco Fossati, Emilio Dorigatti, and Claudio Giuliano. *N-ary Relation Extraction for Joint T-Box and A-Box Knowledge Base Augmentation*. Tech. rep. URL: https://plus.google.com/.

[17] Matthew Gamble. “Modelling and computing the quality of scientific information on the web of data”. PhD thesis. the University of Manchester, 2014. URL: https://www.research.manchester.ac.uk/portal/files/54551459/FULL_TEXT.PDF.

[18] Matthew Gamble and Carole Goble. *Quality, Trust, and Utility of Scientific Data on the Web: Towards a Joint Model*. Tech. rep. 2011. URL: http://www.ebi.ac.uk/GOA/.

[19] Shrinath Gupta and Himanshu Kumar H.K. Singh. “A semiautomated method for classifying program analysis rules into a quality model”. In: *22nd International Conference on Program Comprehension, ICPC 2014 - Proceedings*. Hyderabad, India, 2014, pp. 266–270. ISBN: 9781450328791. DOI: 10.1145/2597008.2597808. URL: http://dx.doi.org/10.1145/2597008.2597808.

[20] Vishal Gupta and Gurpreet S Lehal Professor. “A Survey of Text Mining Techniques and Applications”. In: (). URL: http://www.jetwi.us/uploadfile/2014/1230/20141230112729939.pdf.

[21] Melissa A Haendel, Nicole A Vasilevsky, and Jacqueline A Wirz. *Community Page Dealing with Data: A Case Study on Information and Data Management Literacy*. Tech. rep. URL: http://scientificdatasharing.com/..

[22] Tom E Hardwicke et al. “Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition”. In: (2018). DOI: 10.31222/OSF.IO/39CFB. URL: https://osf.io/preprints/bitss/39cfb/.

[23] Matthew J. Harvey, Andrew McLean, and Henry S. Rzepa. “A metadata-driven approach to data repository design”. In: *Journal of Cheminformatics* 9.1 (Dec. 2017), p. 4. ISSN: 1758-2946. DOI: 10.1186/s13321-017-0190-6. URL: http://jcheminf.springeropen.com/articles/10.1186/s13321-017-0190-6.

[24] Alan R Hevner. *A Three Cycle View of Design Science Research*. Tech. rep. URL: http://community.mis.temple.edu/seminars/files/2009/10/Hevner-SJIS.pdf.

[25] Diane I Hillmann. "Cataloging & Classification Quarterly Metadata Quality: From Evaluation to Augmentation". In: 46.1 (2008), pp. 65–80. ISSN: 1544-4554. DOI: 10.1080/01639370802183008. URL: http://www.tandfonline.com/action/journalInformation?journalCode=wccq20.

[26] Indurkhya N Damerau F. *NATURAL LANGUAGE PROCESSING SECOND EDITION*. Vol 2. CRC Press, 2010. URL: https://karczmarczuk.users.greyc.fr/TEACH/TAL/Doc/Handbook%20Of%20Natural%20Language%20Processing,%20Second%20Edition%20Chapman%20&%20Hall%20Crc%20Machine%20Learning%20&%20Pattern%20Recognition%202010.pdf.

[27] Ian Jacobi et al. "Rule-Based Trust Assessment on the Semantic Web". In: *Springer* 6826 (2011), pp. 227–241. URL: http://www.foaf-project.org/%20http://www.foaf-project.org/.

[28] Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. "Understanding and predicting Web content credibility using the Content Credibility Corpus". In: *Information Processing & Management* 53.5 (Sept. 2017), pp. 1043–1061. ISSN: 0306-4573. DOI: 10.1016/J.IPM.2017.04.003. URL: https://www.sciencedirect.com/science/article/pii/S0306457316306471.

[29] Jens Lehmann et al. "DeFacto - Deep Fact Validation". In: (2012). DOI: https://doi.org/10.1007/978-3-642-35176-1{_}20. URL: http://aksw.org.

[30] Thomas Margaritopoulos et al. "A Fine-Grained Metric System for the Completeness of Metadata". In: Springer, Berlin, Heidelberg, 2009, pp. 83–94. DOI: 10.1007/978-3-642-04590-5{_}8. URL: http://link.springer.com/10.1007/978-3-642-04590-5_8.

[31] Barend Mons et al. "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud". In: *Information Services & Use* 37.1 (Mar. 2017), pp. 49–56. ISSN: 18758789. DOI: 10.3233/ISU-170824. URL: http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/ISU-170824.

