A research data management framework for a South African university-based research entity

TV Bester

orcid.org 0000-0003-2194-322X

Mini-dissertation submitted in partial fulfilment of the requirements for the degree *Master of Business Administration* at the North-West University

Supervisor: Mr JA Jordaan

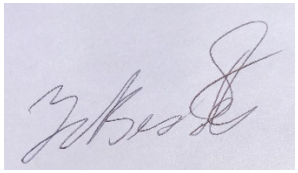
Graduation ceremony: May 2018

Student number: 10728929

DECLARATION

I, Tertius Vorster Bester, student number 10728929, declare that the research *"Research data management framework for a South African university-based research entity"* is my own work and all the references used are acknowledged in the reference list.

The research complies with the research ethical standards of the North-West University.

Signature: 

Date: 20 November 2017

ACKNOWLEDGEMENTS

A word of sincere gratitude to:

- My creator, the Lord Jesus Christ, for the mercy, the opportunity and guidance.
- Mr J Jordaan, study supervisor, for his positive support despite life challenges.
- Mrs Wilma Pretorius from the MBA office at the School of Business and Governance.
- Language editing by Ms Tanya-Lee Stewart for language editing.
- The AUTHeR-team for sharing their real-life research data challenges despite busy working schedules.
- My family for their continuous support and for spontaneously understanding that this study was something that had to take priority over so many other family activities.

I dedicate this research to all the role players associated with research data management in South Africa and the researchers, our unsung heroes that relentlessly continue to explore and describe and understand this complex world called life.

LIST OF ABBREVIATIONS

API	Application programming interface
ARPANET	Advanced Research Projects Agency Network
DataONE	Data Observation Network for Earth
DCC	Data Curation Centre
DMP	Data management plan
DOI	Digital object identifier
eRIC	eResearch – communication and infrastructure
FOI	Freedom of information
ICT	Information and communications technology
IoT	Internet of things
IR	Institutional repository
IRBs	Institutional review boards
IT	Information technology
LIS	Library and information science
METS	Metadata encoding and transmission standard
NeDICC	Network of data and information curation communities
NIH	National Institutes of Health (United States of America)
NRF	National Research Foundation
NSB	National Science Board
NSF	National Science Foundation (United States of America)
NWU	North-West University
OECD	Organisation for Economic Co-operation and Development
OMB	Office of Management and Budget
RD	Research data
RDM	Research data management

RDMP	Research data management plan
ORCID	Open researcher and contributor identification
PI	Principal investigator
PREMIS	PREservation Metadata: Implementation strategies
PURE	Prospective Urban and Rural Epidemiology
RDA	Research Data Australia
TOGAF	The Open Group Architecture Framework
UNF	Universal numeric fingerprint
URI	Uniform resource identifier
USA	United States of America

ABSTRACT

Research data management (RDM) at universities is complex and lacks standards between scientific communities. Globally, research data is viewed as a valuable commodity. The demand for optimal research RDM at universities, globally, is an area of specialisation. The push and pull for optimal research quality, open access and funders' policy frameworks require a new perspective on information technology infrastructure and processes at universities. An exploration into the challenges and realities of RDM experienced by a typical research unit at a South African university, highlighted the gap in RDM infrastructure, systems and processes within the ethical-legal context. Two questions emanated: "What are the realities of RDM within a South African university-based research entity as perceived by researchers?"; "What are the components of a RDM framework for a South Africa university-based research entity?" The research explored and described national and international theories, models and frameworks on RDM and the realities of RDM as experienced by research team members working within a research entity at a South African university. The final objective was to propose a RDM framework applicable to South African university-based research entities

Being an evolving phenomenon, a qualitative, explorative, descriptive and contextual design deemed appropriate. This research focused on a transdisciplinary research unit within a South African university as a typical research unit. Three phases were conducted in the research process, i) literature review; ii) empirical evidence; iii) proposed RDM framework. The literature review highlighted RDM as evolving, lacking a golden standard. The empirical phase led to a purposive sampling of participants based on inclusion criteria. Prior to data collection, ethical clearance was obtained followed by the goodwill permission by the primary investigator (PI). The PI acted as gatekeeper and later also as mediator. Eight (8*n*) semi-structured, individual interviews were digitally voice-recorded and transcribed where after the five steps of interpretive analysis followed. Results were confirmed through a consensus discussion with a co-coder. The researcher kept field notes and adhered to the strategies of trustworthiness.

From the empirical phase, eight themes emanated. RDM is a comprehensive system extending beyond researchers' skills. There is a need for storage and access solutions. Data security lacks despite researchers' awareness of the necessity thereof. Researchers are responsible to preserve, share and disseminate quality research data. An organisational fragmentation regarding RDM and the changing higher education landscape were described. RDM is everybody's responsibility, requiring a deliberate allocation of appropriate resources. Participants identified risks and opportunities for RDM.

The following steps led to the formulation of the framework: i) identified and selected key concepts; ii) defined and evaluated relevant concepts, theories, models; iii) added additional elements; iv) formulated the framework following the structure of why (defined research), what (listed building blocks of the RDM framework), with (assessed capabilities and defined gaps) and how (defined the RDM programme). The preliminary framework is a starting point to enable research entities to strategically build capacity to enhance RDM practices and enable South African-based universities to identify building blocks and methodologies used to accelerate the delivery of RDM services.

Key terms: Data, dataset, research data, metadata, research data management, RDM, data management, data curation, research data plan, research data curation, higher education, universities, framework.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: OVERVIEW TO RESEARCH

1.1 Introduction

This study presents a preliminary research data management (RDM) framework developed for a research entity within a South African university. RDM is a national and international challenge within the Higher Education space (Webster & Moyo, 2016. Technological advances have solved some of the current digital curation challenges within the RDM domain but are also contributing to new challenges. The rapid adoption of the Internet of things (IoT) is leading to new challenges regarding the amount and frequency of new data created. This study approaches RDM from the realities perceived by members in a research team and concludes with a proposed RDM framework. Chapter 1 provides a brief overview of related literature and argues the appropriate methodology that was followed.

1.2 Background

In 1665, the Royal Society in London created the very first scientific journal titled the Philosophical Transactions of the Royal Society (Briney, 2015). Prior to the existence of scientific journals, the method of communicating scientific results was direct communication between experts in the same field and via written letters between relevant parties. Originally, scientists were opposed to the idea of publishing their findings in scientific journals because of the intense competition between researchers. This competition caused researchers to not share their data with their peers until academic journals were generally accepted as a medium to communicate scientific findings.

At present, multiple research funders are now insisting on a data management plan that must include a section on research data preservation, sharing and reuse (Michener, 2015:1). This requirement has led to a change in the data lifecycle in that data was previously seen as only a by-product for a publication in an academic journal (see figure 1.1). The only reason for sharing data in the past was for peer review and to verify the credibility of the findings.

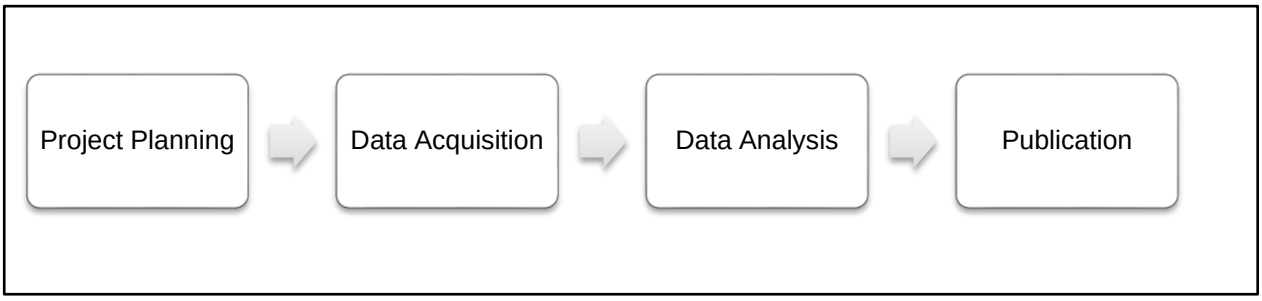


Figure 1.1: The old data lifecycle (Briney, 2015)

According to Briney (2015), data has become an important research product in its own right. New research topics and fields are emerging between the boundaries of traditional disciplines, and the

questions that investigators can address has expanded rapidly (Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, 2009:ix). A change in perception of the value of research data complements the practical need for data management. It has come to be viewed as an asset that should be managed to sustain its value (Borgman, 2012:1071; Carlson & Garritano, 2010:5; Lavoie, 2012:70). Government agencies, global organisations, funders, research institutions and researchers have gone through a dramatic change in how they view data. It is clear that more value was attached to data in the research process, and more focus was placed on the data lifecycle. Briney (2015) proposed the new data lifecycle as proposed in figure 1.2 below.

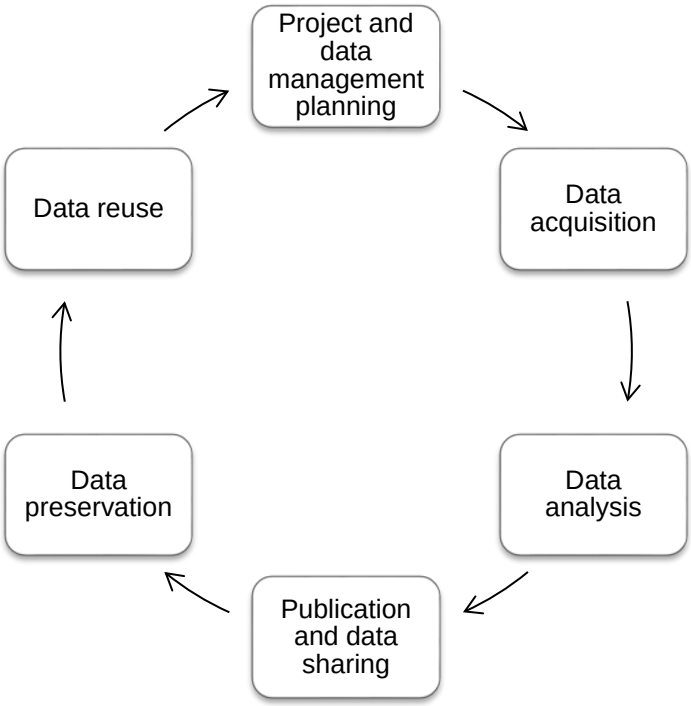


Figure 1.2: The new data lifecycle (Briney, 2015:331)

1.3 Problem statement

Research institutions are suddenly forced to put policies, procedures and services in place to assist researchers in the RDM process. Funders' expectations have caused various challenges for research institutions because of the complexities of aspects such as the lack of standards within specific research disciplines and fields, limited standards in terms of communication, and the sharing and re-use of data. Another question that was prevalent from the beginning of digital research data management pertains to who is responsible for the management of RDM within an institution. The question poses challenges regarding funding and building, and providing new capabilities within the institution. Various role players and stakeholders were confused and, to some extent, are still confused about the roles and responsibilities in terms of RDM within institutions. Due to the complexity of RDM, various role players have had to contribute

collaboratively over a long period of time in order to set up RDM services. Some of the general role players identified include the institutional research support office, library services, information technology (IT) support services, researchers and research support staff. Various role players and stakeholders disagreed about who should be responsible for RDM within institutions. Initially, library services took a leading role with their strong knowledge base pertaining to describing and preserving artefacts, but with the exponential growth of digitisation and the impact of the Internet on research, Information Technology (IT) departments also started to play a more prominent role.

One of the fundamental challenges in RDM is to conform to policies and procedures that are constantly changing and evolving. This is because some of the building blocks in RDM have not been standardised. A standardised method of describing each data object, dataset, data collection, data transformation process, workflow and storage and preservation action must be used to enable data to be described, captured, used, analysed, re-used and shared. The current prescribed method makes use of metadata to describe the above objects, workflows and processes. The main challenges are the availability of various metadata standards to describe even the most basic requirements like provenance (United Nations Archives and Records Management Section [ARMS], 2004:14). Provenance describes how an object came into existence or extant whilst PREMIS (acronym for PREservation Metadata: Implementation Strategies) is an attempt at specifying the semantic units needed to support core preservation functions (The Library of Congress, 2017). PROV, a specification that provides a vocabulary to interchange provenance information from the semantic web, is another standard that could possibly be used (Provenance Working Group, 2013). There are currently groups working on other potential standards, yet it remains a challenging task to decide which standard to follow.

Once a metadata standard has been selected to describe data, the next challenge is to standardise which ontology, thesauri, controlled vocabulary and taxonomy to use for the specified research field or area. This is an especially complex task for multi-disciplinary research since different/multiple research areas may have different vocabularies to describe the same term or observation. This poses challenges to how researchers describe research data and may force them to use multiple terms to describe the same data object. The use of multiple terms may subsequently pose problems when a dataset is shared and described on a publicly available repository.

The South African reality is also very different from countries in the European Union, the United States of America (USA) and Australia. Aspects such as financial constraints, political shifts and negative political sentiment towards research in South Africa have meant that financial resources and support for RDM is limited within research institutions. Of the 23 universities in South Africa, only the University of Pretoria (UP) currently has a formal RDM policy (Woolfrey, 2014:4). Lotter (2014:11-15) reports, however, that more universities are now getting involved in varying degrees

in analysing researcher awareness of RDM, engaging with management about RDM and taking part in learning activities related to RDM, such as conferences and workshops. As government, funders, and research institutions in South Africa become more involved in RDM, those involved in the actual research process such as researchers, research offices, ethics committees, IT departments, and libraries will have to be made aware of the potential benefits of RDM and also of RDM processes and requirements. There has already been some activity in RDM awareness and capacity-building in South Africa. The Network of Data and Information Curation Communities (NeDICC), for example, arranges seminars, workshops and a conference to promote awareness about digital (including data) curation aimed at practitioners and managers involved with digital object management, and encourages the growth of knowledge in this area (Khan *et al*., 2014:297).

The combination of the abovementioned factors places a research entity at a South African university in a vulnerable position with limited or no institutional support. No policies, procedures or guidelines are in place on an institutional level. This is also the case in support regarding IT infrastructure, the generation of metadata or training related to RDM topics or processes on an institutional level. The responsibility lies with the researchers and the research entity itself. The research entity also faces increasing expectations and pressure from global partners and funders of various projects. An additional challenge for the research entity will be to advise all research disciplines regarding controlled vocabularies and taxonomies that describe the data. A governance framework and careful planning will be required to achieve strategic research data management goals in the short term, medium term and long term.

1.4 Research questions

From the background and problem statement expounded above, the research questions asked were:

- What are the realities of RDM within a South Africa university-based research entity as perceived by researchers?
- What are the components of a RDM framework for a South Africa university-based research entity?

1.5 Research aim and objectives

The aim of this research is to facilitate RDM for research entities within South African universities aligned with international best practices. The following objectives where, however, vital to obtain the stipulated aim:

- To explore and describe national and international theories, models and frameworks on RDM by means of a literature review.

- To explore and describe the realities of RDM as experienced by research team members working within a research entity at a South African university.
- To propose a RDM framework applicable to South African university-based research entities.

1.6 Central theoretical statement

RDM is a concept used not only throughout research entities in South Africa, but also internationally. It is also used within non-academic environments. Despite various literature on RDM, there is no framework for researchers working in multi- and transdisciplinary research and within the South African university context to optimise their research. A qualitative exploration into and description of current, relevant national and international literature on RDM theories, models and frameworks as well as insight into researchers' perceptions, can assist the researcher to formulate a preliminary RDM framework that is applicable within a South African university-based research entity context.

1.7 Core concepts and definitions

The following concepts are central to this study and defined briefly. Please refer to Chapter 2 for an in-depth discussion of these concepts.

1.7.1 Research data management (RDM)

RDM in this study refers to the holistic process to identify the stakeholders, programme components, drivers, and influencing factors currently present in a research entity at a South African university. The definition also builds on definitions found in the literature which defines RDM as the organisation of data from its entry point to the research cycle, through to the dissemination and archiving of valuable results. RDM consists of many different activities and processes associated with the data lifecycle that involve the design and creation of data, storage, security, preservation, retrieval, sharing, and reuse, all of which take into consideration technical capabilities, ethical considerations, legal issues and governance frameworks. Precisely what these activities and processes are may be radically different in different contexts.

1.7.2 Framework

In this study a framework is defined as a broad overview, outline or skeleton of interlinked items, which support a particular approach to RDM in a research entity at a South African university. It serves as a flexible guide that can be modified as required by adding and deleting items. It also builds on the definition of an architecture framework, which is a foundational structure or set of structures that can be used for developing a broad range of different architectures. The framework should describe a method of designing a target state of RDM within the research entity in terms of a set of building blocks, and should moreover indicate how the building blocks fit together. It

should contain a set of tools and provide a common vocabulary. The framework should also include a list of recommended standards and compliant products that can be used to implement the building blocks. In this study the researcher presented a preliminary framework for RDM.

1.7.3 University-based research entity

An entity positioned with the primary focus on conducting research. The following characteristics best describe a university-based research entity applicable to this research: all research-related activities fit into the strategic goals and priorities of the university; has support from relevant departments, schools and faculties; income to be obtained from a variety of sources; critical mass and substantial organisational grouping; recognised core staff, equipment, space and facilities; own accounts and cost centres (North-West University [NWU], 2008:10).

1.8 Research methodology

The research methodology is presented as the design and methods.

1.8.1 Research design

The research design followed was qualitative, explorative, descriptive and contextual. This research was qualitative since there remains little known (Botma *et al.,* 2010:182) about RDM and the realities thereof as perceived by members of a research team within a South African-based research entity. A qualitative design was appropriate since it enabled the researcher to gain more insight into the realities of RDM in the real-life environment of a research team and the meaning that these team members attached to RDM expressed in words and through literature, by means of an in-depth exploration. An exploration and description were appropriate since the researcher could explore RDM within literature and as experienced by a research team within a South African-based research unit by investigating the research phenomenon and reporting the characteristics thereof (Botma *et al.,* 2010:50-51). This research was also contextual as it explored and described RDM as perceived by a research team within their natural, non-manipulated environment, namely within a research unit based within a South African-based university. This research results are therefore not generalisable but should be understood within this specific context. Please refer to 1.8.2.2 for an outline of the research context and the setting in which data collection was conducted.

1.8.2 Research method

The research method occurred within three phasesa s described next.

1.8.2.1 Phase 1: Literature review

During phase 1 the researcher aimed to explore and describe national and international theories, models and frameworks on RDM by means of a literature review.

Population

The population was comprised of all available national and international literature on RDM. This literature entailed primary and secondary sources of academic work as well as non-academic literature such as policies and guidelines, also known as grey literature.

Search strategy

Keywords were formulated and used as a search strategy on selected databases and search engines. These keywords were used for Boolean searches:

- “data”;
- “dataset”;
- “research data”;
- “metadata”;
- “Research data management” AND/OR RDM AND/OR “research data manag*”;
- “curation” and/or “data curation”;
- “Research data plan” AND/OR “Data Plan”.

The following search engines and databases were accessed through the Ferdinand Postma Library of the North-West University: LexisNexis, EBSCOhost, Emerald Insight Journals, Google Scholar, JSTOR, Sabinet Online, SAePublications, ScienceDirect, Scopus and the NWU Institutional Repository of thesis and dissertations. The literature review was conducted from January 2016 to April 2017. During the literature review, the researcher searched for titles and then scoped the abstracts for applicability. The researcher had a predetermined structure in place that assisted with the critical, analytical synthesis of the literature. Firstly, literature was assessed to understand the basic definitions used within RDM and looked at terms such as data, research data, data management and curation, metadata and datasets. Secondly, literature related to theories, models and frameworks for RDM was sourced. Thirdly, the researcher accessed literature pertaining to RDM within the university-based research context. In all instances, the literature was examined from both an international and national perspective. The literature review is presented in Chapter 2.

1.8.2.2 Phase 2: Empirical evidence

The second phase aimed to explore and describe the realities of RDM as experienced by researchers within a research entity at a South African university.

Population

The researcher was approached by a South African university-based research entity specialising in transdisciplinary research. This research entity is based within the Faculty of Health Sciences of a university in the North West Province. The population therefore referred to all the role players within a transdisciplinary research team, N36.

Context and research setting

The context is a research unit within the Faculty of Health Sciences of a North West Province-based university that was activated in 2005. The unit's focuses on transdisciplinary health research and embraces health sciences research from a holistic perspective, acknowledging that the world of health is complex and dynamic. Transdisciplinary health research entails the holistic, integrated approach to research that transcends disciplinary boundaries and focuses on the contribution in order to find solutions to complex real-life challenges in a promotive manner. It is collaborative and innovative, and ensures partnership engagement. It brings the richness of mindfully and insightfully, identifying, what is in the best interest of the person when various disciplines share and integrate skills and knowledge, to promote health and enhance quality of life (North-West University [NWU], 2017). The research unit's core competence and competitive advantages are (NWU, 2017):

- Research across disciplinary boundaries.
- Quality, relevant and applied research.
- Empowering communities.
- Research with a multidimensional impact.

The unit's staff complement for 2016-2017 is eight (8) primary researchers, nine (9) secondary researchers, and five (5) postdoctoral fellows. There are three (3) research interns, five (5) permanent support staff, and two (2) temporary support staff. By the end of 2016, the research unit had four (4) National Research Foundation (NRF) rated researchers. The unit presents three masters' degrees and one doctoral degree. In 2016, the research unit published 32 articles in accredited journals and participated in nine (9) conference oral presentations (NWU, 2017).

Research projects within the research entity are conducted in collaboration with the following faculties: Engineering, Economic and Management, Natural Sciences, Arts, Law and Theology. The research focus is on transdisciplinary health research within health promotion, consumer sciences, food security, epidemiology and chronic diseases, positive psychology and community engagement. Since 2005, this research unit has partnered with 27 high-, medium-, and low-income countries in a longitudinal research project referred to as PURE (Prospective Urban and Rural Epidemiology) and therefore houses mega data sets in different formats. The research unit's financial status is dependent on second, third and fifth sources of income. It is anticipated that the research unit will have a decrease in subsidy from 2017 to 2019 since it has activated new research programmes (NWU, 2017). The research director acted as gatekeeper and referred the researcher to the primary investigator (PI) of PURE to act as the mediator for access to understand the data sets and biobanks of this large project. The interviews were conducted during office hours within the offices of the participants, except for one participant whose office wasn't

private and therefore participated in another private office within the same building. The setting was private and comfortable. While there were background sounds as the offices are split with rhino board, participants and researcher were comfortable during data collection.

Sample, sampling technique and sample size

Participants from the participating research unit were selected by means of purposive sampling based on inclusion criteria. The inclusion criteria were:

- Being involved in transdisciplinary research projects for the past two years.
- Willingness to participate after signing voluntary informed consent.
- Willingness to participate in semi-structured individual interviews in either English or Afrikaans, that were digitally voice recorded.

After eight (n=8) individual, semi-structured interviews were conducted, sufficient data saturation (Brink *et al.,* 2010:135) was reported. When no more themes emerged during interviews, the final the research sample size was established.

Data collection

Prior to data collection, the researcher obtained ethics clearance (see Addendum A) followed by goodwill permission by the primary investigator (PI) from the participating research entity. The PI acted as gatekeeper and later also as mediator since the research director requested the researcher to explore this research theme and wanted to minimise conflict of interest. Thereafter, the researcher made appointments with participants and obtained informed consent (see Addendum B, after which he/she conducted semi-structured individual interviews (Grove et al., 2013:271) that were digitally voice-recorded. Interviews were appropriate to gain a better understanding of the research team's experiences of the realities of RDM. Each interview followed four phases as indicated by Welman *et al.* (2012:167-169), summarised as:

- *Phase 1: Preparation*

During the preparation phase, the researcher developed an interview schedule that started with demographic data, listed as: participant code, position in unit, working years' experience in unit, projects involved in (project name, role(s) in project, project start and end dates, data formats, date for data destruction). This demographic data enabled the researcher to gain a better understanding of each participants' role within the research team.

- *Phase 2: Pre-interview*

In the pre-interview phase, the researcher received the names of interesting participants from the mediator and made appointments with them to conduct the interviews.

- *Phase 3: The interview*

During each interview, the researcher followed the predetermined interview schedule (see addendum C). Interviews were conducted within the offices of the participants, on the premises of the research entity. Each interview lasted between 30 to 60 minutes and were digitally voice recorded. The researcher had sufficient time during each interview to clarify uncertainties, summarise content, and allow participants to elaborate on their answers.

- *Phase 4: Post-interview*

After each day's interviews were completed, the researcher downloaded the interviews from the digital voice-recorder to his personal, password protected laptop. The researcher also wrote down any personal, methodological and theoretical field notes obtained during the interviews. He then handed all the voice recordings to a transcriber to type.

Data analysis

Interpretive analysis of Terre Blanche, Durrheim and Kelly (*in* Botma *et al.,* 2010:226-227) was conducted with all the interviews. This type of data analysis required that the researcher engage with data analysis from an empathic understanding, with a continuous inductive and deductive reasoning of analysis and interpretation. The following five (5) steps were followed during data analysis:

- *Step 1: Familiarisation and immersion:* The researcher conducted the interviews and kept field notes to enrich the data analysis and results. The researcher also conducted a literature review and familiarised himself with RDM. In the first step of data analysis, the researcher paged through the transcriptions and started to identify links between responses, even before in-depth analysis started.
- *Step 2: Development of themes:* As the researcher studied the transcriptions, themes and sub-themes emerged and these preliminary themes were written in the participants own words. During this step, the emerging themes were clustered together as they were directed during the interview schedule, namely: what is RDM; why is RDM necessary; what RDM components are in place; what factors influence RDM and who are the major stakeholders for RDM.
- *Step 3: Coding:* During coding, the researcher linked coded data to identified themes. The unit of analysis was the verbatim words of the participants. The researcher preferred not to use any coding software, but conducted coding by copying and pasting codes into the five clusters of themes as listed in step 2 above.
- *Step 4: Elaboration:* In step 4 the researcher reviewed the coded themes and the meaning of words to identify similarities and integration of themes. During the process of

elaboration, the researcher read through the themes, spending time on ascertaining whether there were more subtle, implicit meanings in the words of the participants.

- *Step 5: Interpretation and checking:* The final step was to summarise the research themes by revisiting and interpreting each one to gain a deeper understanding, and clearly describe their meaning to the reader.

After the data analysis was completed, the researcher conducted a consensus discussion with a co-coder before the research results were presented in themes, sub-themes and categories as presented in Chapter 3.

1.8.2.3 Phase 3: Formulation of a preliminary RDM framework

The final phase of this study was to formulate a RDM framework. This framework is based on the context of a research unit within a South African-based university and is a preliminary framework since an operational framework is beyond the scope of this research. The following steps, as described by Vinz (2016), were followed in the formulation of the framework:

- *Step 1: Selected key concepts:* The key concepts were identified in the formulation of the research protocol as research data management (RDM); framework; and university-based research entity. Understanding the concept RDM required an in-depth exploration into the building blocks that made up research data.
- *Step 2: Defined and evaluated relevant concepts, theories and models:* The first phase of the methodology was a literature review (see Chapter 2), utilising a clear search strategy and exploring all available literature on models, theories and other relevant aspects associated with RDM.
- *Step 3: Adding additional elements to the framework:* The second phase of the methodology involved gaining additional insight into RDM by focusing mainly on one aspect of RDM reported minimally in the literature, namely the real-life realities of RDM as experienced by members of a typical research team. This was conducted by means of semi-structured interviews and the results are presented in Chapter 3. In step 3, the researcher analysed the similarities and differences (Vanz, 2017) identified between the literature review and interview results, and thus acquired a better insight into the context in which this framework would be functional. The combination of a literature review with interviews integrated into a framework is argued to be a contribution to the body of knowledge.
- *Step 4: Formulation of the framework:* The researcher planned to formulate a RDM framework with the following structure in mind: why, what, with, how (refer to figure 1.3).

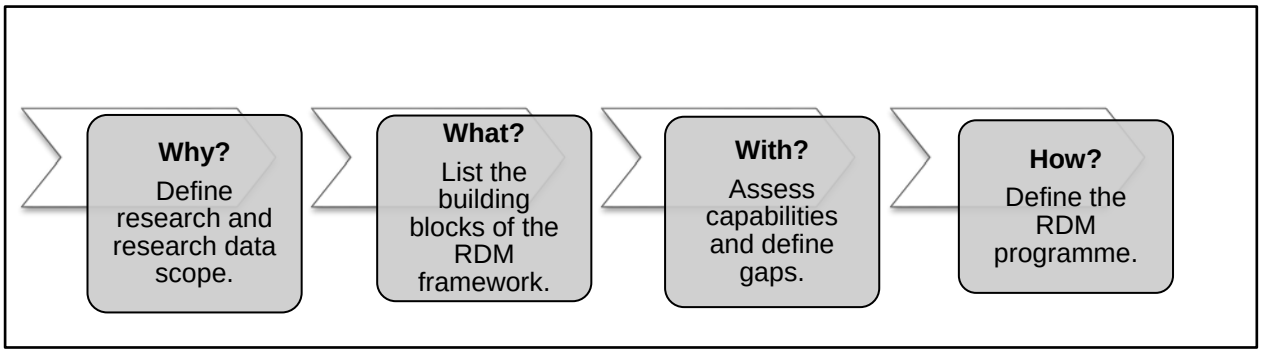


Figure 1.3: The preliminary RDM framework structure

1.9 Rigour through trustworthiness

There are multiple critiques against qualitative research. For example, some take the stance that qualitative research is too subjective and lacks rigour. Therefore, the researcher deployed strategies to improve the trustworthiness of the research. The strategies adhered to are tabled by Botma *et al*. (2010:232), and are based on the original work of Krefting (1991) and Lincoln and Guba (1985). Whilst Krefting, Guba and Lincoln initially formulated the four dominant strategies of trustworthiness, namely truth value, applicability, consistency, and neutrality, Botma et al. (2010:232) added authenticity as a fifth strategy. These strategies were applied to this study as follows:

- *Truth value was increased through credibility:* The researcher collected all the data himself and therefore engaged extensively with the literature, the participants and the construction of the framework. Data triangulation was done between the literature review and the themes from the interviews.
- *Increased applicability through transferability:* The researcher obtained the literature by means of a search strategy, and purposely selected and sampled the recruited research team for interviews.
- *Enhanced consistency through dependability:* The researcher reported the research process to provide an audit trail.
- *Neutrality through confirmability:* The researcher aimed to remain objective about this research by having regular discussions of the literature and interviews with peers and consulting information managers in the formulation of the preliminary RDM framework.

1.10 Ethical considerations

The researcher adhered to specific ethical considerations. He first obtained ethics clearance from the NWU's Faculty Research Meeting (see Addendum A) and then, before commencing with the interviews, obtained goodwill permission from the PI on behalf of the research director and informed consent from all participants. Through an appropriate research proposal, the researcher justified the significance and feasibility of the research. This research is significant to the participants and the research unit as it strengthens their knowledge of RDM and may support

them with a preliminary RDM framework. The RDM framework could also be beneficial for any research entity and the larger university. This research was feasible because the data was collected in time and there were no unrealistic financial expenses required to obtain or process data. The researcher was invited to conduct this research within the participating research entity and therefore had buy-in. Furthermore, he was able to do the literature review and interviews with both supervision and support.

The researcher aimed to preserve the anonymity of the participating research unit by not providing any information that could lead to its identification. The participants' names were replaced by codes and therefore no results can be traced back to any specific participant. During the research process, the researcher never revealed the identities of participants and their identities will remain confidential. The researcher has the only master list that can link participants with research results, but it is saved in digital format on the researcher's password protected computer.

The researcher showed respect for the research participants by allowing them at least 24 hours before an interview to decide whether they wanted to participate. Interviews were scheduled to suit the participants' programmes and were conducted in their own respective offices. This ensured that participants were not stressed about other commitments but could plan their participation. It also ensured that they did not have to travel anywhere, but remained comfortable within their offices. Interviews where done in private offices and participants had the freedom to share their experiences. The researcher confirmed with them their right to end their participation at any time.

The researcher established that this research presented as low risk for harm because the nature of the research was not personal or emotional. It is argued that the possible risks of emotional discomfort or the frustration to participate in research is outweighed by the benefit that this mini-dissertation can provide the research unit. It was predetermined with participants that their participation in this study was voluntary and did not hold any direct benefits for them. This was also clarified verbatim with the mediator and gatekeeper. The indirect benefits were however identified, namely that the research unit could access this mini-dissertation after its completion and utilise the preliminary RDM framework. Participants did not receive any reimbursement for their participation and there were not any anticipated risks associated with this research.

The researcher, along with his study supervisor, remains responsible to safeguard all the data generated through this research. All hard documents will be kept, locked away, in the office of the study supervisor for at least five years after the completion of this research. In addition, all the digital data will remain on a password-protected computer of the supervisor for five years. The researcher will hand over all the hard and digital sets of data to the supervisor after this research

has been completed. The destruction of the research data will be conducted by the supervisor, according to the NWU's record-keeping protocol.

The dissemination of the research results will be done by having the mini-dissertation available on the NWU's online repository of thesis and dissertations through the university's library. Should this research lead to an additional research publication or conference proceeding, it will be deemed as an additional research output.

The researcher declared his role throughout this research as:

- Conducting the literature review and completing the research proposal.
- Formulating the interview schedule from the literature review content.
- Obtaining ethical clearance and goodwill permission.
- Obtaining informed consent from participants.
- Conducting the individual interviews and data analysis, and participating in a consensus discussion about the research results.
- Formulating the preliminary RDM framework.
- Completing the research report by means of this mini-dissertation.

The researcher utilised the support of a gatekeeper and mediator to access the research unit and to make appointments with participants. Furthermore, the researcher outsourced the transcribing of the interviews and the transcriber signed a confidentiality agreement, as did the co-coder.

The researcher wishes to declare his conflict of interest with the research entity. The participating research unit was familiar to the researcher and approached the researcher with this research problem. However, during the early phase of the research process, the research director referred the researcher to the PI of a large project within the entity to minimise bias and conflict of interest.

1.11 Outline of mini-dissertation

The mini-dissertation's outline is described as follows:
Chapter 1 serves as an introduction to the research problem, the motivation of the research methodology, and the ethical considerations adhered to. In Chapter 1 the reader gains a better insight into RDM and why a RDM framework is required within the context of a South African university-based research entity. The research methodology is described from a qualitative perspective and the three phases of the research methods are outlined. The strategies to strengthen trustworthiness are also declared.

Chapter 2 presents the first objective of this research, namely a comprehensive literature review on RDM from both a global and national perspective. In this literature review the researcher

explored and described all available literature to obtain theories, models and frameworks for RDM.

In Chapter 3 the researcher declares the results of the interviews conducted with role players in a research team, and aims to explore and describe these participants' experiences of the realities of RDM. Chapter 3 is also aligned with phase 2 of the research and refers to the empirical evidence.

Chapter 4 presents the preliminary RDM framework as the last phase of the planned research. The researcher formulates the conclusions, evaluate the research and provides recommendations.

1.12 Summary

RDM is a growing phenomenon within the research domain. It forces all associated stakeholders to review the traditional research results communication process and requires a new view of the use and reusability of research results. There are many gaps in the RDM of a typical research entity at a South African-based university, one of which was voiced as a lack of institutional support, not because of unwillingness, but rather absence of policies. This study proposed a qualitative exploration into RDM based on a literature review, followed by semi-structured individual interviews that lead to the formulation of a RDM framework. The methodology, strategies to enhance trustworthiness and ethical considerations were declared. In Chapter 2 follows a literature review on RDM.

CHAPTER 2: LITERATURE REVIEW ON RESEARCH DATA MANAGEMENT WITHIN HIGHER EDUCATION

2.1 Introduction

To glean a comprehensive understanding of research data management, the specific components which form part of the research data management process must be thoroughly understood. A clear understanding and description of all the building blocks, as well as how they fit together is essential to appreciate the complexity and challenges regarding RDM. Building blocks can also be combined to form larger building blocks in RDM.

2.2 Defining research data

2.2.1 Data

It is important to define data as a concept and understand the main research approaches and the data lifecycle, to fully comprehend research data and the research data management process. Data is one of the cornerstones on which science is built. We must all accept that science is data and that data is science and thus provide for, and justify the need for the support of much-improved data curation (Hanson, Sugden & Alberts 2011:649). Data is distinct pieces of information, usually formatted in a special way. Strictly speaking, data is the plural of datum, a single piece of information. In practice, however, people use data as both the singular and plural form of the word (Grammarist, 2014). According to Baltzan (2013:10), data is raw facts that describe the characteristics of an event. Higman and Pinfield (2015:2) state that in a commonly-cited and wide-ranging definition, “data” is characterised as “facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors”.

2.2.2 Dataset

According to Borgman (2012:1061), the term dataset is sometimes confused with the notion of data. The integration of heterogeneous data in varying formats from diverse communities requires an improved understanding of the concept of a dataset and of key related concepts, such as format encoding, and version. A normative formal framework of such concepts is required to support the effective curation, integration, and use of shared multi-disciplinary scientific data. In order to develop framework (Renear, Sacchi & Wickett, 2010:1), the researcher reviewed the definitions of dataset found in technical documentation and the scientific literature. Four basic features can be identified as common to most definitions: grouping, content, relatedness, and purpose. A conceptual map of dataset features indicated by words and phrases in definitions in the literature was produced as indicated in Figure 2.1 on the following page:

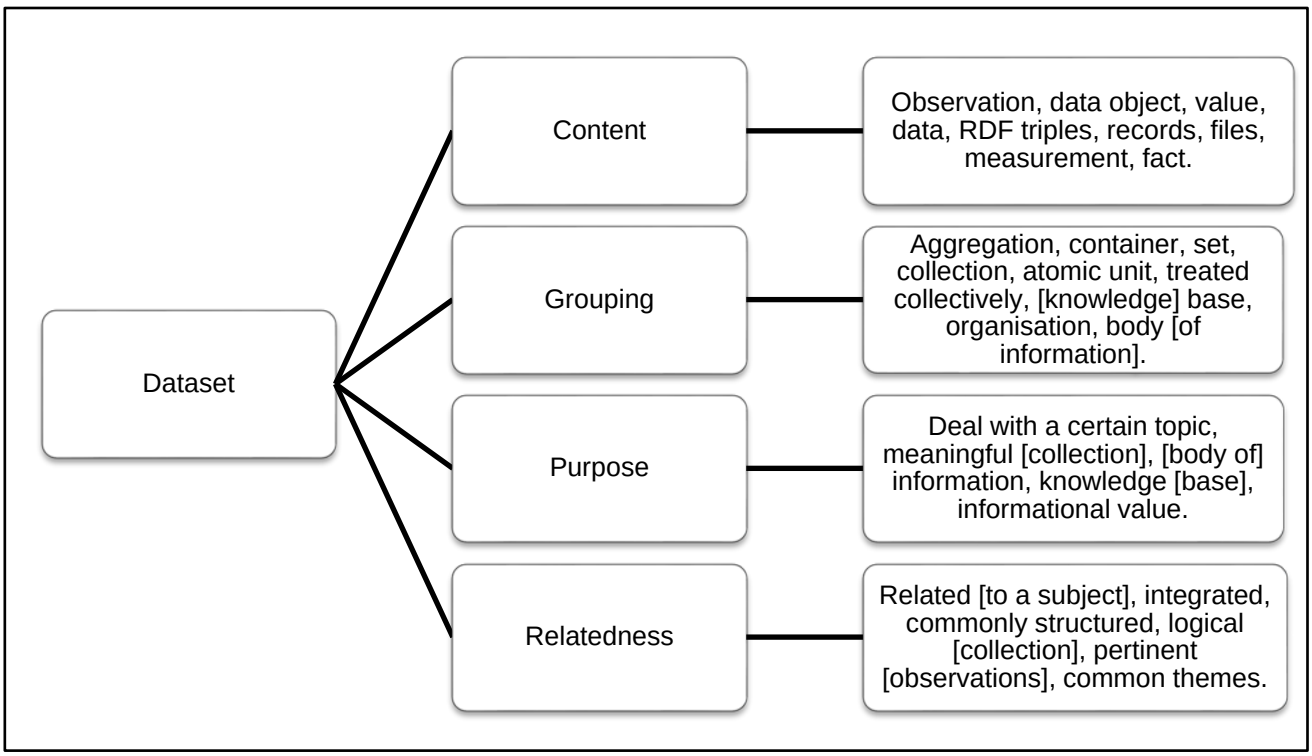


Figure 2.1: Conceptual map of dataset features indicated by words and phrases in definitions in the literature (Renear, Sacchi & Wickett 2010:1)

2.2.3 Communities and data

A specific community of interest could create a thesari, ontology, controlled vocabulary or a taxonomy to describe not only data artifacts but also processes and workflows involved in the creation of data artifacts. Some of the methods and terms of description could be available in formal publications or be agreed upon by members within a community of interest such as a reseach discipline. If no formal standards are available, the most widely used method of verification and standardisation of terms used is the peer review process. The description of data, data sets, and methods used to derive data such as workflows and processes must not be left to the researcher or investigator alone. This could lead to confusion and to challenges in verifying results and also make it difficult to re-use and share data. According to Borgman (2012:1061), an investigator who may be part of multiple overlapping communities of interest should clearly identify the appropriate community or communities, not only for funding purposes, but also for re-usability. This requirement could pose major challenges to principal investigators of multi-disciplinary projects (Borgman, 2012:1061).

2.2.4 Categories of data

The value of data can be linked to a specific moment in time and also the duration of the value of data. Specific data types can be more or less valuable with immediate effect, or they could become more valuable over time. The value of data types that cannot be easily recreated must also be recognised. Observational data cannot be repeated within a specific context in many instances. A researcher cannot go back in time and measure the temperature or air pressure again. If the data is lost or damaged, the process cannot be repeated. Some categories of data could be repeated, but it might be costly. Executing a computer simulation or model might require long and costly processing times as well as expensive data transfer costs for large data sets. Repeating such processes could prove to be costly in both monetary and human capital terms. The National Science Board (NSB) (2005) categorises data in the following categories:

- Observational data includes weather measurements and attitude surveys, either of which may be associated with specific places and time, or may involve multiple places and times.
- Computational data results from executing a computer model or simulation, whether for physics or cultural virtual reality.
- Experimental data includes results from laboratory studies such as measurements of chemical reactions, or from field experiments such as controlled behavioral studies.
- Records from government, business, and public and private life also yield valuable data for scientific, social scientific, and humanistic research.

2.2.5 Data lifecycle

An important factor in better understanding data and research data is to consider the data lifecycle. Briney (2015) states that the so-called “data lifecycle” is common within data management as it helps identify the role data plays at different points in a research project. The current data lifecycle has been in existence since the publication of research articles became the standard almost 400 years ago. The cycle starts with project planning, continues with data acquisition and analysis, and concludes with the publication of research results (see figure 1.1 in Chapter 1).

This simple view of the research process helps to frame data’s role within research. Data occupies an important place in the middle of this process, with acquisition and analysis being very data-centric activities. However, it also plays a small part in the other stages of project planning and publication. Taken as a whole, this presents a picture of data as a means to an end: namely article publication. This cycle does not reward the use of data for much beyond analysis, and so data can be seen as a by-product of research instead of an important research product, such as articles. One of the biggest indicators that data is not viewed as a research product is that data is usually lost after the end of the study (Briney, 2015:321). According to Viney *et al*. (2014:2), the

major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or stored on inaccessible storage media. In the case of papers in which authors reported the status of their data, the odds of the data being extant decreased by 17% per year. Briney (2015:321) argues that the research data lifecycle worked well for hundreds of years, but the prevalence of digital data in research means that more can be done with research data beyond losing it at the end of the project. According to Briney (2015:329), the new data lifecycle adds data sharing, data preservation, and data reuse as steps in the research process. Overall, the cycle includes: project and data management planning; data acquisition; data analysis; article publication and data sharing; data preservation; and data reuse (see figure 1.2 in Chapter 1).

Briney (2015:329) also states that this lifecycle assigns greater importance to research data than the previous cycle by making data an actual product of research. The new lifecycle is also a true cycle, in that data from a previous project can feed into a new project and cause the cycle to begin over again. Data does not default to being lost at the end of the project and instead is preserved and reused. DataONE (the abbreviation for Data Observation Network for Earth), a collaborative initiative from the USA's National Science Foundation (NSF) (2017), defines the data lifecycle as in Figure 2.2 below:

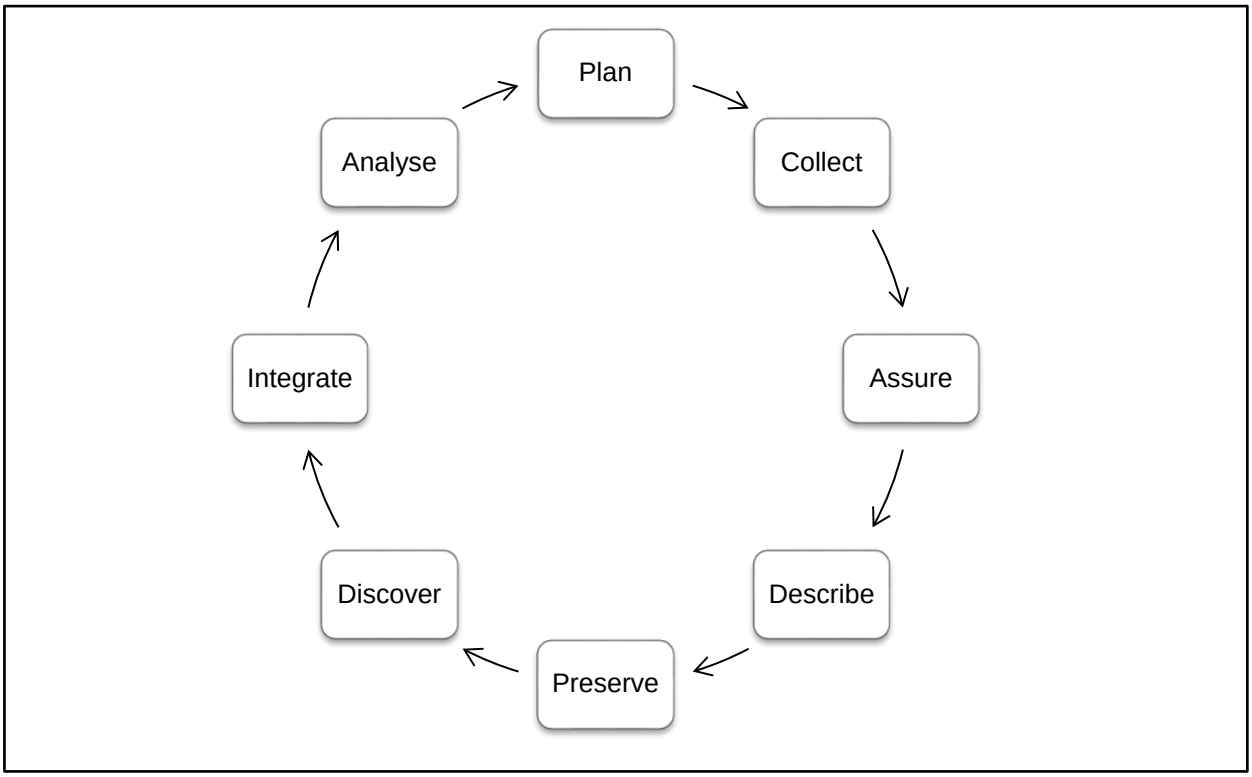


Figure 2.2: Data lifecycle according to DataONE (NSF, 2017)

The content and actions within each step of the data lifecycle according are presented in Chapter 4.

2.3 Research data

Defining research data is challenging since data by its very nature is heterogeneous. Research fields are diverse and even specific sub-fields use a huge variety of data types (Briney, 2015). According to the literature, there also seems to be conflicting views on what research data really is. Various authors and institutions view research data as final data sets necessary to verify and support research finding, thereby ignoring the context in which the data was collected and the processes followed to obtain the data. According to Briney (2015), in the USA, research data created under federal funding falls under the definition of data in the Office of Management and Budget (OMB) Circular A8-81: "*…the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analysis, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues*." This "recorded" material excludes physical objects (e.g., laboratory samples). Furthermore, Briney (2015) points out that globally, the Organisation for Economic Co-operation and Development (OECD), which consists of 34 member-nations, provides a similar definition in their Principles and Guidelines for Access to Research Data from Public Funding. Research data is defined as factual records (numerical scores, textual records, images, and sounds) used as primary sources for scientific research, and that is commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated. The term does not cover the following: laboratory notebooks, preliminary analysis, and drafts of scientific papers, plans for future research peer reviews, or personal communication with colleagues or physical objects (e.g. laboratory samples, strains of bacteria and test animals such as mice) (OECD, 2007).

Many research institutions have a much broader view of what research data is, which includes data related to the context within which the it was obtained and the processes that were followed throughout the research process. According to the Boston University Libraries (2017), research data is data that is collected, observed, or created for purposes of analyses to produce original research results. Research data can be generated for different purposes and through different processes, and can be divided into different categories. Each category may require a different type of data management plan, as listed below:

- *Observational:* data captured in real-time, usually irreplaceable. For example, sensor data, survey data, sample data, neurological images.
- *Experimental:* data from lab equipment, often reproducible, but may be expensive. For example, gene sequences, chromatograms, toroid magnetic field data.
- *Simulation:* data generated from test models where model and metadata are more important than output data. For example, climate models, economic models.

- *Derived or compiled:* data is reproducible but expensive. For example, text and data mining, compiled database, 3D models.
- *Reference or canonical:* a (static or organic) conglomeration or collection of smaller (peer-reviewed) datasets, most probably published and curated. For example, gene sequence data banks, chemical structures, or spatial data portals

Research data may therefore include all of the following formats: text or word documents, spreadsheets; laboratory notebooks, field notebooks, diaries; questionnaires, transcripts, codebooks; audiotapes, videotapes; photographs, films; test responses; slides, artefacts, specimens, samples; collection of digital objects acquired and generated during the process of research; data files; database contents, including video, audio, text, images; models, algorithms, scripts; contents of an application such as input, output, log files for analysis software, simulation software, schemas; methodologies and workflows; standard operating procedures and protocols. Furthermore, the following research records may also be important to manage research data both during and beyond the life of a project (Boston University Libraries, 2017): correspondence, including electronic mail and paper-based correspondence; project files; grant applications; ethics applications; technical reports; research reports; master lists; and signed consent forms.

2.4 Understanding metadata

Metadata refers to “data about data” (Anon, 2017). Metadata is an added value, usually required to interpret data. Metadata is not only a digital preservation technique per se, but is also required to correctly apply other techniques. For instance, emulation and migration require highly detailed metadata. With respect to digital preservation, we can define different classifications of metadata. Barateiro *et al*. (2010:13) propose the following classifications:

- *Descriptive metadata: is* information describing the content of a specific digital object. In domains like digital libraries and archives, descriptive metadata standards are broadly used.
- *Technical metadata:* focuses on the characterisation of the technological context (specific software and hardware) used in the generation of digital objects, describing, for instance, the format, format-specific technical characteristics, and so forth.
- *Structural metadata:* provides information to establish relationships between different digital objects in order to create a logical unit.
- *Preservation metadata:* are metadata elements that could be used explicitly for preservation. The PREMIS dictionary of preservation metadata relies upon the concept of Intellectual Entity, Object, Rights, Agent and Event, to prove authenticity and integrity of digital contents.
- *Rights metadata*: are used to characterise and define rights of digital contents. Some standards have been developed, like copyrightMD and METSrights.

Qin, Ball and Greenberg (2012:63) address two questions fundamental to metadata for scientific data, namely which functional requirements should metadata standards for scientific data support, and how should metadata standards for scientific data be modelled to support functional requirements? Although research in scientific data management recognises the role and importance of metadata, there are gaps between the properties of scientific data required in the e-science environment and functional requirements of metadata. The authors begin with the requirements for scientific data: which properties are expected of scientific data in the cyberinfrastructure-enabled research environment and how such expectations affect metadata modelling. Next, they seek the methodology and conceptualisation in developing metadata for scientific data. Metadata models for scientific data must not only be "scientifically" built but also easy to use and useful, and should allow the data to be cited (Smith, 2009).

Metadata for scientific data can be considered as mission-critical in scientific data discovery, use, and citation. Research conducted in the cyberinfrastructure environment needs scientific data to have the following properties (Qin, Ball & Greenberg, 2012:63):

- *Verifiable:* datasets should bear provenance metadata that allow researchers to trace them back to the raw data for quality control and data reuse purposes. The verifiability ensures the validity of research and allows researchers other than the data owner to repeat the study using the same data.
- *Interworkable:* datasets should contain sufficient metadata to facilitate data discovery, selection, aggregation or filtering, and reuse. The interworkability of data types and related metadata should be built to accommodate researchers generating data from the very beginning and throughout the research lifecycle.
- *Analysable:* datasets should be in a state that requires minimal data manipulation in order to proceed with science research. This property implies that the data management system prepares the data to be ready for analysis based on science requirements for one or more research communities. Such analysis-ready datasets would be of appropriate type and include necessary documentation and/or metadata delivered as part of infrastructure services.
- *Interoperable:* datasets should conform to standards so that they can be communicated and processed by different systems and software tools. Such interoperability ensures that the verifiability, interworkability, and analysability of data will not only transcend space and time, but also reach across the practices of research communities that are required to use or reuse them (Qin, D'Ignazio & Baldwin, 2011).

The abovementioned properties of scientific data can be translated into functional requirements for metadata used to describe and represent scientific datasets, which Greenberg (2009) summarises as resource discovery and use; data interoperability; automatic and semi-automatic metadata generation; linking of publications and underlying datasets; data/metadata quality

control and data security. Furthermore, Qin, Ball and Greenberg (2012:65) incorporate the "e-science properties" of scientific data and functional requirements for scientific metadata, and developed four (4) areas of requirements for scientific description and representation as illustrated in Figure 2.3 below:

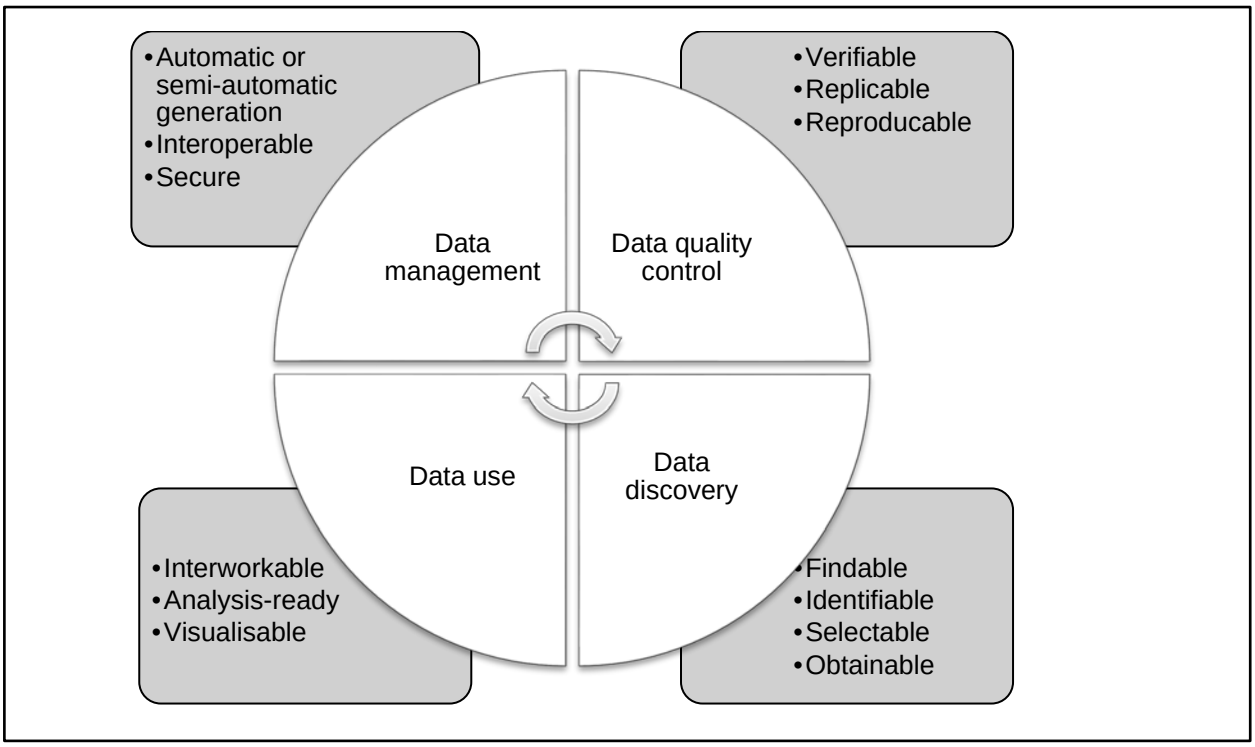


Figure 2.3: Metadata requirements for scientific data in support of data management, data quality control, data discovery, and data use (Qin, Ball &Greenberg, 2012:65)

According to Qin, Ball and Greenberg (2012:65), the architectural view is less frequently mentioned. This so-called "architectural view" of metadata views metadata attributes as building blocks that form a comprehensive representation of data or information objects. The architectural view of scientific data is illustrated in Figure 2.4 below.

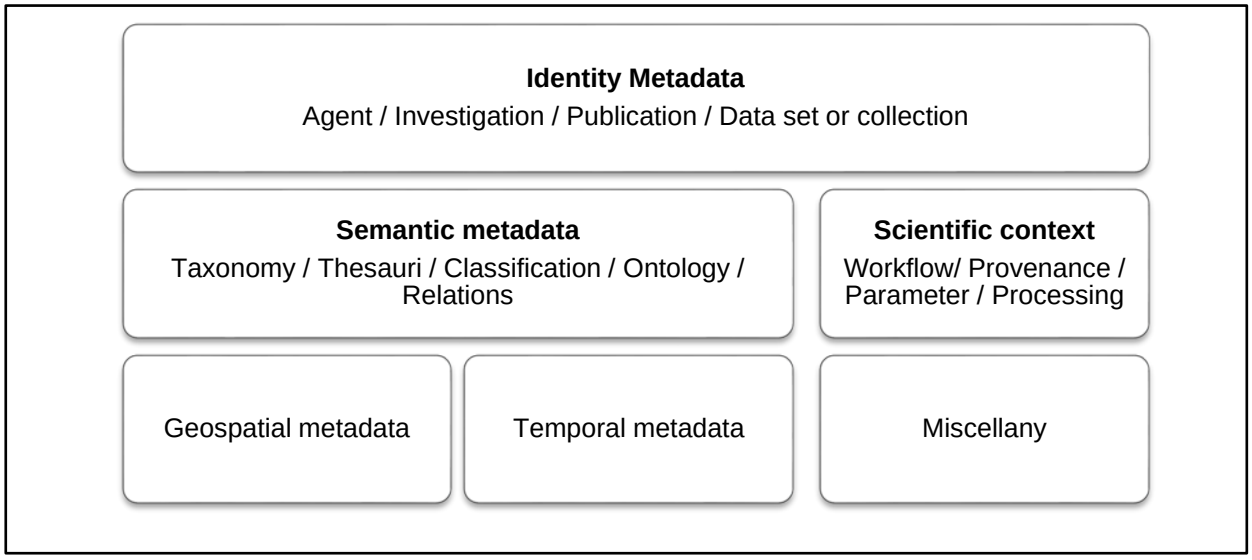


Figure 2.4: An architectural view of metadata requirements (Qin, Ball & Greenberg, 2012:65)

A common approach to model metadata schemas is entity-based modelling. This method focuses on identifying entities and relationships in a domain. Typical entities include agent or person/corporate body, event, place, and object while “is-a”, “is-part-of” and “contains” are examples of general relations (Qin, Ball & Greenberg, 2012:63). In the science metadata domain, the important entities are those related to investigation (study or project), investigator, topic and keyword, publication, sample, dataset, data file, parameter and authorisation (Matthews *et al*., 2010). Each of these entities has its own set of metadata elements for description purposes, for example, a person entity has a name, role, and affiliation, contact information and may be identified by a standard identifier system such as ORCID and ResearcherID. An event entity has a name, time and place of occurrence, description, type, keywords, and other attributes. It would also have an identifier conforming to some standard system. The unique identifiers may come from standard identifier systems such as Digital Object Identifier (DOI), Uniform Resource Identifier (URI), Handle System, and/or Universal Numeric Fingerprint (UNF). Another example is data citation in which only identity metadata is required in the actual citation, with the identifier or identifiers pointing to where the dataset or data collection is located. The DataCite Metadata Schema, a Dublin Core compliant metadata schema, is designed for just this purpose. Identity metadata forms the basis for making data identifiable and readily findable when such identities are known (Qin, Ball & Greenberg, 2012:66).

Semantic metadata for scientific data plays two roles: firstly, as the subject identifier for data and, secondly, as the subject grouping criteria and linking mechanism for data with similar subject content. Scientific context, geospatial, and temporal metadata fulfil the requirements for data verifiability, replicability, and reproducibility. The miscellany metadata includes elements that do

not fit into any of the other blocks: file size, storage medium and dissemination medium (for offline data) are typical examples (Qin, Ball & Greenberg, 2012:66).

Based on a typical research life cycle; Qin, Ball and Greenberg (2012:68) defined 10 data user tasks as presented in Table 2.1 below. The authors view the metadata required to perform the user tasks from two perspectives: the function embedded in a type metadata, and the architectural building block of metadata attributes needed to support the metadata function. This sense, they divide data user tasks into:

- *Generic tasks:* discovery, identity, selection, obtain.
- *Scientific tasks:* verify, analyse.
- *Data tasks:* manage, archive.
- *Dissemination tasks:* publish, cite.

Table 2.1: Mapping data user tasks with metadata functions and architectural building blocks (Qin, Ball & Greenberg 2012:66)

Data user tasks	Metadata function	Architectural
Discover	Descriptive metadata	Identity and semantic metadata
Identify	Descriptive metadata	Identity metadata
Select	Descriptive, technical metadata	Identity, semantic, scientific context, geospatial, temporal, miscellany metadata
Obtain	Descriptive metadata	Identity metadata
Verify	Descriptive metadata	Scientific context metadata
Analyse		Scientific context, geospatial, and temporal metadata
Manage	Descriptive, administrative, structural, and technical metadata	Identity, semantic, scientific context, geospatial, temporal, miscellany metadata
Archive	Descriptive, administrative, structural and technical metadata	Identity, semantic, scientific context, geospatial, temporal, miscellany metadata
Publish	Descriptive metadata	Identity, semantic, scientific context, geospatial, and temporal metadata
Cite	Descriptive metadata	Identity metadata

2.5 Defining digital preservation techniques

Barateiro *et al.* (2010:10) presents the most relevant techniques and strategies that can be used to handle digital preservation vulnerabilities and threats, increasing the probability of matching digital preservation requirements. These techniques are listed in Table 2.2 below:

Table 2.2: Digital preservation techniques according to Barateiro *et al.* (2010:10)

Redundancy	If data is stored in a single component, it will be lost when that component fails, which is very likely to happen in the long term. Therefore, digital preservation systems can take advantage of a basic attribute of digital information: it can be copied without any loss of information. This means that several copies of the data can be stored across many components (Barateiro *et al*., 2010:10). One of the major problems of systems with replication capabilities is the maintenance of coherence between copies.
Migration	The goal of migration is to keep digital objects in recent media formats. Lossless migrations of data maintain exactly the same contents as the original version, whilst loss migrations may imply the loss of some information in the process. Several techniques can be used in migration processes (Barateiro *et al*., 2010:10): • Analogue media: converting digital media back to analogue formats, such as paper or microfilm. • Version update: converting data from an old format version to a new one. • Conversion to other formats: some formats are closed, making conversion possible only when software vendors support the conversion between the format's versions. The ideal is to convert to an open format. • Normalisation: In the scope of digital preservation, normalisation consists of reducing the number of different formats, in order to reduce the complexity of migration tasks.
Emulation	Emulation is the simulation of the original hardware and/or software conditions of execution for which the information objects were initially conceived (production environment) in more recent systems. This can sometimes be a very complex strategy to implement, since it requires not only the preservation of the original objects but also detailed knowledge of the original systems (Barateiro *et al.,* 2010:10).
Refreshing	The goal of the refreshing technique is to keep the system infrastructure updated with the most recent technology, consisting of the replacement of components by more recent ones. The refreshing of components can be used to prevent failures and obsolescence in the infrastructure's components. Media refreshing for more reliable, durable and less expensive technology can also be seen as a viable option (Barateiro *et al*., 2010:11).

Diversity	System failures are far from independent. Diversifying the properties of the components can limit the number of simultaneous failures in the system and can be used to design a replication strategy, which is more likely to survive a large correlated failure, as in the case of a worm outbreak. Important properties that can be diversified are (Barateiro *et al.,* 2010:11) physical location; software; hardware; administration; storage; funding.
Inertia	A system that works quickly also fails quickly, especially if it is subject to deliberate attack. Since digital preservation systems usually do not require speed, they can be designed to not change rapidly. Consequently, the system is less likely to fail abruptly. A good example is limiting the rate at which an administrator can delete files. The system should implement this strategy through rate-limiting techniques, which allows it to limit the rate at which an attacker can make progress (Barateiro *et al*., 2010:12).
Auditing	Auditing supports the detection of latent faults, allowing the system to recover faster and reduces the chance of losses. For example, faults that cause data loss may only be detected when data are accessed. This can be done by auditing the system periodically. Auditing is especially important in digital preservation systems with few data accesses. When data ingestion is performed by a third party, it may also be crucial to audit these systems to check if the data are property ingested (Barateiro *et al*., 2010:12).

2.6 Defining research data management

Research data management (RDM) is defined by Whyte and Tedds (2011:2) as ‘‘*the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results*’’. RDM consists of several different activities and processes associated with the data lifecycle, involving the design and creation of data, storage, security, preservation, retrieval, sharing, and reuse, all taking into consideration the technical capabilities, ethical considerations, legal issues and governance frameworks. Precisely what these are may be radically different in different contexts (Cox & Pinfield, 2014:300). RDM is part of the research process, aims to make the research process as efficient as possible, and meet expectations and requirements of the university, research funders, and legislation (University of Leicester, 2017). It is concerned with how one creates data and plans for its use; organises, structures and names data; keeps secure, provides access, stores and backs it up; finds information resources, and shares it with collaborators and more broadly, how it publishes data and gets cited (University of Leicester, 2017).

On a more practical level, Briney (2015) states that RDM involves many practices. Data management includes data management planning, documenting your data, organising it, improving analysis procedures, securing sensitive data property, having adequate storage and

backups during a project, sharing data effectively, and finding data for reuse in new projects. Such a wide range of practices means that data management is something you do before the start of a research project, during the project, and after the project's completion.

The Mercury Project Solutions (2013:2) outlines the key steps research institutions are required to consider in research data management before, during and after a research project. The steps are divided into pre-research, research and post-research steps. See Figure 2.5 below.

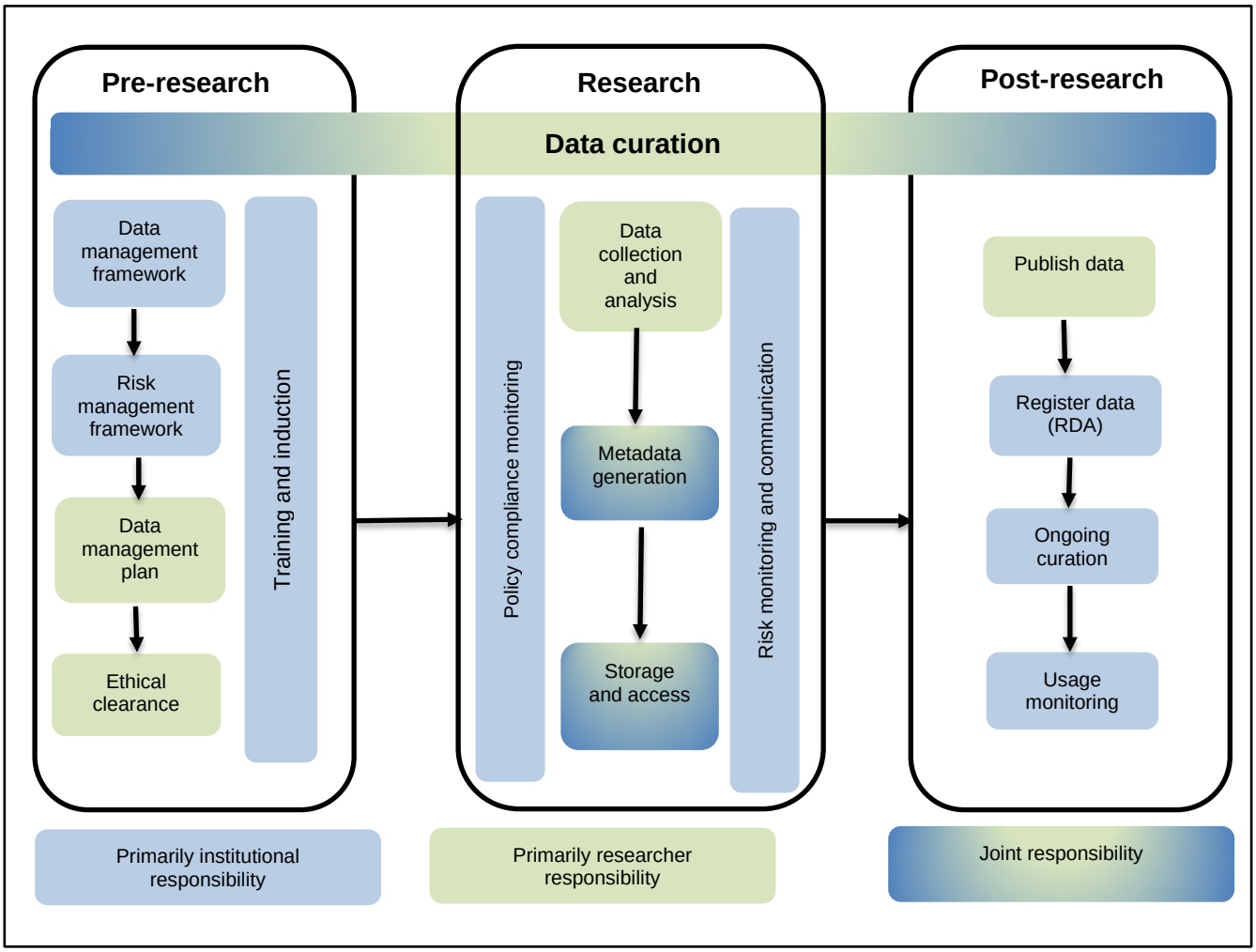


Figure 2.5: Key steps in research data management (Mercury Project Solutions, 2013:2)

Each of the building blocks forms an important part of a holistic approach to research data management on an institutional level. Each of the components will be further described and analysed from the literature.

2.6.1 Research data curation

Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle. The active management of research data reduces threats to its long-term

research value and mitigates the risk of digital obsolescence. As well as reducing duplication of effort in research data creation, curation enhances the long-term value of existing data by making it available for further high-quality research (Digital Curator Centre [DCC], 2011a). With traditional publication, most curation activities occur at the end of the research cycle. In contrast, digital curation of data is characterised by activities planned for from the outset and occurring throughout the data lifecycle (Mercury Project Solutions, 2013:2). These will include, for example, the processes outlined in figure 2.6 (below).

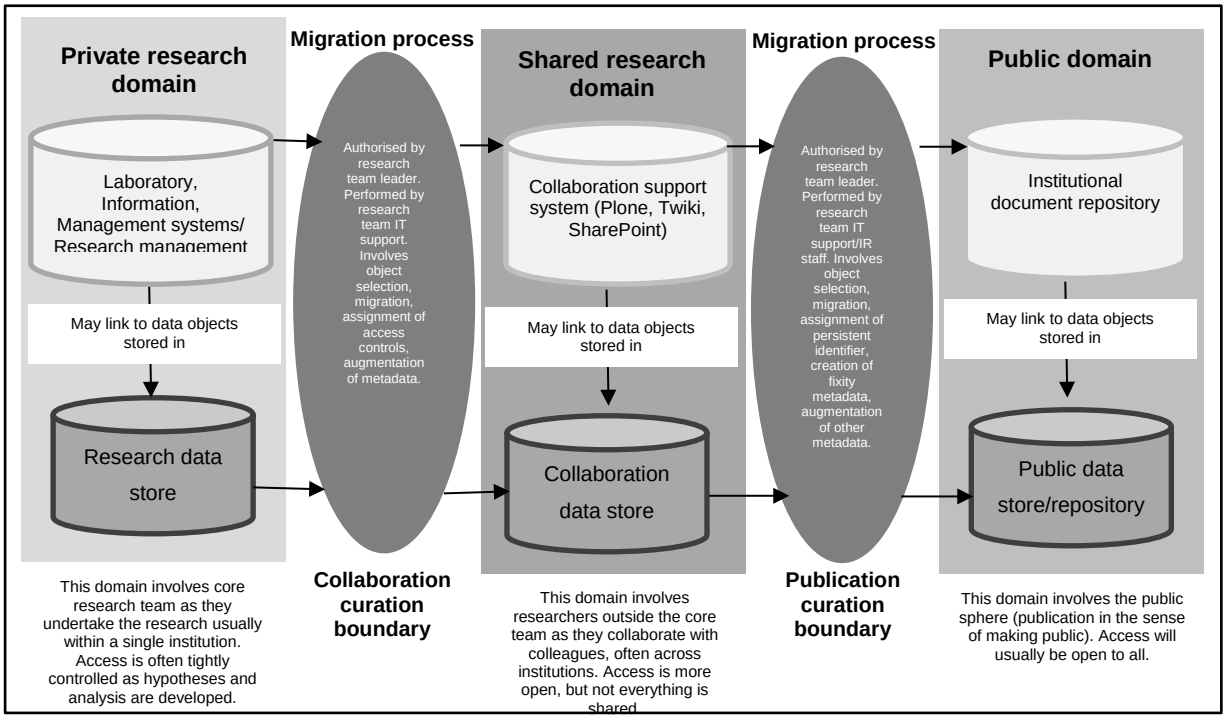


Figure 2.6: Research data curation continuum (Mercury Project Solutions, 2013:3)

Figure 2.6 illustrates that research data must be curated and managed in the private research domain, shared research domain and the public domain. This is aligned with the more expansive new data lifecycle. The Digital Curation Centre (DCC) is an internationally-recognised centre of expertise in digital curation focused on building capability and skills for research data management. The DCC provides expert advice and practical help to research organisations wanting to store, manage, protect and share digital research data (DCC, 2011b). Digital curation and data preservation are ongoing processes, requiring considerable thought and the investment of adequate time and resources. One must be aware of, and undertake, actions to promote curation and preservation throughout the data lifecycle. The digital curation lifecycle comprises of the following steps (DCC, 2011a):

- *Conceptualise:* conceive and plan the creation of digital objects, including data capture methods and storage options.

- *Create:* produce digital objects and assign administrative, descriptive, structural and technical archival metadata.
- *Access and use:* ensure that designated users can easily access digital objects on a day-to-day basis. Some digital objects may be publicly available, whilst others may be password protected.
- *Appraise and select:* evaluate digital objects and select those requiring long-term preservation. Adhere to documented guidance, policies and legal requirements.
- *Dispose:* rid systems of digital objects not selected for long-term curation and preservation. Documented guidance, policies and legal requirements may require the secure destruction of these objects.
- *Ingest:* transfer digital objects to an archive, trusted digital repository, data centre or similar, again adhering to documented guidance, policies and legal requirements.
- *Preservation action:* undertake actions to ensure the long-term preservation and retention of the authoritative nature of digital objects
- *Reappraise:* return digital objects that fail validation procedures for further appraisal and reselection.
- *Store:* keep the data in a secure manner as outlined by relevant standards.
- *Access and reuse:* ensure that data are accessible to designated users for first time use and reuse. Some material may be publicly available, whilst other data may be password protected.
- *Transform:* create new digital objects from the original by, for example, migration into a different form.

The Curation Lifecycle Model from the DCC provides a graphical, high-level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt through to iterative curation cycle as depicted in Figure 2.7 below:

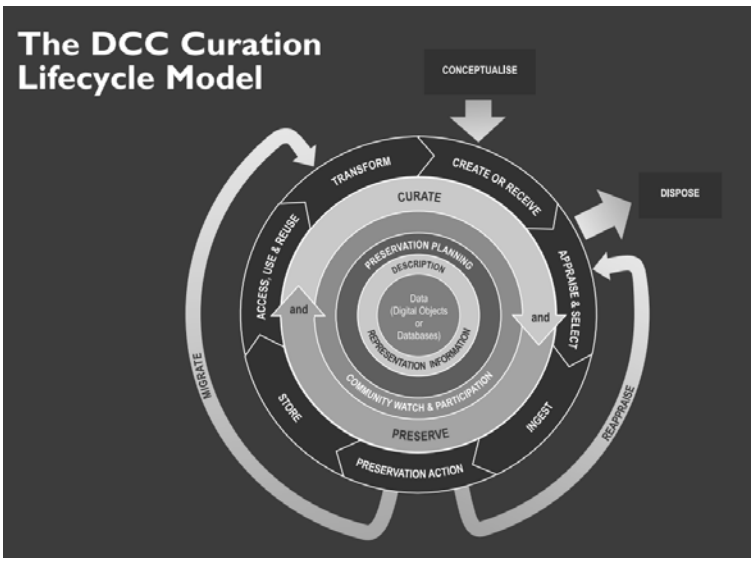


Figure 2.7: Key elements of the DCC curation lifecycle model (DCC, 2011c)

Data is at the centre of the Curation Lifecycle. The next layers represent full lifecycle actions and include description and representation information; preservation planning; community watch and participation; curate and preserve. The outer layer represents the sequential actions which consist of the following actions: create or receive; appraise and select; ingest; preservation action; store; access, use and reuse; transform. Furthermore, the following occasional actions are presented as to dispose, reappraise and to migrate.

2.6.2 Data management framework

According to Mercury Project Solutions (2013:3), the research institution is responsible for providing an adequate research data management framework that can provide the basic elements required within an institutional context to support effective research data management. These elements comprise of four categories:

- *Institutional policy and procedures:* these should be up to date, addressing data-related issues, and be publicised to all those who have a data creation and/or management role.
- *IT infrastructure:* the hardware, software and other facilities which underpin data-related activities, as well as identity management and access control.
- *Support services:* people and other means of providing advice and support, such as web-pages.
- *Metadata management:* so that data records can be used for both internal and external purposes.

Pinfield *et al*. (2014:8) identified several key components of an institutional RDM programme including the following:

- *Strategies:* defining the overarching vision for research data management within the institution and how it relates to the institutional mission and priorities, and outlined major developmental goals and principles which inform activity.
- *Policies:* specifying how the strategies are to be operationalised through regular procedures, including an RDM policy and a set of complementary policy frameworks covering issues such as intellectual property rights and openness that may be relevant.
- *Guidelines:* providing detail on how the policies will be implemented, often written from the point of view of a particular user group (such as those within a particular disciplinary area) and defining specific activities, and roles and responsibilities.
- *Processes:* specifying and regulating activities within the research data life-cycle, including research data management planning for individual projects, data processing, ingesting data into central systems, selecting data for preservation, etc., and involving the use of standards and standardised procedures wherever possible.
- *Technologies:* underpinning processes with technical implementations, including data repositories and networking infrastructures allowing for storage and transport of data.
- *Services:* enabling end-user access to systems and providing support for research data life-cycle activities (including supporting the creation of data management plans, providing skills training, and delivering helpdesk services).

One of the respondents in the work from Pinfield *et al*. (2014:9) summarised a framework as follows: "*You want some framework, so you do need a policy, and a road map, a sense of where you are going. You do need to inform longer term strategic investment, particularly in key areas like storage, equipment, and staffing for services. And you need something which outlines roles and responsibilities in terms of research data management, and this has to in some way be endorsed from the top.*"

2.6.3 Risk management plan

Digital preservation aims to keep digital objects accessible over long periods of time, ensuring the authenticity and integrity of these digital objects. In such complex environments, Risk Management is a key factor in assuring the normal behaviour of systems over time (Barateiro *et al*., 2010:4). A number of standards have been developed worldwide to help organisations implement risk management systematically and effectively. These standards seek to establish a common view on frameworks, processes and practice, and are generally set by recognised international standards bodies or by industry groups. A commonly used standard is ISO 31000 2009 – Risk Management Principles and Guidelines (Institute of Risk Management, 2017). It aims to define the principles and implementation of risk management to control the behaviour of an organisation with regard to risk, and is based on the principle that Risk Management is a process operating at different levels, as shown in figure 2.8 below (Barateiro *et al*., 2010:6).

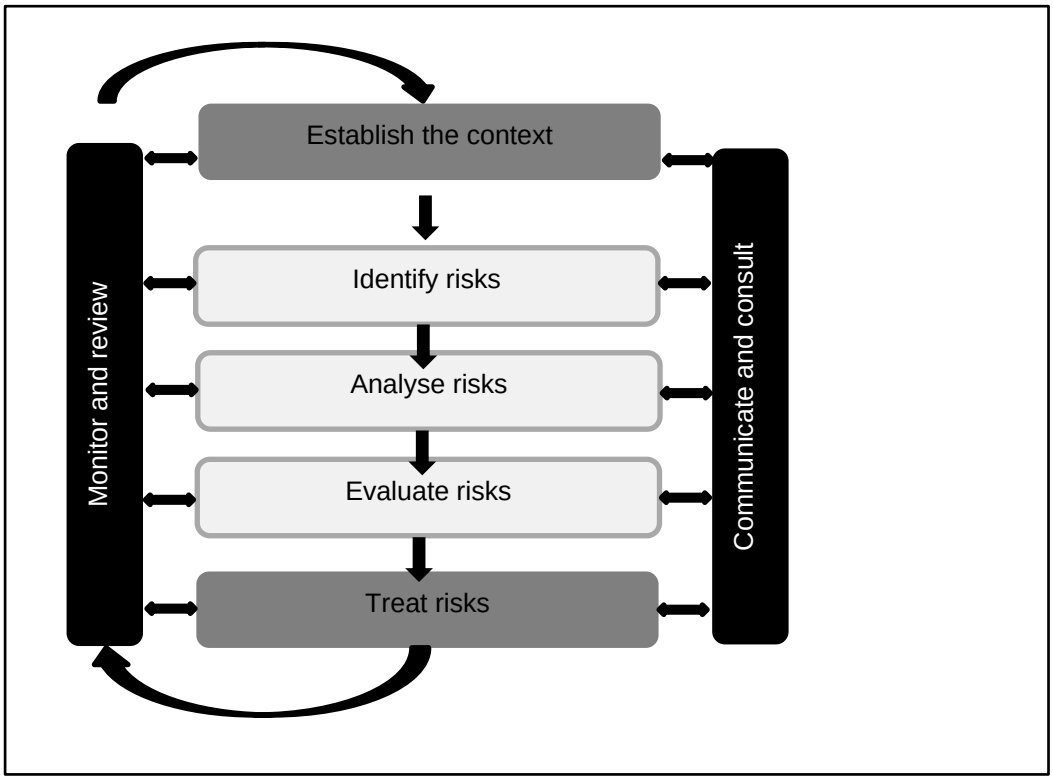


Figure 2.8: Risk management process (Institute of Risk Management, 2017)

The risk management process includes the limitation of the context, risk assessment (identification, analysis and evaluation of risks) and risk treatment. This process requires a continuous monitor and review activity to audit the behaviour of the whole environment allowing, for instance, the identification and treatment of an unexpected vulnerability (Barateiro *et al*., 2010:6). Barateiro *et al*. (2010:7) furthermore proposed a three (3) step framework based on risk management that can be used to assess existing solutions which comprises of the following steps:

- Establish digital preservation requirements (context and strategic objectives).
- Identify digital preservation vulnerabilities and threats.
- Address digital preservation threats and vulnerabilities (treat risks).

The first digital requirement is reliability. Digital preservation requires that a copy of any preserved digital object survives over the system's lifetime. The second digital preservation requirement is to enable the user to assess if the information is sufficiently trustworthy. This requires the authenticity assurance of digital objects. The provenance of digital objects should be required, especially their creator or the entity responsible for them. It is crucial to ensure the integrity of digital objects and guarantee that their information content has not been modified. Thirdly, digital preservation requires that future consumers are able to obtain the preserved information as its creators intended, dealing with obsolescence threats. This requirement involves several challenges, since a digital object to be explored, requires a technological context defined by specific software, and in some cases, even by specific hardware. Finally, dynamic collections and

environments for digital preservation require technical scalability to face technology evolution that allow, for instance, the addition of new components through incremental updates (Barateiro *et al*., 2010:7).

Table 2.3 below presents a taxonomy, which is based on the Risk Management terminology, considering vulnerabilities and threats to digital preservation. Vulnerabilities are weaknesses (potential points of failure) in the environment and threats are events that affect the normal behaviour.

Table 2.3: Taxonomy of vulnerabilities and threats to digital preservation (Barateiro *et al*., 2010:9)

Vulnerabilities	Process	Software faults Software obsolescence
	Data	Media faults Media obsolescence
	Infrastructure	Hardware faults Hardware obsolescence Communication faults Network service failures
Threats	Disasters	Natural disasters Human operational errors
	Attacks	Internal attacks External attacks
	Management	Economic failures Organisational failures
	Legislation	Legislative changes Legal requirements

The digital preservation techniques described in paragraph 2.5 will be applied to moderate the various threats and vulnerabilities identified in table 2.4 below.

Table 2.4: Addressing digital preservation threats and vulnerabilities (Barateiro *et al*., 2010:14)

Threats and vulnerabilities			Techniques							
			Redundancy	Migration	Emulation	Refreshing	Diversity	Inertia	Metadata	Auditing
Vulnerabilities	Data	Media faults	R	-	-	r	-	-	R	R
		Media obsolescence	-	r	r	-	-	-	R	R
	Infrastructure	Hardware faults	-	-	-	r	r	-	-	R
		Hardware obsolescence	-	-	-	r	r	-	-	R
		Communication faults	-	-	-	r	r	-	-	R
		Network service faults	-	-	-	r	r	-	-	R
	Process	Software faults	-	-	-	r	r	-	-	R
		Software obsolescence	-	-	-	r	r	-	-	R
Threats	Disasters	Natural disasters	-	-	-	-	r	-	-	-
		Human operational errors	-	-	-	-	r	r	R	R
	Attacks	Internal attack	-	-	-	-	r	r	R	R
		External attack	-	-	-	-	r	r	R	R
	Management	Economic failures	-	-	-	-	r	-	-	R
		Organisational failures	-	-	-	-	r	-	-	R
	Legislation	legislative changes	-	-	-	-	r	-	r	-

r=: reduces the risk of the threat/vulnerability; R=: required for recovery; -: does not fit

2.6.4 Data management plan

Funding bodies increasingly require grant-holders to develop and implement Data Management and Sharing Plans (DMPs). Plans typically state what data will be created and how, and outline the plans for sharing and preservation, noting what is appropriate given the nature of the data and any restrictions that may need to be applied (DCC, 2011d). Development of a RDM Plan is a critical aspect of the pre-research stage. It lays out what data will be created, what policies will apply to the data, who will own and have access to the data, what research data management practices will be used, what facilities and equipment will be required, and who will be responsible for each of these activities. It will include activities such as (Mercury Project Solutions, 2013:4) data organisation and storage; metadata standards and guidelines; backups; archiving for long-term preservation; version control and derived data products; data sharing or publishing intentions; ensuring security of confidential data; data synchronisation and governance, roles and

responsibilities. The plan usually defines all RDM-related activities during and after the research activity. The following sections should be described in a RDM plan:

2.6.5 Ethical clearance

Sharing of research data that relates to people can often be achieved using a combination of obtained consent, anonymising data and regulating data access. Research data, even sensitive and confidential data, can be shared ethically and legally if researchers pay attention to four important aspects from the beginning of their research (Mercury Project Solutions, 2013:4): namely including provision for data sharing when gaining informed consent; protecting people's identities by anonymising data where needed; consider controlling access to data; applying an appropriate license.

2.6.6 Training and induction

Effective implementation of an institutional research data management framework requires that all institutional staff receive adequate training. All staff involved in the research project will benefit from up-to-date training. Furthermore, for project specific issues such as risk management and the implementation of the research data management plan, all researchers and support staff will need to be inducted. Inductions take place not only at project commencement, but also be presented to any personnel getting involved during the project. For longer projects refresher training and induction should be considered (Mercury Project Solutions, 2013:5).

2.6.7 Policy compliance monitoring

Acknowledging the policies and guidelines as defined in the institution's research data management framework at the start of a research project is of little use in itself. Demonstrating compliance through review or audit frameworks allows non-compliance to be identified early and corrective action to be taken. This allows the organisation to respond to compliance breaches in a systematic, rather than ad hoc, fashion. Policy compliance is normally dealt with as part of the review of the research data management plan (Mercury Project Solutions, 2013:5).

2.6.8 Risk monitoring and communication

Continuous monitoring and review are vital components of an effective risk management process. In terms of the research project, review of risk would normally be incorporated into the review of the research data management plan. The primary purpose of monitoring and review is to determine whether risks still exist, whether new risks have emerged, and to reassess the risk priorities (Mercury Project Solutions, 2013:5).

2.6.9 Research data collection and analysis

Throughout the research cycle data will be collected. Data analysis will generate derived data and in many cases data will be shared between researchers and institutions for collaboration

purposes. Data at this point may be made available to collaborators, as specified in the research data management plan (Mercury Project Solutions, 2013:5).

2.6.10 Metadata generation

Data discovery and access is dependent on the availability of rich metadata. Metadata is collected on both collection and object level and can be stored separately, or embedded in the data collection. Collection level metadata is generated by the researcher as part of the research process and supplemented by object level metadata for data publication, in most cases, by librarians or other data professionals. Good metadata creation can be supported by tools designed to simplify metadata input and to enhance interoperability. Metadata is also useful in tracking the history of derived data products (Mercury Project Solutions, 2013:5).

2.6.11 Storage and access

The choices regarding the approach to data storage have implications in terms of cost, security, and future access. It is an institutional responsibility to ensure that adequate and appropriate storage facilities are available. The goals of “reusing and sharing data more often” are met by storage that make data discoverable and accessible in the long term, which means the tendency should be more metadata rich, curated stores with a wide community scope (Mercury Project Solutions, 2013:6).

2.6.12 Publishing research data

There is an increasing expectation that the outputs of publicly funded research, including the data, will be made available for others to use. That means published data should be well-described (metadata), citable, discoverable and re-usable wherever possible. Potential re-users of research data have to have clear guidance about what they can and cannot do with the data: this is normally achieved by means of a licence. Research data can be published in the form of collection descriptions, citable and online accessible data elements, or citable other objects such as web services, Application Programming Interfaces (APIs), or concept definitions. At times, merely the existence of data collections is published; this occurs when data cannot be accessed, or accessed under strict conditions. Online data publication for download or web-service access is desirable for those kinds of data that is not restricted (Mercury Project Solutions, 2013:6).

2.6.13 Register research data

The process of registration and citation depends on the type of data published and how it is published. Data can be considered as ‘published’ when it is generally discoverable. An example of where research data could be registered is Research Data Australia (RDA). Research Data Australia (RDA) is designed to expose the description and existence of data collections (Mercury Project Solutions, 2013:6).

2.6.14 Ongoing curation

Research data that is not shared and not (immediately) required for further research should be properly archived, or disposed of at an appropriate time. Retention periods and regulatory requirements must be carefully assessed and considered (Mercury Project Solutions, 2013:7).

2.6.15 Usage monitoring

One of the great benefits of publishing research data with proper identifiers is the ability to track usage and citation statistics. It is recommended that institutions monitor data usage on a regular basis. Data citation tracking supports research evaluation (Mercury Project Solutions, 2013:7) and provides data citation metrics.

2.7 The need for RDM

The following requirements for proper RDM are listed and motivated by literature:

2.7.1 Volume of generation and complexity

The development and rapid advance of digital technologies have enabled immense quantities of data to be created, processed, and disseminated around the world. This data can capture the characteristics of phenomena in far greater detail and with a dynamic verisimilitude never before possible. Data is the foundation on which scientific, engineering, and medical knowledge is built. The generation, analysis, communication, and preservation of data are currently undergoing a period of profound change, and research is being similarly transformed (Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age 2009:ix). Researchers in Higher Education (HE) are producing ever-increasing quantities of born digital data during the course of their work which have to be managed for both immediate and long-term use (Highman & Pinfield, 2015:2; Pinfield, Cox & Smith; 2014:1). According to Borgman (2012:1), researchers are producing an unprecedented deluge of data by using new methods and instruments. Scientists are struggling with huge amount, complexity and variety of data that is now being produced. The scientific community strives to meet its basic responsibility toward transparency, standardisation and data archiving (Hanson, Sugden & Alberts 2011:1). This creates a need for active data management before datasets deteriorate (Highman & Pinfield, 2015:2). Qin, Ball and Greenberg (2012:1) states that we are facing a proliferation of scientific data and increased challenges relating to management and curation. Another challenge is that resources, infrastructure, policy and governance structures are still in flux (Cox, Verbaan & Sen, 2012).

2.7.2 Technological changes

New technologies have vastly increased the ease of data collection and, consequently, the amount of data collected, whilst also enabling data to be independently mined and reanalysed by others (Hanson, Sugden & Alberts, 2011:1). Tools, services and standards are emerging to help researchers manage their research assets, and make more widely available the evidence,

including raw and processed data that underpins their research articles (Whyte & Tedds, 2011:1). The open sharing of data, tools, and services over the Internet is creating new ways of conducting research and forming new relationships among researchers (Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, 2009:ix). Digital technologies have fostered a new world of research characterised by immense datasets, unprecedented levels of openness among researchers, and new connections amongst researchers, policy makers, and the public (Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, 2009:1).

Even as these new capabilities are expanding the influence and reach of research, they are raising complex issues for researchers, research institutions, research sponsors, professional societies, and journals. Digital technologies can complicate the process of verifying the accuracy and validity of research data, in part because of the enormous rate at which data can be generated and the intricate processing the data undergoes. The high rate of innovation in digital technologies, a lack of standards, and issues such as privacy, national security, and possible commercial interests can inhibit the sharing of data, which may reduce the ability of researchers to verify results and build on previous research (Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, 2009:1). The importance of managing research data has been emphasised by the government, funding agencies, and scholarly communities. Increased access to research data increases the impact and efficiency of scientific activities and funding. Thus, many research institutions have established or plan to establish research data curation services as part of their Institutional Repositories (IRs) (Lee & Stvilia, 2017:1).

2.7.3 New value in data

Effective management provides institutions with new ways to form synergies across research groups, producing new knowledge by engaging a broader range of stakeholders, and enabling wider reuse of data in teaching and learning, commercial exploitation and policy developments (Whyte & Tedds, 2011:1). New research topics and fields are emerging amid the boundaries of traditional disciplines, and the questions that investigators can address are rapidly expanding (Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, 2009:ix). Complementing the practical needs for data management is a change in perception of the value of research data. It has come to be viewed as an asset that should be managed to sustain its value (Borgman, 2012:1071; Carlson & Garritano, 2010:5; Lavoie, 2012:70). Data is the currency of research; but analogue and digital data generated within academia has largely been an invisible resource utilised within the research unit and shared with a select group of trusted colleagues, and consequently, their management are poorly understood (Koopman & de Jager, 2016:1).

2.7.4 Greater good and transdisciplinarity

Society now relies on scientific data of diverse kinds; for example, in responding to disease outbreaks, managing resources, responding to climate change, and improving transportation. It

is obvious that making data widely available is an essential element of scientific research (Hanson, Sugden & Alberts, 2011:1). Koopman and de Jager (2016:1) state that climate change research has alerted governments and researchers to the value of long term ecological studies. As science becomes more collaborative, data-intensive, and computational, academic researchers are faced with a range of data management needs. When these requirements are combined with funding directives that necessitate data management planning, there is both a need and an imperative for research data services in colleges and universities.

2.7.5 Risk avoidance – integrity research

Highman and Pinfield (2015:3) state that more negative concerns related to risk avoidance can also drive RDM developments, including those regarding researchers complying with Freedom of Information (FOI) legislation and the potential costs of not doing so (Whyte & Tedds, 2011:2). The changes in the nature and conduct of research are greatly enhancing the capabilities of researchers. However, these changes also are posing challenges, and in some cases, they have had negative consequences (Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, 2009:ix). Vines *et al*. (2013:1) investigated how research data availability changes with article age. In the case of papers in which authors reported the status of their data, the odds of the data being extant decreased by 17% per year. The prevention of data loss should be mitigated by proper RDM.

2.7.6 Funding

The finding of funding has become an extremely competitive exercise and major funders want evidence that research has not previously been undertaken, that the data will be preserved, and that the research will be open to scrutiny (Koopman & de Jager, 2016:1). As science becomes more collaborative, data-intensive, and computational, academic researchers are faced with a range of data management needs. As already mentioned, when these needs are combined with funding directives that require data management planning, there is both a need and an imperative for research data services in colleges and universities. Funding is often a preliminary barrier for organisations that wish to provide research data services to their researchers, particularly since the cost of handling supplementary materials such as data sets is not well known (Tenopir, Birch & Allard, 2012:9). Van Wyk and van der Walt (2014) state that by managing research data, institutions will meet funding body grant requirements, (e.g. NSF, NIH).

2.7.7 An international drive towards research data access

According to Borgman (2012:1059), there are four rationales for sharing data: to reproduce or to verify research; to make results of publicly funded research available to the public; to enable others to ask new questions of extant data; and to advance the state of research and innovation. The most effective method to describe the tremendous drive towards research data access is to identify all the relevant stakeholders and the benefits they hope to gain from increased access to

research data. According to Koopman and de Jager (2016:42), there are defensible reasons for data sharing, listed as:

- Creating opportunities for further integrated research.
- Contributing to global research initiatives.
- Preventing expensive duplication of research.
- Verifying research findings.
- Sharing data to make research more efficient and to ensure continuation of research.
- Making research transparent.
- Improving researchers' international profiles.

The pressure to share data comes from many quarters: funding agencies, both public and private; policy bodies such as national academies and research councils; journal publishers; educators; the public at large; and from researchers themselves. These stakeholders each have their own reasons for requiring or encouraging data sharing (Borgman, 2012:1066). In addition, Borgman (2012:1067) presents a model encompassing rationales for sharing data as illustrated in Figure 2.9 below.