[32] Amihai Motro. "Integrity = Validity + Completeness". In: (). URL: https://cs.gmu.edu/~ami/research/publications/pdf/tods89.pdf.

[33] M Natarajan. "Role of Text Mining in Information Extraction and Information Management". In: (2005). DOI: 10.14429/dbit.25.4.3663. URL: https://www.researchgate.net/publication/228673138.

[34] Martin Joseph O'connor et al. "An Open Repository Model for Acquiring Knowledge About Scientific Experiments BioPortal: Manangement and Dissemination of Scientific Ontologies View project Stanford View project An Open Repository Model for Acquiring Knowledge about Scientific Experiment". In: (). DOI: 10.1007/978-3-319-49004-5{_}49. URL: https://www.researchgate.net/publication/309694700.

[35] Stephen Pinfield, Andrew M. Cox, and Jen Smith. "Research Data Management and Libraries: Relationships, Activities, Drivers and Influences". In: *PLoS ONE* 9.12 (Dec. 2014). Ed. by Pascal Launois, e114734. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0114734. URL: http://dx.plos.org/10.1371/journal.pone.0114734.

[36] Laura Plaza and Alberto Díaz. "Using Semantic Graphs and Word Sense Disambiguation Techniques to Improve Text Summarization". In: (2011). URL: http://www.d.umn.edu/.

[37] Graham Pryor and Martin Donnelly. *58 Skilling Up to Do Data The International Journal of Digital Curation Skilling Up to Do Data: Whose Role, Whose Responsibility, Whose Career?* Tech. rep. 2009. URL: http://www.cilip.org.uk/publications/updatemagazine/archive/archive2008/june/Interview%20with.

[38] J Qin et al. "Functional and architectural requirements for metadata: Supporting discovery and management of scientific data". In: *dcpapers.dublincore.org* (). URL: http://dcpapers.dublincore.org/pubs/article/download/3660/1883.

[39] Gollam Rabby et al. "A Flexible Keyphrase Extraction Technique for Academic Literature". In: *Procedia Computer Science* 135 (Jan. 2018), pp. 553–563. ISSN: 1877-0509. DOI: 10.1016/J.PROCS.2018.08.208. URL: https://www-sciencedirect-com.proxy.library.uu.nl/science/article/pii/S1877050918314984?via%3Dihub.

[40] Arcot K. Rajasekar and Reagan W. Moore. "Data and Metadata Collections for Scientific Applications". In: Springer, Berlin, Heidelberg, 2001, pp. 72–80. DOI: 10.1007/3-540-48228-8{_}8. URL: http://link.springer.com/10.1007/3-540-48228-8_8.

[41] Leon Reznik and Sergey Edward Lyshevski. *Sensors & Transducers Data Quality Indicators Composition and Calculus: Engineering and Information Systems Approaches*. Tech. rep. 2015, pp. 140–148. URL: http://www.sensorsportal.com.

[42] Victoria L. Rubin and Yimin Chen. "Information manipulation classification theory for LIS and NLP". In: *Proceedings of the American Society for Information Science and Technology* 49.1 (2012), pp. 1–5. ISSN: 00447870. DOI: 10.1002/meet.14504901353. URL: http://doi.wiley.com/10.1002/meet.14504901353.

[43] Susanna-Assunta Sansone, Patricia Cruse, and Mark Thorley. "Comment: High-quality science requires high-quality open data infrastructure". In: (2018). DOI: 10.1038/sdata.2018.27. URL: https://doi.org/10.1101/225490.

[44] F Sartori, MA Sicilia, and N Manouselis. *Metadata and Semantic Research: Third International Conference, MTSR 2009, Milan, Italy, October 1-2, 2009. Proceedings*. 2009. URL: https://books.google.nl/books?hl=en&lr=&id=NTWcE5sdklMC&oi=fnd&pg=PP2&dq=Metadata+and+Semantic+Research+Third+International+Conference,+MTSR+2009+Milan,+Italy,+October+1-2,+2009+Proceedings&ots=bfLv5c7jOl&sig=NSBdgQKH-H3z0yXKit6lVxXkQEE.

[45] Kathleen Shearer. *Comprehensive Brief on Research Data Management Policies*. Tech. rep. 2015. URL: https://portagenetwork.ca/wp-content/uploads/2016/03/Comprehensive-Brief-on-Research-Data-Management-Policies-2015.pdf.