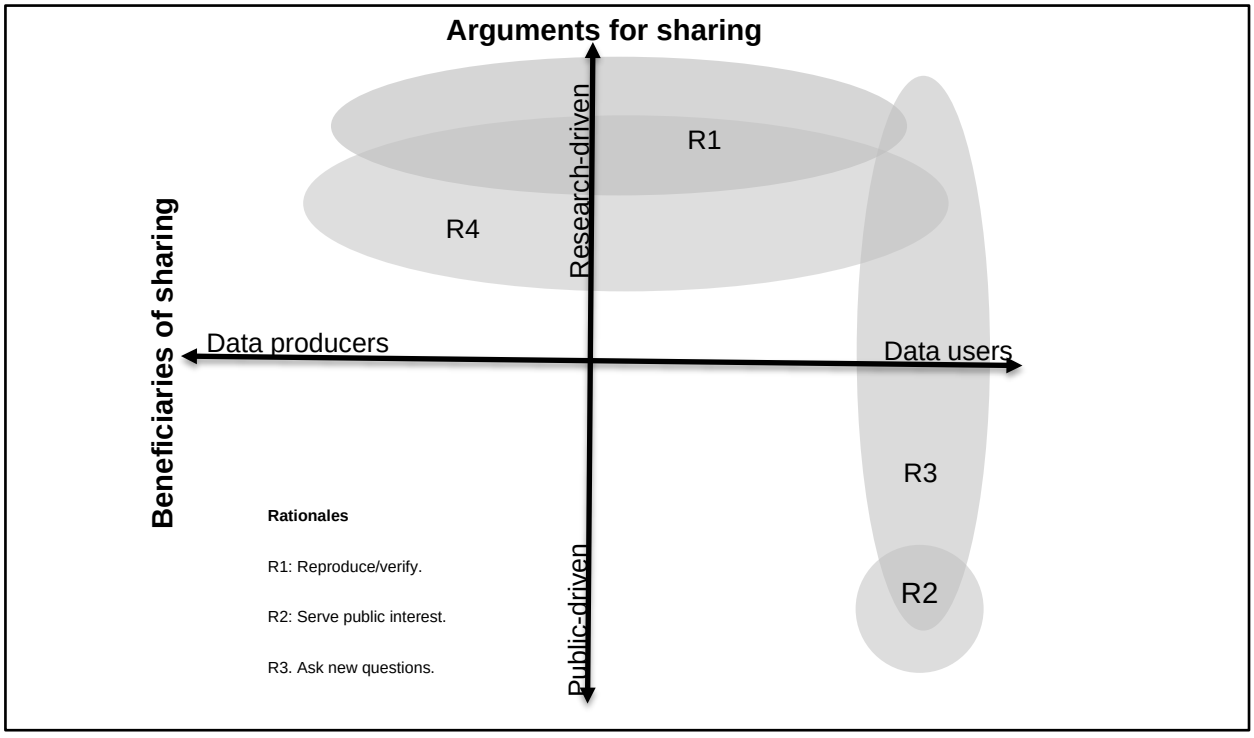


Figure 2.9: Rationales for sharing research data (Borgman, 2012:1067)

A rationale is an explanation of the controlling principles of opinion, belief, or practice. Motives and incentives underlie these rationales, whether stated explicitly or left implicit. A motivation is something that causes someone to act, whereas an incentive is an external influence that incites

someone to act. Rationales for sharing data also include beneficiaries, whether stated or implicit. A beneficiary in this case is an individual, agency, community, or other stakeholder who receives a benefit from the act of sharing data, such as the use of those data for a particular purpose (Borgman, 2012:1067). In addition, Borgman (2012:1067) presents four rationales in Figure 2.8 above, positioned on the two axes. The four rationales are: to reproduce or verify research; make results of publicly funded research available to the public; enable others to ask new questions of extant data; and advance the state of research and innovation. The dimensions on which these rationales are positioned are arguments for sharing and beneficiaries of sharing. The model is not exhaustive, either in terms of rationales or dimensions, but is offered as a useful framework for examining the complex interactions of players, policies, and practices involved in sharing research data. The arguments dimension (vertical axis) positions the rationales by their emphasis on the needs of the research community or the needs of the public at large. The beneficiaries dimension (horizontal axis) positions rationales by their emphasis on benefits to researchers who produce the data or benefits to those who might use research data. Subtle distinctions in the rationales for data sharing may lead to markedly different policies, economic models, research practices, curation practices, and degrees of compliance.

Reproducibility or replication of research is viewed as the “gold standard” for science (Jasney, Chin, Chong & Vignieri, 2011), yet it is the most problematic rationale for sharing research data. This rationale is fundamentally research driven, but can also be viewed as serving the public good. Reproducing a study confirms the science, and in doing so confirms that public monies were well spent (Borgman, 2012:1067). Public sentiment for sharing research data is based largely on the rationale that tax monies should be leveraged to serve the public good. In this view, data produced with public funds should be available for use and should not be hoarded by researchers. Borgman, (2012:1069) states that this public good argument is implicit in the OECD principles to which several of the funding agency policies refer, namely, that open access to research data is a means to leverage public investment in research (OECD, 2007). The OECD document also builds on an earlier U.S. study, explicitly quoting the passage: “The value of data lies in their use. Full and open access to scientific data should be adopted as the international norm for the exchange of scientific data derived from publicly funded research” (National Research Council [NRF], 1997). Borgman (2012:1070) states that a more focused rationale is that sharing data enables others to ask new questions, whether from an individual dataset or by combining multiple sources. This framing has two strands, one for the benefit of researchers and one for the general public. Researchers have argued that open access to data encourages meta-analysis: the ability to combine data from multiple sources, times, and places in order to ask new questions (Whitlock, 2011:61). The rationale for sharing data that resonates with the widest array of stakeholders is that research and innovation can be advanced more effectively (Borgman, 2012:1071). This is the claimed “fourth paradigm” in which computational science constitutes a

new set of methods beyond empiricism, theory, and simulation. One distinction between the “ask new questions” and “advance research” rationales is that the latter is wholly motivated by research interests. It goes beyond asking new questions of extant data; it addresses the need for more data and for curation of existing data in ways to ensure their usefulness. Simply put, “science depends on good data” (Whitlock *et al*., 2010:145).

For the last twenty-five (25) years, the need to share research data has been declared an urgent problem. Nonetheless, the discussion continues, policies proliferate, and evidence of data sharing is apparent in only a few research fields. Sharing research data is clearly a conundrum: an intricate and difficult problem. Acknowledging that data sharing is difficult does not mean abandoning all hope that data will be shared with some people some of the time. The challenge is to understand which data might be shared, by whom, with whom, under what conditions, why, and to what effect (Borgman, 2012:1072).

2.7.8 The Internet as RDM enabler

The Advanced Research Projects Agency Network (ARPANET) was established in 1969, specifically to enable researchers to share data between laboratories in geographically distant locations. ARPANET was the template upon which the Internet was subsequently built. The ubiquity of the Internet was the cornerstone of the open access initiative, which raised the question of universal access to research, particularly public funded research (Koopman & de Klerk, 2016:1). Initially, the Internet was designed for research purposes, as was the World Wide Web although society deviated from this intended use and as such, many aspects of our daily lives have changed drastically over the past 20 years. A culture that grew after the first scientific revolution some 300 years, which has brought humanity quite far is on the verge of its second profound metamorphosis. It is likely that the way that researchers publish, assess impact, communicate, and collaborate will change more within the next 20 years than it did in the past 200 years (Bartling & Friesike, 2014:v).

2.8 Researchers sceptical about RDM

Borgman (2012:1066) states that it is evident that investigators (and their collaborators, students, and staff) devote a massive amount of physical and intellectual labour to collecting, managing, and analysing their data and publish their results. Although data is the lifeblood of research in any field, it raises the question of variations, by purpose, approach, instrumentation, community, and many other local and global considerations. Some of the data may be in shareable forms, others not. Some researchers wish to share all their data all the time, whilst some wish never to share any of their data, and most are willing to share some of their data some of the time.

2.9 Legal and ethical polarity

Tsoukala *et al*. (2013:17) analysed the legal and ethical issues related to sharing data and the concept of open data. The project examined in particular, both legal issues such as intellectual property rights, privacy and data protection and open access mandates; and ethical ones, including the unintended secondary use, misappropriation and commercialisation of research data, unequal distribution of scientific results and disproportionate impacts of scientific freedom. The project considered how these different issues impact on a range of different stakeholders such as policy-makers, researchers, repository managers, and institutional representatives. Tsoukala *et al*. (2013:17) demonstrated in their analysis that intellectual property rights, especially in relation to data purchased from commercial organisations or cultural data, can act as a significant barrier to providing open access to research data, since the data creators may sometimes not hold the intellectual property rights to the material they collect, and to which they seek to provide access. Similarly, research participants, rather than researchers, institutions, repositories and other stakeholders, have primary control over the use of personal information for research purposes, which can limit the extent to which data can be made available in open access. These legal regimes often create a complex landscape, with real consequences for researchers, organisations and institutions. Open access mandates from governments and funders may place researchers and institutions in a situation where they are pressured to provide open access to data, despite the fact that intellectual property rights or data protection rights specifically and explicitly limit their ability to do so.

2.10 RDM in South Africa

According to Koopman and de Jager (2016:1), digital data archiving and research data management have become increasingly important for institutions in South Africa, particularly after the announcement by the National Research Foundation (one of the principal South African academic research funders), recommending these actions for the research that they fund. The researchers found in their study that while some researchers were already engaged in digital archiving in repositories, neither researchers nor the university had implemented systematic research data management. The international focus on research data makes it important for South African researchers and policymakers to ensure that data is managed in a way that enables long-term security and accessibility. In South Africa, those involved in research are becoming increasingly aware of the importance and value of curating and sharing the research data produced through public funding and thus, RDM policies are emerging (Kahn *et al*., 2015:296).

It is encouraging to observe that some of the Higher Education institutions have made steady progress in terms of RDM. Chiware (2015:11) indicates that the Cape Peninsula University of Technology (CPUT) has strategic partnerships in place, an eResearch Service Model, working groups, active pilot projects, and an partnership with Technical University Munich regarding the

eResearch (Communication and Infrastructure (eRIC) solution),. The eRIC initiative aims to develop an integrated communication and data management infrastructure for accompanying the complete life cycle of scientific knowledge generation and transfer. eRIC attaches particular importance to the analysis and development of models and tools for communication between researchers and libraries, both amongst eRIC project partners and within research teams. In cooperation with an international team of scientists and library staff, they will develop suitable communication and decision-making structures, establish a research accompanying consultation infrastructure, and set up teams to implement required software tools (Mitscherling, 2014:1).

In South Africa, certain libraries are beginning to provide frameworks for these services with some degree of success as policies are being formulated, infrastructure set up, library staff trained, and awareness and advocacy campaigns held with academic staff and researchers. Challenges being faced include the availability of resources and infrastructures and limited data management skills amongst library staff (Chiware & Zanele, 2016:1). As government, funders, and research institutions in South Africa become more involved in RDM, those involved in the actual research process, such as researchers, research offices, ethics committees, IT departments and libraries, will need to be made aware of the potential benefits of RDM and its processes and requirements. There has already been some activity in terms of awareness and capacity-building in South Africa. The Network of Data and Information Curation Communities (NeDICC), for example, arranges seminars, workshops and a conference to promote awareness around digital (including data) curation aimed at practitioners and managers involved with digital object management and encourages the growth of knowledge in this area (Kahn *et al*., 2015:298).

van Deventer and Pienaar (2015:1) state that the most important lesson being learnt is that one can learn many wonderful and valuable RDM lessons from the international trend setters, but in the end, one has to get one's "hands dirty" and do the work oneself. Therefore, one must, within the set parameters, implement the RDM practice that is both appropriate and acceptable for and to one's own set of researchers, who may be conducting research in a context very dissimilar to that of international peers.

2.11 Summary

The original purpose of the Internet, or the Advanced Research Projects Agency Network (ARPANET), was originally to enable scientist to share data. The exponential growth and adoption of information technology has a direct impact on the scientific community. Unimaginable amounts of data can be stored and analysed in real-time through large data centres all over the world. (Bartling & Friesike, 2014:v) is of the opinion that the way researchers publish, assess impact, communicate, and collaborate will change more within the next 20 years than it did in the past 200 years. The management of research data for various stages during the research process has

become a challenges for researchers due to the rapid expansion in volume and the forever changing technology landscape. Research data captured on a specific device using specific software could be redundant and unusable within a year or two if vendors go out of business. Adding to the research data management challenges of researchers is growing pressure from intergovernmental agencies, governments, funders, councils and the general public to re-use and share funded research data objects. Research institutions are facing various challenges due to a lack of international standards regarding RDM. The lack of standards start with the most basic of elements, the data element in RDM. Stakeholders and role players define data and research data using varying definitions. Most stakeholders and role players are beginning to acknowledge the long term value of research data and the need to curate research data during the pre-research, during research and post-post research phases. The sharing and re-use components pose specific challenges due to the lack of provenance, context and workflow description in current data sets. Metadata is the unsung hero in describing data objects. Various metadata standards are competing to become the dominant standard in the RDM domain. Another challenge to transdisciplinary research entities is how to describe items and events. Ontologies, taxonomies and controlled vocabularies differ between research communities. Multiple ontologies and taxonomies might need to be included in the metadata to enable sharing and re-use across research community boundaries. With the ability to share and re-use research data many ethical and risk questions and issues should be addressed. Research participants, researchers and institutions could become vulnerable if ethical standards are not adhered to and research data are used for unintended and even malicious purposes. It is critical that RDM should become imbedded in ethical clearance policies and procedures and for a risk management framework to be implemented to mitigate risks. A research data management plan should form part of the initial research planning process and should be updated throughout the research process. It became clear in the literature that the majority of researchers are willing to share research data, but feel uncertain due to inadequate institutional policies, procedures and support services. The South African context is different than the first world global context where most of the research and publications regarding RDM has been done. RDM is still in its infancy at South African universities with the primary focus on planning and the creation of awareness through the conduct of workshops. It is clear from the literature that RDM will become a critical component in conducting successful research in the near future.Chapter 3 follows, describing the realisation of the data collected and analysed and discussing the research results that emanated from the semi-structured, individual interviews.

CHAPTER 3: RESEARCH RESULTS

3.1 Introduction

In Chapter 2 the development of and need for RDM were discussed after a fundamental outline of the building blocks of RDM, such as data and datasets, were described. In Chapter 3, the realisation of the second phase of the research process from planned interviews to declared themes is formulated.

3.2 Realisation of data collection and analysis

Interviews were conducted during March 2017. The mediator recruited prospective participants and first established their interest in participation. Once participants indicated their willingness to participate, their names were given to the researcher who then made an hour's appointment with each one. Semi-structured, individual interviews were done in the participants' offices, although one participant's office wasn't private and thus this particular interview was done in a separate, private office. At the beginning of the interviews, the researcher introduced himself and first obtained written informed consent. Thereafter, the researcher activated the digital voice-recorder and conducted the interviews using the prepared interview schedule. The first part of the interview was to obtain demographic information. After each day's interviews, the researcher downloaded the interviews from the voice recorder onto his laptop, ensuring that the interviews were deleted from the voice recorder. After eight (n=8) interviews, the researcher obtained data saturation since the participants started to repeat information and no new themes emerged. All the interviews were handed to a transcriber on an external hard drive. Transcribed interviews had a code that replaced the participants' names and transcriptions were anonymised. The transcribed interviews went through a process of interpretive analysis (see the steps in Chapter 1, 1.8.2.2.4). After a consensus discussion was held with a co-coder, the research results were presented as themes, sub-themes and categories.

3.3 Demographic data

The demographic data of the participants are presented in Table 3.1.

Table 3.1: Demographic data of participants (N=36, n=8)

Participant	Age and gender	Years' experience in research team	Position in research team	Amount projects involved
PAR001	32, female	7	Researcher	4
PAR002	28, female	2	Research assistant	5

PAR003	32, female	9	Project manager	All projects
PAR004	47, female	4	Assistant	All projects
PAR005	48, female	6	Research assistant	2
PAR006	42, female	8	Researcher	5
PAR007	35, female	7	PI, researcher	2
PAR008	42 years, female	2½	PI, senior researcher	5

Of the eight (n=8) participants, all were female and the average years' work experience between them in a research team was 5.7 years. Four (n=4) participants were researchers, of which two of the four were also primary investigators. Four (n=4) participants were working in areas of research assistance and/or project management.

3.4 Interview results

Eight (8) research themes emerged from the data analysis process. These themes are presented in Table 3.2 (below) and described thereafter.

Table 3.2: Research themes and sub-themes that were identified in individual interviews

Research theme	Sub-themes
1st theme: RDM is a comprehensive system.	• RDM is a comprehensive process from inception to final destruction of data. • RDM requires curatorship of data. • Necessary for optimal access, safety and use.
2nd theme: RDM for storage and access solutions.	• User-friendly storage solutions of data in multiple formats. • Storage and archiving as an organisational and ethical requirement. • Storage solutions that enable immediate access. • Compliance with standards of funders and ethics.
3rd theme: Data security lacking, but essential.	• Awareness of the responsibility to keep data confidential and anonymous. • Security mechanisms are poor and primitive. • Necessary to safeguard data against hackers and against unskilled researchers.
4th theme: Researchers are responsible to preserve, share and disseminate quality data.	• Researchers must give data back to society. • Preservation necessary to curb dishonesty. • Quality data should represent the truth. • Sharing of data is a complicated process.
5th theme: Organisational fragmentation of RDM.	• Ethics and scientific committees are prominent components in basic RDM. • Gap between researchers and information management.
6th theme: A changing higher education landscape, from an inclusiveness to competition and regulation.	• Higher education changes. • Change management, acceptance and incentives necessary. • Increased demand for RDM. • Different approaches to research and RDM. • Conflicts arise at the publication phase of research outputs.

7th theme: RDM is everyone's responsibility and requires resource allocation.	• Although the PI is accountable, RDM requires a team effort. • Training needs regarding RDM and organisational skills. • Dedicated human capital and operational resources required. • Unaware of larger university initiatives for RDM.
8th theme: The risks and opportunities related to RDM for the research unit.	• Inefficiency: focused on data collection and not data dissemination. • Inaccuracy of data. • An appropriate and functional RDM system can improve the research process and can be outsourced to other research units.

3.4.1 RDM is a comprehensive system

Participants described RDM as a comprehensive system of multiple processes. The RDM system is complex and includes the planning of data collection, writing the protocol, actual data collection, data capturing, analysis and dissemination, ethics clearance pathways, data storage, data sharing, data mining, data archiving and destroying of data. One participant voiced "*a very comprehensive package"*. This system should ideally commence at the inception of a research idea, and remain in place until the data is finally discarded and destroyed. RDM entails curatorship of data, acknowledging and implementing the responsibility to control and keep track of it. Participants also described RDM as necessary to ensure optimal access to data, keep data safe, and enable optimal use of the data.

3.4.2 RDM for storage and access solutions

Participants explained that RDM is necessary for storage and access solutions. They were aware of the difficulty in accessing data in the absence of a proper storage and archiving process. One participant stated that "st*orage is, on the one hand, the robust storing of large sets of data but on the other hand, the nimbleness to access and use information*". Data storage is defined as being able to accommodate multiple data formats, from digital and hard copies to different types of data. This is an especially big challenge in a transdisciplinary research unit where researchers from different disciplines utilise both quantitative and qualitative data. Data storage is not only a practical and operational process, but also an organisational and ethical requirement. Participants were informed of the ethical requirements for data storage for five years on the premises of the university under the supervision of the project owner, sharing as stipulated by the Health Research Ethics Committee and the document management policy of university in general. Storage solutions in RDM should go together with immediate access to data. This entails the ability to know where what data is and how to obtain specific data, and should be aligned with the standards of funders, ethics and the institution.

3.4.3 Data security lacks but essential

Participants were aware of their responsibility towards research participants to ensure confidentiality and to keep data anonymous. They agreed that data security is a mechanism to safeguard data and to maintain control, as one participant stated, "…*safeguarding the participant and safeguarding the researcher and the university"*. Nonetheless, participants declared that their security mechanisms were poor and primitive as they used external hard drives, kept data on their own laptops with general university network access control, or even kept hard copies in locked cupboards, having fire distinguishers available. Some participants perceived Dropbox as a secure storage space and one participant voiced her frustration with using Alfresco (open source document management system) as a back-up mechanism. In general, participants acknowledged that they could not confirm that they possessed the most recent data as there was no audit trail available to secure access authorisation. Researchers explained that data security did not only refer to keeping data safe from hacking and ensuring confidentiality, but was also necessary to safeguard data against unskilled team members. Team members that are unfamiliar with technology and ignorant of the value of data, hold the risk of destroying data, which can be a crisis if there are not sufficient and secure back-up systems available.

3.4.4 Researchers are responsible to preserve and disseminate quality data

One participant stated that "*Information is expensive, and you have the responsibility to provide information back to society".* It was clear that all the participants shared an awareness that data should be disseminated to society. Dissemination can also occur through data sharing and is not only limited to the traditional research outputs of publications. Nevertheless, participants also stated that researchers as accountable scientists should give data back to society, referring to all data in its purest and most honest format. Data should assist the researchers to represent the truth and therefore, RDM is necessary to ensure optimal data preservation to curb dishonesty. "*It is part of good research practice to keep your data available to society…. If there is any uncertainty about the trustworthiness and validity…."* One participant explained that researchers have, in the past, lied about research results and that data preservation could assist in curbing dishonesty since it serves as an opportunity to access sufficient evidence. Participants explained that data sharing is a necessary, but complicated process. Data sharing should be planned and executed in a thorough manner at the inception phase of a project and should be declared to the appropriate ethics committee. Data sharing and dissemination should be viewed as essential steps in the project's lifecycle, and be actively managed and monitored.

3.4.5 Organisational fragmentation of RDM

The fifth theme referred to participants' description of the prominent role that applicable ethics– and scientific committees played in basic RDM. Participants could list standard operating procedures for data storage and RDM as communicated from the Health Research Ethics

Committee and the awareness of RDM voiced by scientific committees. These committees provided only basic RDM solutions with their main focus on ethical considerations. Unfortunately, they did not provide practical and technological solutions for RDM. Some participants were vaguely aware of organisational policies on access to data. Only one participant could voice the link between RDM and collaborations with librarians. Participants did not describe the information management of the university as an active partner to researchers in RDM, and one participant explained that she was not clear who to contact for RDM at the university. In general, participants experienced that they themselves, within their research unit, were fully responsible for RDM, irrespective of the larger organisational information technology architecture and referred only to Alfresco as an available mechanism for RDM.

3.4.6 A changing higher education landscape, from an inclusiveness to competition and regulation

As the sixth theme, participants described a research unit today, within a higher education institution, as a stressful working environment where there is a significant shift from research projects that were open and inclusive towards competition for research output, exclusiveness and increased regulation over research projects and outputs. One participant explained that researchers must work harder for less subsidy, whilst another described how research projects had no RDM plan in place, but are now regulated in terms of data storage, sharing and archiving. This retrospective management of research data causes conflict and is time consuming since data is not presented in a similar format. To align existing projects to RDM standards will be costly and will require skilled human capital and such support is not in place. As researchers had to move from a less-regulated research context to a more regulated one, there is resistance to change and to adopting RDM principles. Nonetheless, participants explained that change management is necessary for RDM implementations, which will be accepted once researchers can see the benefit thereof, can understand that RDM can imply less administrative effort, and when there are sufficient incentives available. Incentives are subjective but, according to the participants, researchers' access to personal research funds is a powerful incentive.

The participants were aware of an increased demand for RDM, although this demand presented itself in the form of crisis interventions. Three participants explained that a current, large, longitudinal data set presented complaints from various researchers that it took too long to share data, that data was captured into different software packages, and that the PI didn't have the most recent data. In this particular research unit, the demand for RDM arose from the following factors: a growing pool of data sets; increased needs for data sharing; the absence of a proper audit trail; and lack of a technology standard. Furthermore, participants voiced their opinions that working in a transdisciplinary research context implied diversity, where researchers hold different ideas to research and RDM. Eventually, conflicts arise too late amongst researchers at the publication

phase of research output when there is not a clarification of exact roles and responsibilities throughout the research process, and confirmed that this conflict could be addressed through efficient RDM.

3.4.7 RDM is everybody's responsibility and requires resource allocation

RDM is described as a comprehensive system that can only be obtained through team work with shared responsibility, although the PI remains accountable for it. For successful RDM, a collaborative team effort between all members of the research team, academic and non-academic, is necessary. All the role players concerned with a research project and researchers from participating research units, statisticians, ethics committees and scientific committees were listed as stakeholders in RDM. According to the participants, funders played varying roles in RDM. Yet, despite this, research team members still need training in RDM, including training to improve their organisational skills. However, RDM implementation is perceived as an additional duty and system that must be integrated into the research unit, requiring a dedicated allocation of adequately trained staff and appropriate operational resources such as software packages, hardware, networks, connectivity and information technology infrastructure as well as physical office space. As one participant stated that "*nobody ever sat me down and said these are your responsibilities in RDM…. hands-on learning".* Participants, ranging from senior researchers to research assistants, were unaware of the RDM initiatives available to them or the research unit in the larger university context.

3.4.8 The risks and opportunities related to RDM for the research unit

The final theme referred to the risks and opportunities related to RDM as voiced by participants. Participants explained that the absence of RDM implies too much input into data collection, causing insufficient research output. When PIs have to spend too much time on the administration and logistics of data collection, analysis, storage, sharing and archiving amidst insufficient RDM infrastructure, critical time is wasted since these researchers do not have time to disseminate the results and contribute to research output. This makes research activities within the research unit inefficient. A second risk identified by participants was that of current software and the absence of a proper RDM system that possibly leads to poor control over data, absent audit trails, manual calculations, and therefore increases the risk of inaccurate data.

Participants described the opportunities for the research unit once an appropriate and functional RDM system is set in place. Firstly, RDM can cause a direct improvement in the total research process. As one participant stated, "*….it (RDM) was never before a well thought through process*", which can be integrated into the research process. Secondly, a functional RDM system does not have to be implemented for each research entity, but can be outsourced to other research entities through a service level agreement.

3.5 Discussion

Both the literature and the research results indicate that RDM is a comprehensive system consisting of multiple processes and building blocks. Many of the steps in the data lifecycle mentioned in figure 1.2 were identified by the participants. Specific steps that were mentioned were planning, collection, assurance, preservation, discovery and analysis of research data. Steps mostly omitted by the participants were description, discovery and integration of research data. Most of the participants voiced the opinion that compliance to the requirements from the ethical committee will solve many of the data assurance steps, but that they found that no proper guidelines were in place or followed ethical clearance. Most of the participants agreed with Whyte and Tedds (2011:2) that RDM is defined as, "…*the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results*". The majority of participants agreed with the findings of Borgman (2012:1071), Carlson and Garritano (2010:5) and Lavoie (2012:70) that data is viewed as an asset that should be managed to sustain its value. It became clear from respondents' answers that there was uncertainty regarding how research data should be curated during the research process due to a lack of policies, procedures and guidelines.

The research results were consistent with the literature regarding concerns of data loss and inaccessible storage media (Viney *et al*., 2014:2). Multiple respondents indicated the desire for a "user-friendly" storage solution, and having difficulty and frustrations using prescribed solutions from the internal IT support department. The end result of this is that none of the respondents are actively using the available institutional storage solution due to the complex naming schema and the poor levels of performance. Most data artefacts are currently stored on local hard disk drives, portable hard disk drives and flash drives. No prescribed standards regarding the storage media or prescribed policies, procedures or guidelines were mentioned by the respondents. The research findings were consistent with the literature regarding security, retrieval and preservation concerns (Cox & Pinfield, 2014:300). Lack of access control and difficulty in sharing data due to no access control were mentioned during the interviews.

Participants felt strongly that the responsibility to preserve, share and disseminate data lay mainly with the researcher. The literature had similar findings for the South African context (Lotter 2014:11; Woolfrey, 2014:4). The Mercury Project Solutions (2013:2) recommend that various steps in the RDM process should be the responsibility of the institution. Most notably, a data management framework, a risk management plan, training and induction, policy compliance monitoring, registration of data with reputable repositories, ongoing curation and usage monitoring. The research results clearly indicated a lack of institutional guidance and/or available services to support the RDM process. The research results indicated that the participants shared

a strong desire to comply with ethical standards regarding research data practices, but that lack of IT support was a prominent challenge.

Both the literature and the research results indicate that RDM should be the responsibility of various stakeholders (NSF, 2017). Furthermore, DataONE (NSF, 2017) states that in addition to primary researcher(s), there may be others involved in the research process that take part in aspects of data management. Several of the roles identified both in the literature and the research results are those of data collector, data analyser, project director, staff responsible for running instruments, and administrative support staff responsible for grant submission. Roles that were mostly omitted in the research results but prominent in the literature were metadata generator, database designer, data modeller, external data centre or archive, and geospatial skilled staff. The research finding that RDM both represent opportunities to share data and new risks is consistent with the findings of Borgman (2012:1072).

3.6 Summary

In Chapter 3, eight (8) themes were identified from the individual interviews with members of a research team (n=8) in a research unit based at a South African university, and were subsequently described and discussed. Chapter 3 concludes that all the themes identified during the interviews were confirmed in literature, strengthening the need to formulate a RDM framework, which will now be provided in the final chapter, Chapter 4.

CHAPTER 4: PRELIMINARY FRAMEWORK, EVALUATION AND RECOMMENDATIONS

4.1 Introduction

It is clear from both the literature and the research conducted that research data management is a complex and evolving concept consisting of multiple stake holders, role players, phases and processes. Most of the participants experience a level of anxiety and uncertainty pertaining to the current and future challenges as related to RDM. A lack of institutional support regarding IT services and general RDM services provided became apparent from the interviews. The desired target state for the research entity is to participate in global research data sharing through supported and approved repositories. The basic building blocks and a collective strategy within the research entity will enable the desired future state. The proposed framework will enable the relevant research entity to start addressing the current and future challenges regarding research data management. The preliminary framework was built using a combination of best practices identified in the literature and applying recommendations regarding policies, procedures and guidelines in the context of a South African university-based research entity. The framework will initially focus on the high-level components required to enable the relevant research entity to follow best practices regarding RDM in their unique context. Information gleaned and current best practices from the literature were used to build the proposed framework.

4.2 Preliminary framework

The preliminary framework consists of various components and guidelines to enable the research entity within a South African university to define and achieve strategic goals regarding research data management. The preliminary framework presented in figure 4.1 below consists of four (4) building blocks, each attempting to address a fundamental question regarding RDM, which will have an impact on the next layer representing the following question. The four questions addressed in the preliminary framework are firstly why the research entity is required to strategically want to invest valuable resources in RDM? The answers and rationale from the why questions will define the question of which building blocks need to be in place. Once the required building blocks have been defined, the next question will address which resources the building blocks will become a reality. This will be followed by the fourth and final question regarding how the required building blocks will be developed with the available resources to achieve the objectives defined in the research and research data scope block.

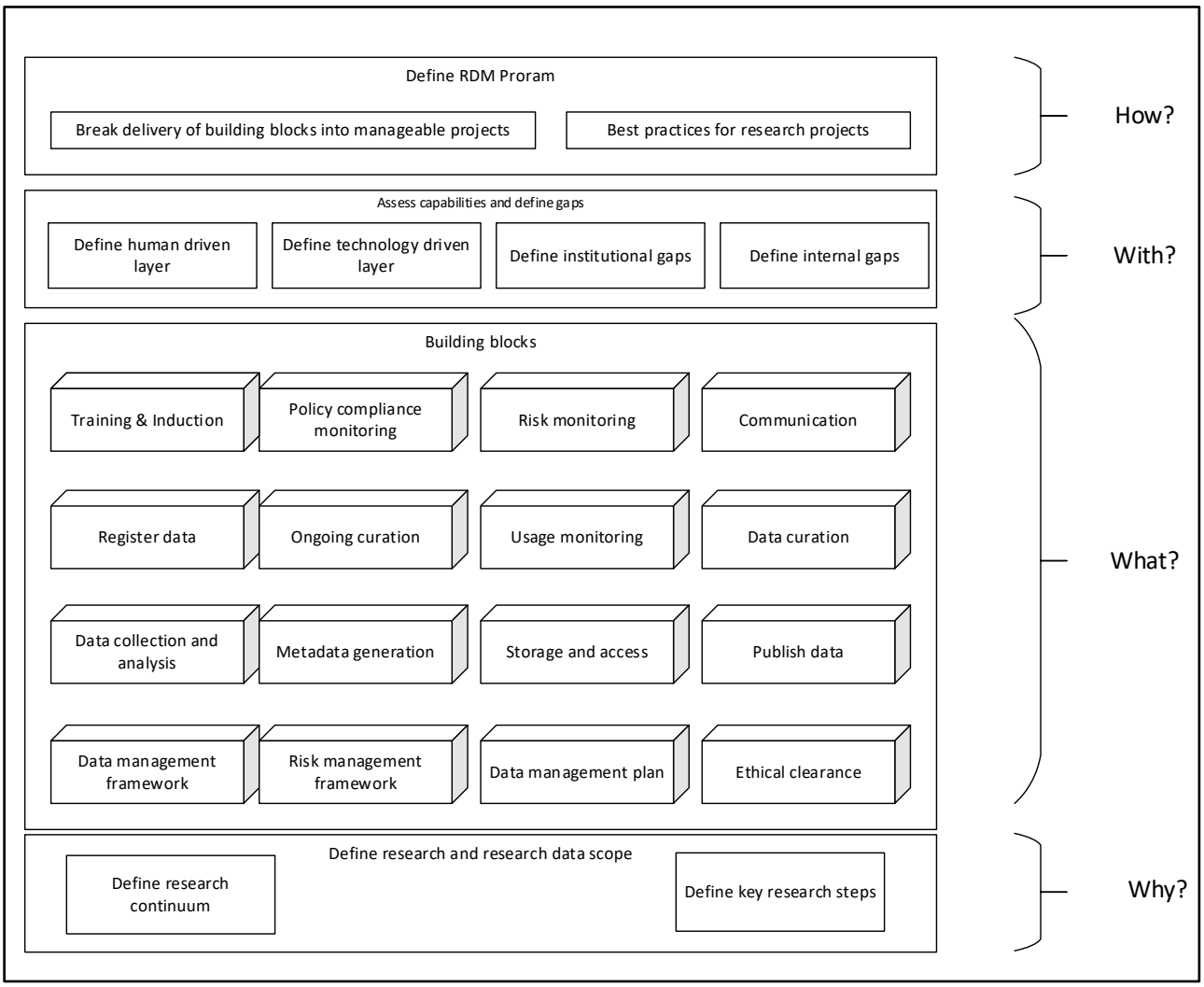


Figure 4.1: Preliminary RDM framework for a South African university-based research entity

4.2.1 Why? – Define research and research data scope

The first question the South African university-based research entity should answer is why it wants to engage in the complexity of RDM? The first pertinent factor to consider is the Research Data Curation Continuum as defined in 4.1 (above). It became clear from the research that the ambition of the research entity is to participate the public domain research point on the continuum. Other research entities initially decided to only focus on the first phase of the continuum, which is the private research domain. This domain involves the core research team at only one institution and is mostly concerned with the initial ingestion of data. The South African university-based research entity is already involved in multiple projects appropriate to the second phase of the continuum, namely the shared research domain. Currently, this collaboration with funders and global research partners presents major challenges due to a lack of building blocks to support the desired collaboration. From examining the research, it became clear that the only building block

in place to support research in the shared research domain on the continuum, was the ethical clearance building block.

The second question which the South African university-based research entity should consider is the key research steps as defined in figure4.2. The research entity aims to strategically align with a future state of managing their data at a post research phase. Strategically, all the building blocks mentioned in the preliminary framework will have to be in place to achieve this ambitious goal. Other research entities with different strategic goals regarding RDM could also use the framework to identify which building blocks would be relevant to their unique situation. All the building blocks have been defined in Chapter 2 of this study.

4.2.2 What? Building blocks

Once the why question has been answered, it will become clear which building blocks will be required to support the strategic decisions regarding the key research steps and the research continuum phase. Each of the components has been described in detail in Chapter 2 of the study. Some of the building blocks will prove to be more relevant, should a research entity be more interested in the post research step as defined in 4.1. Building blocks that should be prioritised are the building blocks stretching across all steps of the key steps in RDM. The most notable is data curation. It is recommended in the framework that data curation should be prioritised and should form part of any research project from the planning phase. Other building blocks that should also be prioritised are the data management framework and the risk management plan. The research results clearly indicate uncertainty regarding policies and procedures, a lack of IT infrastructure, a lack of data management support services and no metadata standards or management. Various datasets are currently at risk due to the absence of a risk management plan. It should be a priority to identify, assess, mitigate and monitor risks related to current data management practices. The research results indicate the presence of a data management plan. The primary purpose of the current data management plans was for funding and therefore the current plans on not on the required standard for proper RDM. Various points that should be focused on urgently in the data management plan are governance, security, version control and backups. The current understanding of the respondents regarding ethical clearance is that data should not be shared. To enable sharing of research data, focus areas regarding ethical clearance should be licencing, access control, informed consent and proper anonymization of research data. The next building block is training and induction. To ensure quality assurance and quality control throughout the research process and research data lifecycle, all staff should be regularly trained and new staff should be trained, regardless of their role in the RDM process. Once policies have been established at either institutional, funder, or national level, or for the research entity itself, compliance must be monitored. A centralised software solution with auditing capabilities could greatly assist in this regard. The following building block is risk monitoring and communication.

RDM is a dynamic process and new risks could be introduced during the lifecycle of research data. Some risks identified earlier may also become irrelevant during the lifecycle of research data. Changes in the risk profile should be communicated to all relevant stakeholders. The next building block is data collection and analysis. During the research process all data artefacts and their path to current state or existence should be described and monitored for quality assurance and quality control. Metadata generation is a critical building block in the preliminary framework. All workflows, data objects, datasets and architecture should be described using metadata. Policies and procedures should guide and enforce the next building block, which is storage and access. The following building block is publishing data. Capabilities and processes should be built to enable the data to be shared to one or more public repositories. Once the data is published, the next important building block is how to register the data. Data should be searchable, identifiable, obtainable and usable. Global identifiers could greatly assist in this regard. The final building block is usage monitoring. All relevant role players and stakeholders related to the data should be acknowledged through proper citation.

4.2.3 With? Assess capabilities and define gaps

The purpose of this step is to define the human driven layer items and technology driven items in order to implement a successful RDM programme. After defining the components on each of the technology- and human layers, an assessment is required to determine whether the components/capabilities are available on institutional level or internally within the research entity. These components/capabilities will be required in order to implement various projects as part of the RDM programme to construct the required building blocks to implement the research data management requirements defined in the first building block answering the ‘Why?’ question. According to McGovern (2016:11), human driven layers consist of governance, collection scope and acquisition. The technology scope consists of workflows, lifecycle storage and monitoring. The technology enabled layers are built on the IT stack, which consists of the following layers: the infrastructure or physical layer; the infrastructure virtual layer; operating system layer; database layer; middleware layer; and the applications layer. Institutional support and availability as related to governance should initially be studied. If any programmes, policies or procedures are available, the research entity should align itself with those. If there are any institutional guidelines regarding compliance and auditing, they should be strictly adhered to. The available IT resources from the institutional IT support department should be analysed for usability to use directly or as tools to help develop building blocks in the ‘Why?’ section. Internal capabilities and resources should also be considered before the RDM programme is launched. A realistic idea of the availability of skills and internal IT resources will assist with the planning and execution of a successful RDM programme.

4.2.4 How? Define RDM programme

The development of RDM will consist of a prioritised list of projects to start developing the building blocks defined in 'What? Building blocks' section. Planning and budgeting will be more accurate after the assessment of capabilities and defining gaps in the previous step in the framework. Decisions that will emerge during the planning phase of the RDM programme include buying versus building, and the potential outsourcing of products and services. With the current availability of various platforms, products, software and complete services on the internet, it may be better to consider such offers, rather than waiting for the institution to offer similar services. The research results indicate that participants had a low level of expectation in terms of institutional research support services and the institutional IT department. The institutional library services are also not currently offering RDM services to research entities within the institution. It is important to do thoroughly investigate options before any services are outsourced to a third party or cloud services are consumed. Typical risks include the possibility of sensitive data moving beyond of the borders of South Africa and not complying to institutional or government legislation regarding the protection of information.

The last component of the framework is the summary of practical guidelines provided by DataONE regarding the best practices during the data lifecycle. The guidelines have been grouped by the researcher in Table 4.1 below to indicate the best practices relevant to each phase of the research data lifecycle. Each of these best practices are summarised in a diagram in the addenda attached at the end of this mini dissertation. The titles have been adjusted to be more meaningful in a general context regarding research data management.

Table 4.1 List of addenda of preliminary RDM framework

Addendum	Best practice / Recommendations / Lessons Learned	Plan	Collect	Assure	Describe	Preserve	Discover	Integrate	Analyse
A1	Use datacasting tools to advertise your data						x		
A2	Data discovery						x		
A3	Give files descriptive names				x		x		
A4	Quality assurance of research data			x					
B1	Backup of research data					x			
B2	The impact of Boyle's Laws					x	x	x	
C1	Ensure accessibility for multiple channels						x		
C2	Enable discovery through standard terminology				x	x			
C3	Data quality communications			x					
C4	Ensure data and metadata are consistent			x	x				
C5	Ensure data can be integrated			x				x	x
C6	Data dictionary creation				x				
C7	Backup policy – Importance of documentation	x				x			

C8	Research data storage solution - Documentation	x							
D1	Tools and services - Considerations						x		
D2	Data preservation – How to decide					x			
D3	Description of data values expected	x							
D4	Data management roles and responsibilities	x							
D5	Data model definition	x			x				
D6	Parameter definitions				x				
D7	Format of spatial parameters				x				
D8	Standardise time and date storage				x				
D9	Describe provenance of data products				x				x
D10	Ensure data contents are clear				x				
D11	Dataset organisation				x				
D12	Research project description guidelines				x				
D13	Dataset spatial extent and resolution				x				
D14	Dataset temporal extent and resolution				x				
D15	Units of measure				x				
D16	Control and assure quality			x					
D17	File format guidelines and documentation					x			
D18	Document steps used in data processing				x			x	x
D19	Taxonomy documentation guidelines				x				
D20	Multi-set data integration							x	
D21	Strategy documentation				x				
D22	Control measures for data entry			x					
D23	Management guidelines for digital preservation an RDM					x			
E1	Metadata improvements				x				
E2	Culture of data sharing		x				x		
E3	Quality control for research data			x					
E4	Guidelines to make datasets reproducible			x			x		x
E5	Web services to make datasets accessible				x	x			
E6	Data backup guidelines					x			
E7	Storage media reliability					x			
F1	Lessons learned regarding data management					x			
H1	Understand reasons for sharing data					x	x		
H2	Data organisation best practices		x						
I1	Metadata standards				x	x			
I2	Identify sensitive data	x				x			
I3	Which data should be preserved for longer?					x			
I4	Standardise on codes for missing values			x					
I5	Guidelines for software identification								x
I6	Guidelines for outliers in datasets			x					x
I7	Guidelines to identify repositories	x				x			
I8	Clarify estimated values			x					x
I9	Understand motivations for sharing data	x	x	x	x				
M1	Data type consistency				x				
M2	Metrics for data usage and citing						x	x	
M3	Flag poor data for quality control			x					

O1	Research process optimisation					x	x	x	
P1	Data management planning – Start early	x				x			
P2	Multi media management planning	x				x			
P3	Store data in its raw format		x			x			
P4	Provenance enable the reproduction of data results			x	x		x	x	x
P5	Provenance and data cite documentation guidelines				x	x			
P6	Guidelines for drawing up a budget	x						x	
P7	Enable community members to tag your data				x				
P8	Register identifier for dataset				x	x			
P9	Versioning of data			x					
R1	Data ownership and recognition					x			
R2	Repeatable software processes to conform data								x
R3	Refer back to RDM plan	x							
S1	Avoid adding data descriptions on data sheets				x				
S2	Guidelines on data precision					x			x
T1	Data discovery and stewardship guidelines		x		x	x	x	x	
T2	Guidelines to make data reproducible			x				x	x
U1	Parameter guidelines for geospatial data							x	x
U2	Guidelines for field delimiters		x		x				
U3	Standardise codes		x		x				

The above table could be used as a reference, and processes and guidelines could be refined throughout the research data lifecycle. Policies and procedures could be incorporated in the diagrams provided to enable awareness and to ensure compliance.

4.3 Evaluation

The research results highlight the challenges researchers are facing within a South Africa university-based research entity. The participants accepted responsibility for the curation for research data and acknowledge that current RDM practices within the research entity pose serious risks. The research results were consistent with other findings in a South African context regarding the lack of institutional policies, procedures and RDM services. International research findings in North America, Europe and Australia indicate more involvement and support from an institutional level. Due to international funding, the expectations from funders on certain projects are RDM practices in line with international standards. This poses a serious challenge to the South African university-based research entity, due to the absence of any support services available from the institution. The research entity has also not included the costs in terms of time and money in the initial research proposals and RDM plans. Certain projects were started more than ten years ago. The RDM plans and funds required to keep participating in the research projects must be revised in order to build capabilities to adhere to funder expectations.

The components of a RDM framework for a South Africa university-based research entity should be the same as for any international research institution. Due to political and financial challenges in the South African context, universities are limited in terms of resources required to build and implement RDM capabilities. The estimated time frame for building such capabilities is not identified for use in the next five years. If the research entity wants to pursue its strategy to be compliant with international practices regarding RDM, it will have to use self-initiative and build internal RDM capabilities. The spin-off for the research entity could be packaging and reselling RDM services to other entities and becoming a valuable specialist regarding RDM on an institutional level. The components were summarised in the preliminary framework provided in paragraph 4.1 above. During the literature review, the researcher found it challenging to find articles describing the reality from the perspective of a research entity. Most of the studies conducted in South Africa and abroad were done on an institutional level and not from the perspective of the researcher. Many articles that were reviewed were written from the perspective of institutional library services. The interview process enabled the researcher to probe deeper in terms of certain topics and to gain insights not previously mentioned in the literature.

The proposal of a RDM framework can only be viewed as a starting point in solving the challenges faced by a South African university-based research entity. It is the sincere objective of the researcher that the framework could be used to promote RDM awareness amongst other research entities in the South African context. It can be argued that the problem defined in the problem statement is real and very challenging for the researchers and support staff involved at the research entity. The methods used enabled the researcher to obtain an overview of international and South African realities, practices and challenges regarding RDM. The objective to propose a preliminary framework could only be identified as a small contribution to the wide field of RDM.

4.4 Limitations

The researcher found it somewhat challenging during the interviews to probe for more details since RDM in South Africa is still in its infancy compared to various first world countries. Most of the literature reviewed was presented from the context of a first world country university where institutional support services could safely be assumed. The research had a few constraints regarding time and budget, and the study could have uncovered more findings if multiple research entities had been involved in the research. The sample size was relatively small, but it became apparent that most of the participants experienced the same realities within the entity regarding RDM challenges. The researcher found it challenging to build a preliminary framework for the South African context. At the time of writing, there are still no institution in South Africa which has implemented complete RDM capabilities to support services. Most institutions are currently in the awareness and planning phases.

4.5 Recommendations

- Due to the increasing importance of research data in the research process, institutions and research entities in the South African context should urgently promote the awareness, planning and implementation of RDM capabilities.
- Due to the complexity of building RDM, the researcher proposes the use of an enterprise architecture framework to assist collaboration between different stakeholders to understand and build RDM capabilities.
- A proposal for further study is the use of frameworks such as the Zachman Framework for Enterprise Architectures or The Open Group Architecture Framework (TOGAF). Both frameworks will enable the different stakeholders and role players on different levels to understand and construct better solutions and capabilities in order to provide RDM services.
- The researcher also proposes the development of a localised curriculum for under-graduate and post-graduate students regarding RDM practices.
- Various international guidelines could be used as a basis to build the content.
- All role players within the research entity should also be skilled or re-skilled to enable them to stay up to date with the latest RDM practices.
- In the current reality of limited institutional support, the researcher also strongly recommends that the research entity take initiative to network with private service providers and other international institutions to outsource services or to get guidance to help building internal capabilities.
- Cloud services could be a viable solution because of their lower initial costs and the lack of internal IT capabilities to support infrastructure.

4.6 Summary

Research data management is a complex challenge because of a lack of and difference in standards between scientific communities. In the South African context, the challenges are even greater because of the lack of institutional support for RDM. Globally, research data is viewed as a valuable commodity that should be curated from the planning process of a project and preserved long past the lifetime of the research project. Many questions remain regarding roles and responsibilities to implement a successful RDM programme on an institutional level. The present approach in South Africa entails various universities collaborating through joint workshops and connection to international institutions that have made progress with their RDM programmes. This lack of momentum leaves South African based research entities and researchers in a very vulnerable situation. According to ethical committees and policies from funders, institutions and councils, it is the responsibility of the researchers and research entity directors to manage research data. This is an impossible task because of the lack of institutional support and the complexities encountered in proper data management. The time frame for institutional support

will not implementable in the foreseeable future, and will only be applicable in five years or even further into the future. Unfortunately, researchers are faced with the challenges regarding RDM demands in the here and now. The preliminary framework provided by the researcher could be used as a starting point to enable research entities to strategically build capabilities to enhance RDM practices. The framework could also enable South African based universities to identify building blocks and methodologies used to accelerate the delivery of RDM services. On a practical level, the diagrams in the addenda could help researchers in the short term to follow best practices and begin building a body of knowledge within research entities and institutions.

References

Anon. 2017. Metadata. https://techterms.com/definition/metadata Date of access: 19 Nov. 2017.

Balzan, P. 2013. Business driven technology. 5th ed. Columbus, OH: McGraw-Hill/Irvin.

Barateiro, J., Antunes, G., Freitas, F., Borbinha, J. 2010. Designing digital preservation solutions: A risk management-based approach. *The international journal of digital curation,* 1(5):7-17.

Bartling, S. & Friesike, S., *eds.* 2014. Opening science: The evolving guide on how the internet is changing research, collaboration and scholarly publishing. New York, NY: Springer International Publishing. http://www.springer.com/gp/book/9783319000251 Date of access: 20 Mar. 2016.

Bartling, S. & Friesike, S. 2014. Towards another scientific revolution. (*In* Bartling, S. & Friesike, S. *eds.* Opening science: The evolving guide on how the internet is changing research, collaboration and scholarly publishing. New York, NY: Springer International Publishing. Volume 445, p. 3 – 14).

Borgman, C.L. 2012. The conundrum of sharing research data. *Journal of the American Society for information science and technology*, 63(6):1059–1078.

Borgman, C.L., Bowker, G.C., Finholt, T.A. & Wallis, J.C. 2009. Towards a virtual organization for data cyberinfrastructure. Proceedings of the 9th annual onternational ACM/IEEE joint conference on digital libraries, Austin, TX, 14-19 June. https://dl.acm.org/citation.cfm?id=1555400 Date of access: 12 June 2016.

Boston University Libraries. 2017. What is research data. http://www.bu.edu/datamanagement/background/whatisdata/. Date of access: 30 Mar. 2017.

Botma, Y., Greeff, M., Mulaudzi, F.M. & Wright, S.C.D. 2010. Research in health sciences. Cape Town: Pearson Education South Africa.

Briney, K. 2015. Data management for researchers: Organize, maintain and share your data for research success [Kindle ed.]. Available: https://www.amazon.com/Data-Management-Researchers-Organize-maintain/dp/1784270113

Brown, M.L. & White, W. 2014. Case study 2: University of Southampton - a partnership approach to research data management (*In* Pryor, G., Jones, S. & Whyte, A. *eds*., Delivering research data management services. London: Facet Publishing. p. 135–162).

Carlson, J. & Garritano, J. 2010. E-science, cyberinfrastructure and the changing face of scholarship: Organizing for new models of research support at the Purdue University libraries. (*In* Walter, S. & Williams, K. *eds*.,The expert library: Staffing, sustaining, and advancing the academic library in the 21st century, association of college and research libraries, Chicago, IL: Association of college and research libraries. p. 234-269).

Chiware, E.R.T. 2015. Research data management services at Cape Peninsula University of Technology. Proceedings of the COAR-SPARC conference on connecting research results, bridging communities and open scholarship, Porto, Portugal, 15-16 April. http://coar-repositories.org/files/3_Chiware-COAR-SPARCPresentation.pdf Date of access: 25 Feb 2016.

Chiware, E. & Mathe, Z. 2016. Academic libraries' role in research data management services: a South African perspective. *South African journal of libraries and information science,* 82(2):1-10.

Cox, A.M. & Pinfield, S. 2014. Research data management and libraries: Current activities and future priorities. *Journal of librarianship and information science*, 46(4):299–316. DOI: 10.1177/0961000613492542

Cox, A., Verbaan, E. & Sen, B. 2012. Upskilling liaison librarians for research data management. http://www.ariadne.ac.uk/issue70/cox-et-al Date of access: 25 Feb. 2016.

Data Observation Network for Earth (DataONE). 2017. What is DataONE? https://www.dataone.org/ Date of access: 25 Feb. 2016.

David, P.A. 2004. Understanding the emergence of ''Open Science'' institutions: functionalist economics in historical context. *Industrial and corporate change*, 13(4):571–589.

Digital Curation Centre (DCC). 2011a. What is digital curation? http://www.dcc.ac.uk/digital-curation/what-digital-curation Date of access: 17 Feb. 2017.

Digital Curation Centre (DCC). 2011b. About the DCC. http://www.dcc.ac.uk/about-us Date of access: 10 Apr. 2017.

Digital Curation Centre (DCC). 2011c. DCC curation lifecycle model. http://www.dcc.ac.uk/resources/curation-lifecycle-model Date of access: 10 Apr. 2017.

Digital Curation Centre (DCC). 2011d. Data management plans. http://www.dcc.ac.uk/resources/data-management-plans. Date of access: 10 Apr. 2017.

Gill, T., Gilliland, A.J., Whalen, M. & Woodley, M.S. 2008. Introduction to metadata, version 3.0. Los Angeles, CA: J. Paul Getty Trust. http://www.getty.edu/publications/intrometadata/introduction/. Date of access: 18 Jan. 2016.

Greenberg, J. 2009. Metadata research supporting the Dryad Data Repository. Presentation at the Metadata Working Group, Cornell University Library. http://dspace.library.cornell.edu/bitstream/1813/12247/1/DryadCornell.pdf. Date of access: 20 Mar. 2017.

Grove, S.K., Burns, N., & Gray, J.R. 2013. The practice of nursing research: appraisal, synthesis, and generation of evidence. 7th ed. St Louis, MO: Elsevier.

Hanson, B., Sugden, A. & Alberts, B. 2011. Making data maximally available. *Science*, 331(6018):649–649.

Higman, R. & Pinfield, S. 2015. Research data management and openness: the role of data sharing in developing institutional policies and practices. *Electronic library and information systems*, 49(4):364-381.

Institute of Risk Management. 2017. Risk management standards. https://www.theirm.org/knowledge-and-resources/risk-management-standards/ Date of access: 7 Apr. 2017.

International Organisation of Standardisation. 2017. ISO 31000:2009. Risk management – Principles and guidelines. https://www.iso.org/standard/43170.html Date of access: 7 Apr. 2017.

Jasney, B.R., Chin, G., Chong, L. & Vignieri, S. 2011. Again, and again, and again. *Science*, 334(6060):1223.

Kahn, M., Higgs, R., Davidson, J. & Jones, S. 2014. Research data management in South Africa: how we shape up. *Australian academic & research libraries*, 45(4):296-308.

Koopman, M.M. & De Jager, K. 2016. Archiving South African digital research data: How ready are we? *South African journal of science,* 2016;112(7/8):1-7, Jul./Aug.

Lavoie, B.F. 2012. Sustainable research data. (*In* Pryor, G. *ed*. Managing research data, London: Facet Publishing. p. 67–82).

Lee, D.J. & Stvilia, B. 2017. Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLoS ONE,* 12(3):e0173987.

Lewis, M.J. 2010. Libraries and the management of research data. (*In* McKnight, S. *ed.* Envisioning future academic library services. London: Facet Publishing. p. 145-168).

Lotter, L. 2014. Reflections on the RDM position in South Africa. LIASA research data management workshop, Cape Town, South Africa, 27 March. http://www.dcc.ac.uk/webfm_send/1631 Date of access: 22 Jul. 2017 [Presentation].

Lynch, C. 2003. Institutional repositories: Essential infrastructure for scholarship in the digital age. Association of Research Libraries. Report No.: No. 226. http://www.arl.org/storage/documents/ publications/arl-br-226.pdf Date of access: 22 Jul. 2017.

Macanda, M., Rammutloa, M. & Bezuidenhout, R. 2015. Research data management at UNISA. http://hdl.handle.net/10500/13907 Date of access: 24 Feb. 2016.

Matthews, B., Sufi, S., Flannery, D., Lerusse, L., Griffin, T., Gleaves, M. & Kleese, K. 2010. Using a core scientific metadata model in large-scale facilities. *The international journal of digital duration,* 1(5):106-118.

Mercury Project Solutions. 2013. Research data management in action. http://www.ands.org.au/__data/assets/pdf_file/0009/394056/research-data-management-in-practice.pdf. Date of access: 27 Mar. 2017.

Michener, W.K. 2015. Ten simple rules of creating a good data management plan. PLoS computational biology, 11(10):e1004525.

Mitscherling, C. 2014. Integrated communication and service infrastructure for libraries. Purdue, paper 2. Proceedings of the IATUL Conference, Aalto University, Espoo, Finland, 2-5 June. http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=2014&context=iatul Date of access: 25 Feb. 2016.

Murray, A. & Wheaton, K. 2016. Welcome to curation 2.0. *Applied science & technology source*, 25(1):28-29.

National Academy of Sciences (US), National Academy of Engineering (US), Institute of Medicine (US) Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. 2009. Ensuring the integrity, accessibility, and stewardship of research data in the digital age. Washington, DC: National Academies Press.

National Research Council. 2009. Ensuring the integrity, accessibility, and stewardship of research data in the digital age. Washington, DC: National Academies Press. http://www.nap.edu/catalog.php?record_id=12615. Date of access: 20 Feb. 2016.

National Research Foundation (NRF). 2015a. Databases. http://www.nrf.ac.za/information-resources/databases Date of access: 24 Feb. 2016.

National Research Foundation (NRF). 2015b. Statement on open access to research publications from the National Research Foundation (NRF)-Funded Research. http://www.nrf.ac.za/media-room/news/statement-open-accessresearch-publications-national-research-foundation-nrf-funded Date of access: 24 Feb. 2016.

National Science Foundation. 2005. National Science Foundation's Cyberinfrastructure vision for 21st century discovery. National Science Foundation Cyberinfrastructure Council. https://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf Date of access: 20 Mar. 2016.

National Science Foundation. 2010. Data management & sharing frequently asked questions (FAQs). https://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp#1. Date of access: 7 Apr. 2017.

Nielsen, M. 2011. Reinventing discovery: the new era of networked science. Princeton, NJ: Princeton University Press.

North-West University (NWU). 2008. Institutional strategy for research and innovation. Potchefstroom: NWU.

North-West University (NWU). 2017. Annual research report 2016. Potchefstroom: NWU.

Organisation for economic co-operation and development (OECD). 2007. Principles and guidelines for access to research data from public funding. https://www.oecd.org/sti/sci-tech/38500813.pdf Date of access: 18 Nov. 2016.

Pinfield, S., Cox, A.M. & Smith, J. 2014. Research data management and libraries: relationships, activities, drivers and influences. *PLoS ONE*, 9(12):e114734. DOI:10.1371/journal.pone.0114734.

Preservation Metadata Maintenance Activity. 2017. Premis resources. https://www.loc.gov/standards/premis/bibliography.html Date of access: 9 Nov. 2017.

Provenance working group. 2014. PROV. https://www.w3.org/2001/sw/wiki/PROV Date of access: 9 Nov. 2017.

Qin, J., Ball, A. & Greenberg, J. 2012. Functional and architectural requirements for metadata: Supporting discovery and management of scientific data. *International conference of Dublin core metadata Application*, 8:62-71, Sept. http://dcpapers.dublincore.org/pubs/article/view/3660 Date of access: 17 Feb. 2016.

Qin, J., D'Ignazio, J. & Baldwin, S. 2011. A workflow-based knowledge management architecture for geodynamics data. A white paper submitted to NSF GEO/OCI EarchCube Charrette meeting. http://earthcube.ning.com/page/whitepapers Date of access: 28 Mar. 2017.

Ray, J. 2014. Introduction to research data management. (*In* Ray, J., *ed.* Research data management: practical strategies for information professionals. West Lafayette, IN: Purdue. p. 1- 23).

Renear, A.H., Sacchi, S. & Wickett, K.M. 2010. Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4.

Smith, V.S. 2009. Data publication: Towards a database of everything. *BMC research notes*, 2:113. http://www.biomedcentral.com/1756-0500/2/113 Date of access: 7 Apr. 2017.

Stier, K. 2015. Data backup in the age of the cloud: best practices for backing up data and retrieving it quickly when necessary. University Business. https://www.universitybusiness.com/article/college-data-backup-age-cloud Date of access: 7 Nov. 2017.

Stvilia, B., Hinnant, C.C., Wu, S., Worrall, A., Lee, D.J., Burnett, K., Burnett, G., Kazmer, M.M. & Marty, P.F. 2015. Research project tasks, data, and perceptions of data quality in a condensed matter physics community. Journal of the Association for Information Science and Technology, 66(2):246- 263.

Tenopir, C., Birch, B. & Allard, S. 2012. Academic libraries and research data services: current practices and plans for the future: an ACRL white paper. Chicago, IL: Association of College and Research Libraries. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf Date of access: 20 Dec. 2016.

Tsoukala, V., Angelaki, M., Kalaitzi, V., Wessels, B., Price, L, Taylor, M.J., Smallwood, R., Linde, P., Sondervan, J., Reilly, S., Noorman, M., Wyatt, S., Bigagli, L., Finn, R., Sveinsdottir, T. & Wadhwa, K. 2013. Policy guidelines for open access and data dissemination and

preservation. RECODE project. http://recodeproject.eu/wp-content/uploads/2015/02/RECODE-D5.1-POLICY-RECOMMENDATIONS-_FINAL.pdf Date of access: 18 Nov. 2016.

Van Deventer, M. & Pienaar, H. 2015. Research data management in a developing country: a personal journey. *International journal of digital curation*, 10(2):33–47 DOI: 10.2218/ijdc.v10i2.380.

Van Wyk, J. & Van der Walt, I. 2014. Going full circle: research data management @ University of Pretoria. Proceedings of the eResearch Africa Conference, University of Cape Town, Cape Town, 23-27 November. http://www.eresearch.ac.za/sites/default/files/image_tool/images/140/Going_Full-Circle_RDM_UP_VanWyk_VanderWalt_26Nov_2014.pdf Date of access: 24 Feb. 2016.

Vines, T.H., Albert, A.Y.K., Andrew, R.L., De'barre, F., Bock, D.G., Franklin, M.T., Kimberley, J.G., Moore, J.S., Renaut, S. & Rennison, D.J. 2014. The availability of research data declines rapidly with article age. *Current biology*, 24:94-97, Jan.

Vinz, S. 2015. The theoretical framework of a dissertation: what and how? https://www.scribbr.com/dissertation.the-theoretical-framework-of-a-dissertation-what-and-how/ Date of access: 10 Nov. 2017.

Webster, L. & Moyo, M. Committee of Higher Education Libraries of South Africa. Costs and challenges associated with maintaining quality academic library and Information Services supporting teaching, learning and research: A presentation to the commission of inquiry into higher education and training. http://www.justice.gov.za/commissions/FeesHET/hearings/set3/set3-27Oct2016-CHELSA.pdf Date of access: 19 Nov. 2017.

Welman, C. Kruger, F. & Mitchell, B. 2012. Research methodology. Cape Town: Oxford University Press.

Witt, M. & Cragin, M. 2008. Introduction to Institutional Data Repositories Workshop. Library research publication, Purdue University. http://docs.lib.purdue.edu/lib_research/83 Date of access: 19 Feb. 2016.

Whyte, A. & Tedds, J. 2011. Making the case for research data management. DCC briefing papers. http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm. Date of access: 31 Mar. 2017.

Woolfrey, L. 2014. UCT research data management policy project: Report. https://www.datafirst.uct.ac.za/images/docs/20140307-uct-rsearch-data-management-woolfrey.pdf Date of access: 26 May 2016.

United Nations Archives and Records Management Section (ARMS). 2004. Standard on recordkeeping metadata. https://archives.un.org/sites/archives.un.org/files/files/Standards/ARMS_Standard_Recordkeeping_Metadata_Final.pdf Date of access: 16 Nov. 2017.

University of Leicester. 2017. What is research data management. http://www2.le.ac.uk/services/research-data/rdm/what-is-rdm Date of access: 31 Mar. 2017.

ADDENDUM A: ETHICAL CLEARANCE

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

Private Bag X6001, Potchefstroom
South Africa 2520

Tel: 018 299-1111/2222
Web: http://www.nwu.ac.za

10728929
BESTER, TV MR
tvbester@gmail.com

Potchefstroom Business School
Tel: 018 299-1419
Fax: 018 299-1416
Email: Wilma.prelorius@nwu.ac.za

11 May 2015

ETHICAL CLEARANCE

This letter serves to confirm that the research project of BESTER, TV has undergone ethical review. The proposal was presented at a Faculty Research Meeting and accepted. The Faculty Research Meeting assigned the project number EMS15/02/31-3/01. This acceptance deems the proposed research as being of minimal risk, granted that all requirements of anonymity, confidentiality and informed consent are met. This letter should form part or your dissertation manuscript submitted for examination purposes.

Yours sincerely

Me MM Heyns

Manager: Research - NWU Potchefstroom Business School

Original details: Wilma Prelorius[12306208] C:\Users\Documents\My Scans\2015\Ethics goedkeuring met nommer.docm
11 Mei 2015

ADDENDUM B: INFORMED CONSENT

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

EMS15/02/31-3/01

PARTICIPANT INFORMATION LEAFLET AND CONSENT FORM FOR RESEACH UNIT STAFF MEMBERS TO PARTICIPATE IN UNSTRUCTURED INTERVIEWS

TITLE OF THE RESEARCH PROJECT: Research data management framework for a South African university-based research entity

REFERENCE NUMBERS: EMS15/02/31-3/01

PRINCIPAL INVESTIGATOR: Tertius Vorster Bester

ADDRESS: 8 Richardson Street, Baillie Park, Potchefstroom, South Africa

CONTACT NUMBER: 082 788 9809

You are being invited to take part in a research project that forms part of my *Master of Business Administration*. Please take some time to read the information presented here, which will explain the details of this project. Please ask the researcher any questions about any part of this project that you do not fully understand. It is very important that you are fully satisfied that you clearly understand what this research entails and how you could be involved. Also, your participation is **entirely voluntary** and you are free to decline to participate. If you say no, this will not affect you negatively in any way whatsoever. You are also free to withdraw from the study at any point, even if you do agree to take part.

This study has been approved by the **Ethics Committee of the Faculty of Health Sciences of the North-West University (NWU-000196-15-S1).** It might be necessary for the research ethics committee members or relevant authorities to inspect the research records.

What is this research study all about?

The objectives of this research are:

- To determine how a Research Data Management framework can be formulated for a South African university based research entity.
- To explore the context of the current research entity within a South African university.

- To explore the aspects to be considered when formulating a RDM framework.
- To explore what stakeholders should be considered when formulating a RDM framework.

Why have you been invited to participate?

You have been invited to participate because you are:

- employed in permanent or temporary position at the relevant Research Entity / Unit.
- are currently playing or played a role in any of the relevant phases regarding Research Data Management (RDM) on one or more research project.
- comfortable to use English when talking to the researcher.
- willing to give voluntary, written informed consent and also willing to spend at least one hour with the researcher for an interview that will be recorded on a digital voice recorder.

What will your responsibilities be?

You will be responsible to spend approximately 45 minutes to one hour of your time with the researcher in an individual interview. During the interview the researcher will ask you to talk about your experiences and current or past involvement regarding Research Data Management. It will be your responsibility to give your honest feedback to the researcher. It will also be your responsibility to first sign this informed consent, voluntary. Please ask all the questions you wish to ask before onset of the interview. If you consent to participate and you cannot adhere to a scheduled interview, it will be your responsibility to inform the mediator that you will not be able to participate. Please state if you wish to stop with an interview or don't want to continue with the research as you can withdraw from the research at any time without discrimination.

Will you benefit from taking part in this research?

- There are no direct benefits for your participation in this study.
- The indirect benefit will be the scientific significance to improve the understanding of how a RDM framework can be formulated for a South African based research entity. The practical implication of the study will be to provide a RDM framework for a South African based research entity, following a deeper understanding of the context, stakeholders, drivers and influencing factors.

Are there risks involved in your taking part in this research?

There are minimal risks involved in taking part in this research and although you might be afraid that punitive action might follow if you provide honest feedback, this will not occur. Should you experience any discomfort, especially emotional discomfort, you can be referred for counseling, free of charge. Please note that your name will be replaced with a code and your response will by no means be linked to you. The researcher will give the final research report to the research entity / unit director. You can therefore terminate your participation at any stage without discrimination.

What will happen in the unlikely event of some form of discomfort occurring as a direct result of your taking part in this research study?

Should you have the need for further discussions after feeling emotional about your experiences; you can be referred to the university's employee assistance programme for counseling, free of charge.

Who will have access to the data?

- Anonymity will be applied as participants will not write their names on the interview schedules or in the transcriptions. Anonymity will be complete because no person other than the researcher will have access to your data and your particulars.
- Confidentiality will be ensured by reporting of findings in an anonymous manner. The researcher will ensure autonomy over the personal information obtained from the participants, meaning the researcher will keep in confidence all personal and other information obtained from the participants.
- The researcher will not allow anybody else to have access to the personal information obtained from the participants. Information will be kept under lock and key, the information will not be linked to the real identities of the participants.
- Only the researchers and the data analysis expert, who will sign a confidentiality agreement, will have access to the data. Data will be kept safe and secure by locking hard copies in locked cupboards in the researcher's office and electronic data will be password protected. As soon as data has been transcribed it will be deleted from the digital voice recorder. Data will be stored for seven (7) years on a password protected computer in a lockable office on the premises of the North-West University's Potchefstroom Campus.

What will happen with the data?

This is a once-off collection and data will be analysed in this study only.

Will you be paid to take part in this study and are there any costs involved?

No, you will not be paid to take part in the study. Although it is planned to conduct interviews at the University, travel expenses will be paid for those participants who have to travel to the site only if applicable. There will thus be no costs involved for you, if you do take part.

Is there anything else that you should know or do?

- You can contact the researcher, Mr Tertius Bester at 082 788 9809 if you have any further queries or encounter any problems.
- You can contact the Faculty of Economic and Management Sciences Ethics Committee via Ms MM Heyns at 018 299 1419, marita.heyns@nwu.ac.za if you have any concerns or complaints that have not been adequately addressed by the researcher.
- You will receive a copy of this information and consent form for your own records.

How will you know about the findings?

The research findings will be reported back to the research entity / unit director. The researcher will also declare his availability to present the results to the research entity / unit's management, and if requested, participants will be invited to the presentation.

DECLARATION BY PARTICIPANT

By signing below, I ..... agree to take part in the research titled *Research data management framework for a South African university-based research entity*

I declare that:

- I have read this information and consent form and it is written in a language with which I am fluent and comfortable.
- I have had a chance to ask questions to both the person obtaining consent, as well as the researcher and all my questions have been adequately answered.
- I understand that taking part in this study is **voluntary** and I have not been pressurised to take part.
- I may choose to leave the study at any time and will not be penalised or prejudiced in any way.
- I may be asked to leave the study before it has finished, if the researcher feels it is in my best interests, or if I do not follow the study plan, as agreed to.

Signed at (*place*) ..... on (*date*) ..... 20...

....

Signature of participant **Signature of witness**

DECLARATION BY PERSON OBTAINING CONSENT

I (*name*) .....declare that:

- I explained the information in this document to .....
- I encouraged him/her to ask questions and took adequate time to answer them.
- I am satisfied that he/she adequately understands all aspects of the research, as discussed above
- I did/did not use an interpreter.

Signed at (*place*) ..... on (*date*) ..... 20...

....

Signature of person obtaining consent **Signature of witness**

DECLARATION BY RESEARCHER

I ..... declares that:

- The mediator did explain the information in this document to ..... after I explained the research in detail to the mediator.
- The participant was encouraged to ask questions to the mediator and myself and adequate time was taken to answer him/her.
- The participant understands all aspects of the research as discussed above, adequately.
- I did/did not use an interpreter.

Signed at (place) ..... on (date) ..... 20...

.....

Signature of researcher

.....

Signature of witness

ADDENDUM C: INTERVIEW SCHEDULE

PARTICIPANT CODE: _____

Demographic data

Position in research unit: _____

Amount of years' experience in research unit: _____

Projects that you are involved in:

	Project 1	**Project 2**
Project name:		
Role(s) in project:		
Project start and end:		

Data formats in project:		
Date for data destruction:		

Semi-structured, individual interview schedule

1. What do you understand about research data management (RDM)?
2. Why do you think RDM is important in a research unit?

   Probing questions about:

   2.1 Storage.

   2.2 Security.

   2.3 Preservation.

   2.4 Compliance.

   2.5 Quality.

   2.6 Sharing.

   2.7 Jurisdiction.

3. What components of RDM do you think this unit have in place?

   Probing questions about:

   3.1 Strategies.

   3.2 Policies.

3.3 Guidelines.

3.4 Processes.

3.5 Technologies.

3.6 Services.

4. What factors do you think are having the most influence on RDM within this unit?

Probing questions about:

4.1 Acceptance.

4.2 Culture.

4.3 Demand.

4.4 Incentives.

4.5 Roles.

4.6 Governance.

4.7 Politics.

4.8 Resources.

4.9 Projects.

4.10 Skills.

4.11 Communications.

4.12 Context.

5. Who would you identify as the major stakeholders in RDM for this research unit?

Probing questions about:

5.1 Library.

5.2 IT services.

5.3 Other academic departments.

5.4 Senior university management.

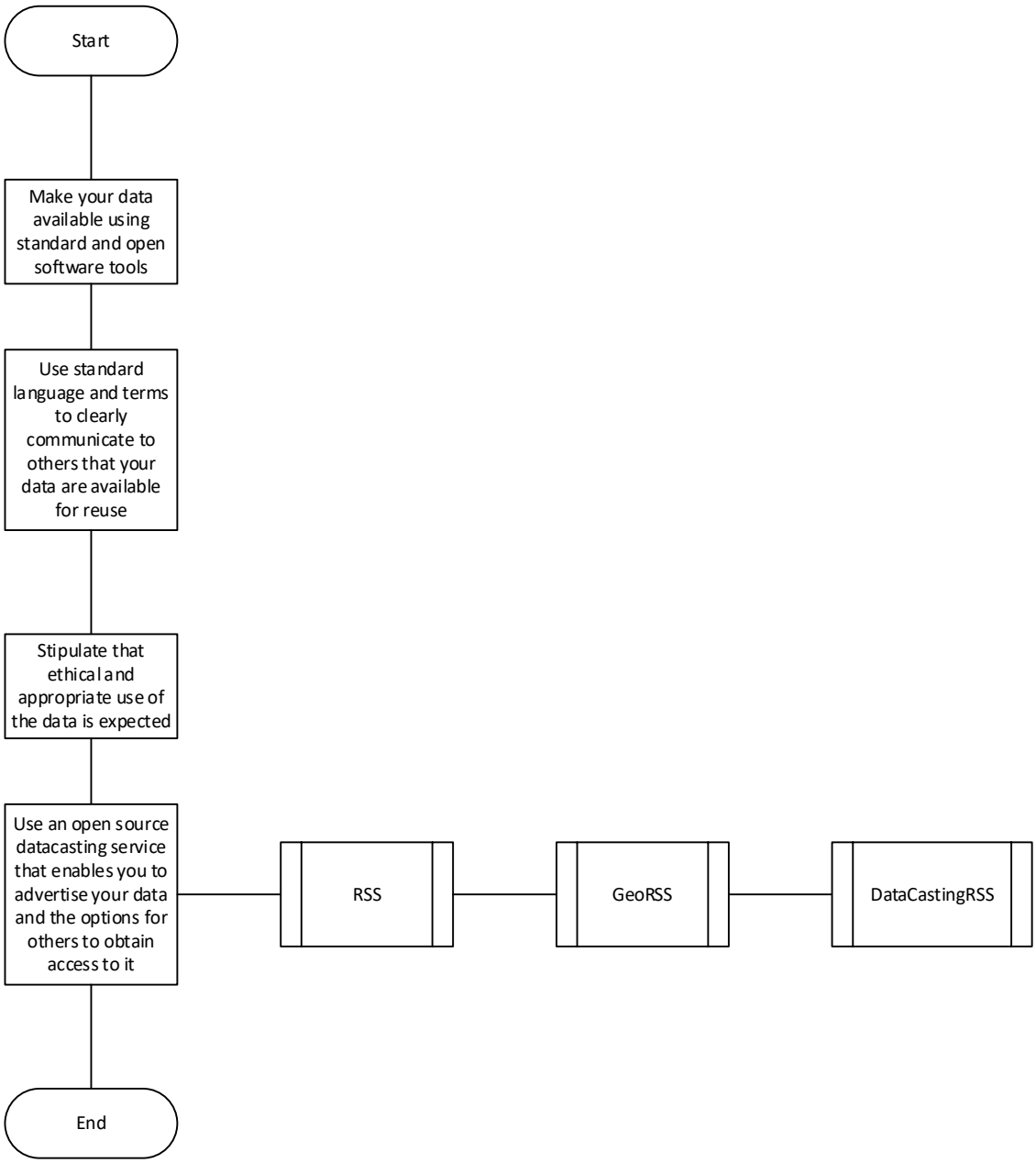
5.5 Research support services.

5.6 Statistical services.

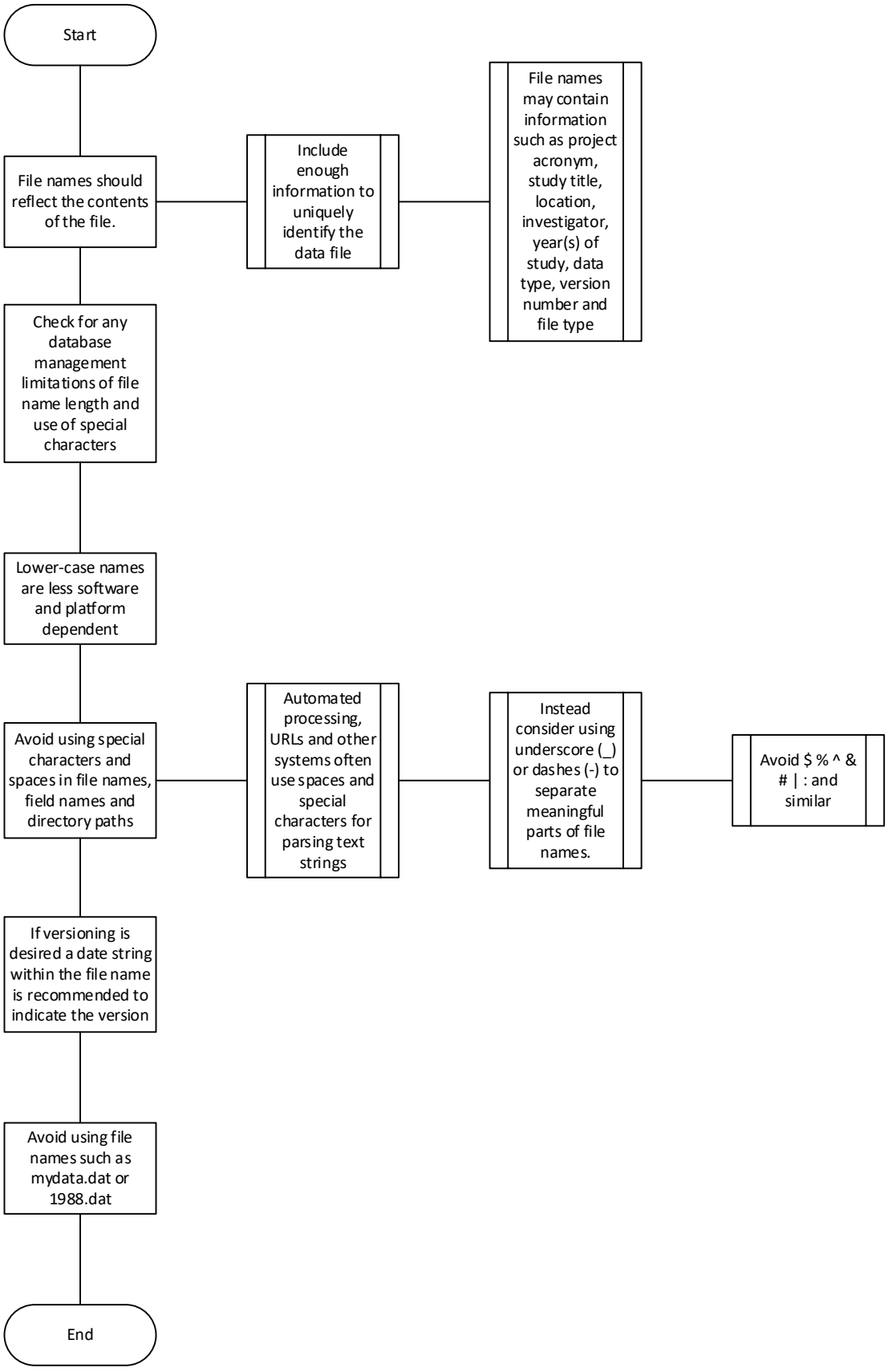
5.7 Other support services.

ADDENDUM D: PRELIMINARY RDM FRAMEWORK BEST PRACTICES, GUIDELINES

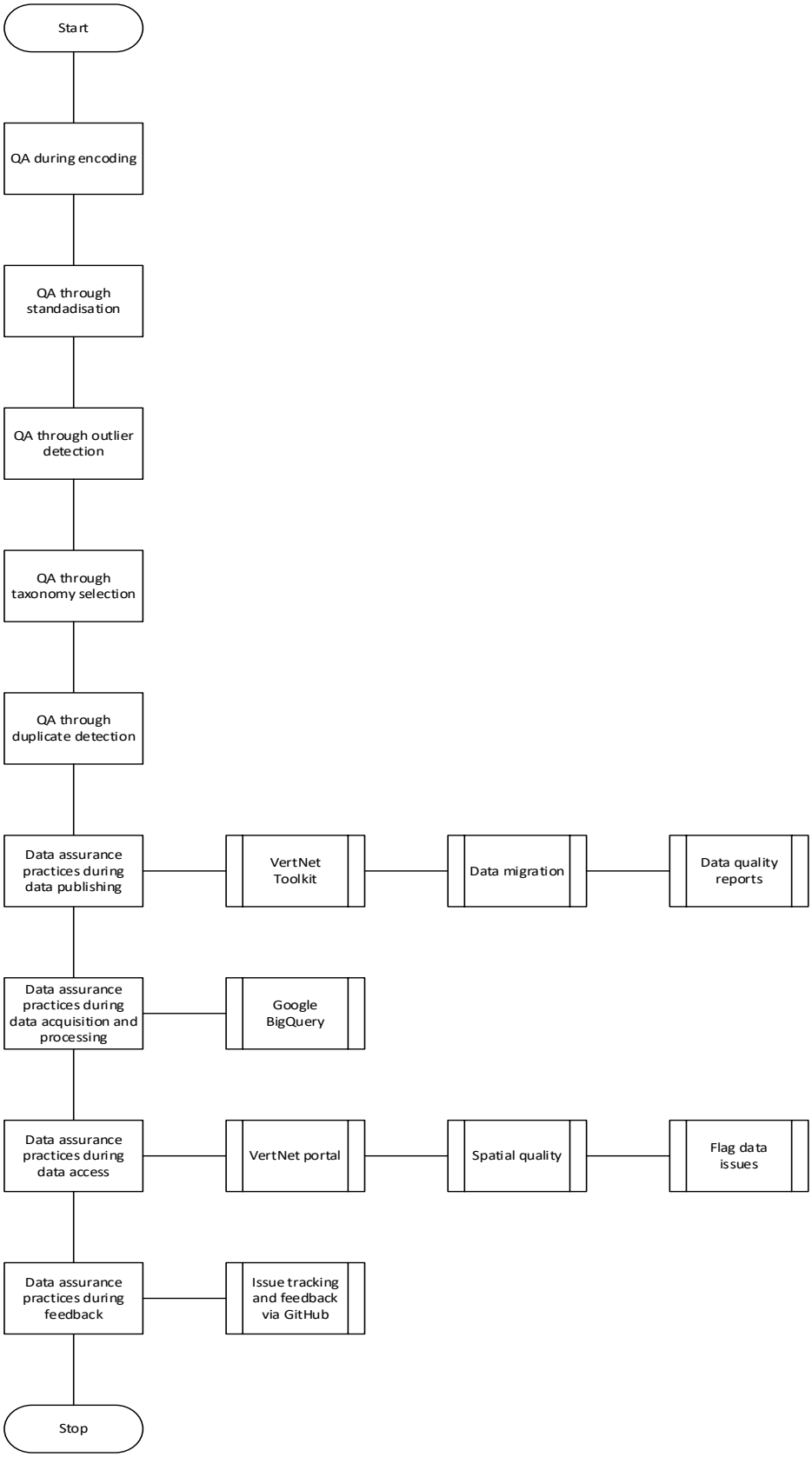
A1 - USE DATACASTING TOOLS TO ADVERTISE YOUR DATA



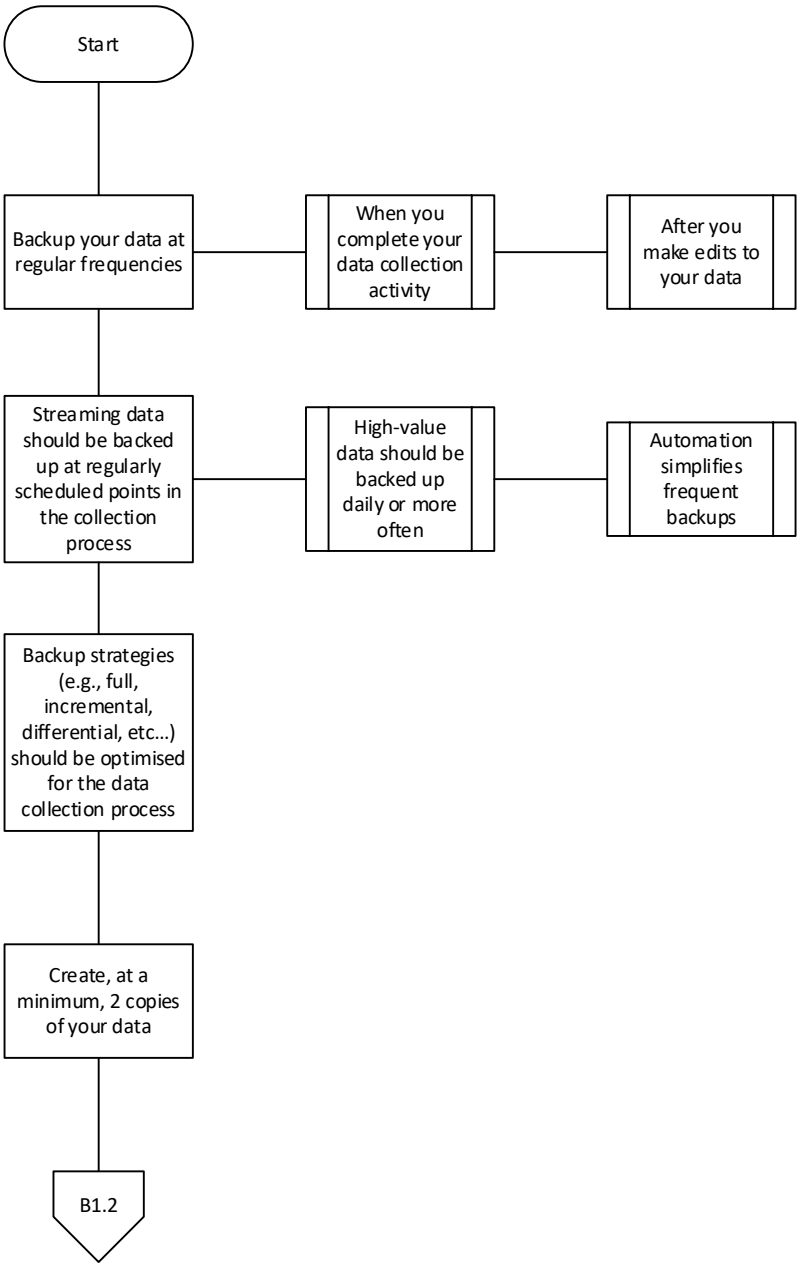
A3 - Give files descriptive names



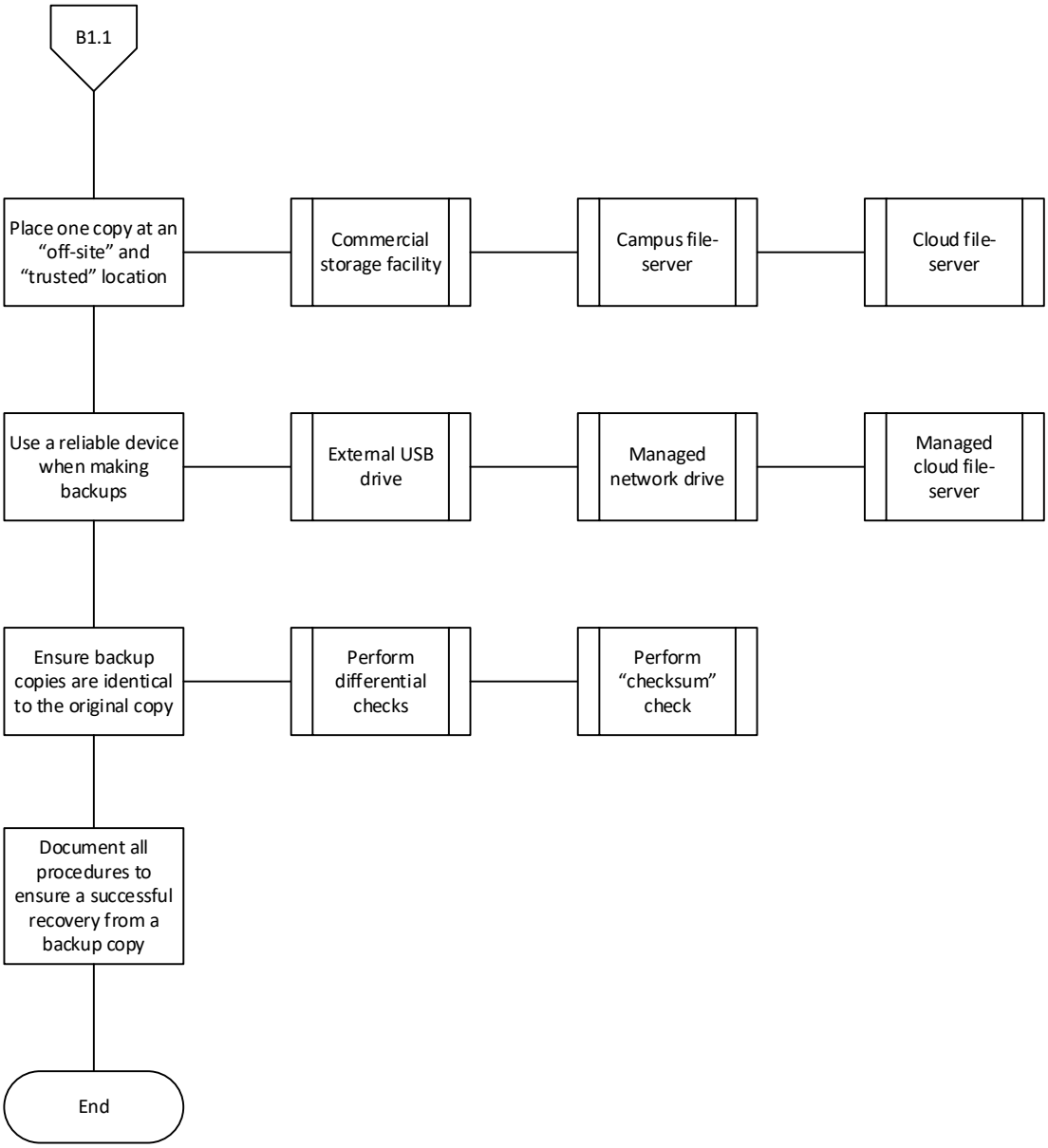
A4 - Quality assurance of research data



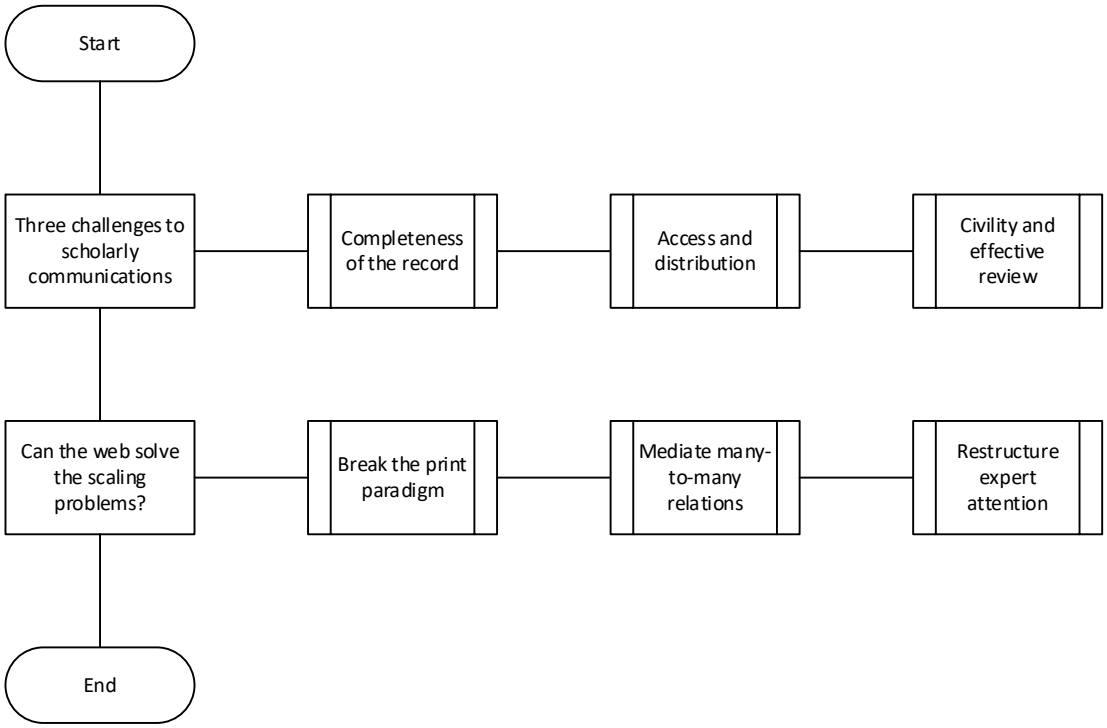
B1.1 - Backup of research data



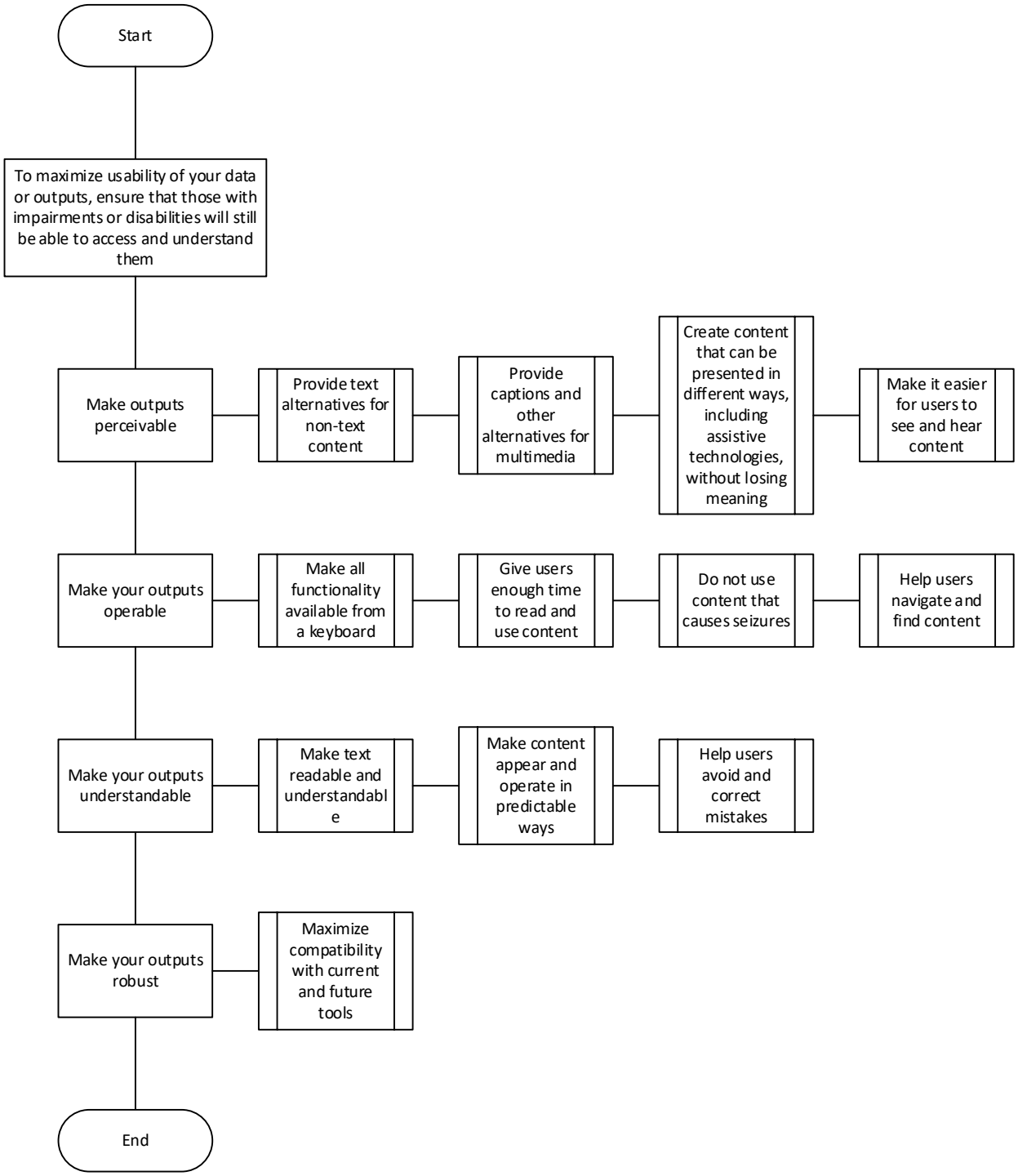
B1.2- Backup of research data



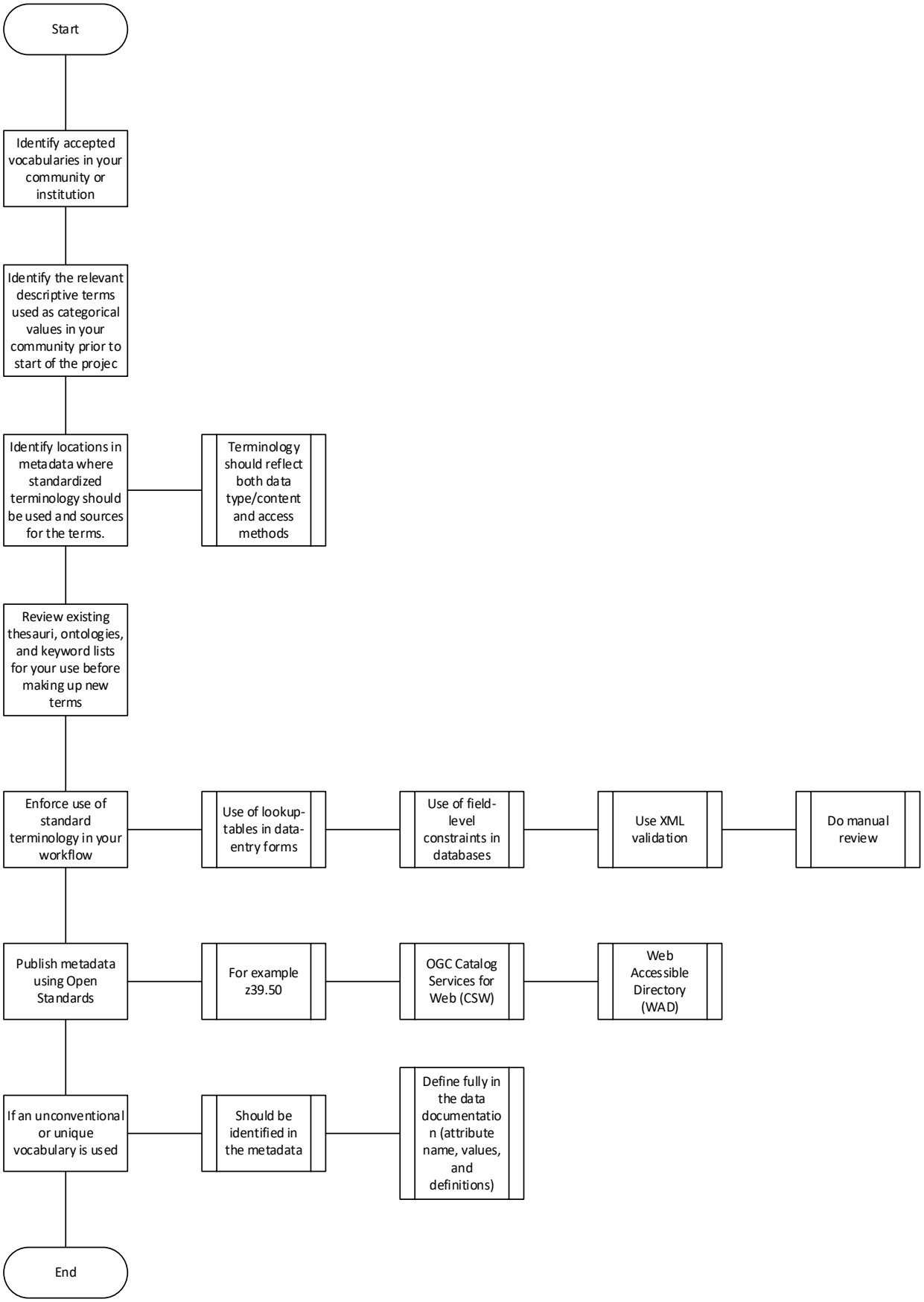
B2 - The impact of Boyle's Laws



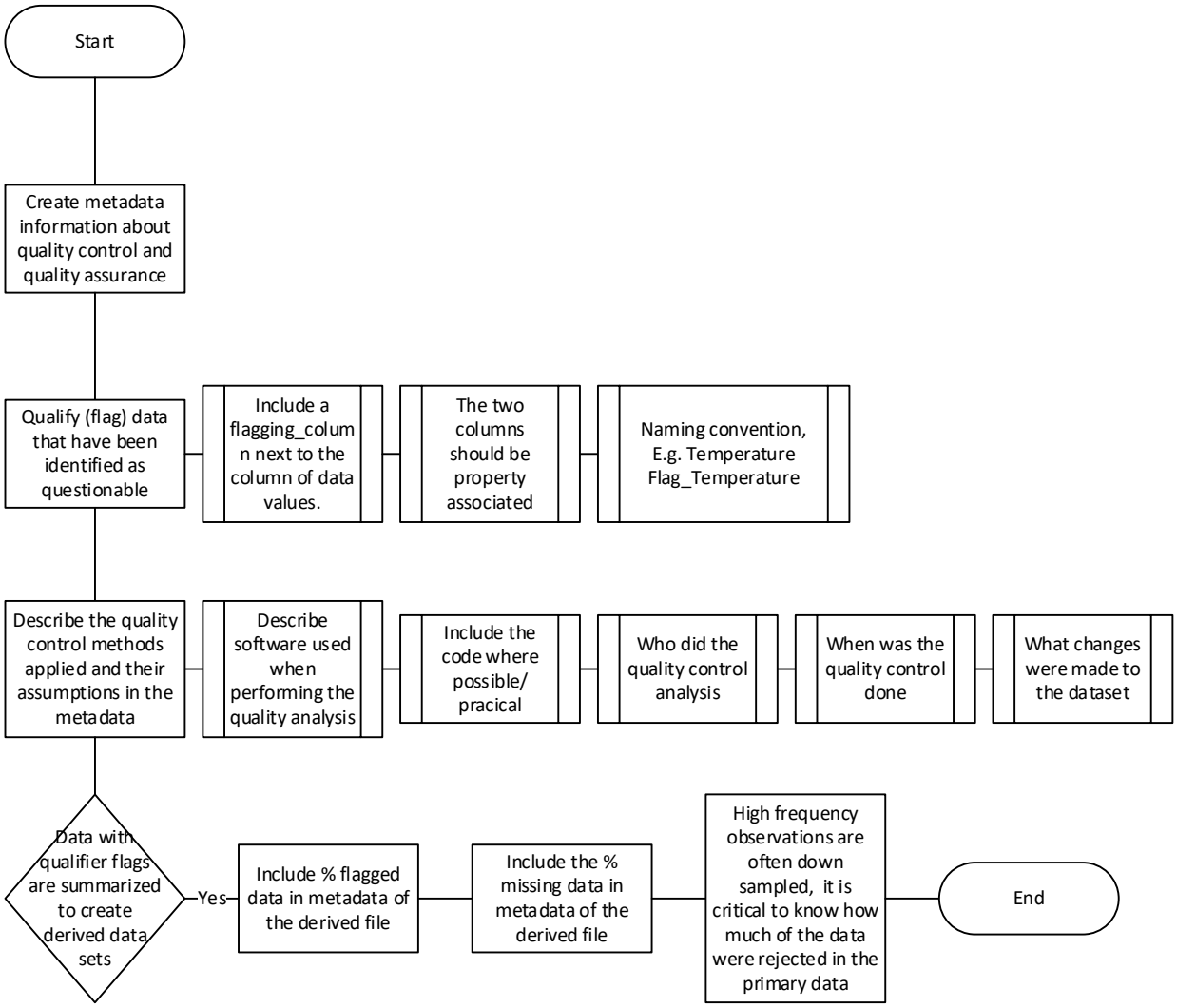
C1 - Ensure accessibility for multiple channels