[46] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. “A survey of data provenance in e-science”. In: *ACM SIGMOD Record* 34.3 (Sept. 2005), p. 31. ISSN: 01635808. DOI: 10.1145/1084805.1084812. URL: http://portal.acm.org/citation.cfm?doid=1084805.1084812.

[47] Issam Souilah. “Provenance in distributed systems : a process algebraic study of provenance management and its role in establishing trust in data quality.” In: (2013). URL: http://eprints.soton.ac.uk/353288/.

[48] Besiki Stvillia and Les Gasser. “Value-based metadata quality assessment”. In: *Library & Information Science Research* 30(1) (2008), pp. 67–74. DOI: https://doi.org/10.1016/j.lisr.2007.06.006. URL: https://ac-els-cdn-com.proxy.library.uu.nl/S0740818807001314/1-s2.0-S0740818807001314-main.pdf?_tid=5ea2d094-99f2-49f3-b1c2-d6b08345b460&acdnat=1536425091_7f30e244d11162be02ac76c19b1178f9.

[49] Carol Tenopir et al. “Data Sharing by Scientists: Practices and Perceptions”. In: *PLoS ONE* 6.6 (June 2011). Ed. by Cameron Neylon, e21101. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021101. URL: http://dx.plos.org/10.1371/journal.pone.0021101.

[50] Cathy Urquhart, Hans Lehmann, and Michael D. Myers. “Putting the ’theory’ back into grounded theory: Guidelines for grounded theory studies in information systems”. In: *Information Systems Journal* 20.4 (2010), pp. 357–381. ISSN: 13501917. DOI: 10.1111/j.1365-2575.2009.00328.x.

[51] Ashish Verma et al. “Building re-usable dictionary repositories for real-world text mining”. In: (). DOI: 10.1145/1871437.1871588. URL: http://wordnet.princeton.edu/.

[52] Oskar Vikholm. “Dealing with unstructured data : A study about information quality and measurement”. In: (2015).

[53] Angus Whyte and Jonathan Tedds. *A Digital Curation Centre Briefing Paper Making the Case for Research Data Management*. Tech. rep. 2011. URL: http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx.

[54] RJ Wieringa. *Design science methodology for information systems and software engineering*. 2014. URL: https://link.springer.com/content/pdf/10.1007/978-3-662-43839-8.pdf.

[55] Manuel Wiesche and Philip W Yetton. “GROUNDED THEORY METHODOLOGY IN INFORMATION SYSTEMS RESEARCH”. In: *MIS Quarterly* 41.3 (2017), pp. 685–701. URL: http://www.misq.org.

[56] Mark D Wilkinson et al. “A design framework and exemplar metrics for FAIRness”. In: (2017). DOI: 10.1101/225490. URL: http://dx.doi.org/10.1101/225490.

[57] Mark D Wilkinson et al. “Comment: A design framework and exemplar metrics for FAIRness”. In: *Nature Publishing Group* 5 (2018). DOI: 10.25504/FAIRsharing.WWI10U. URL: https://doi.org/10.25504/FAIRsharing.WWI10U.

[58] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: 3 (Mar. 2016), p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: http://www.nature.com/articles/sdata201618.

[59] Claes. Wohlin et al. *Experimentation in software engineering : an introduction*. Kluwer Academic, 2000, p. 204. ISBN: 0792386825. URL: https://dl.acm.org/citation.cfm?id=330775.

[60] W. Wruck, M. Peuker, and C. R. A. Regenbrecht. "Data management strategies for multinational large-scale systems biology projects". In: *Briefings in Bioinformatics* 15.1 (Jan. 2014), pp. 65–78. ISSN: 1467-5463. DOI: 10.1093/bib/bbs064. URL: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs064.

[61] François Yergeau and John Cowan. *Extensible Markup Language (XML) 1.1 (Second Edition)*. Tech. rep. URL: http://www.w3.org/TR/2006/REC-xml11-20060816http://www.w3.org/TR/xml11Previousversion:http://www.w3.org/TR/2006/PER-xml11-20060614.

[62] Jing Zhao, Karthik Gomadam, and Ming Hsieh. "Predicting Missing Provenance Using Semantic Associations in Reservoir Engineering Dynamic Graph Analytics for Cyber Physical Security View project Computer Graphics View project Predicting Missing Provenance using Semantic Associations in Reservoir Engin". In: (2011). DOI: 10.1109/ICSC.2011.42. URL: http://en.wikipedia.org/wiki/Oil.