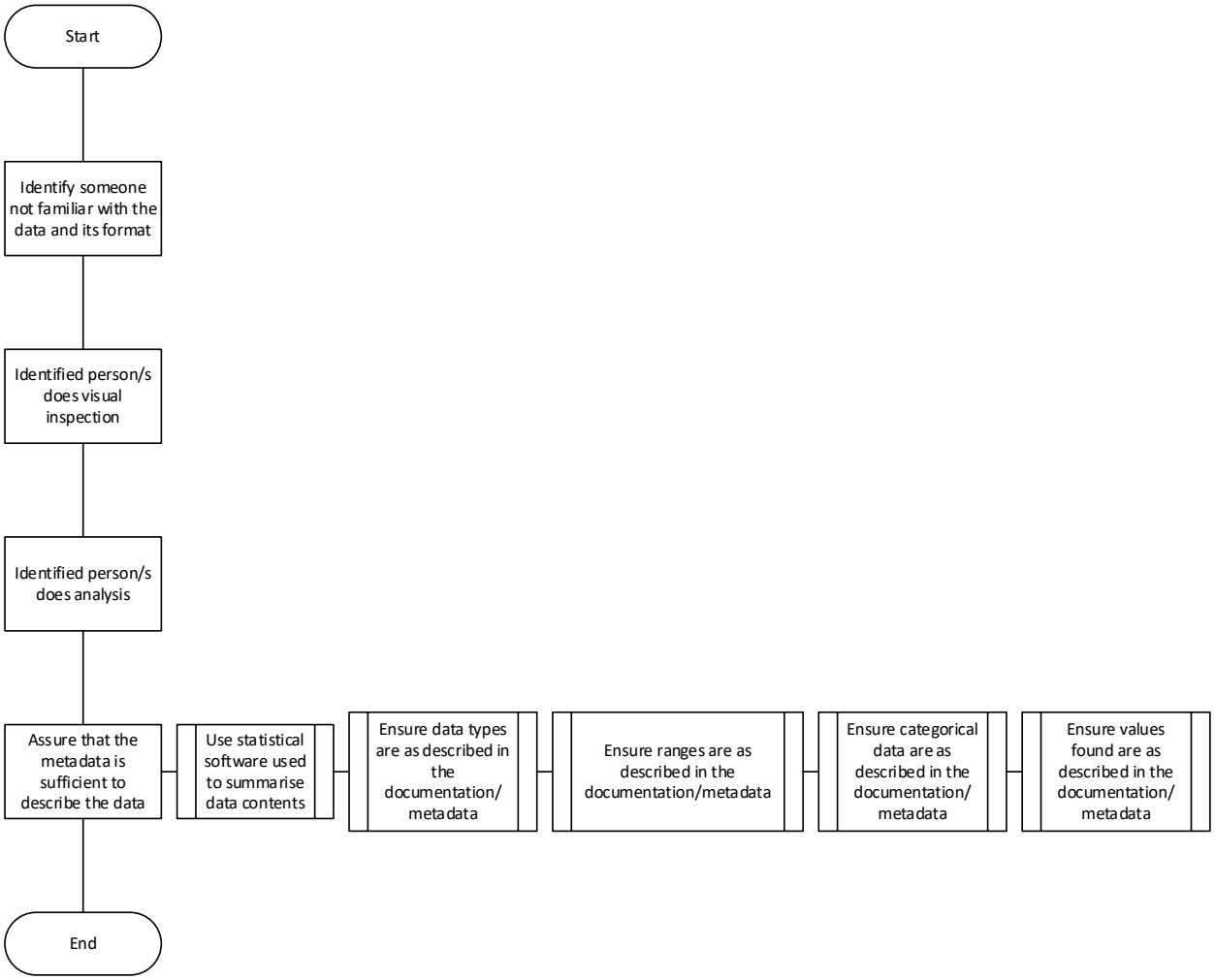
C2 - Enable discovery through standard terminology



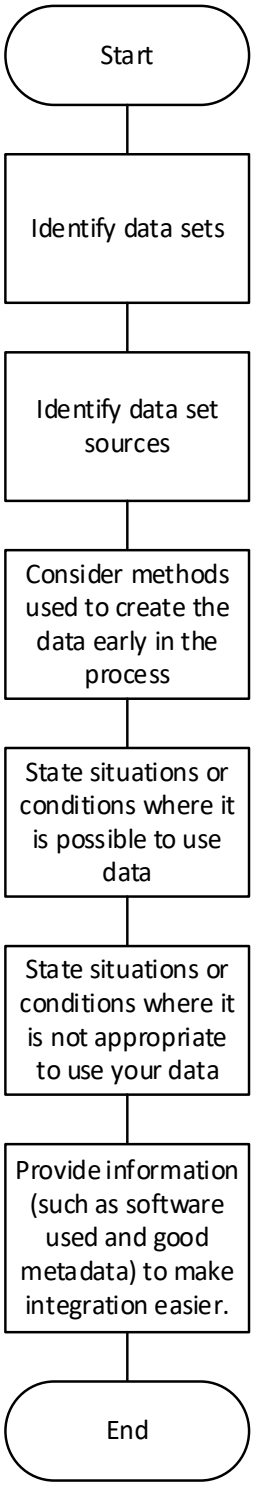
C3 Data quality communications



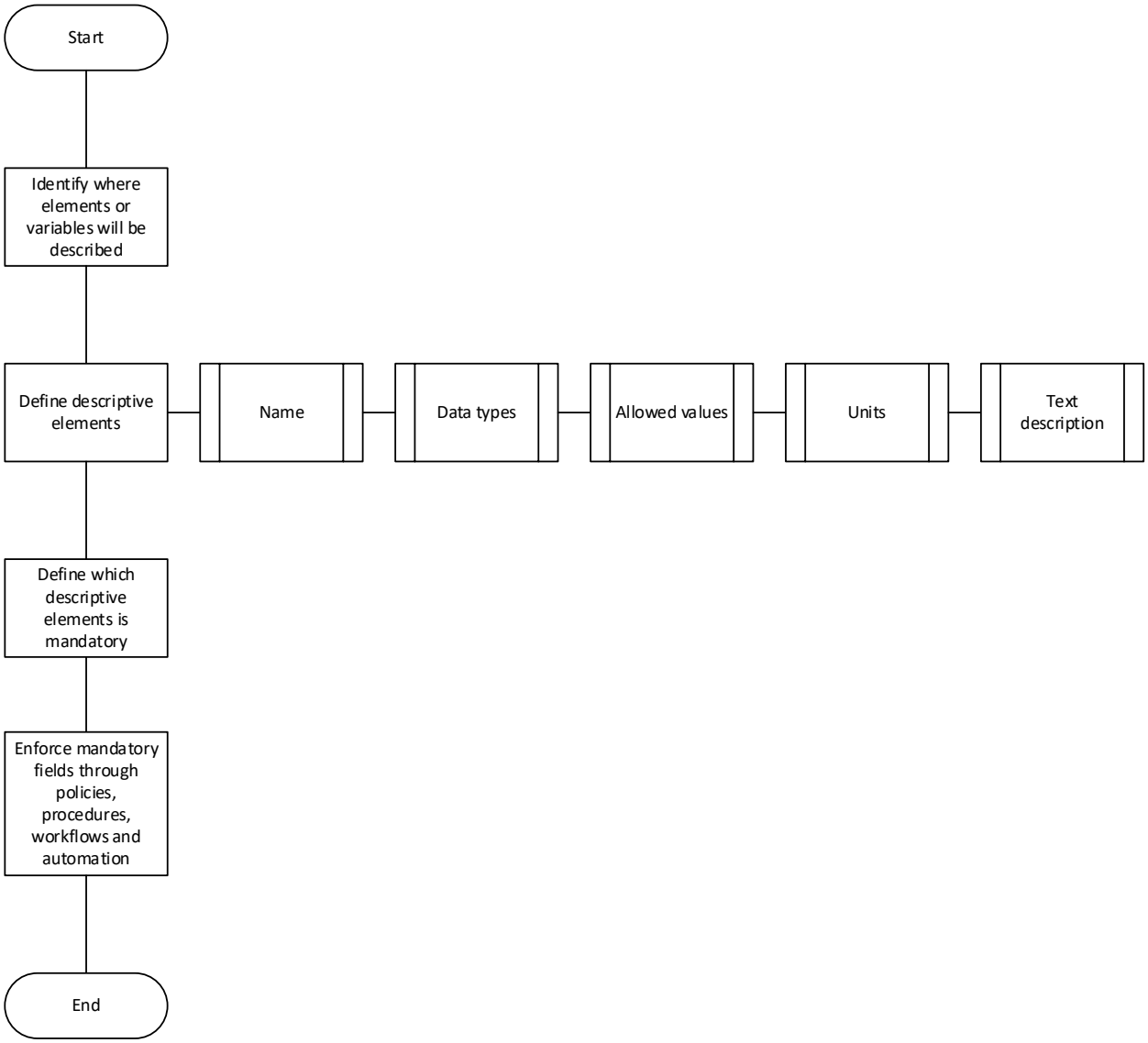
C4 - Ensure data and metadata are consistent



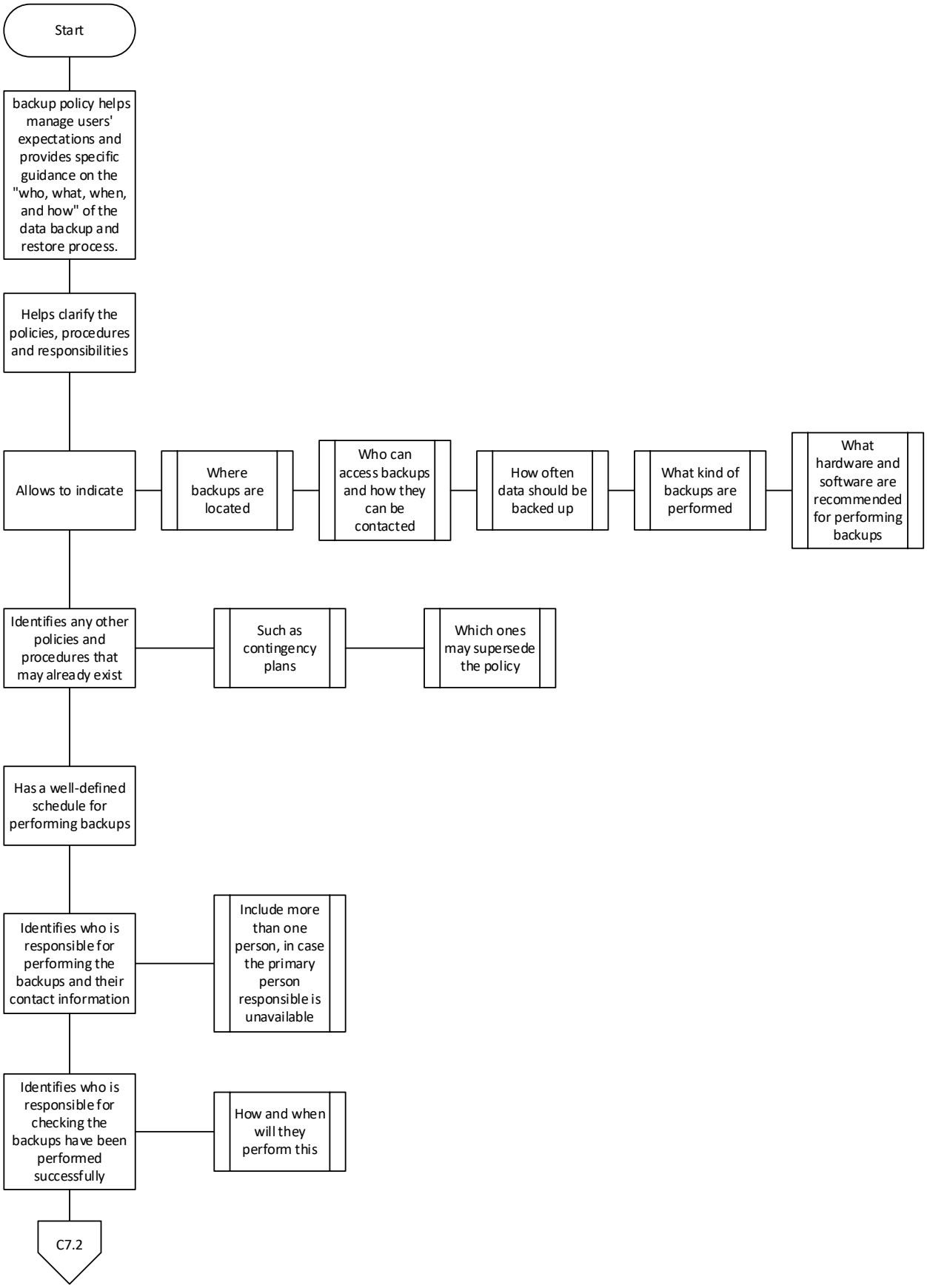
C5 - Ensure data can be integrated



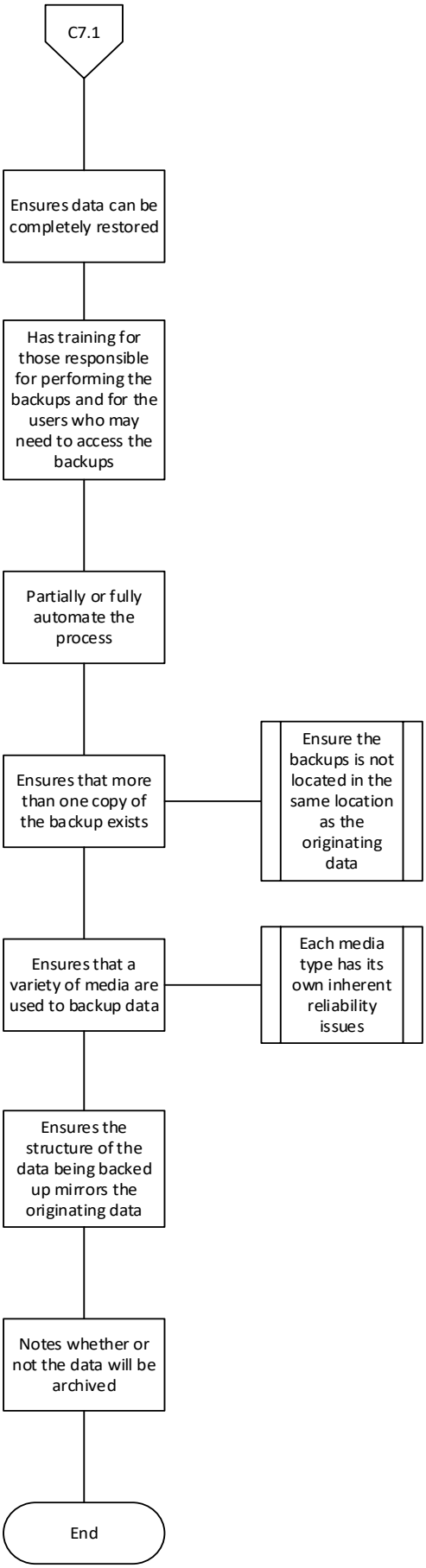
C6 - Data dictionary creation



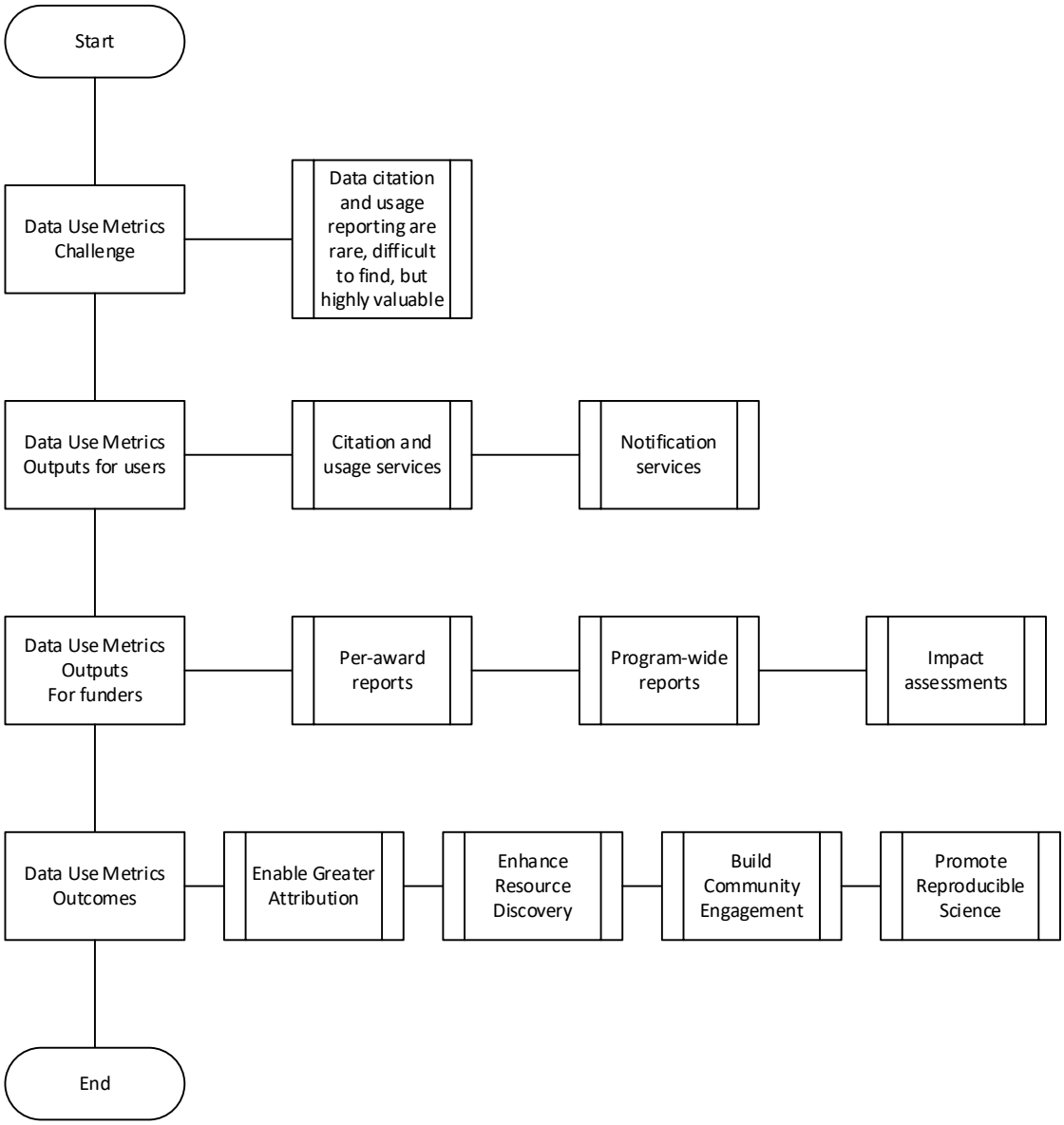
C7.1 - Backup policy – Importance of documentation



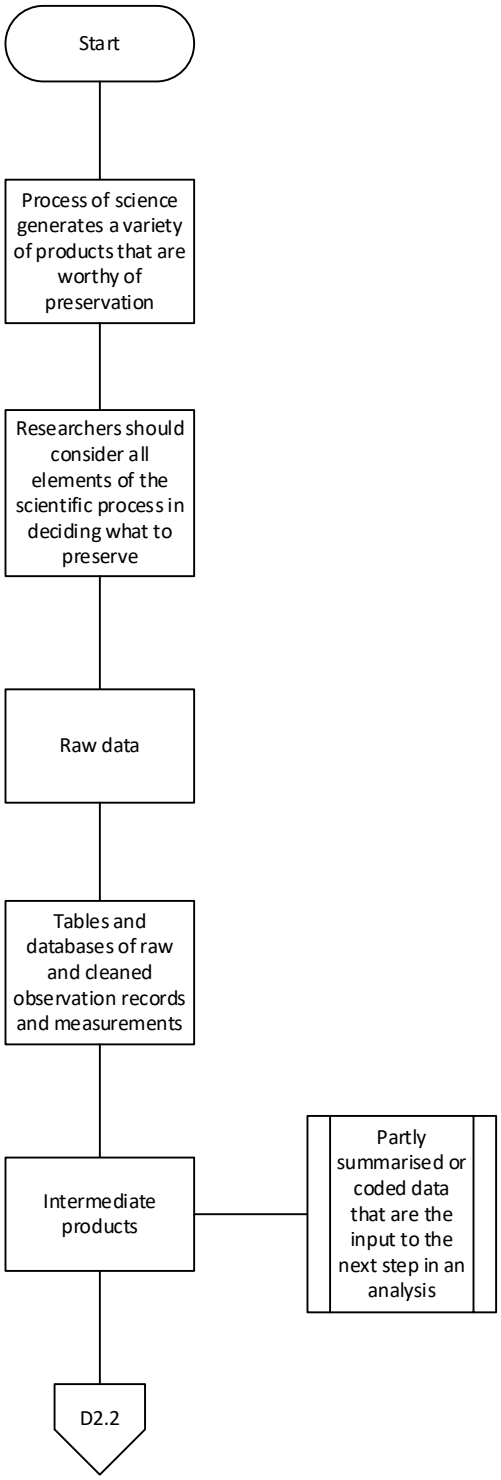
C7.1 - Backup policy – Importance of documentation



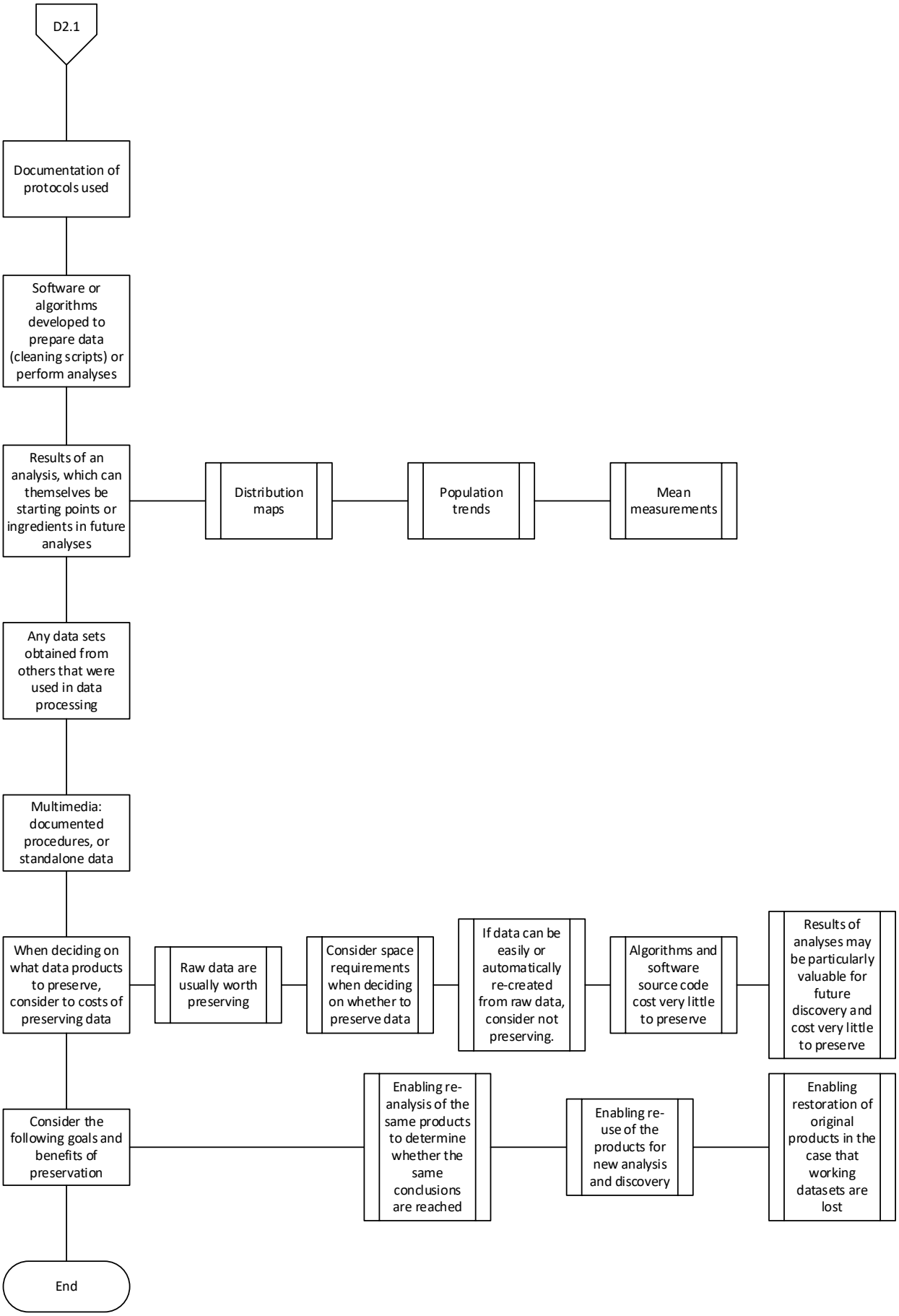
D1 - Tools and services - Considerations



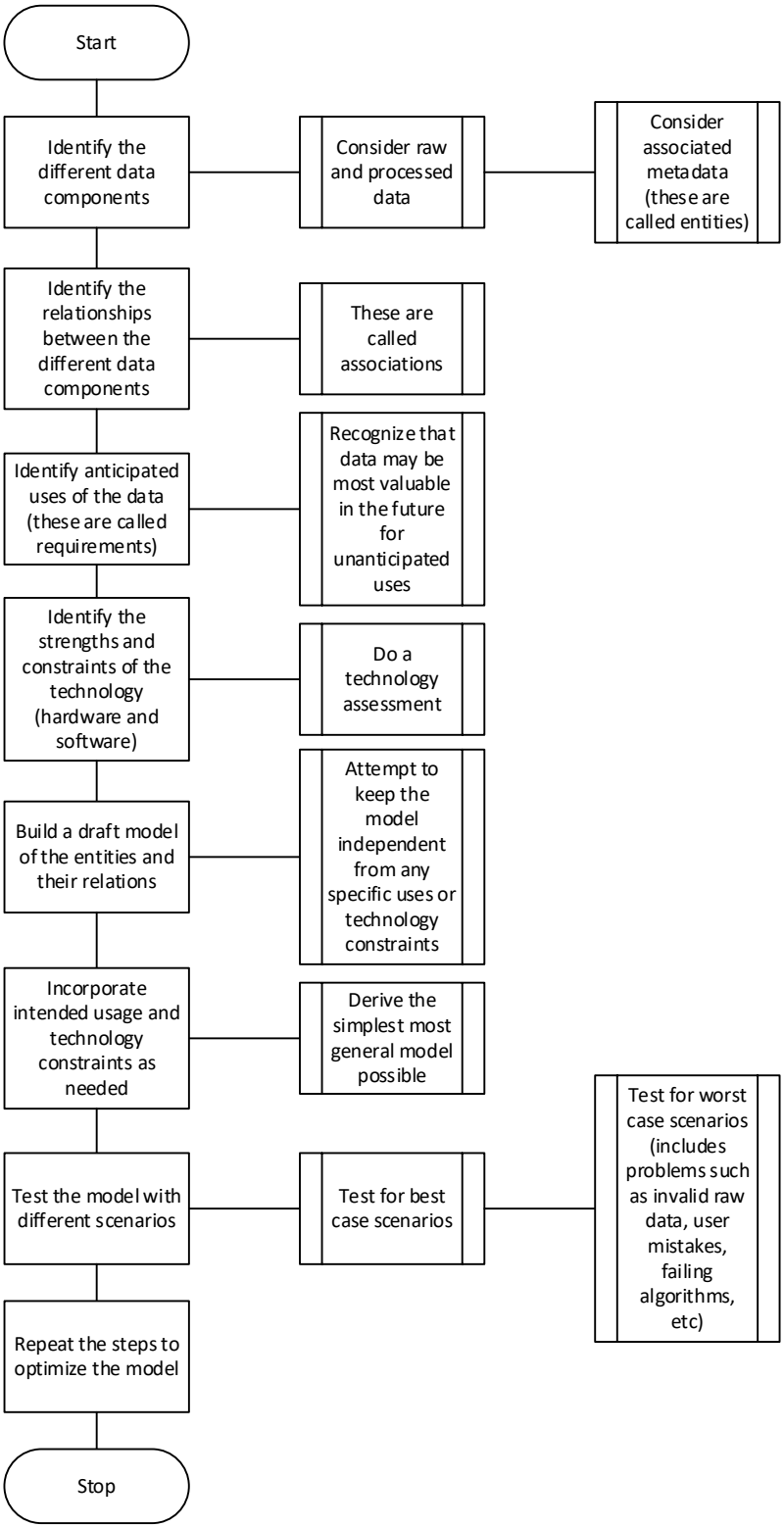
D2.1 - Data preservation – How to decide



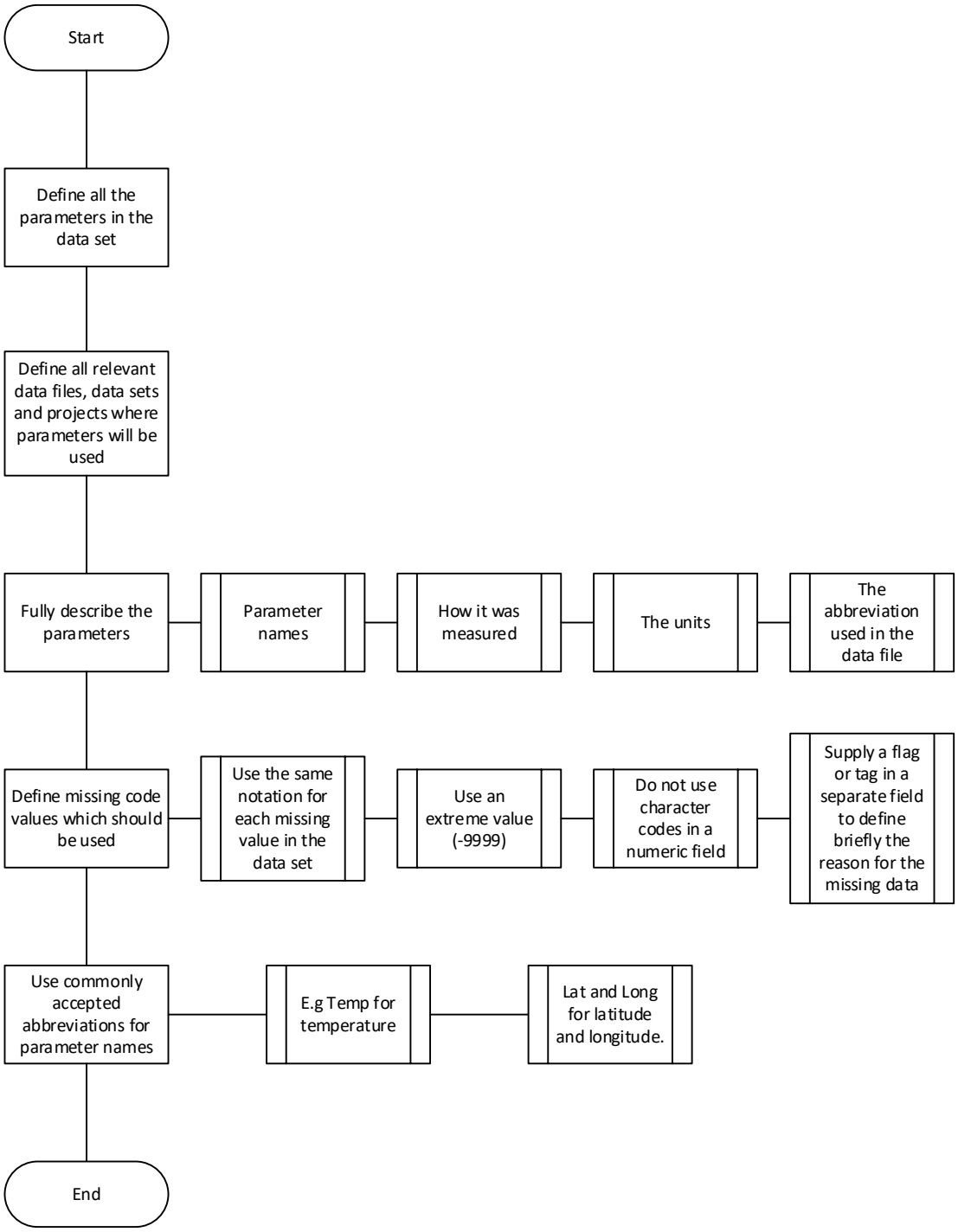
D2.2 - Data preservation – How to decide



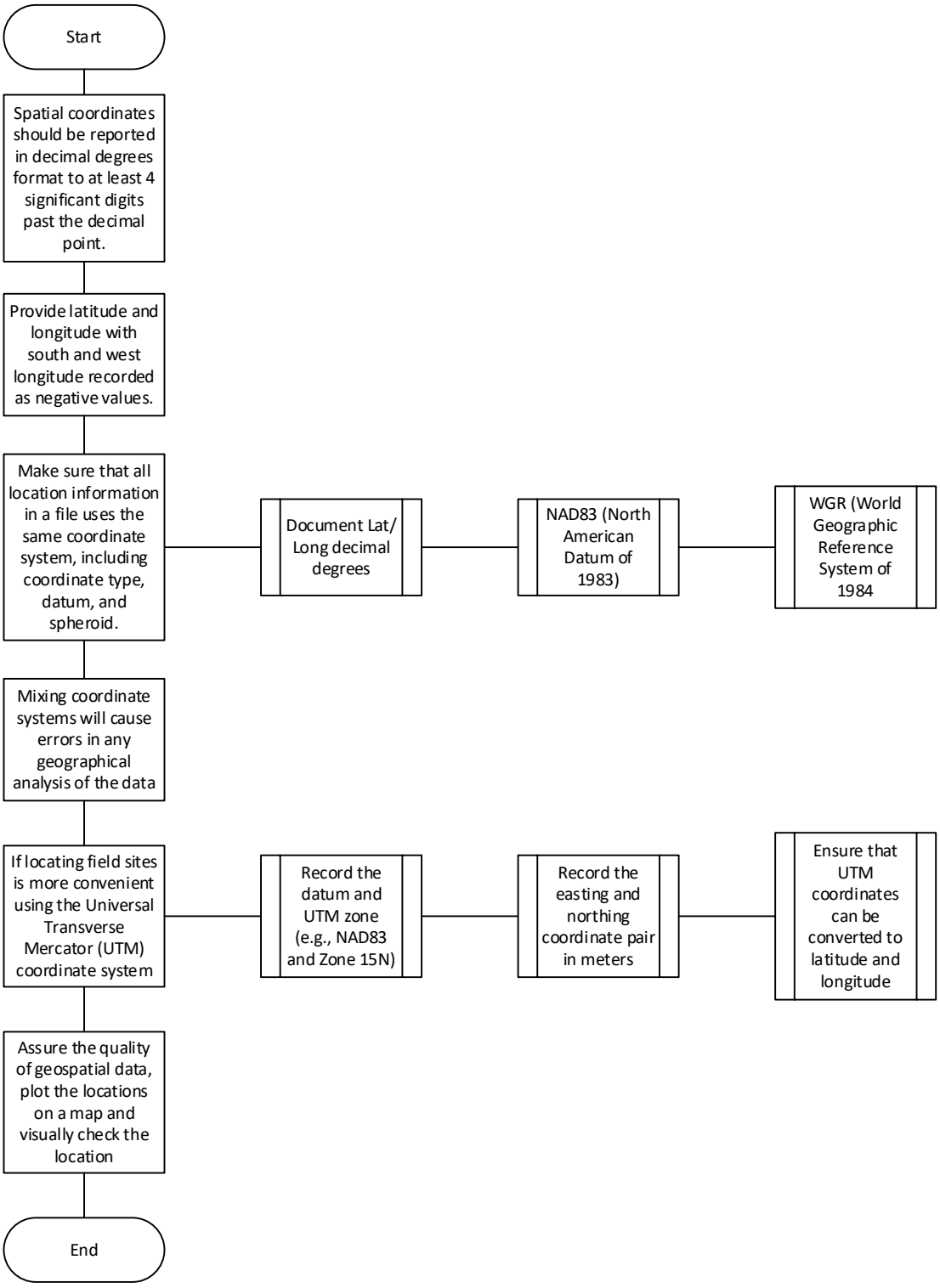
D5 - Data model definition



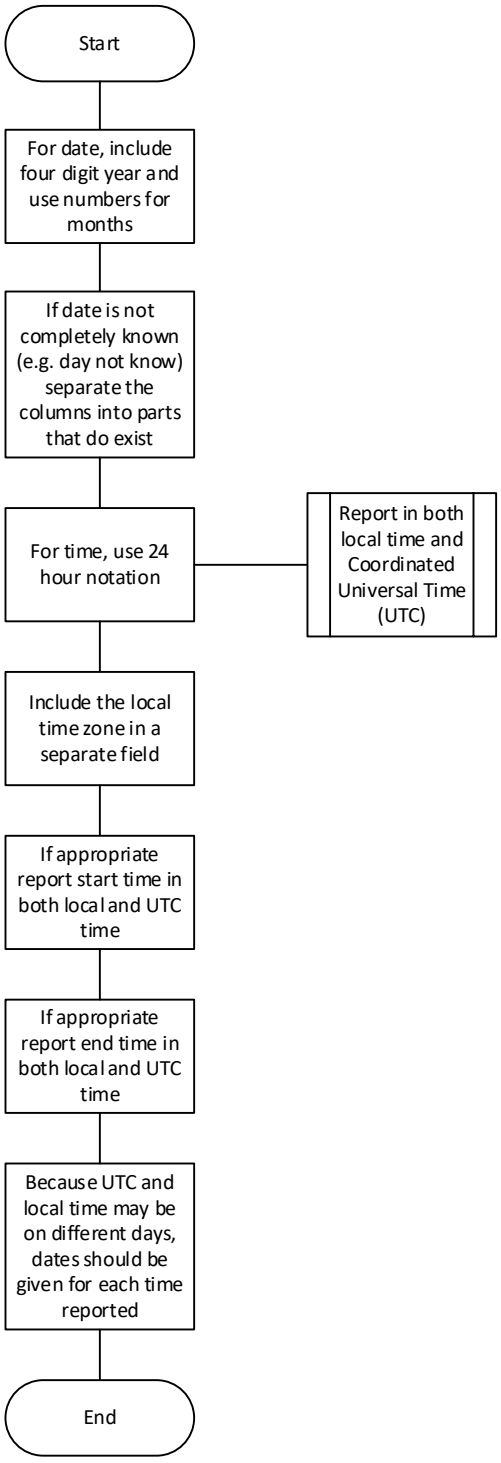
D6 - Parameter definitions



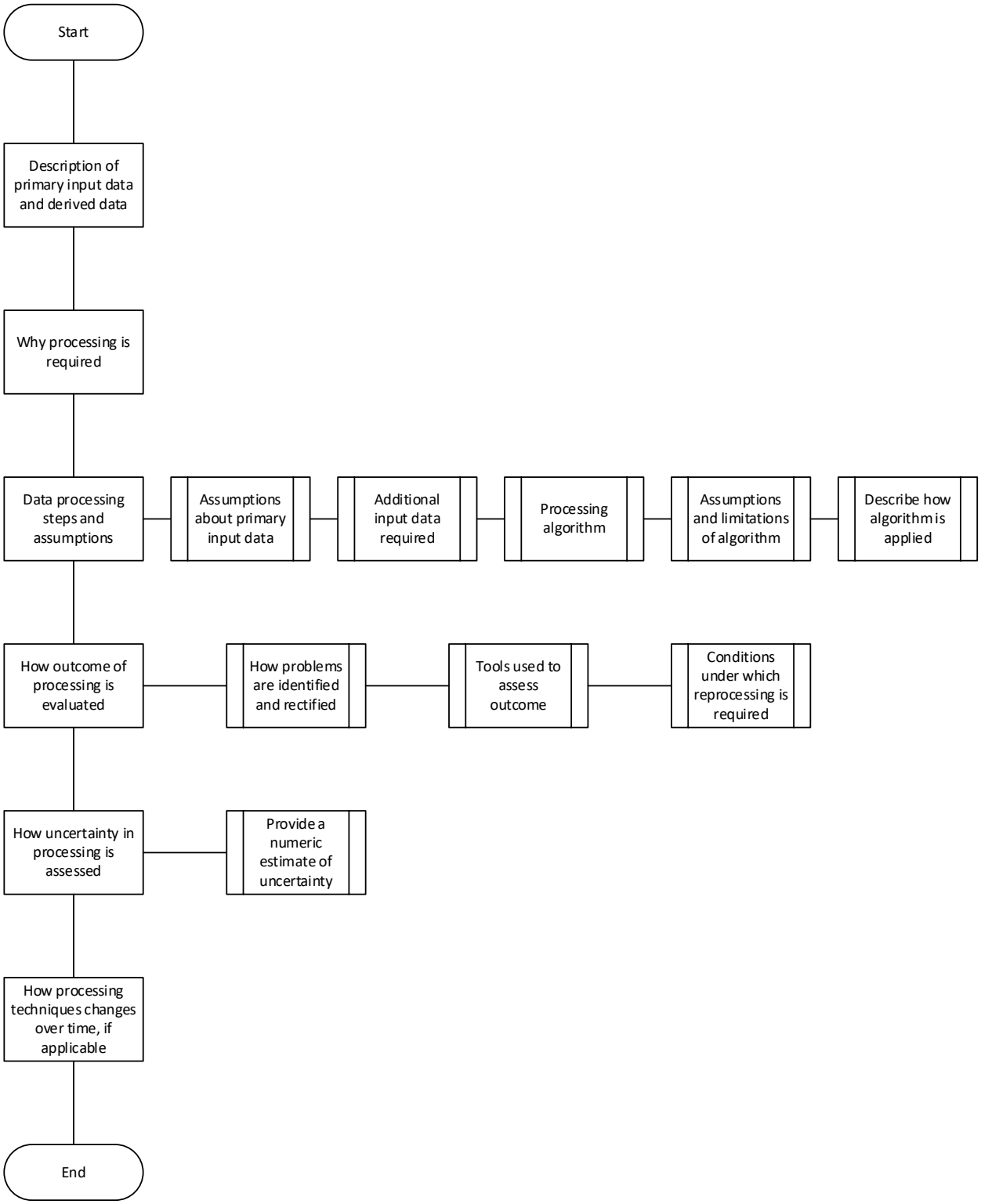
D7 - Format of spatial parameters



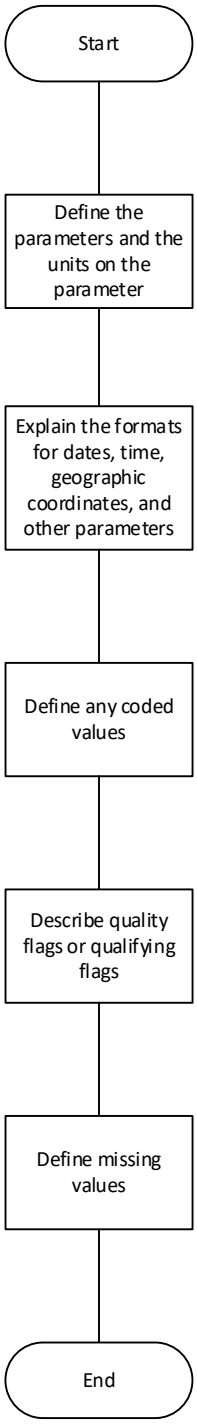
D8 - Standardise on time and date storage



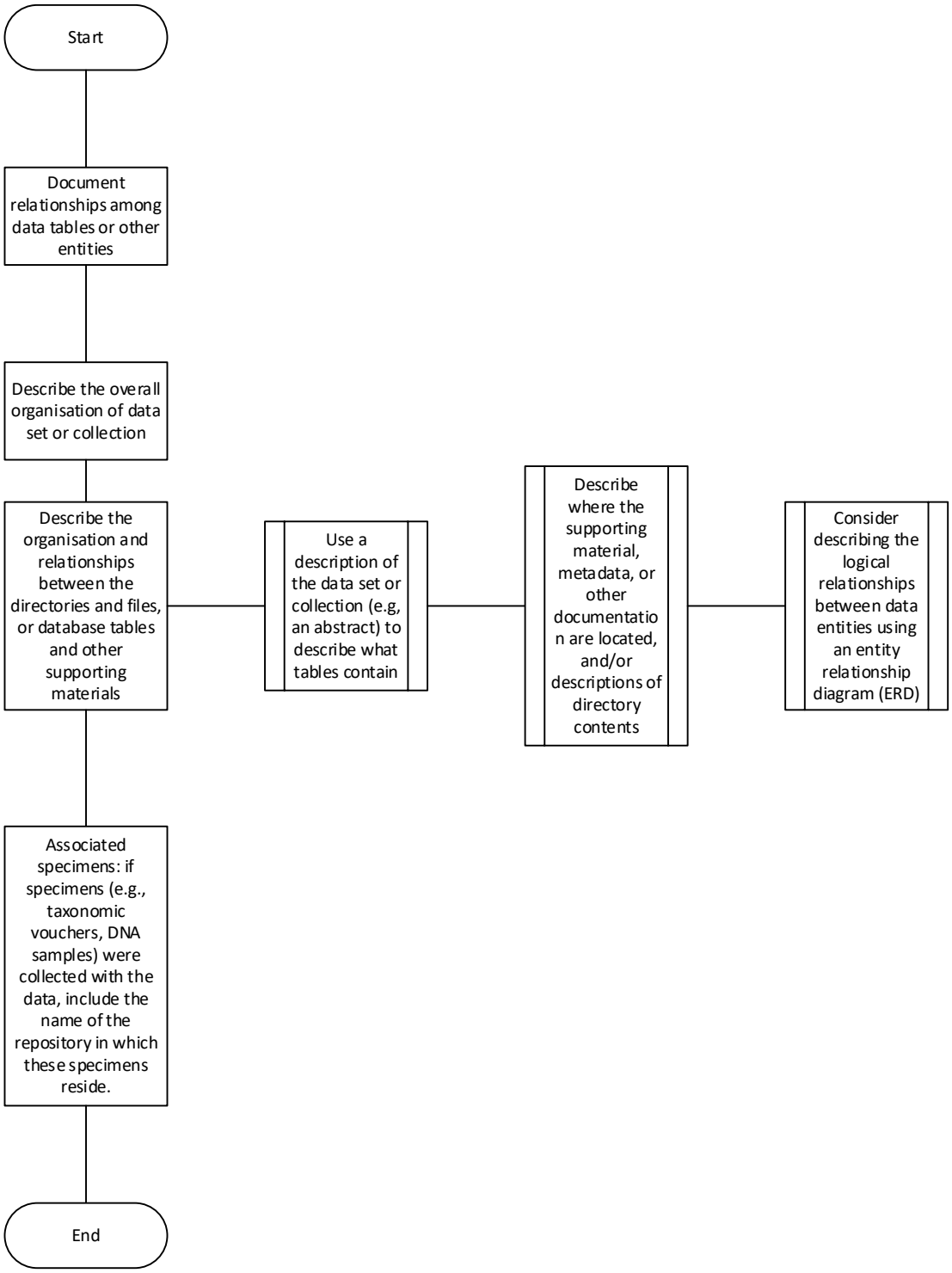
D9 - Describe provenance of data products



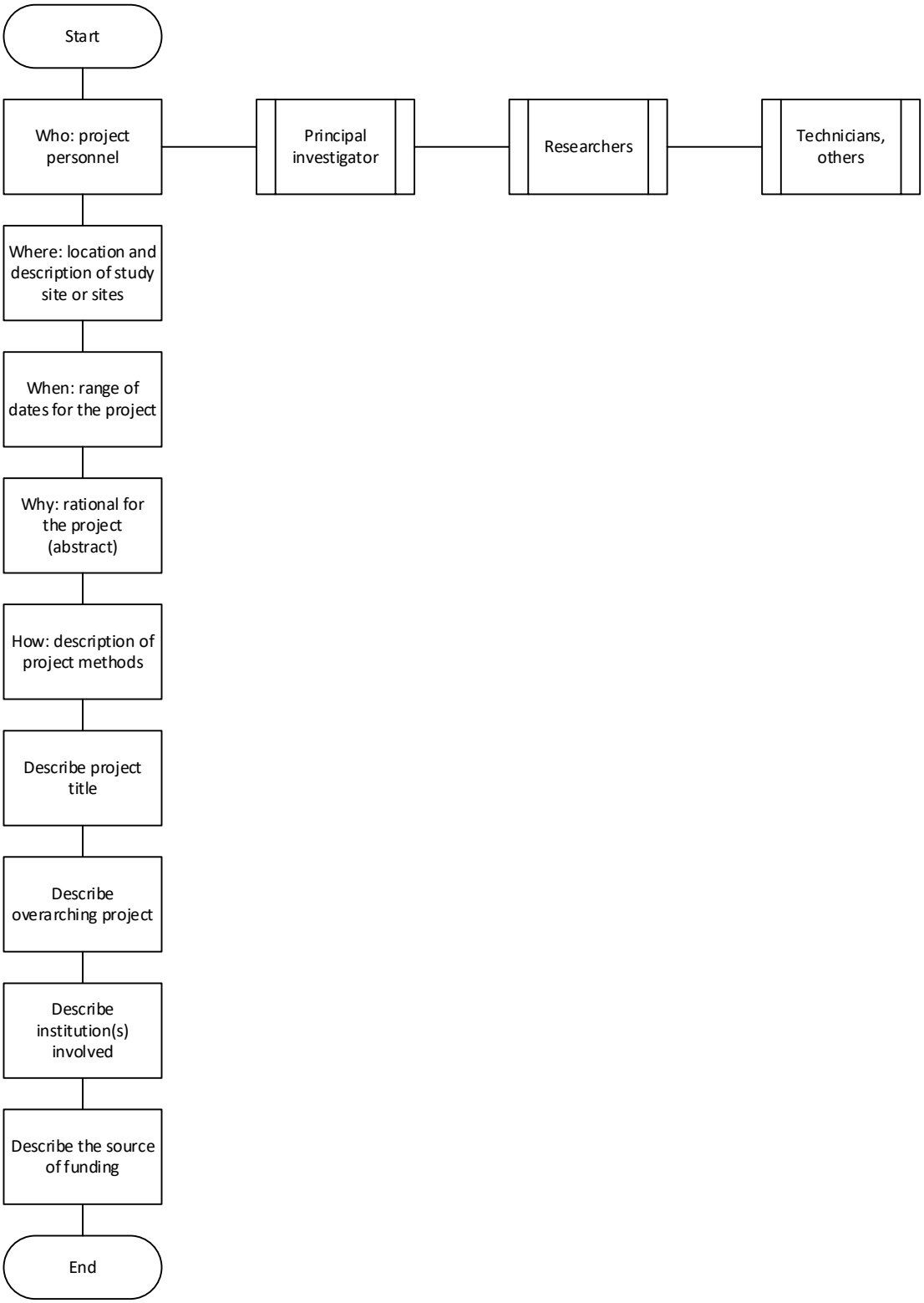
D10 - Ensure data contents are clear



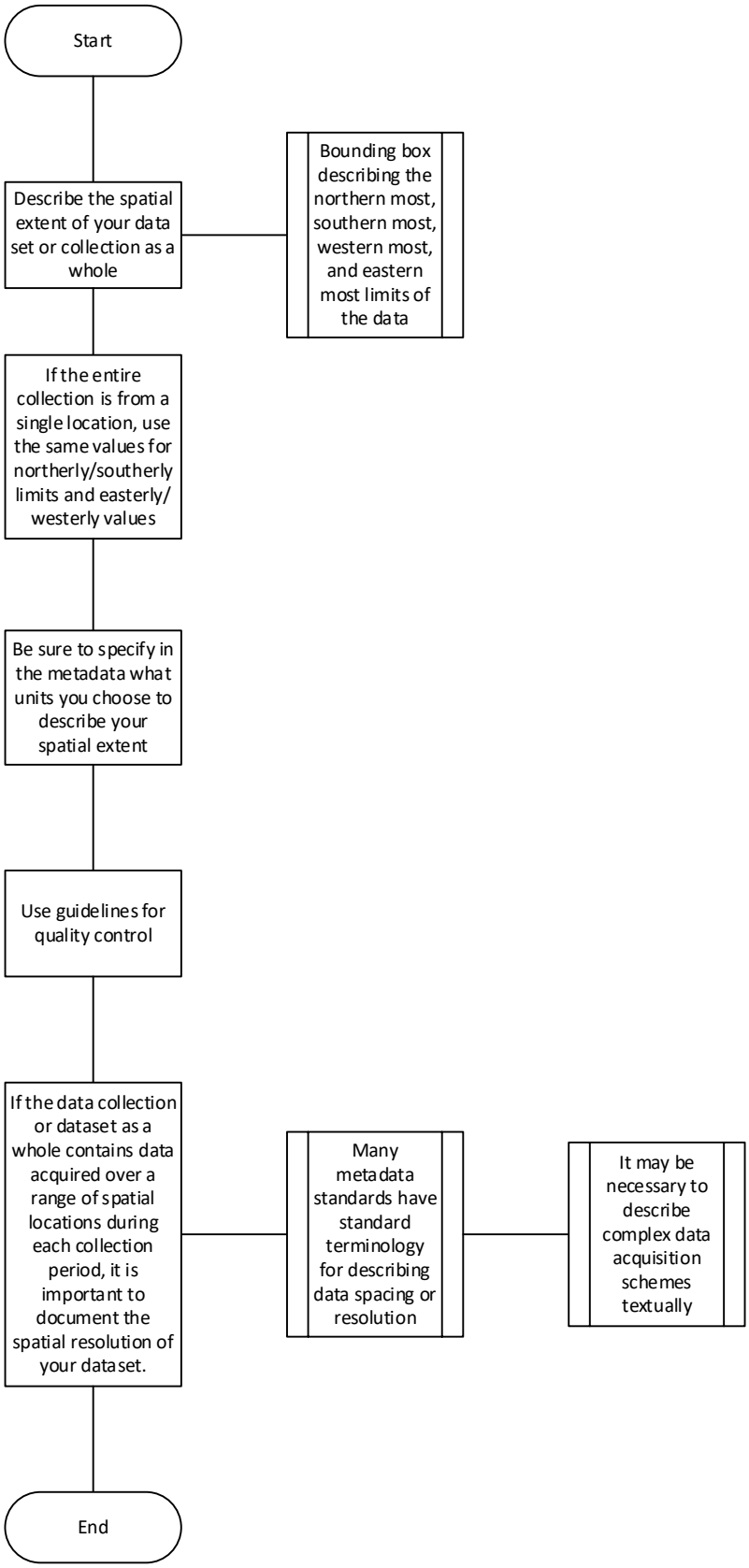
D11 - Dataset organisation



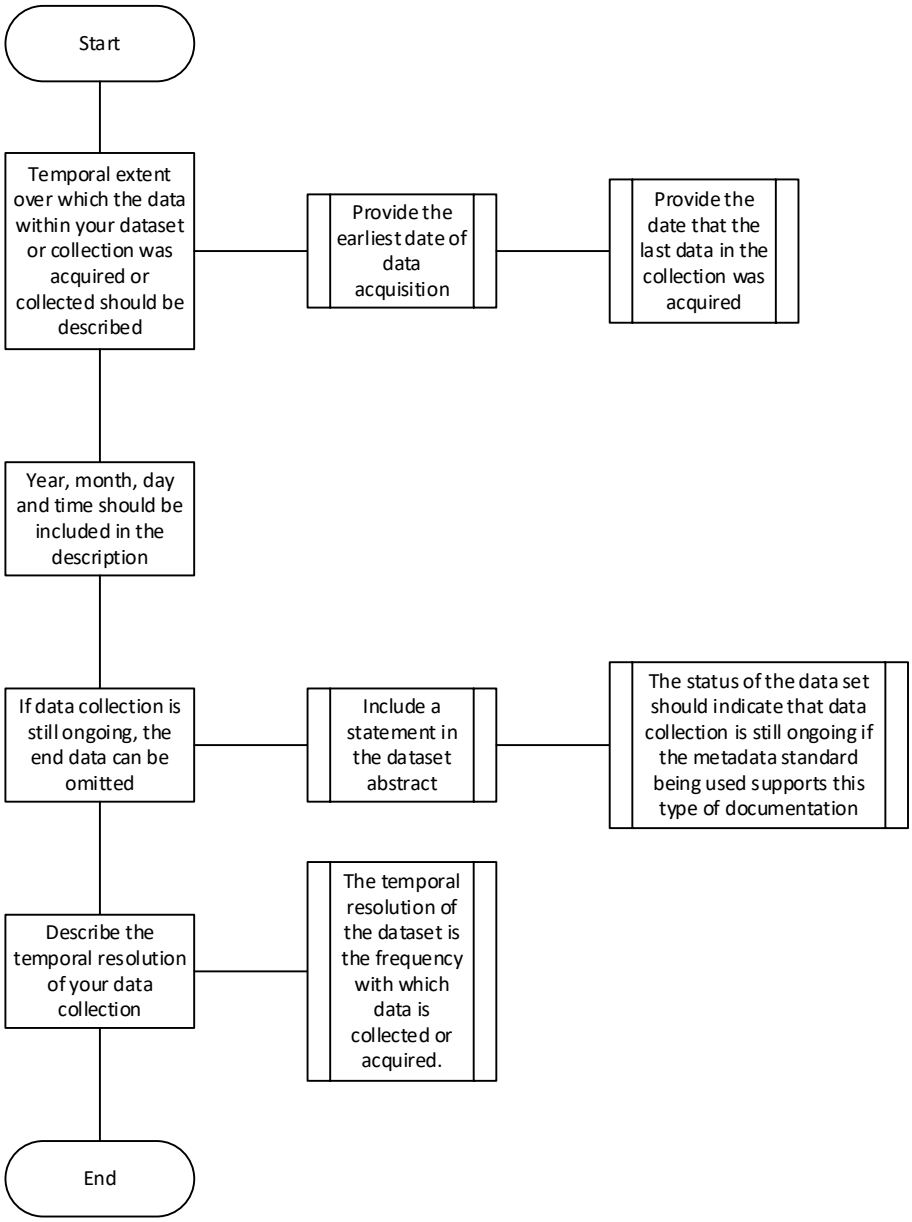
D12 - Research project description guidelines



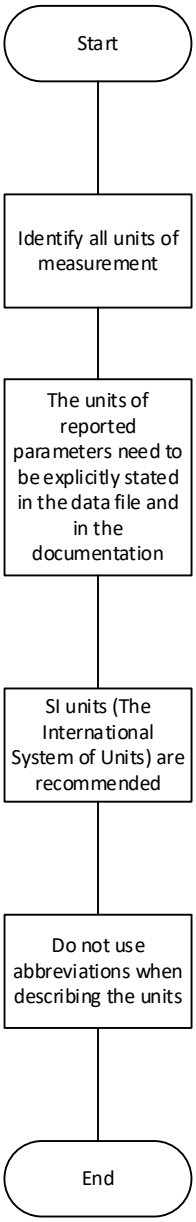
D13 - Dataset spatial extent and resolution



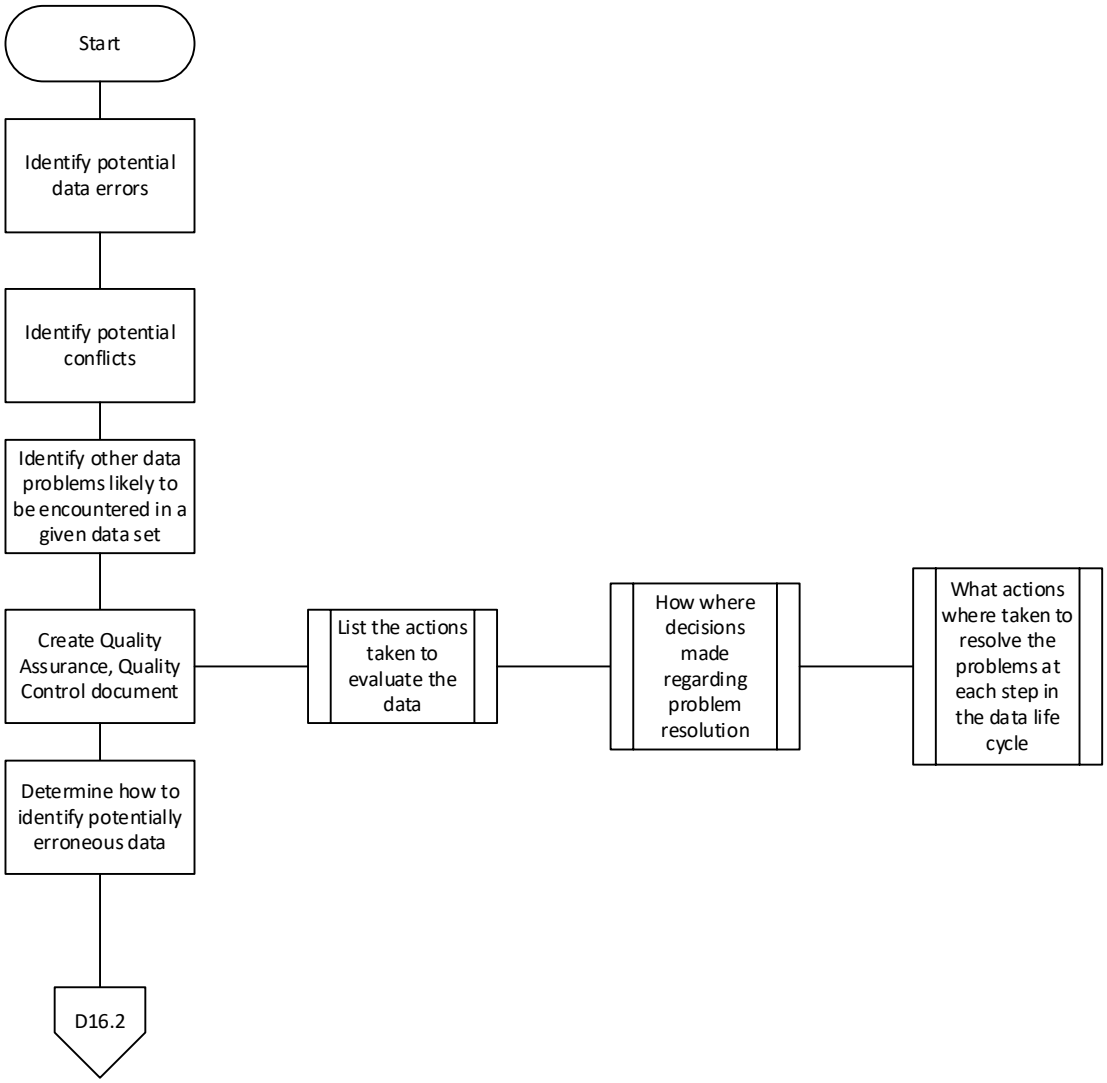
D14 - Dataset temporal extent and resolution



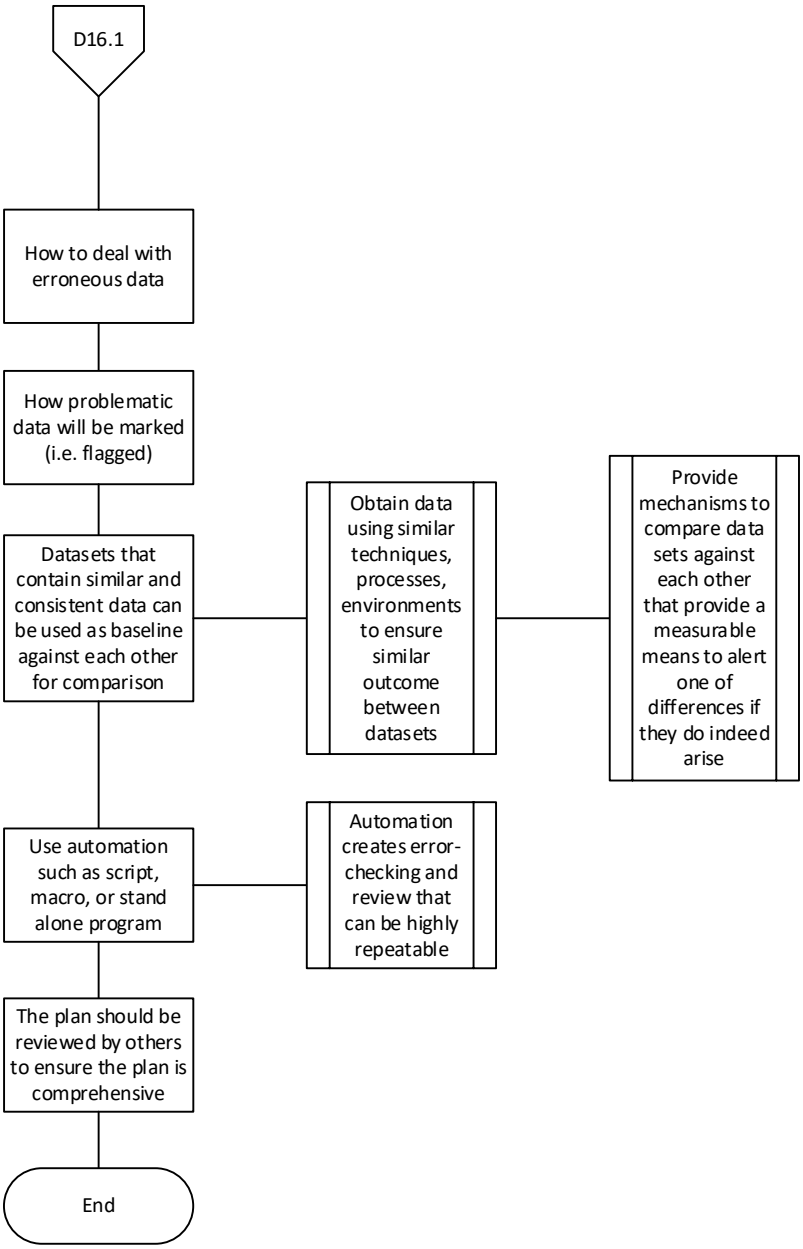
D15 - Units of measure



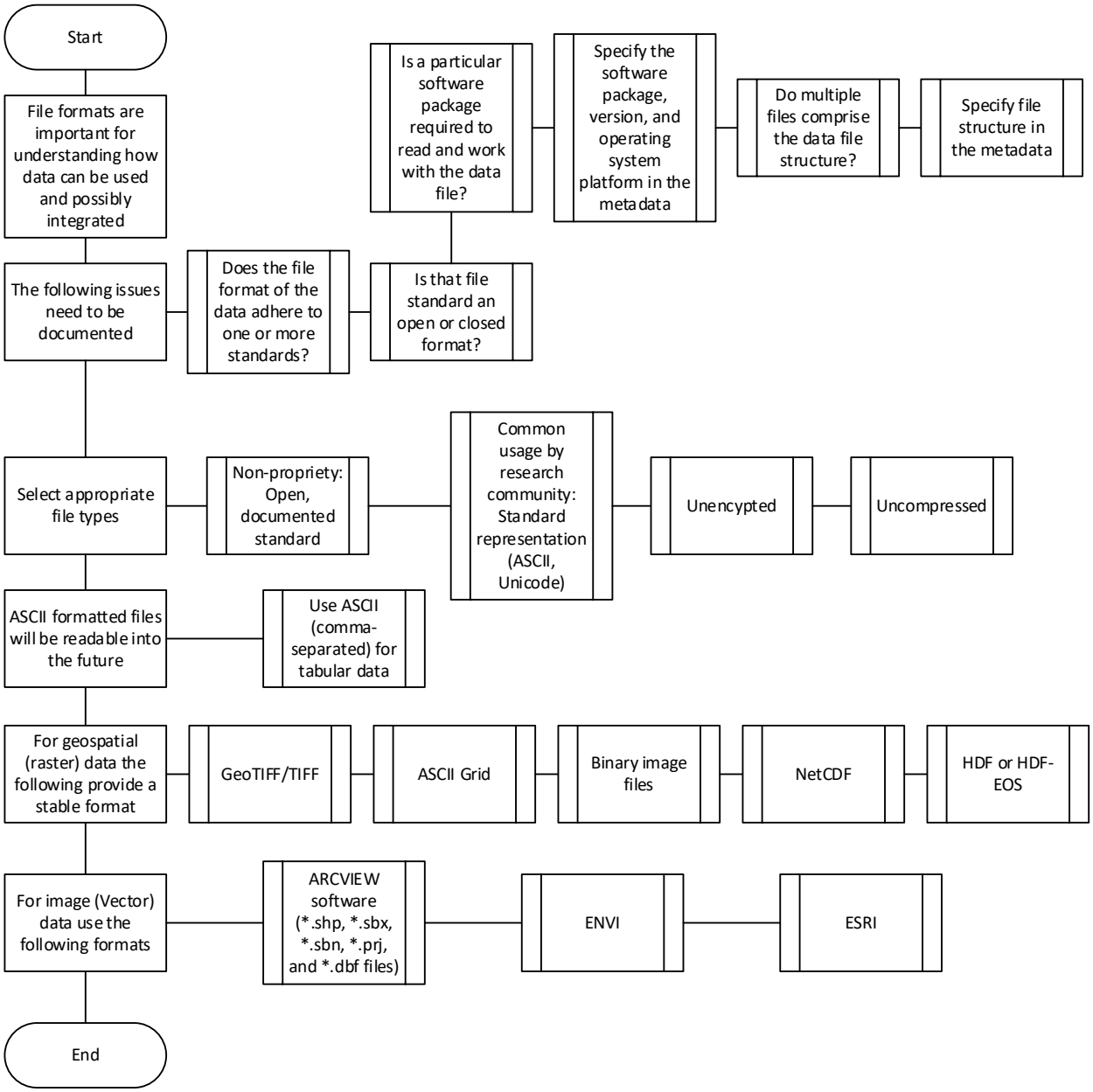
D16.1 - Control and assure quality



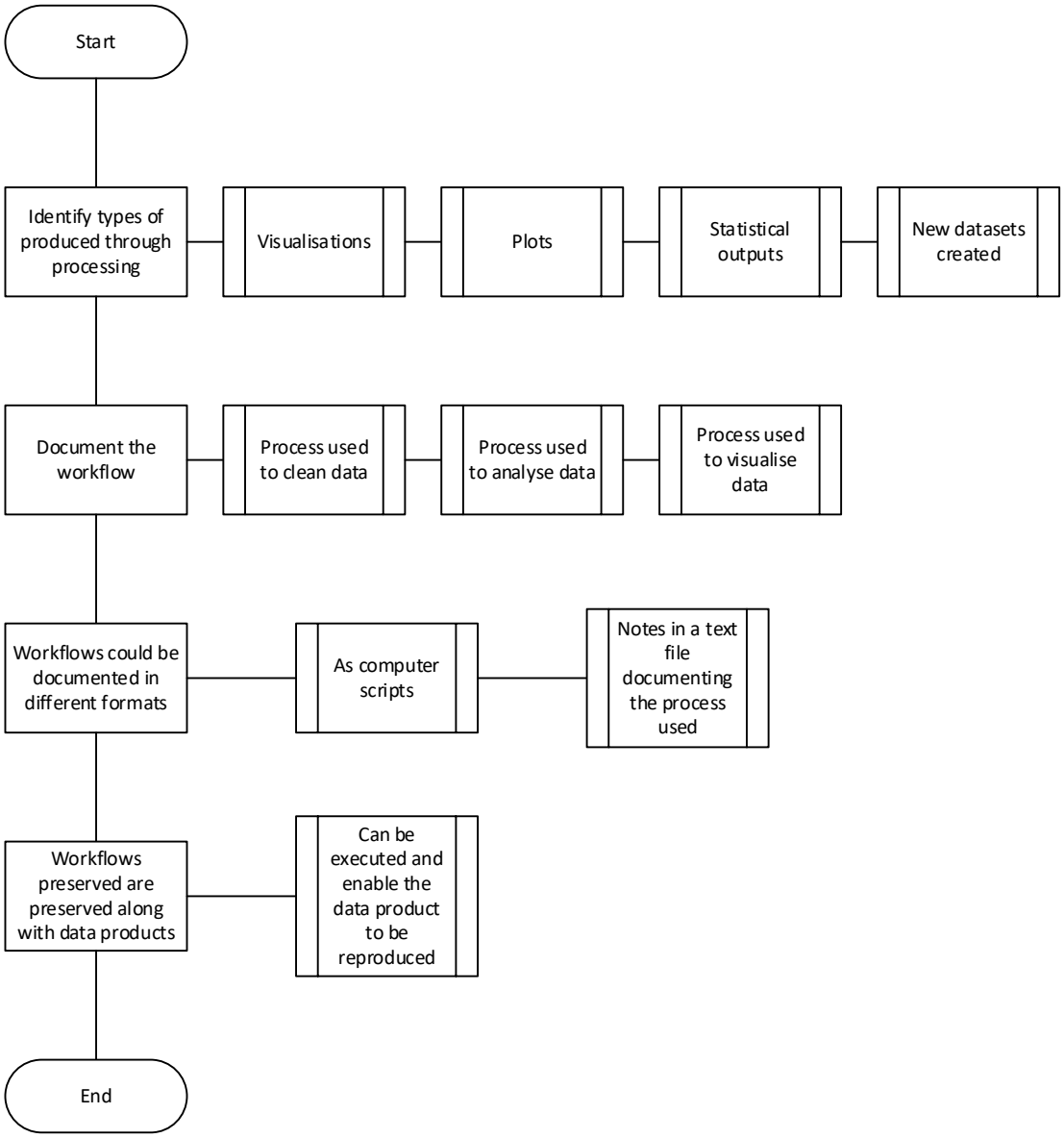
D16.2 - Control and assure quality



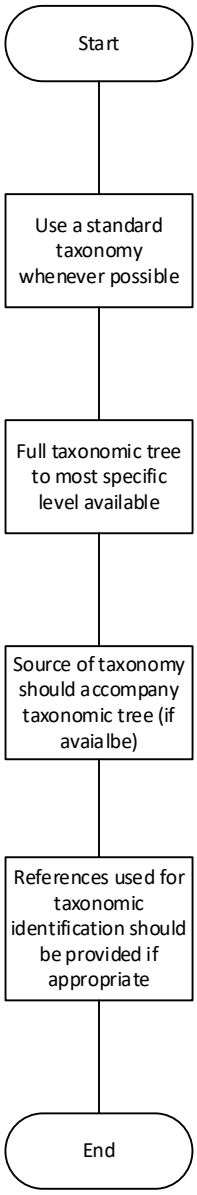
D17 - File format guidelines and documentation



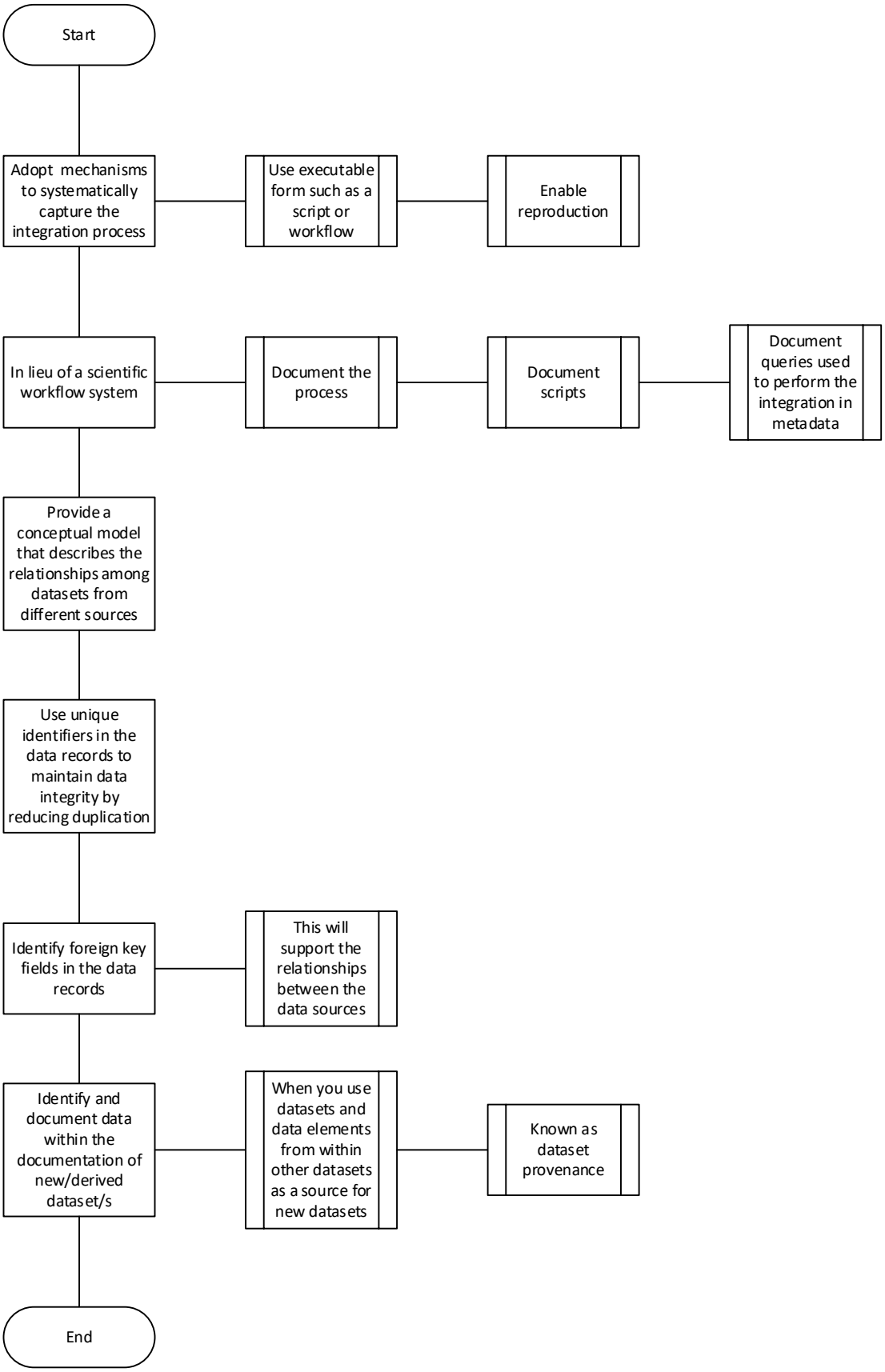
D18 - Document steps used in data processing



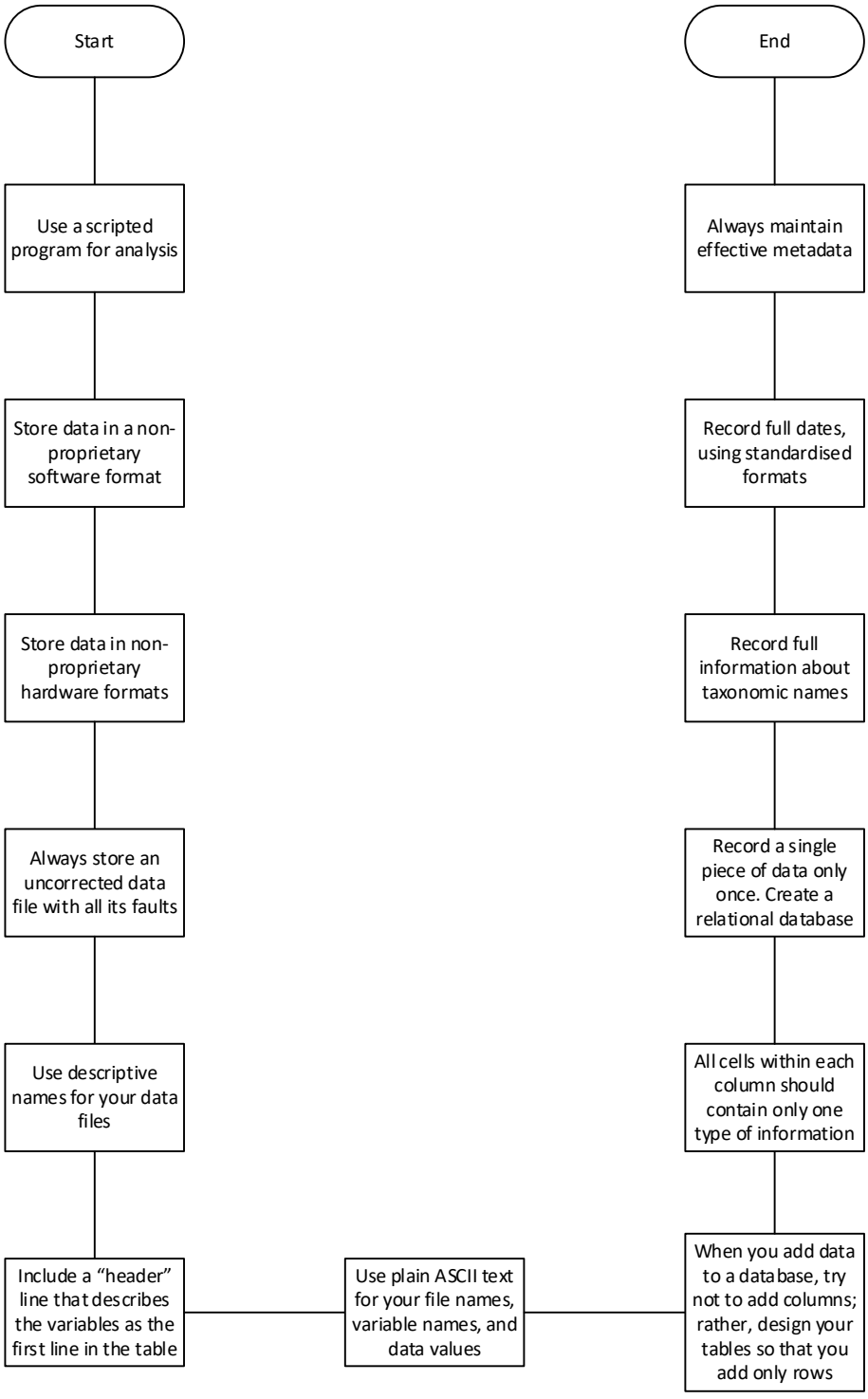
D19 - Taxonomy documentation guidelines



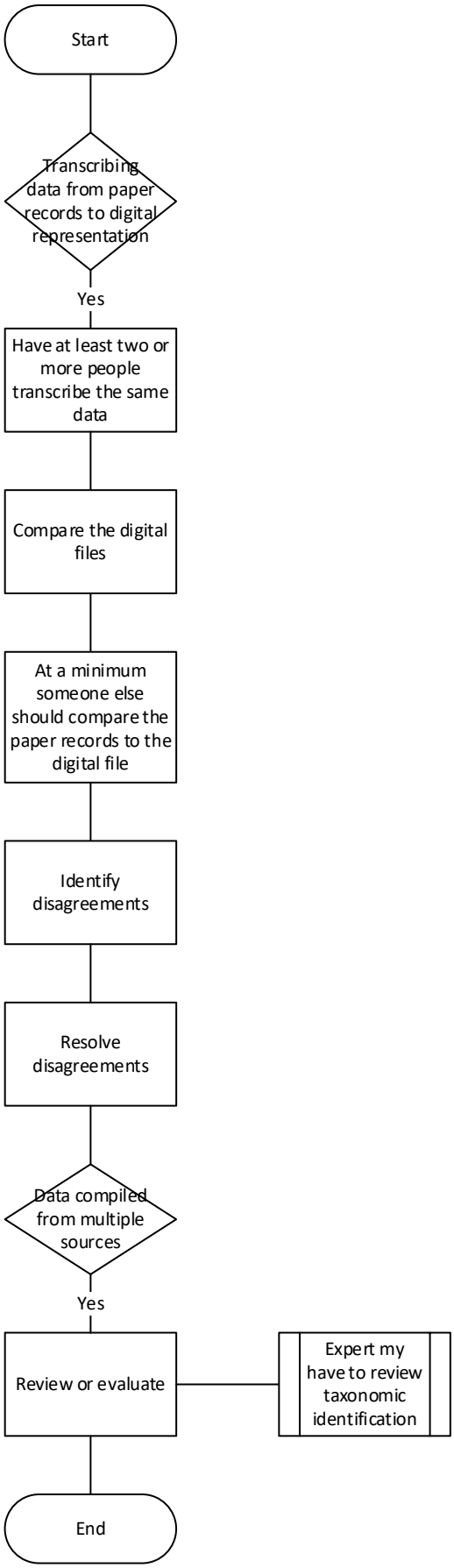
D20 - Multi-set data integration

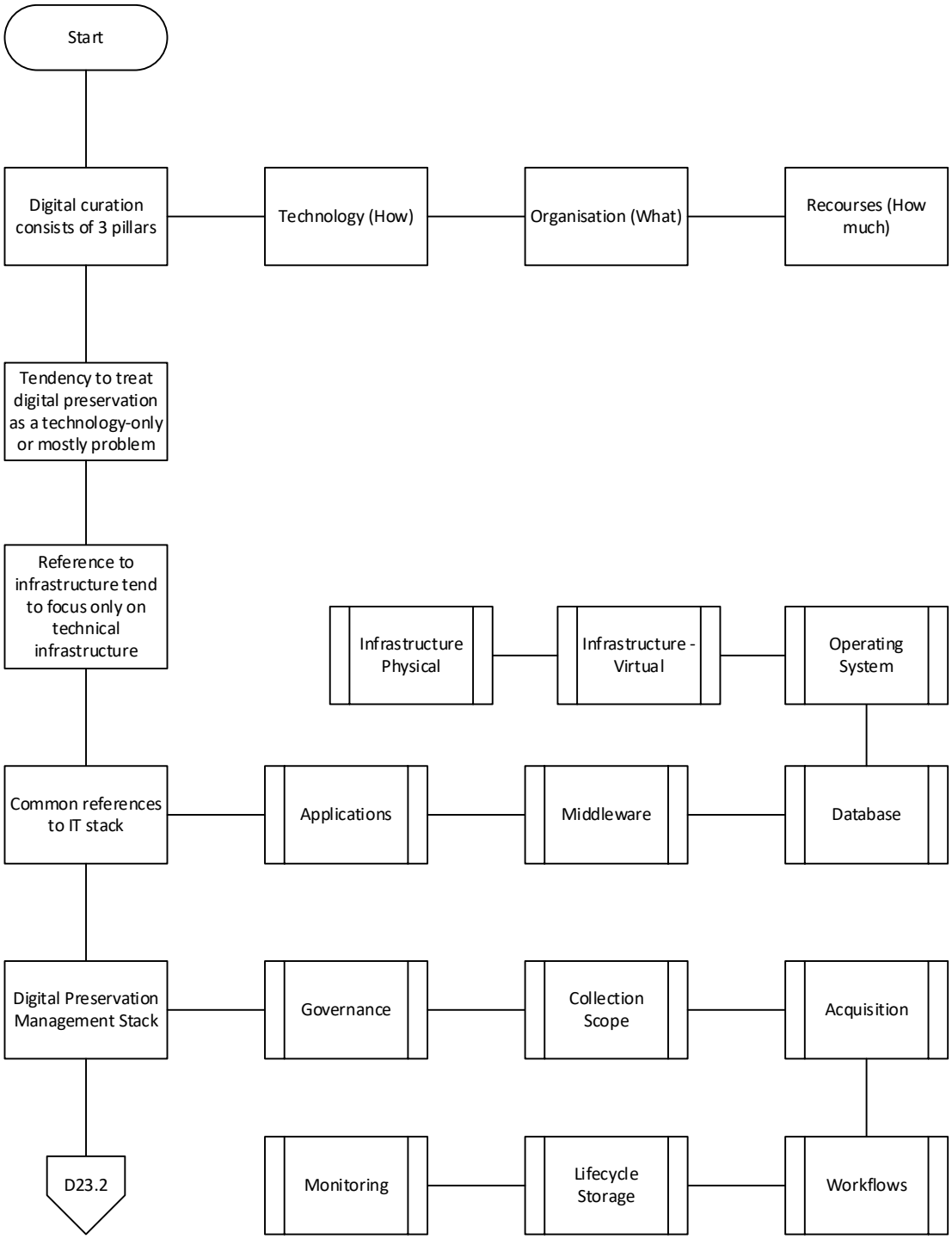


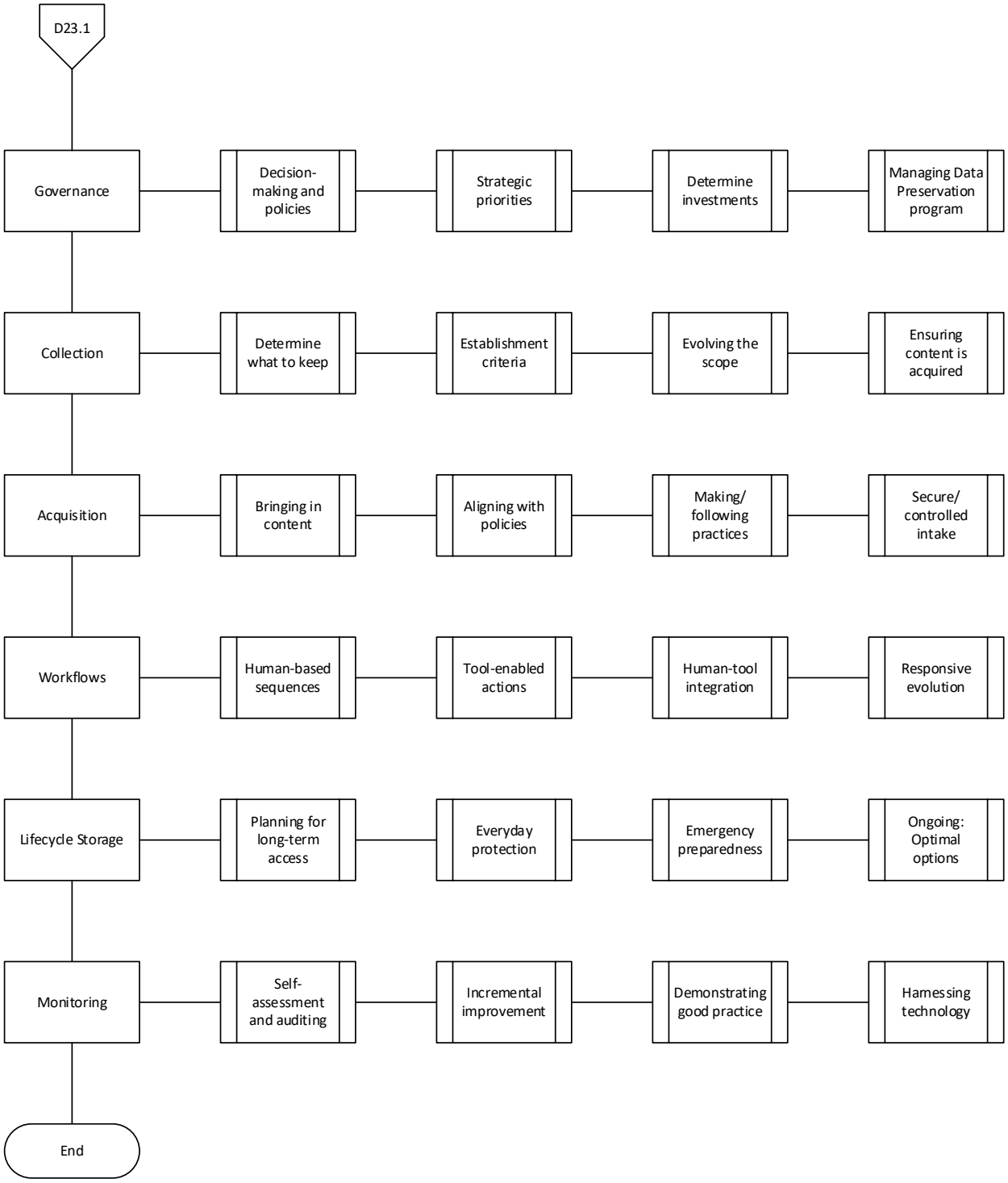
D21 - Data strategy documentation



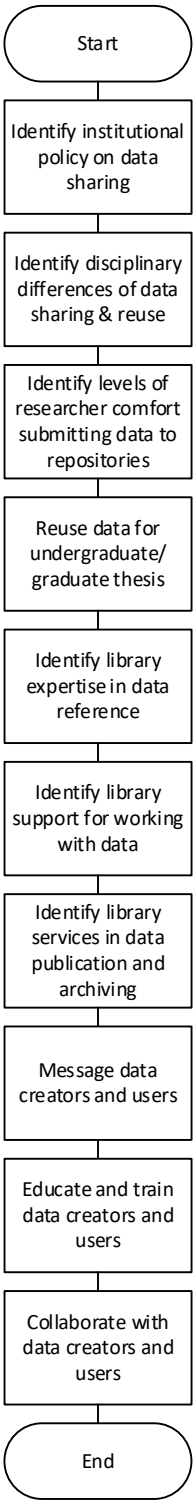
D22 - Control measures for data entry



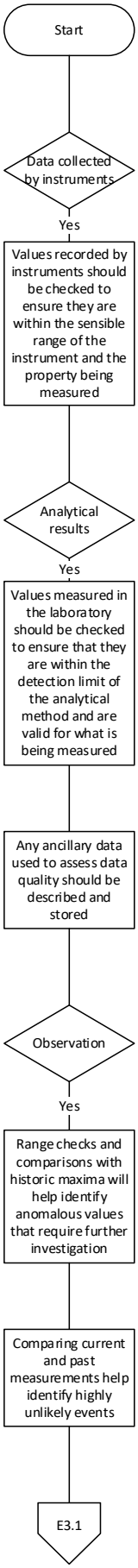

Start
Digital curation consists of 3 pillars
Technology (How)
Organisation (What)
Recourses (How much)
Tendency to treat digital preservation as a technology-only or mostly problem
Reference to infrastructure tend to focus only on technical infrastructure
Infrastructure Physical
Infrastructure - Virtual
Operating System
Common references to IT stack
Applications
Middleware
Database
Digital Preservation Management Stack
Governance
Collection Scope
Acquisition
D23.2
Monitoring
Lifecycle Storage
Workflows


D23.1
Governance
Decision-making and policies
Strategic priorities
Determine investments
Managing Data Preservation program
Collection
Determine what to keep
Establishment criteria
Evolving the scope
Ensuring content is acquired
Acquisition
Bringing in content
Aligning with policies
Making/ following practices
Secure/ controlled intake
Workflows
Human-based sequences
Tool-enabled actions
Human-tool integration
Responsive evolution
Lifecycle Storage
Planning for long-term access
Everyday protection
Emergency preparedness
Ongoing: Optimal options
Monitoring
Self-assessment and auditing
Incremental improvement
Demonstrating good practice
Harnessing technology
End

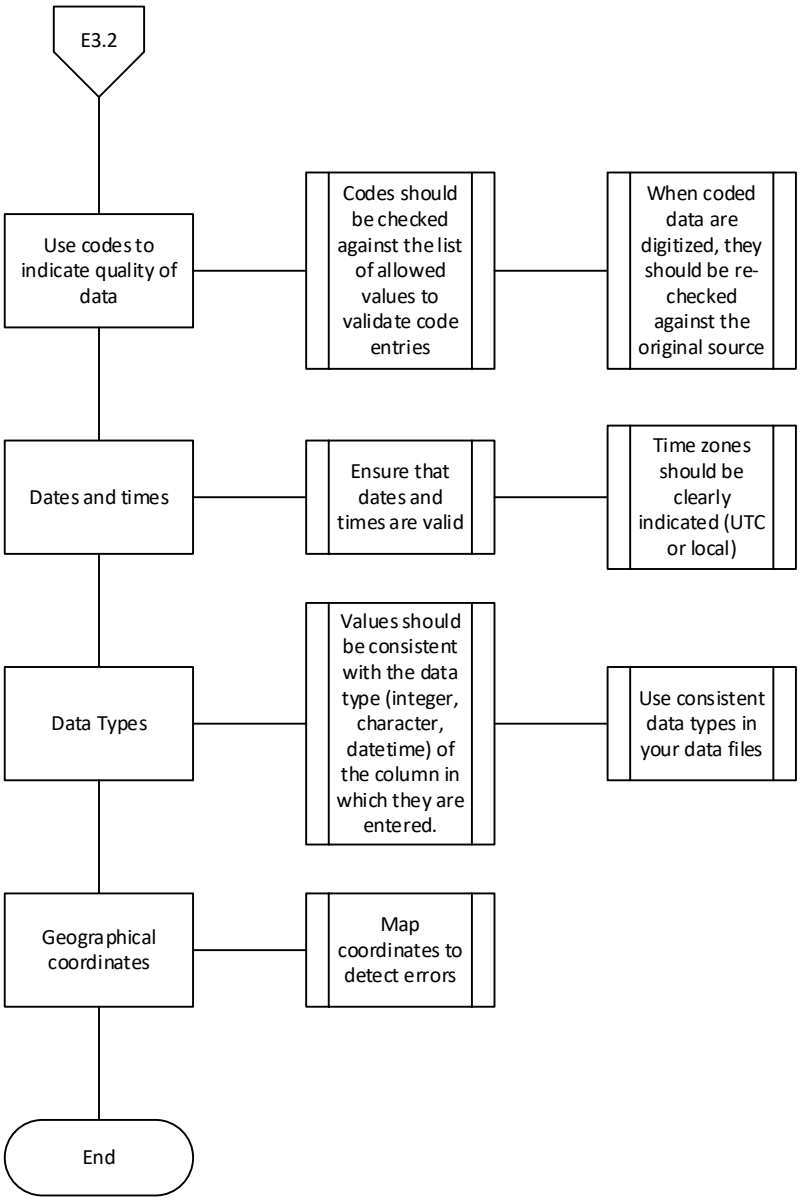
E2 - Metadata improvements



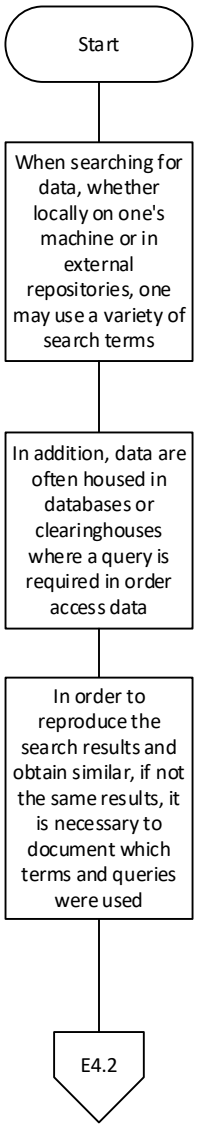
E3.1 - Quality control for research data



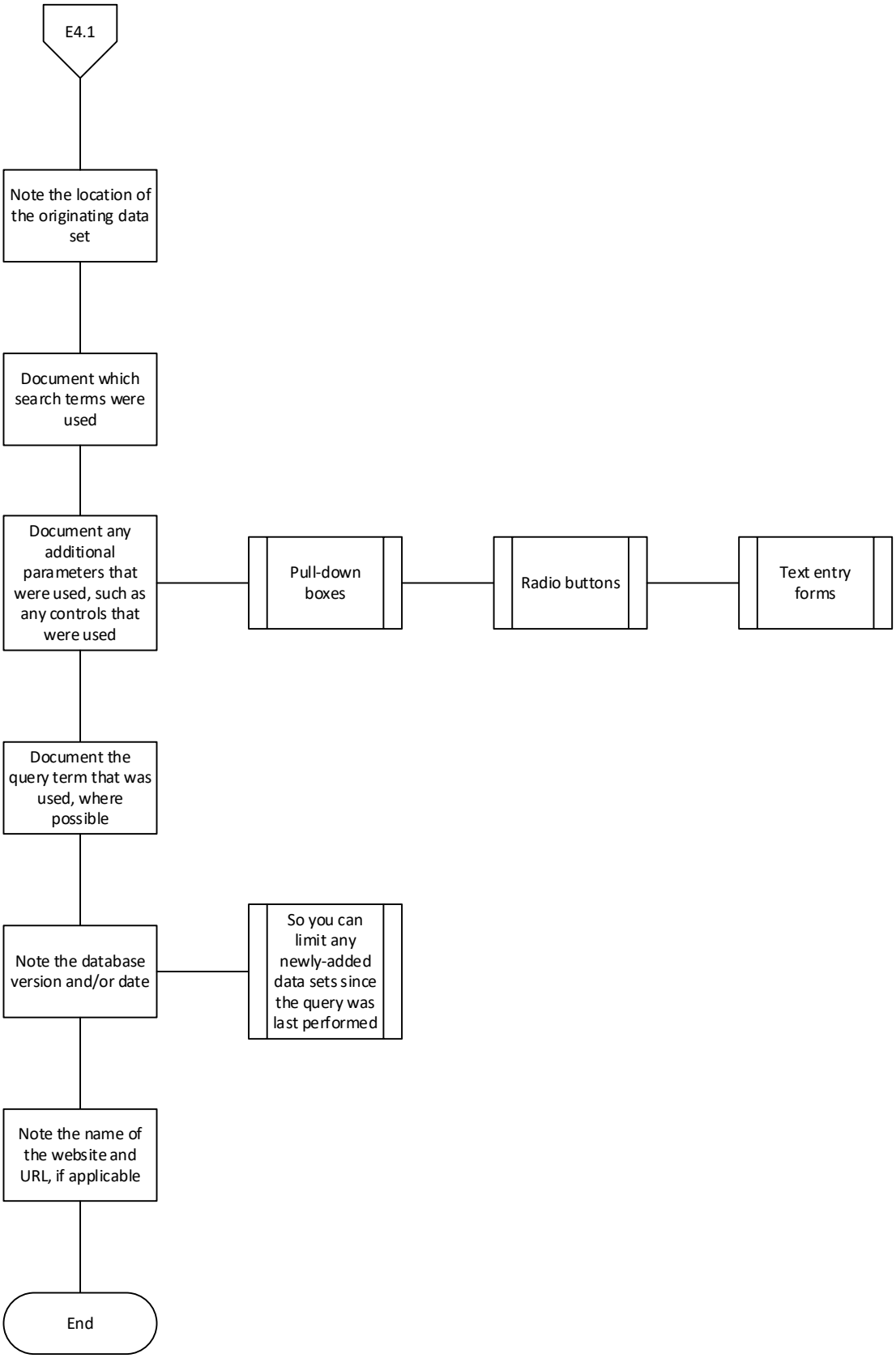
E3.2 - Quality control for research data



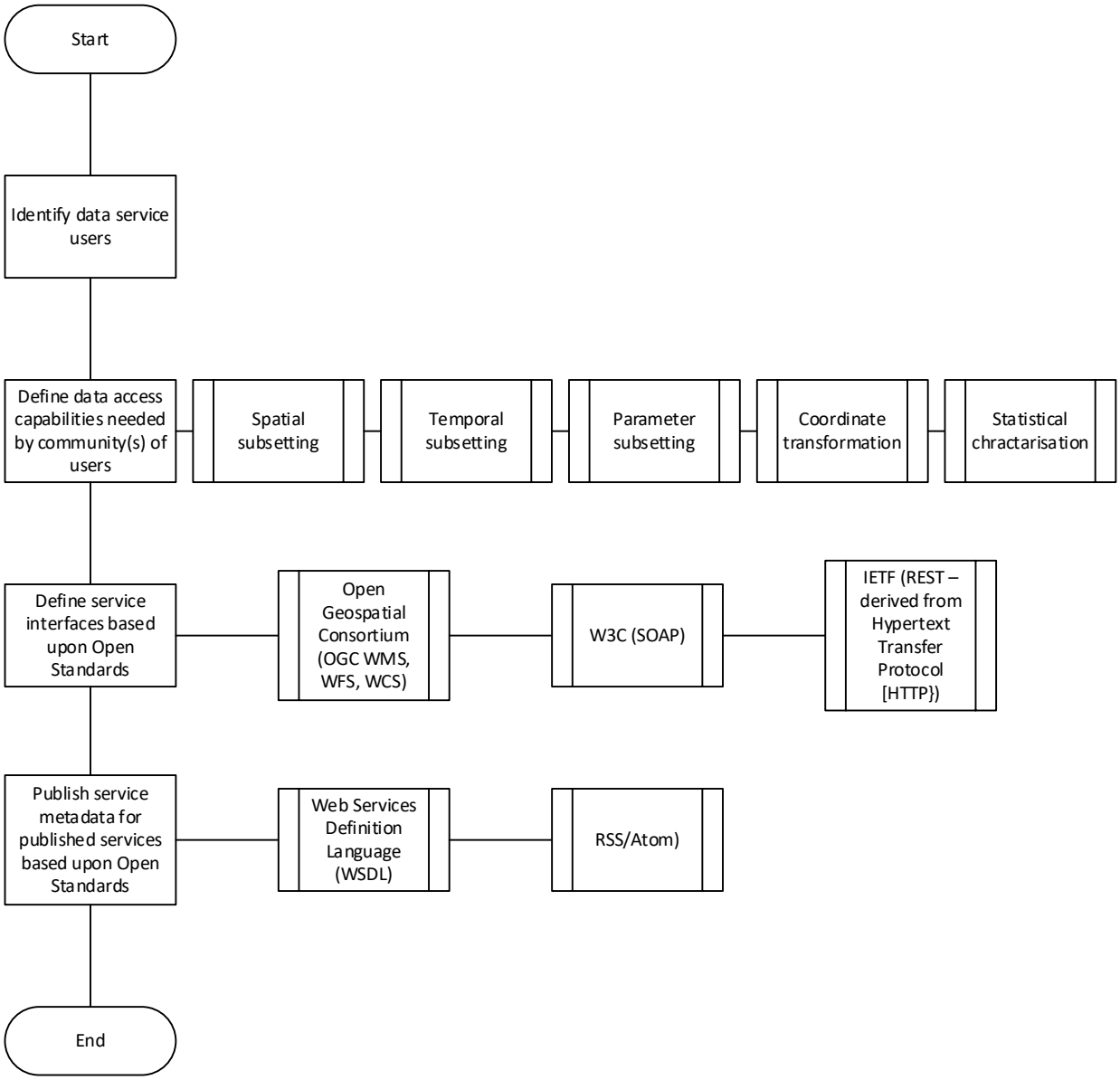
E4.1 - Guidelines to make datasets reproducible



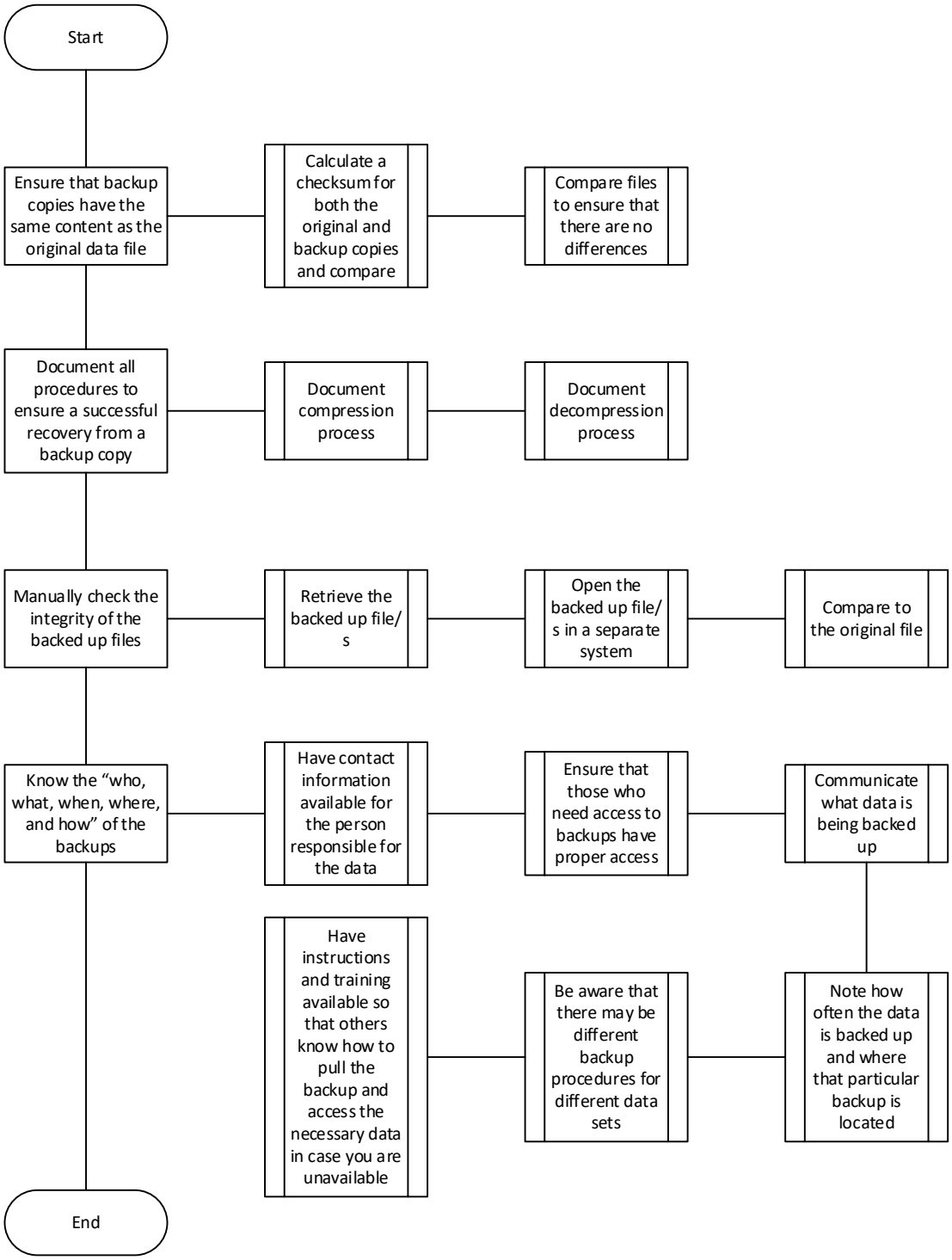
E4.2 - Guidelines to make datasets reproducible



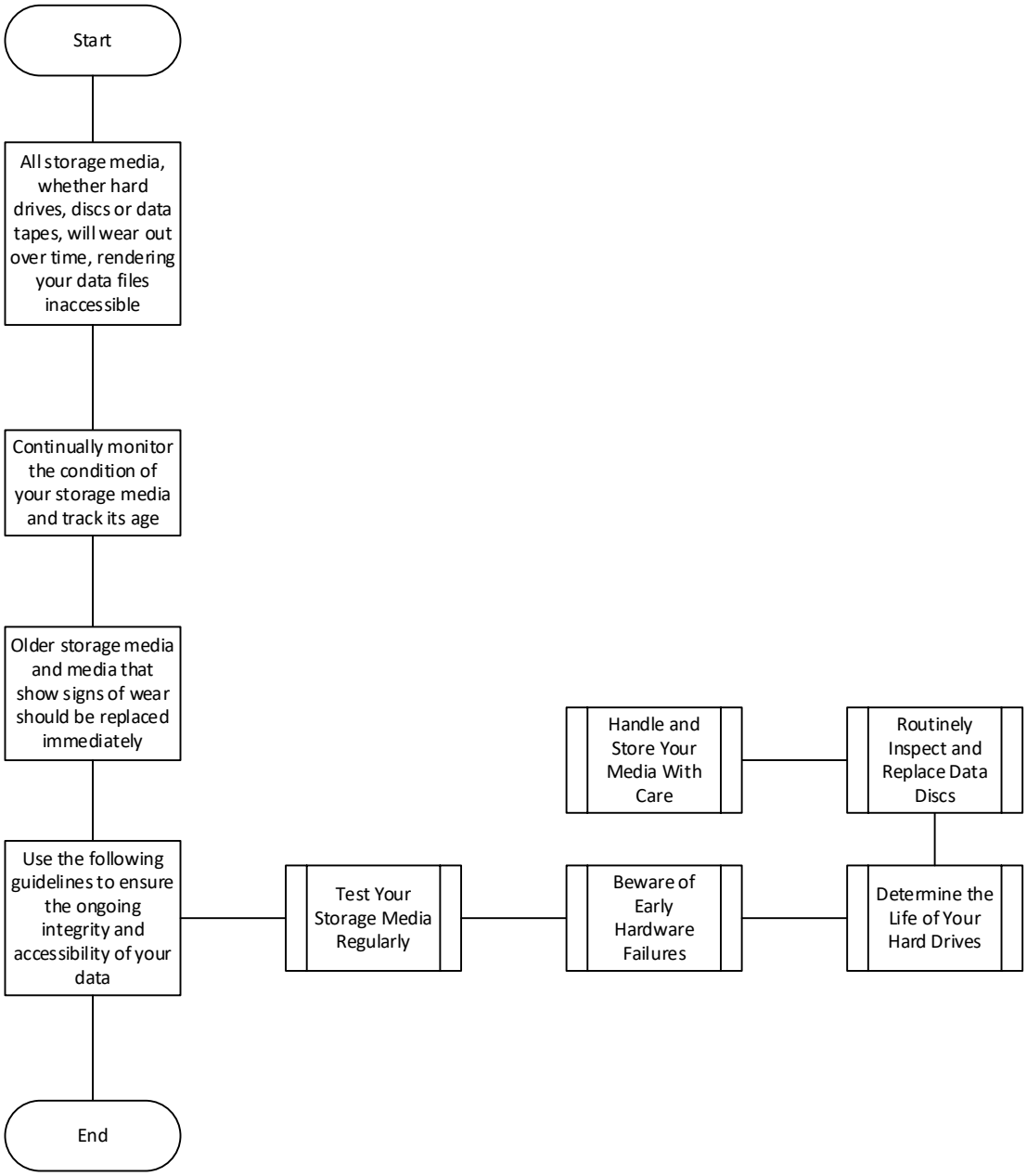
E5 - Web services to make datasets accessible



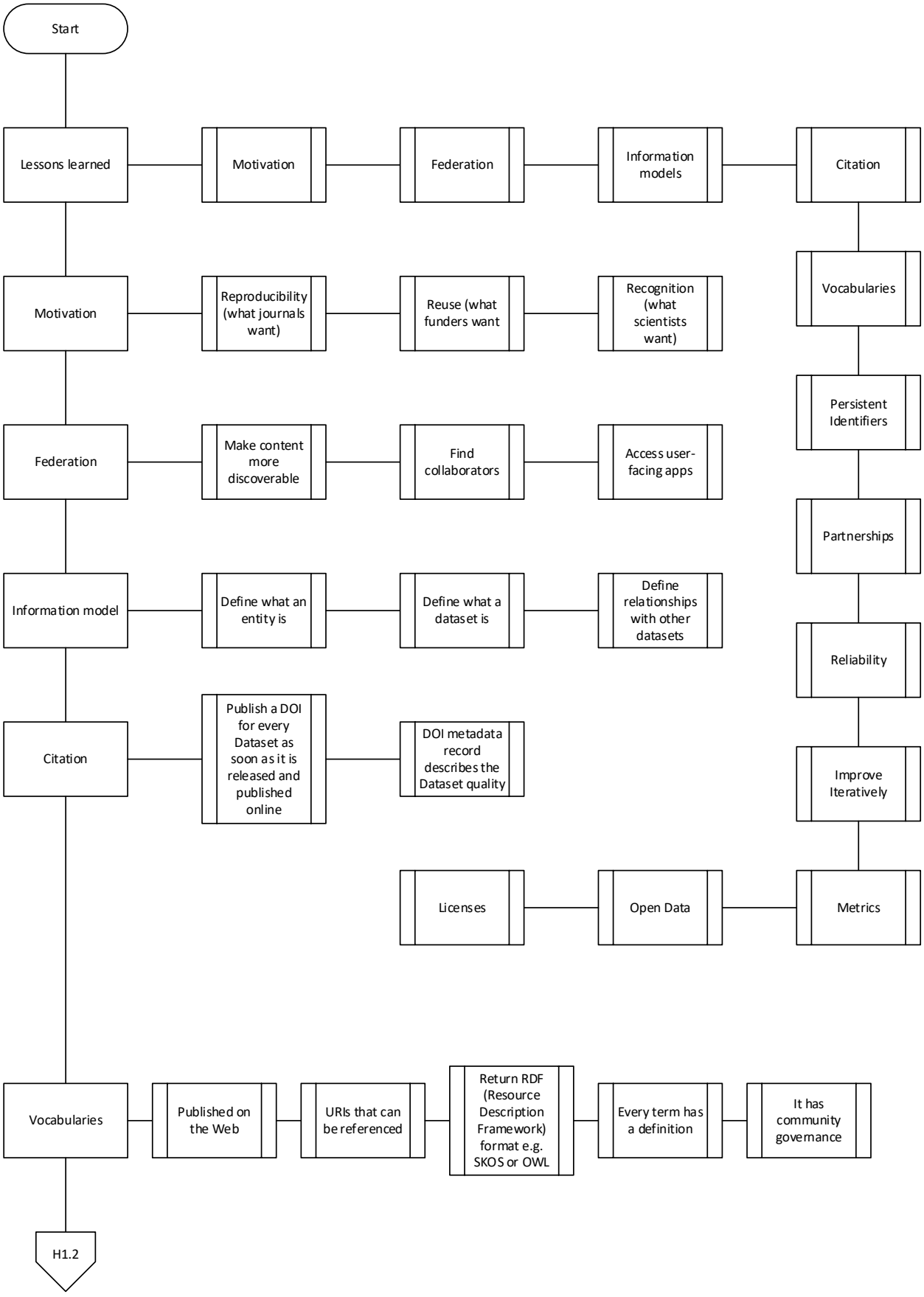
E6 - Data backup guidelines



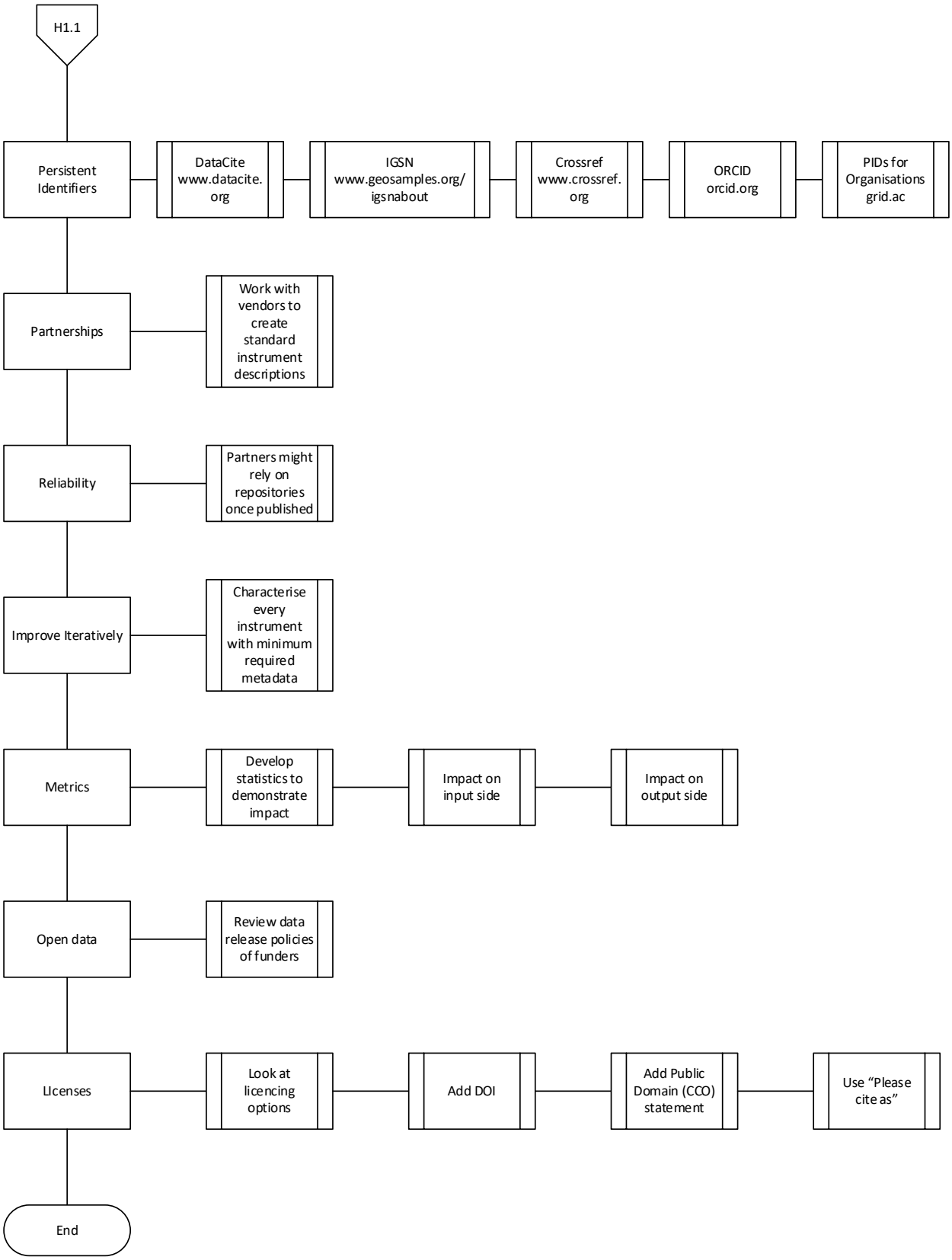
E7 - Storage media reliability



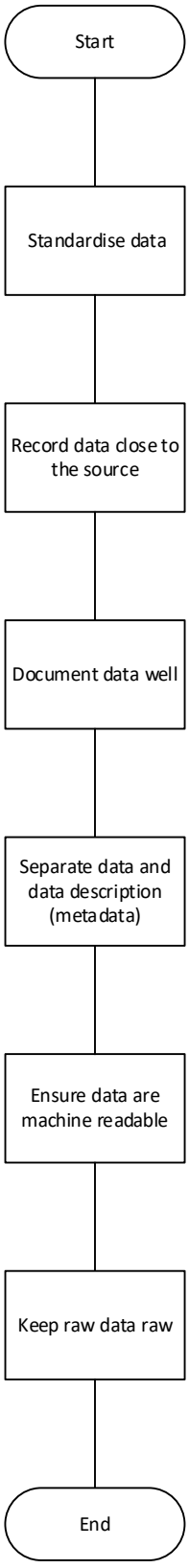
H1.1 - Understand reasons for sharing data



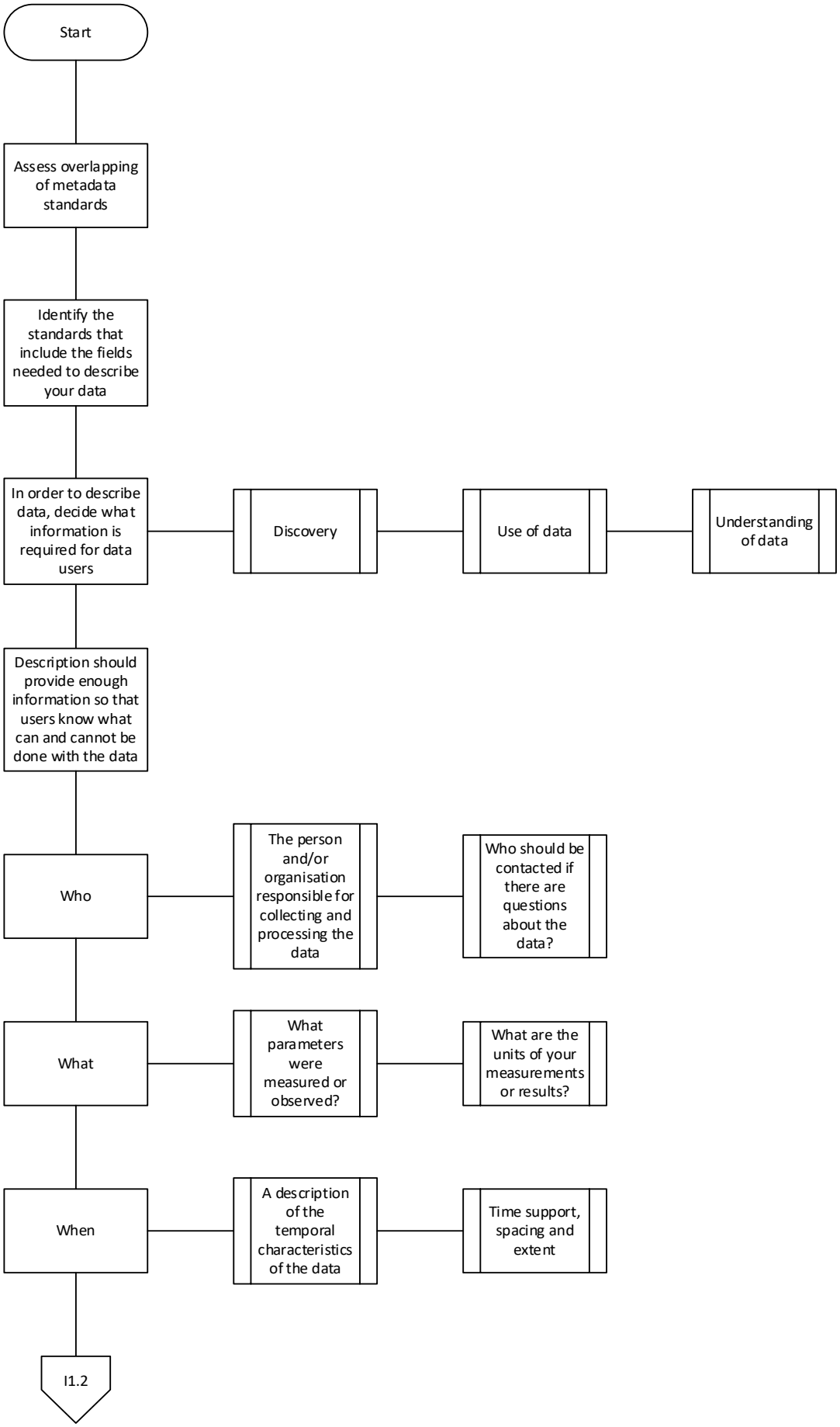
H1.1 - Understand reasons for sharing data



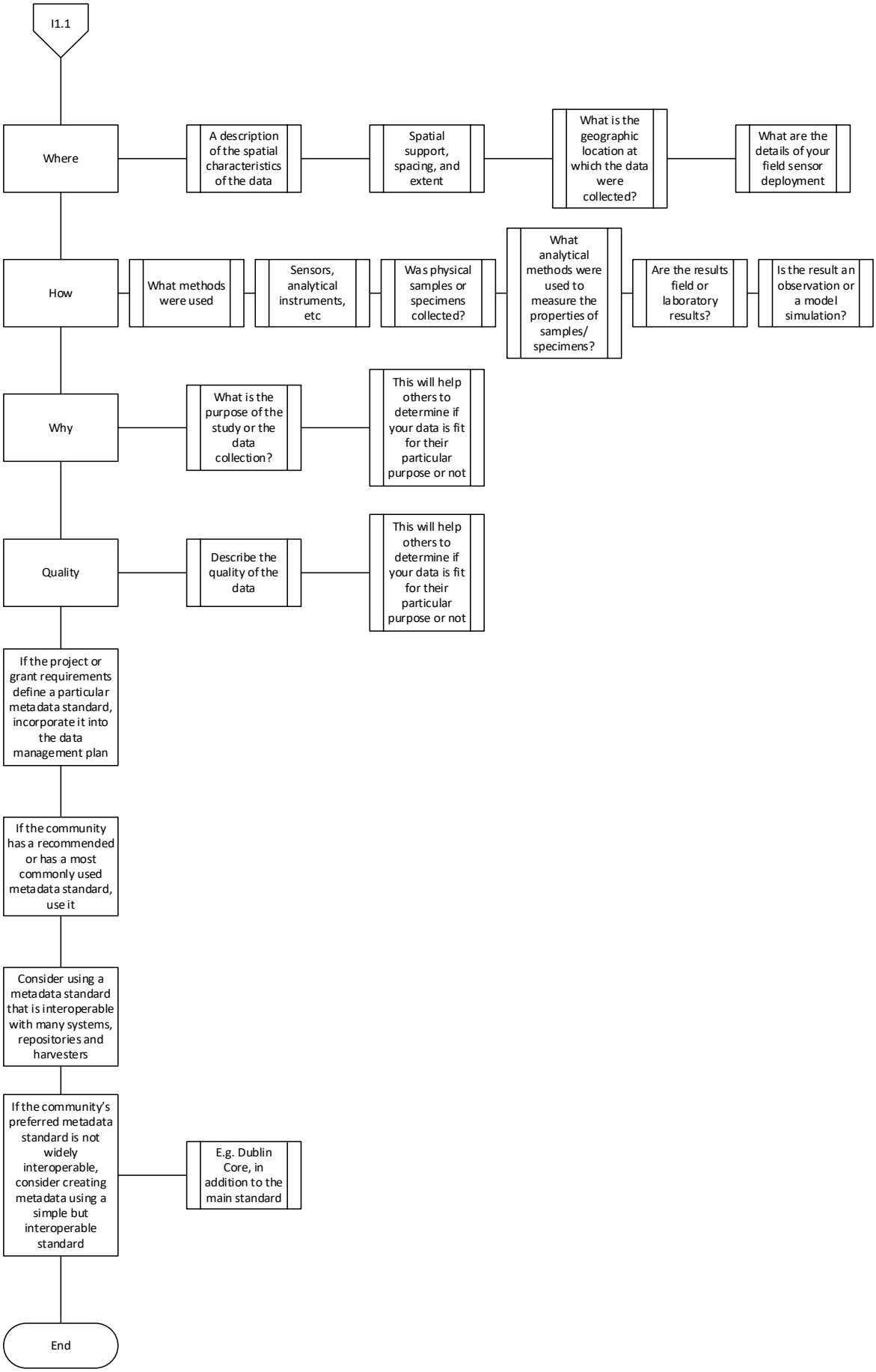
H2 – Data organisation best practices



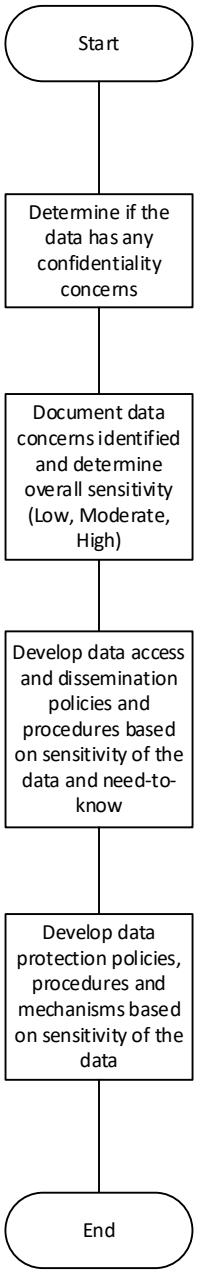
I1.1 - Metadata standards



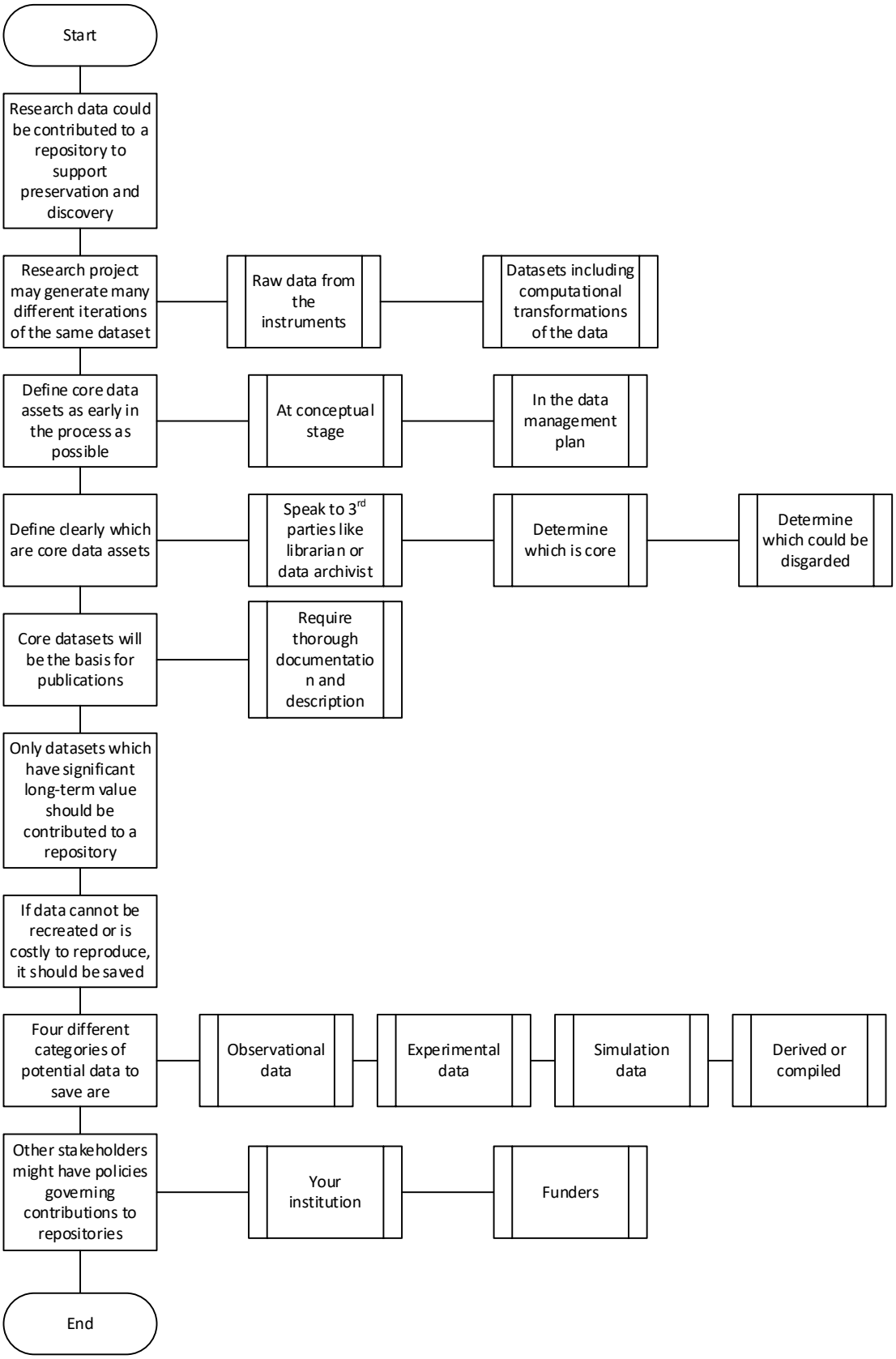
I1.2 - Metadata standards



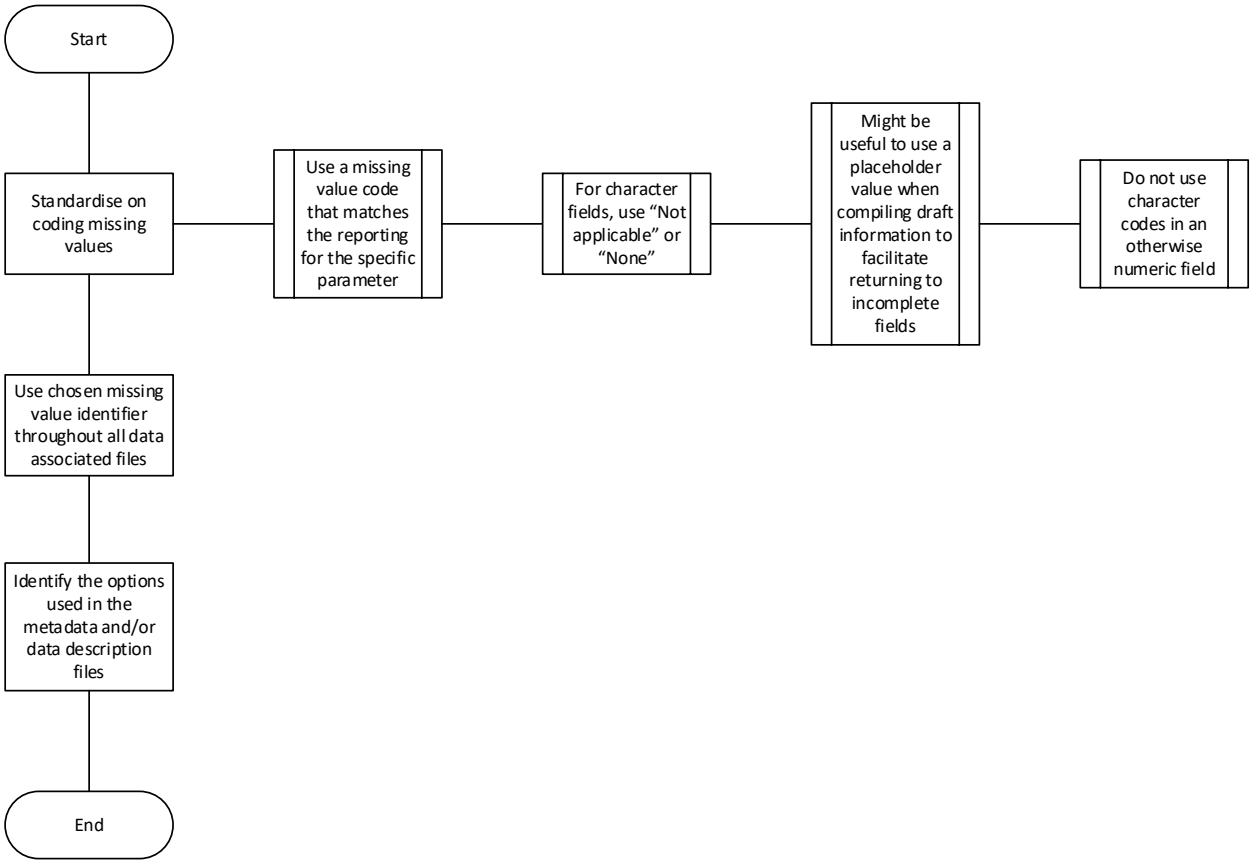
I2 – Identify sensitive data



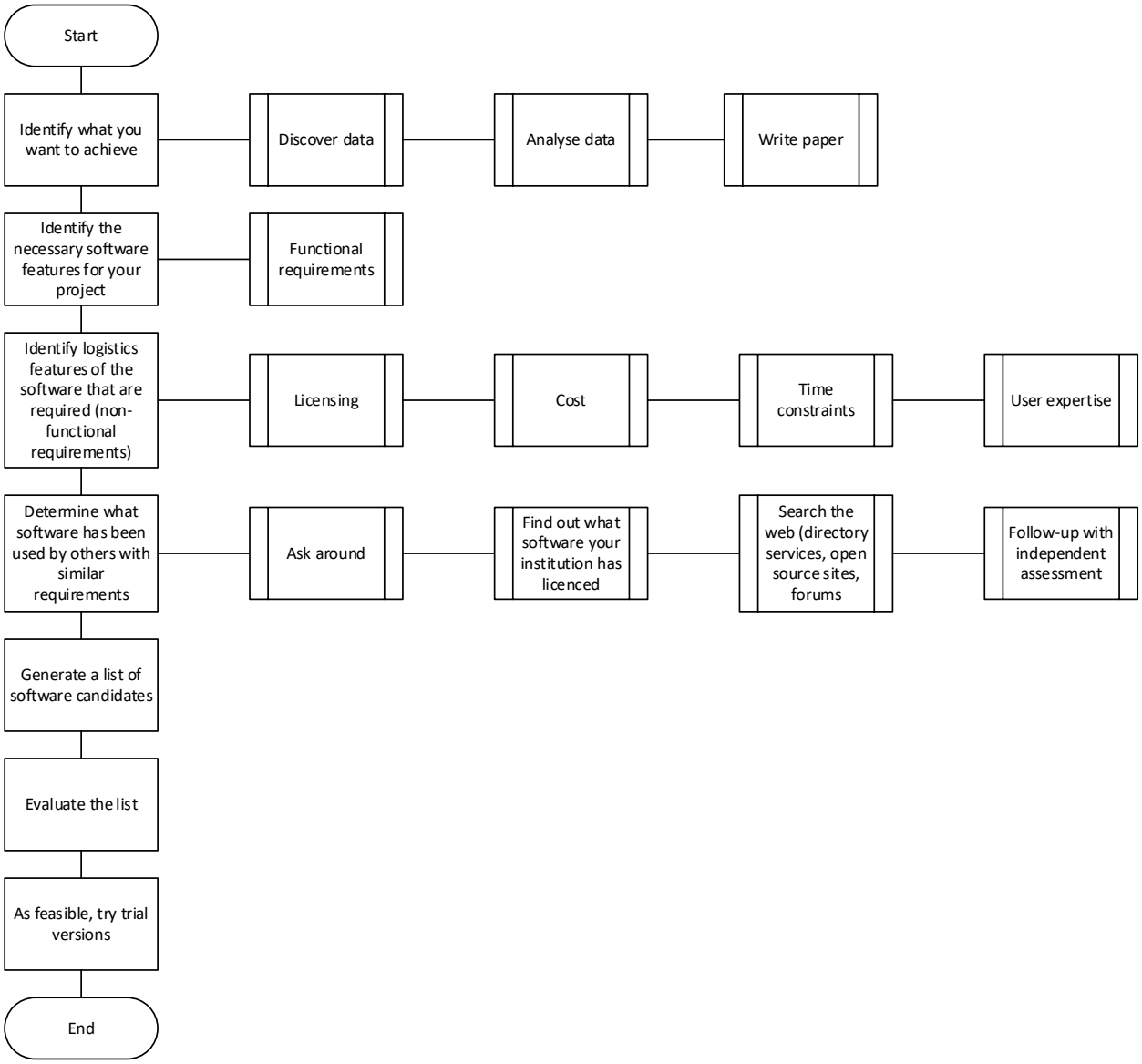
I3 - Which data should be preserved for longer?



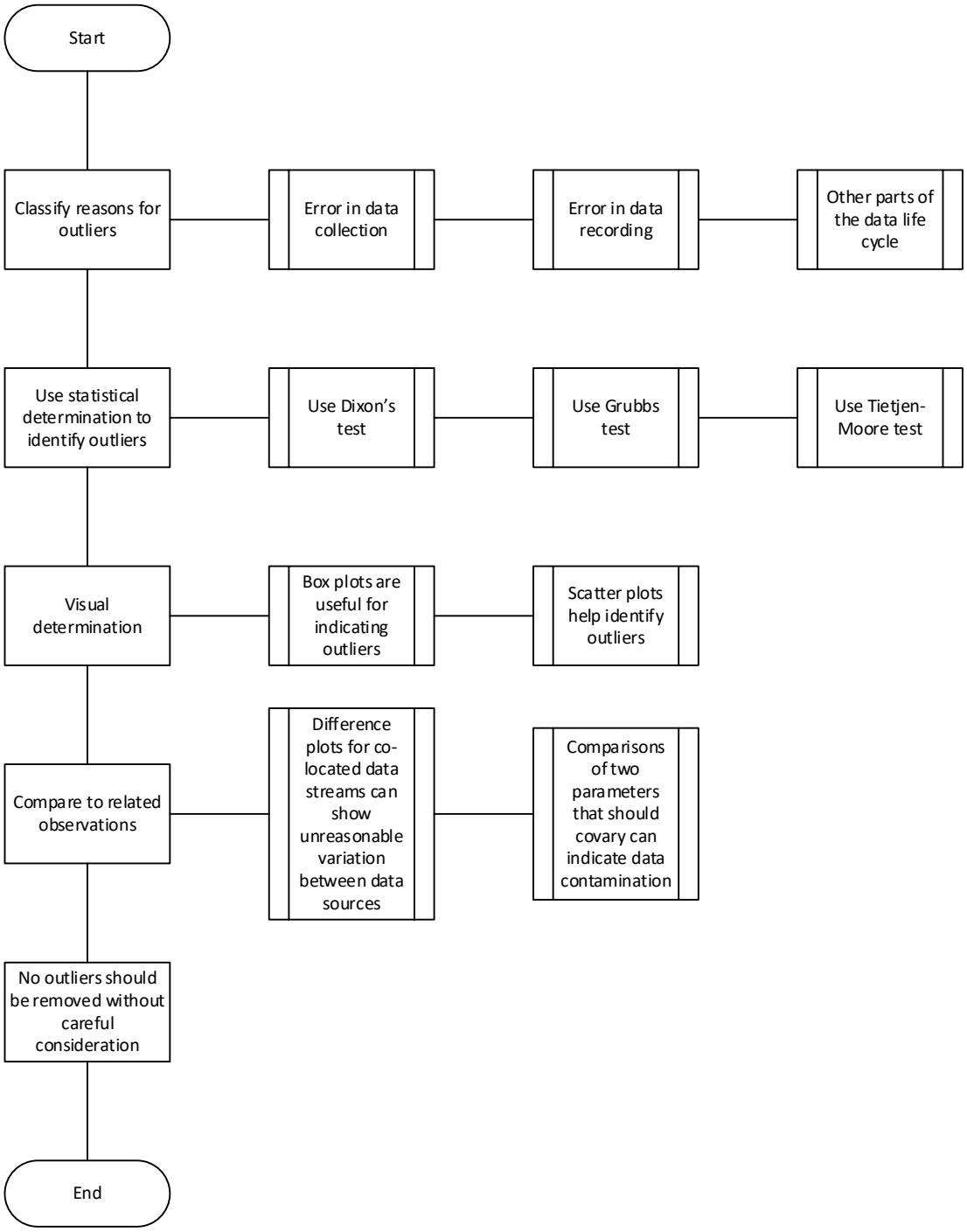
I4 - Standardise on codes for missing values



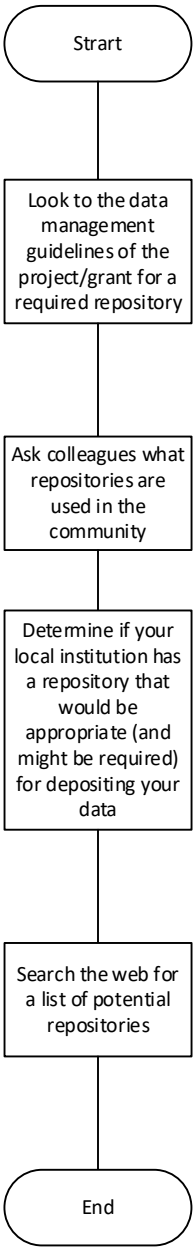
I5 - Guidelines for software identification



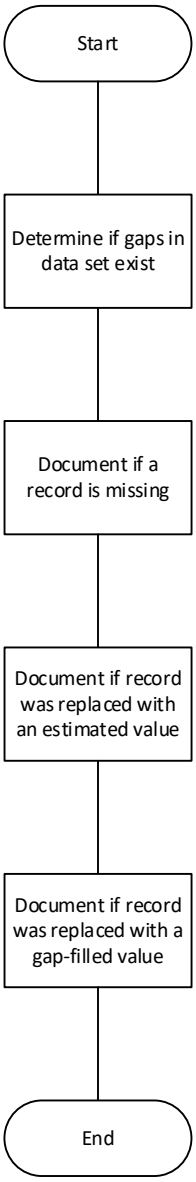
I6 – Guidelines for outliers in datasets



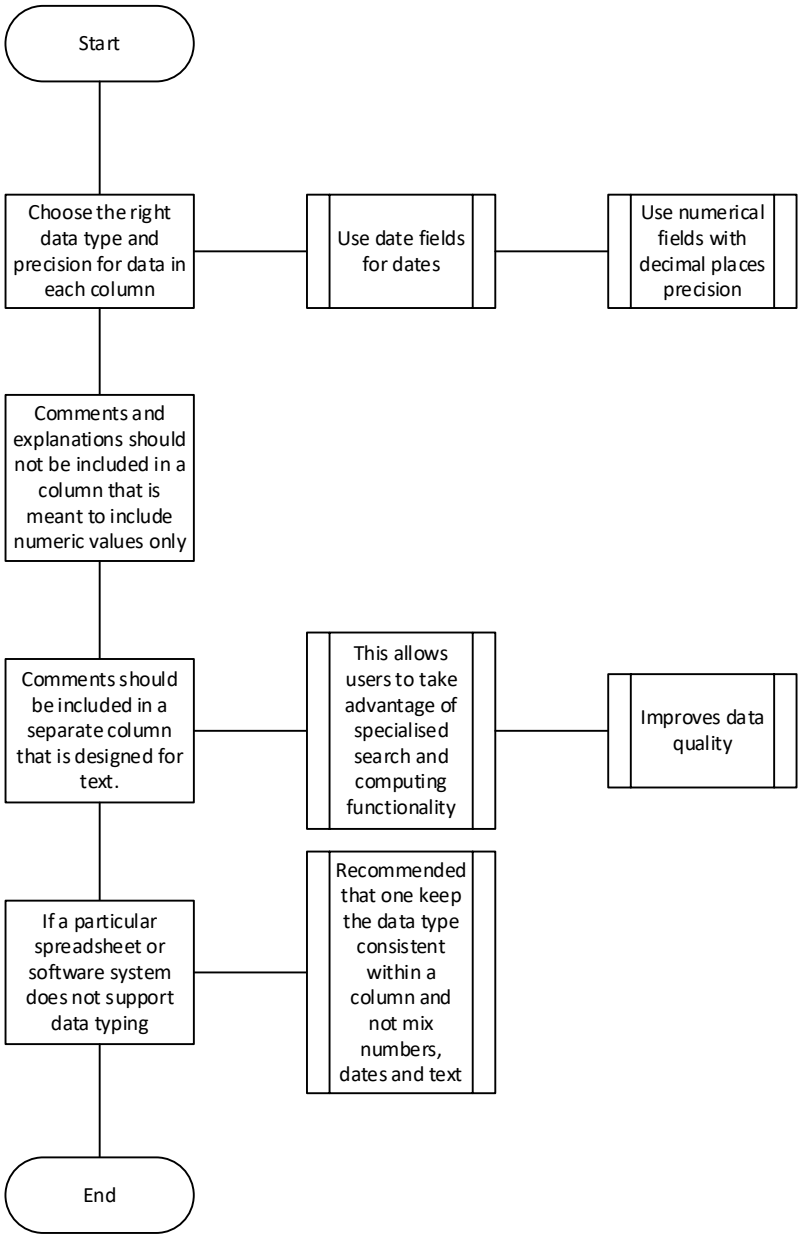
I7 - Guidelines to identify repositories



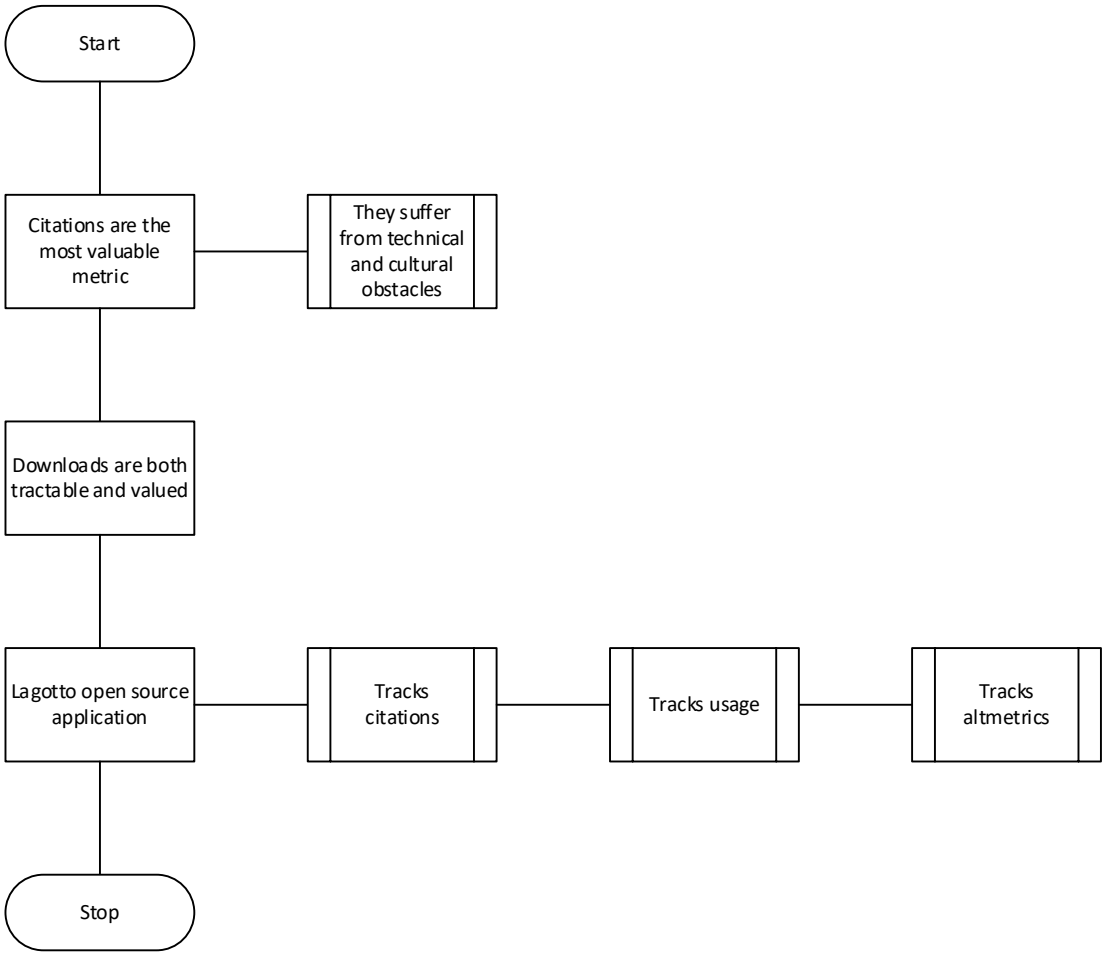
I8 - Clarify estimated values



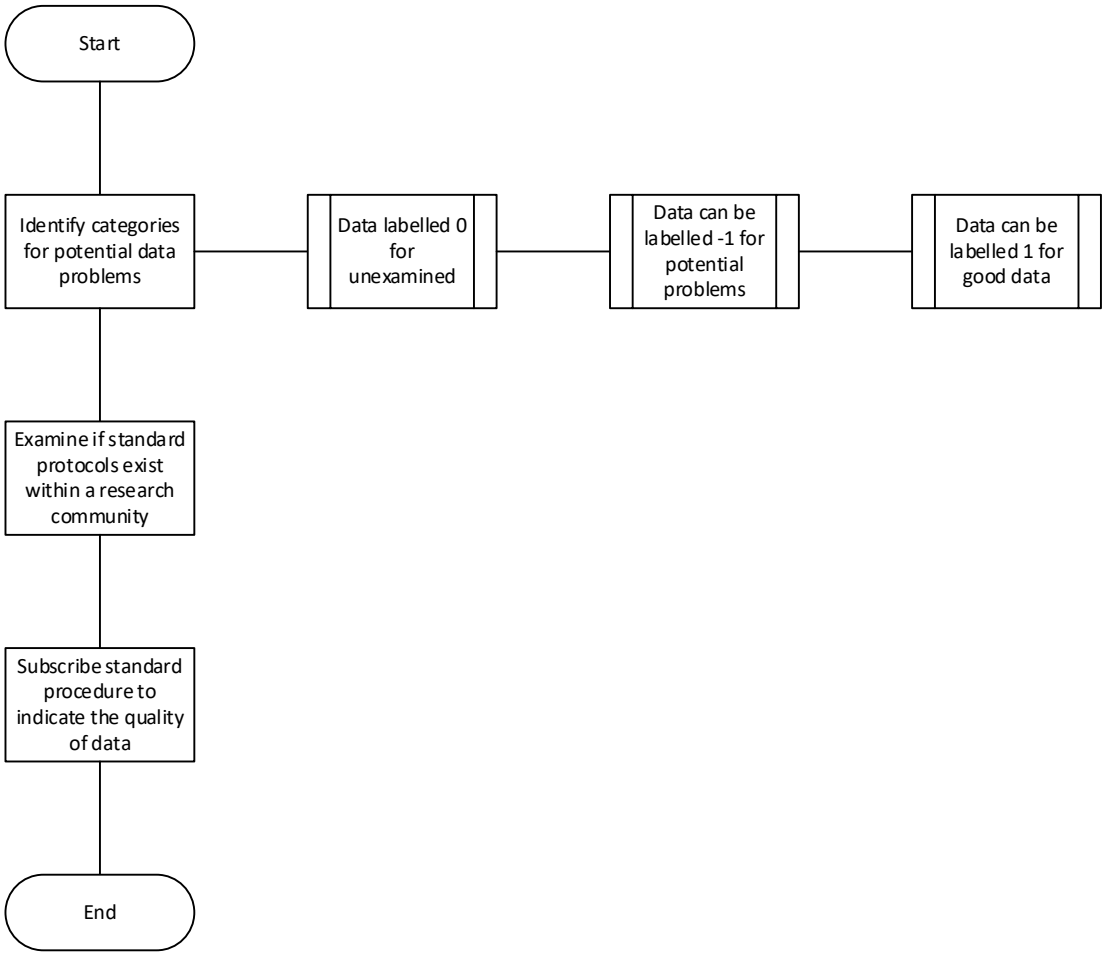
M1 - Data type consistency



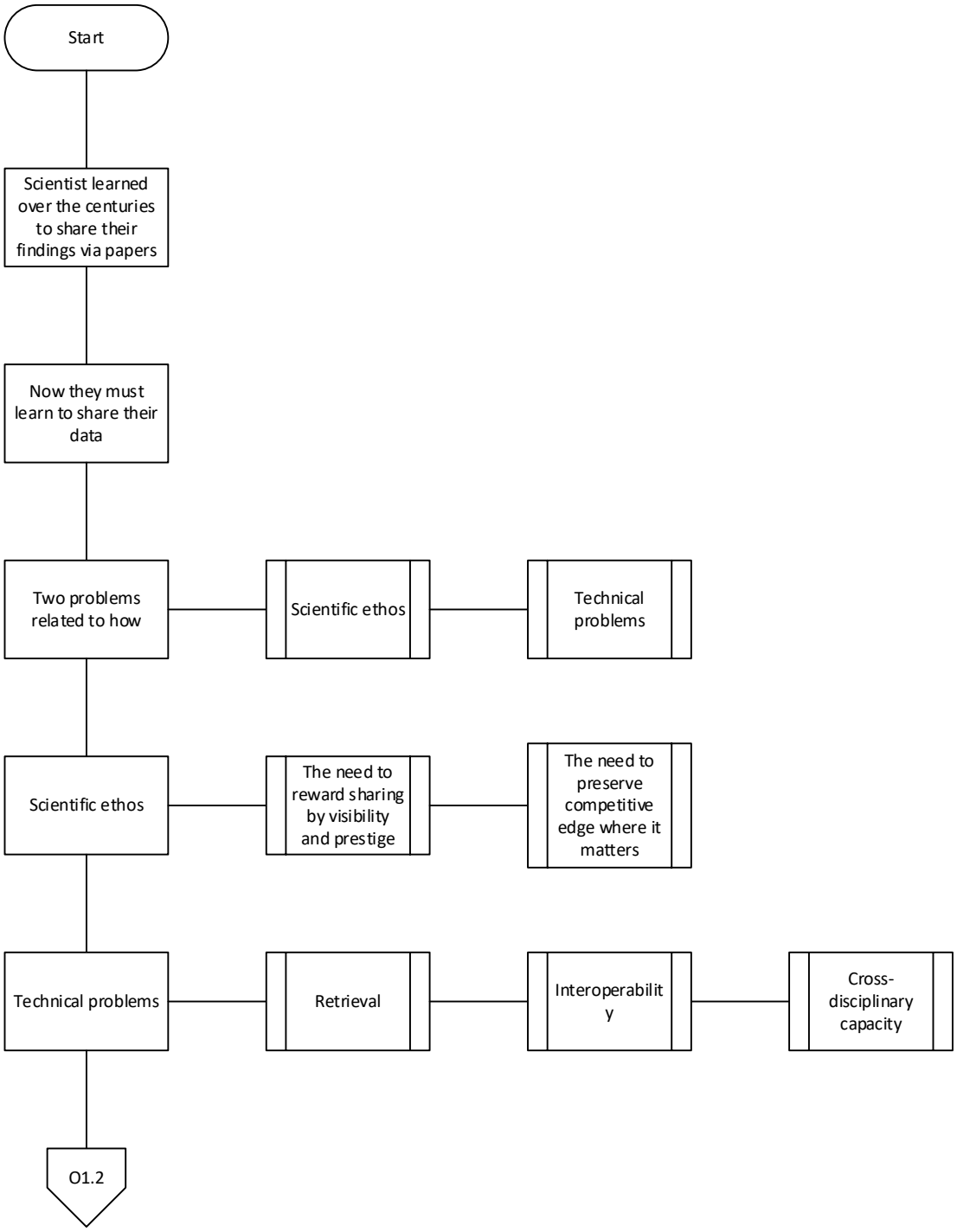
M2 - Metrics for data usage and citing



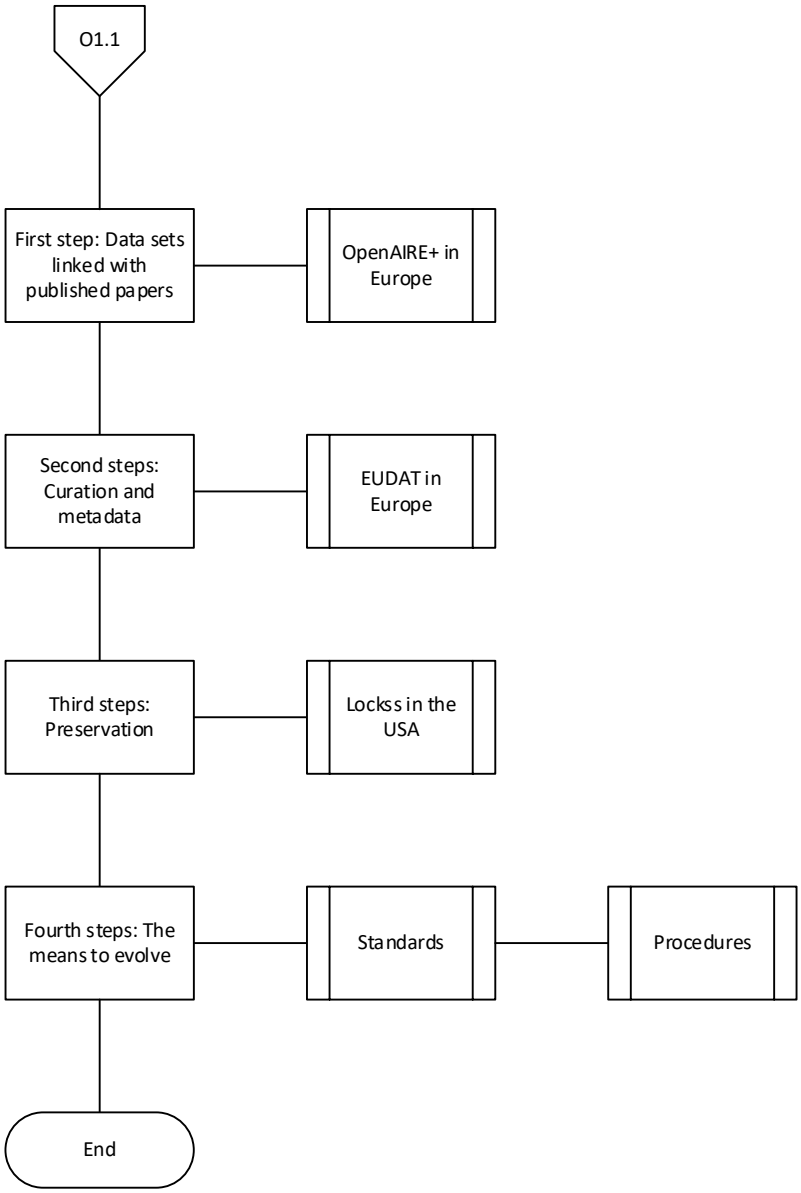
M3 - Flag poor data for quality control



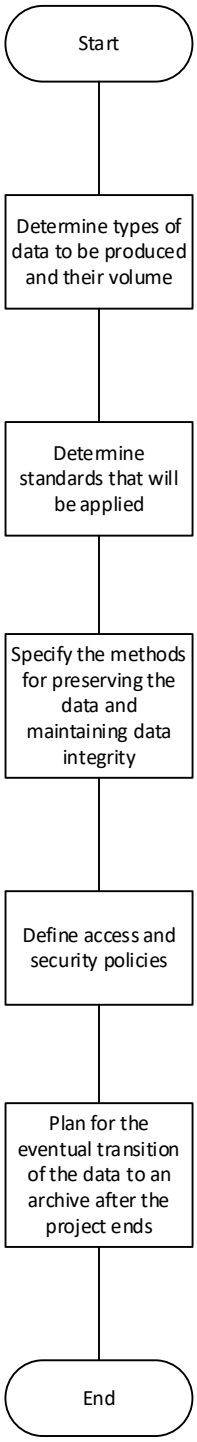
O1.1 - Research process optimisation



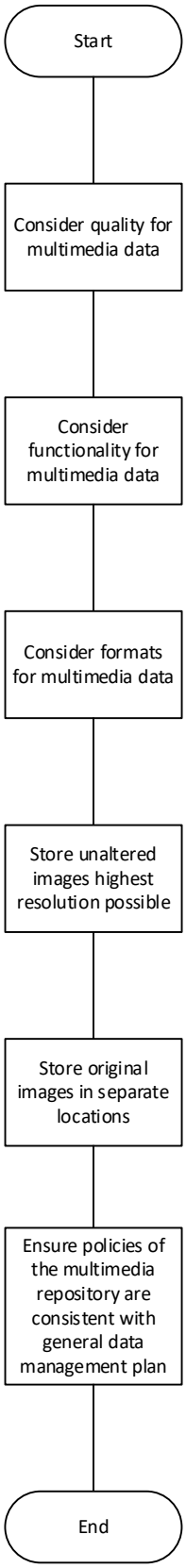
O1.2 - Research process optimisation



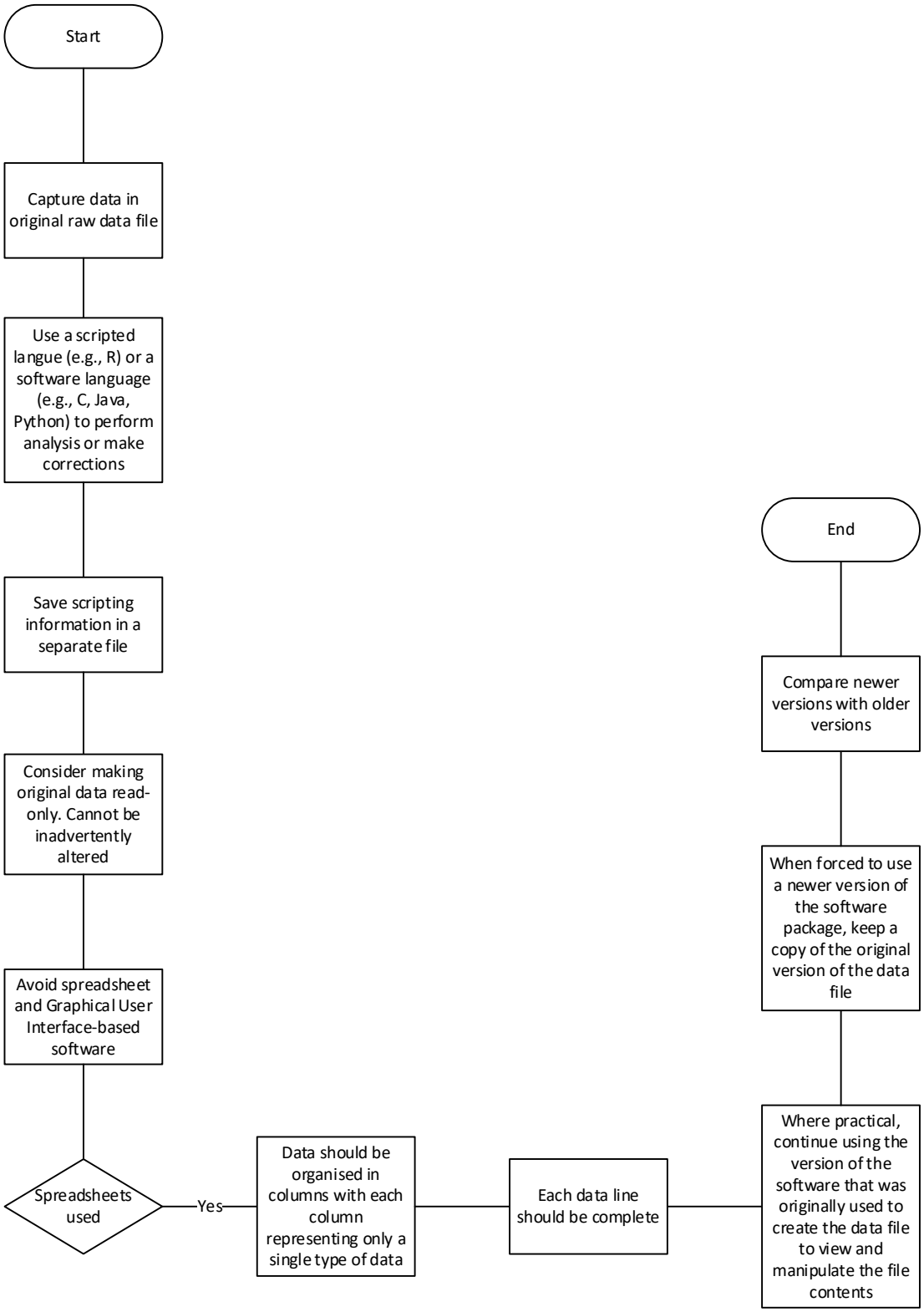
P1 - Data management planning – Start early



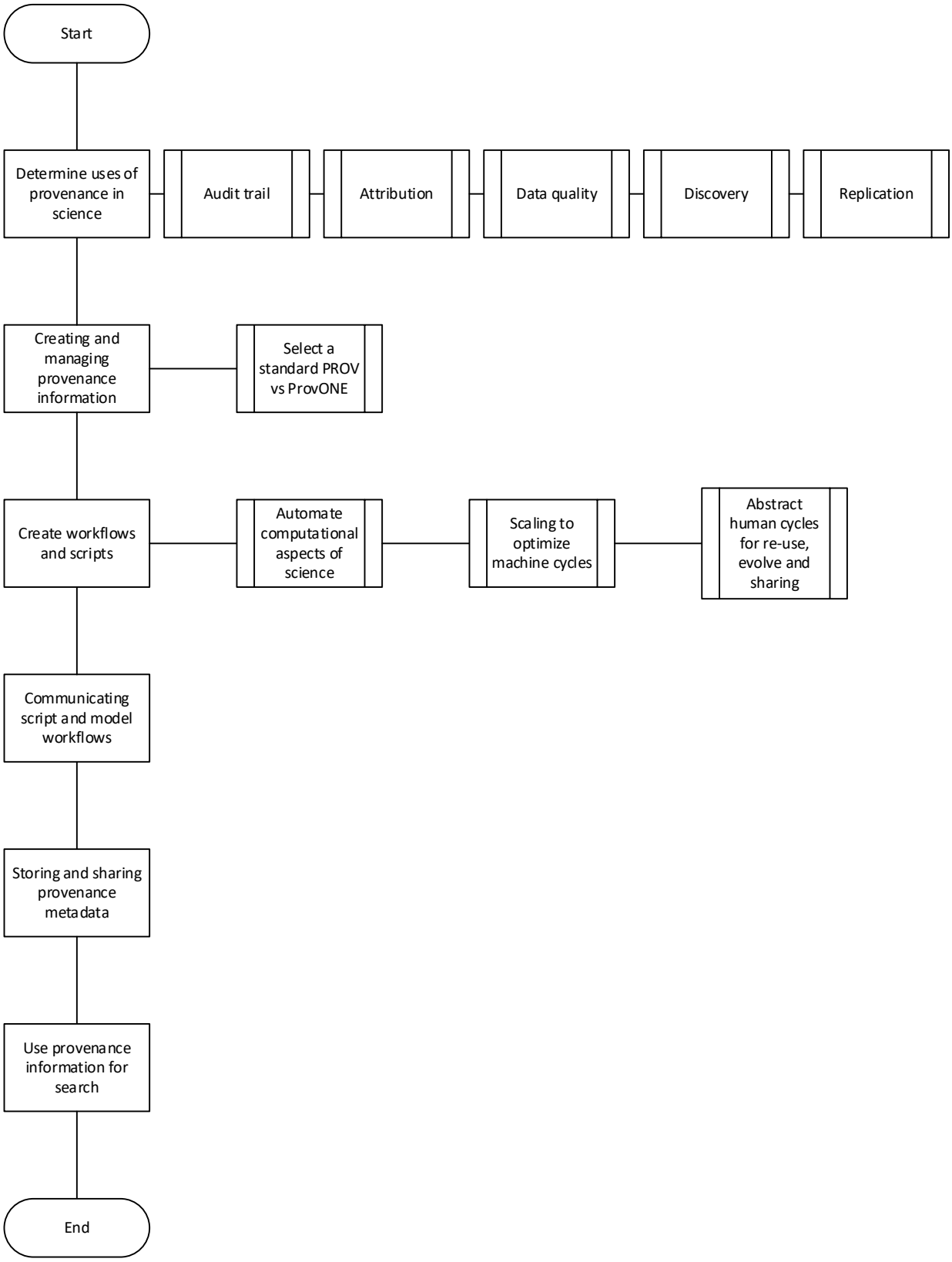
P2 - Multi media management planning



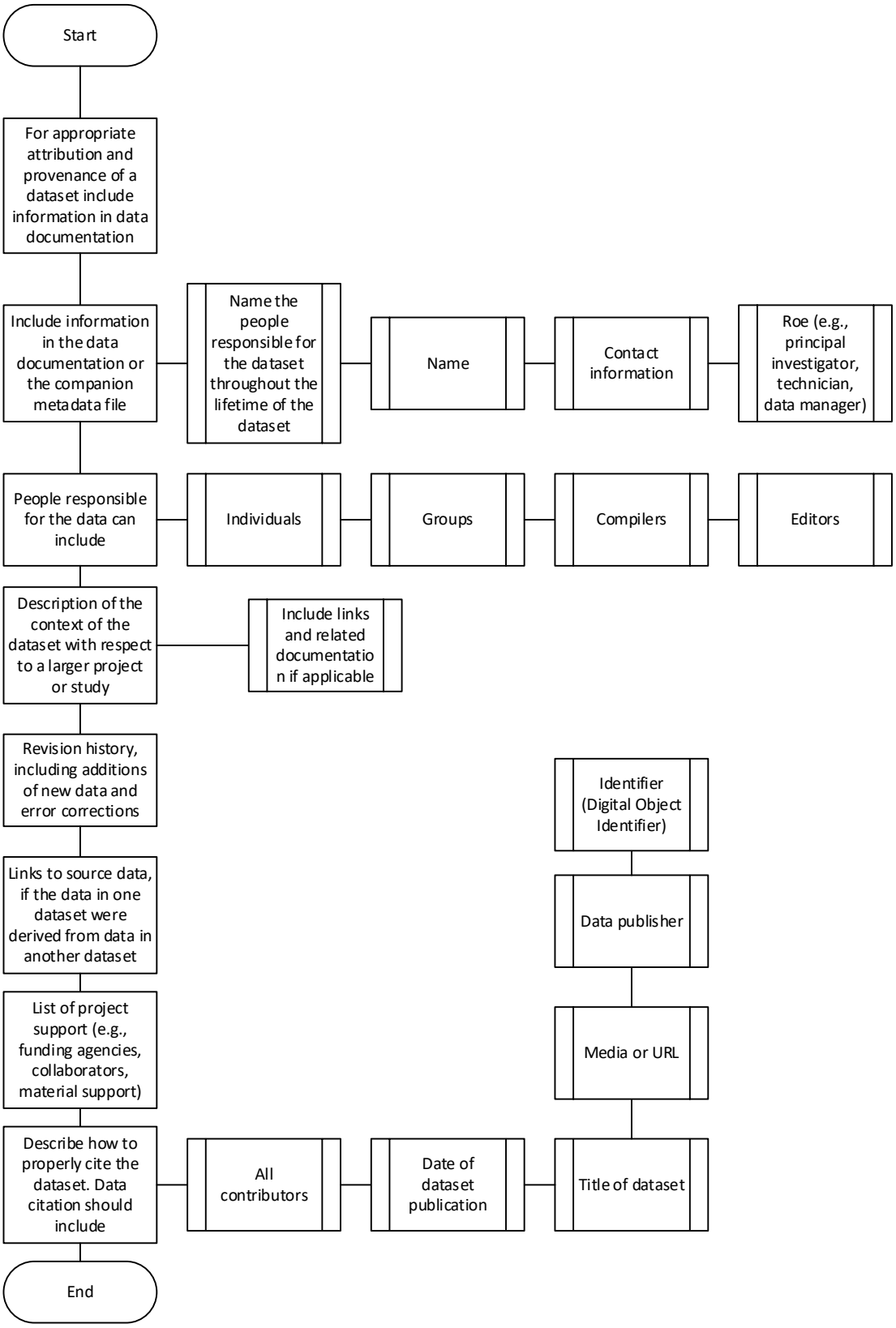
P3 Store data in its raw format



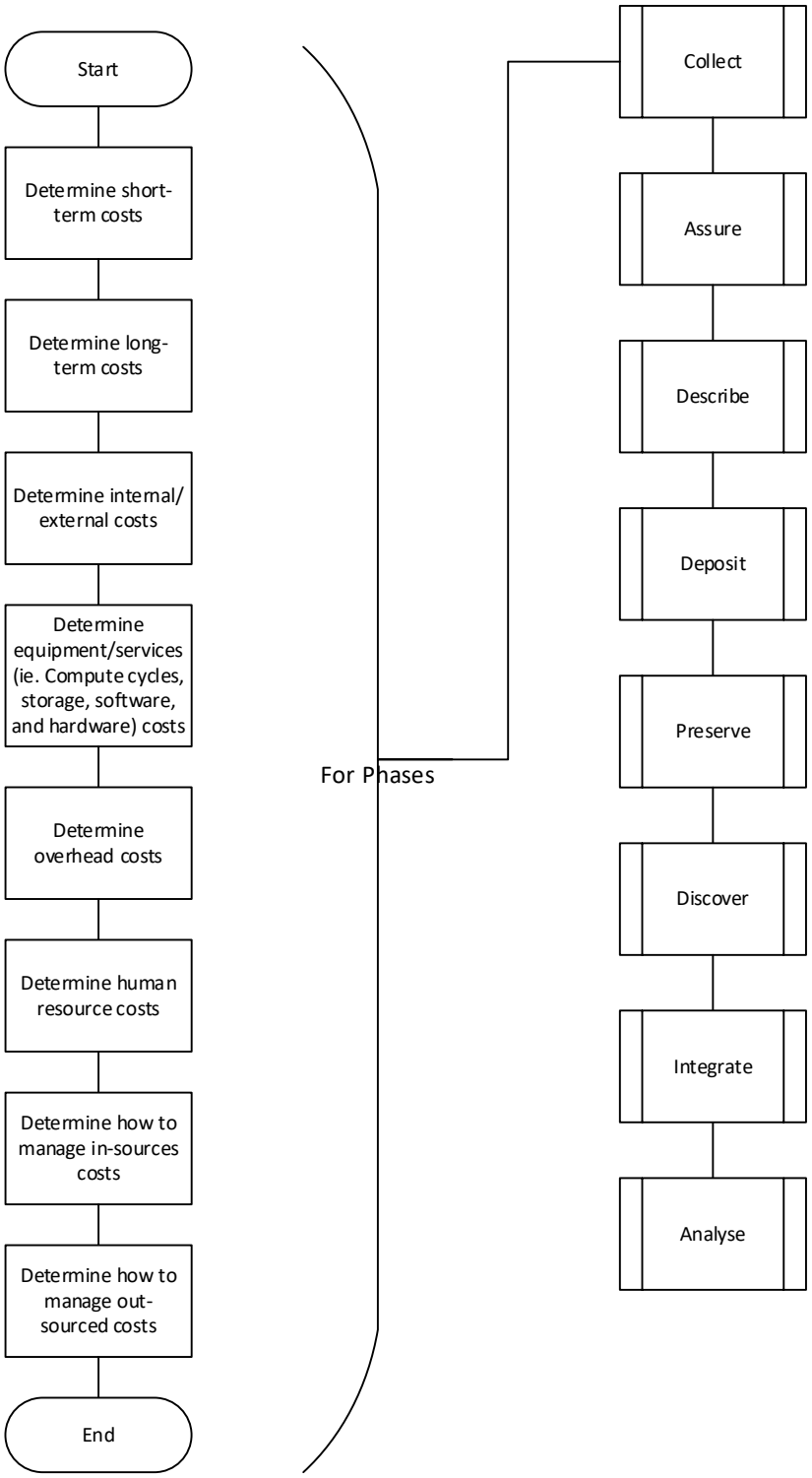
P4 - Provenance enable the reproduction of data results



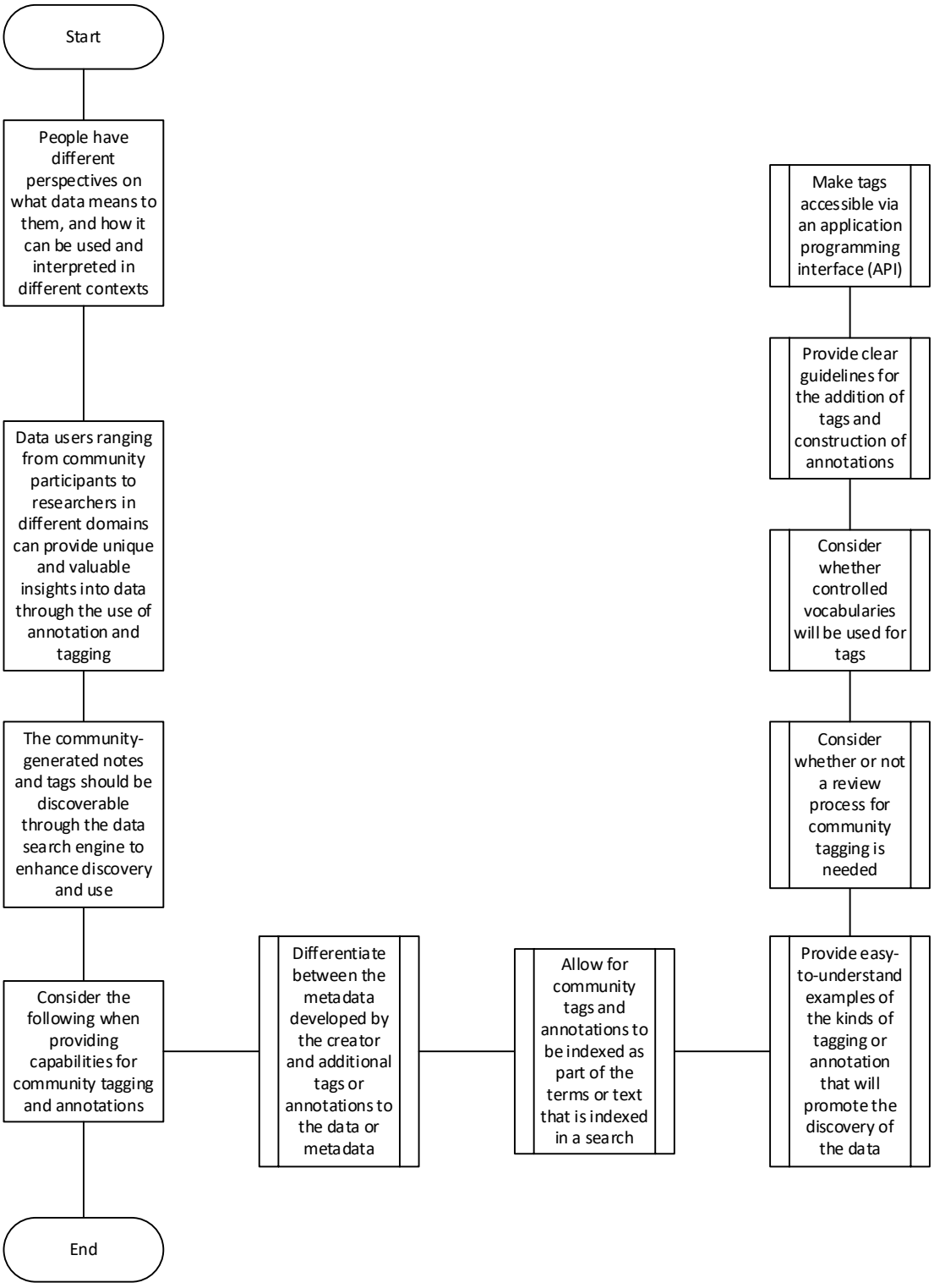
P5 - Provenance and data cite documentation guidelines



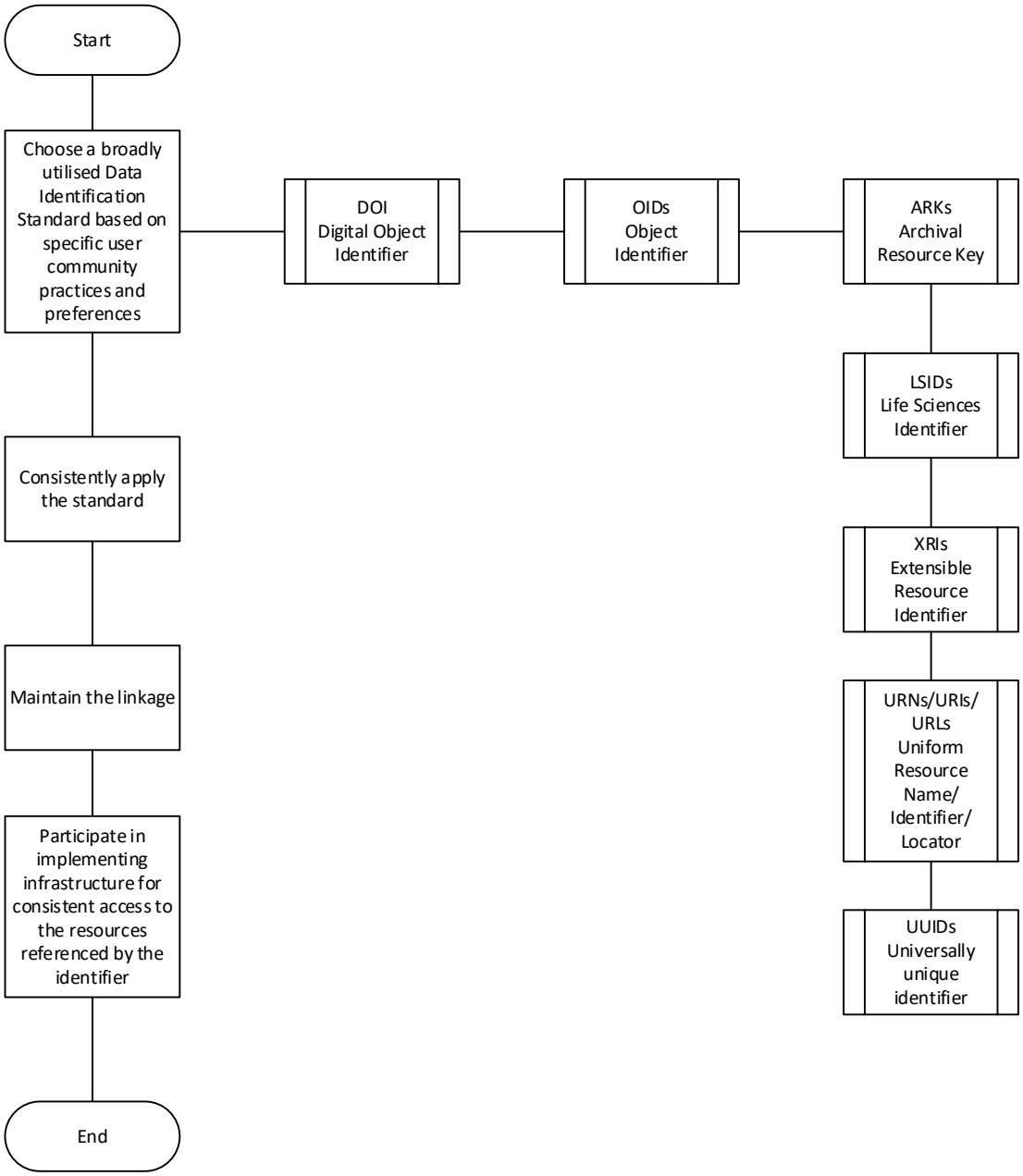
P6 - Guidelines for drawing up a budget



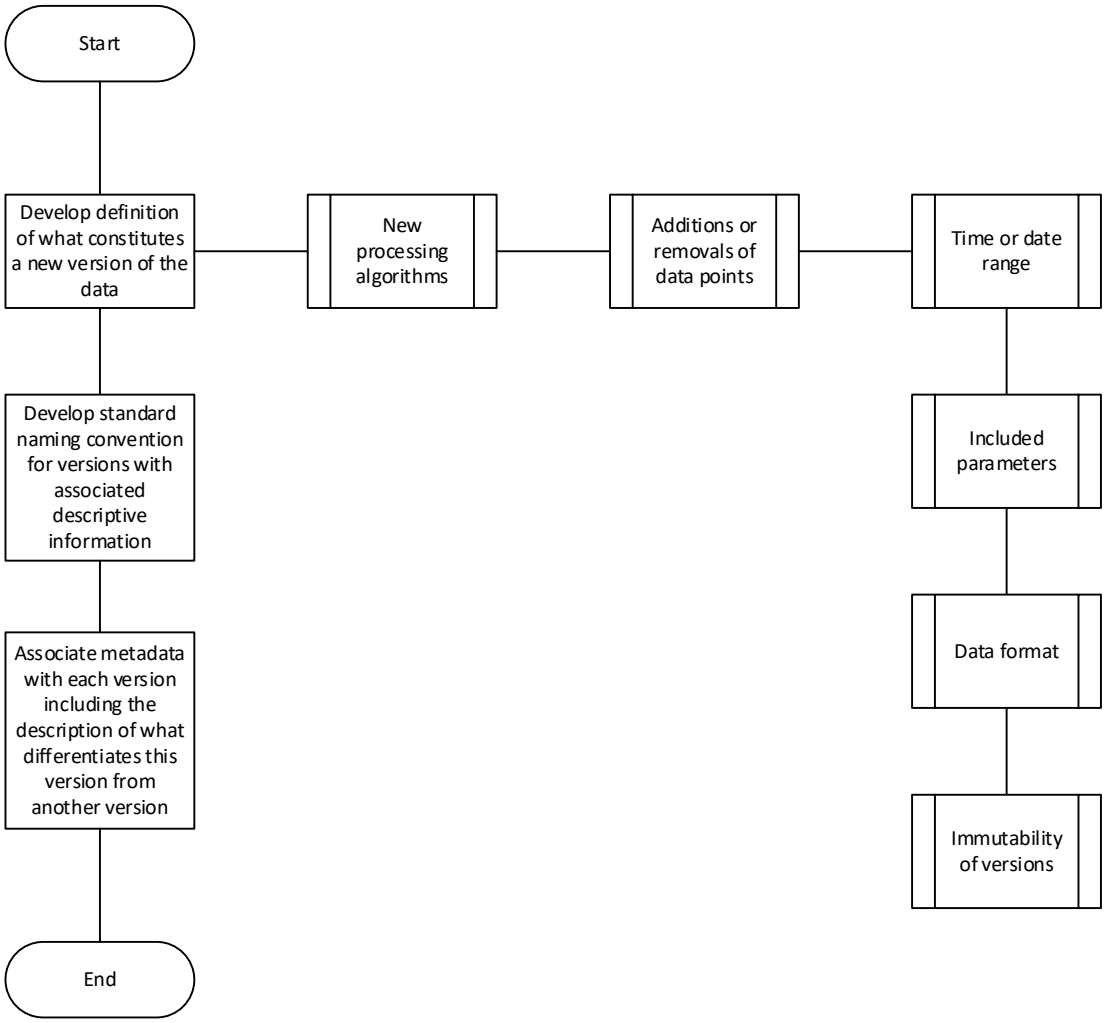
P7 - Enable community members to tag your data



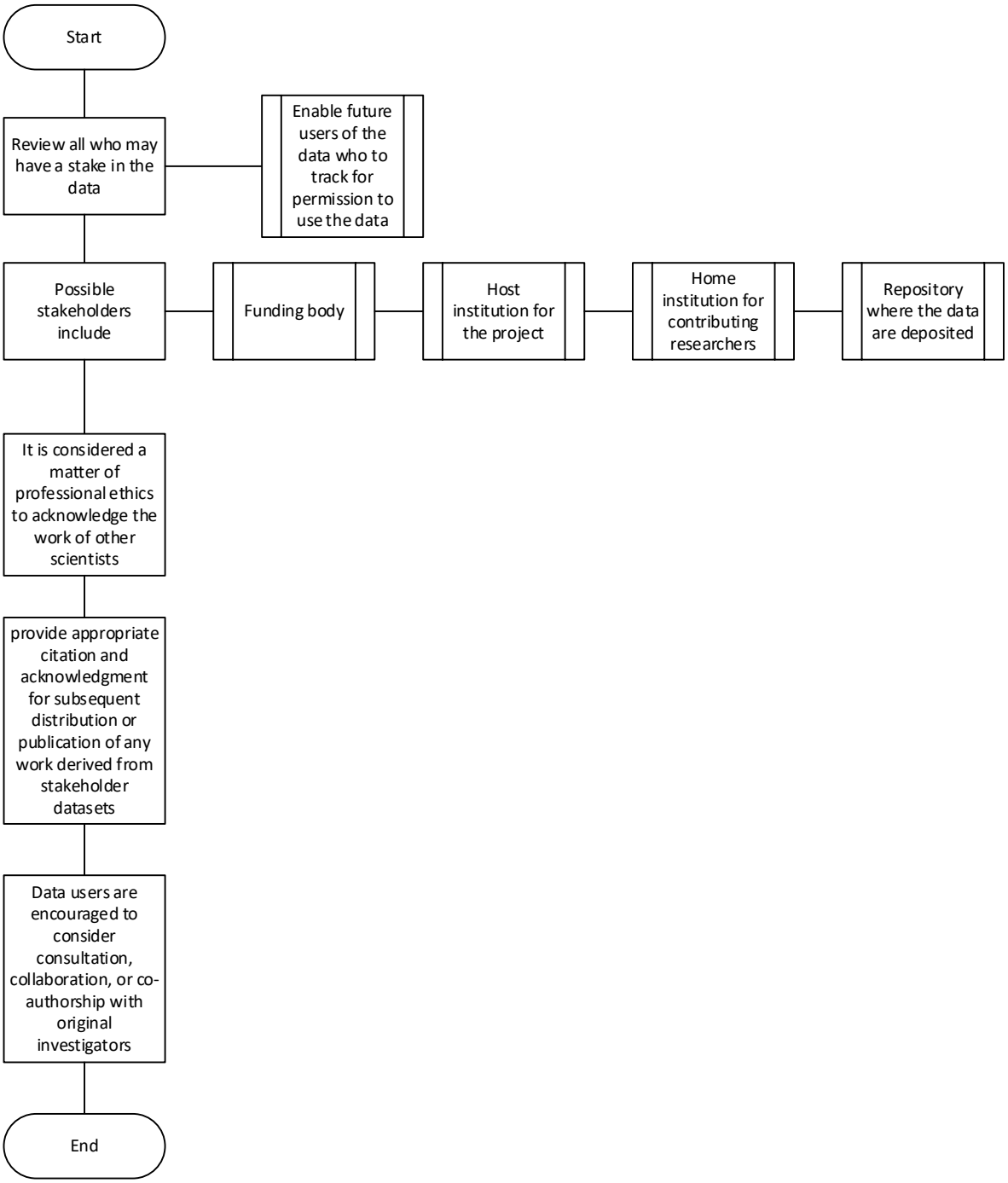
P8 – Register identifier for dataset



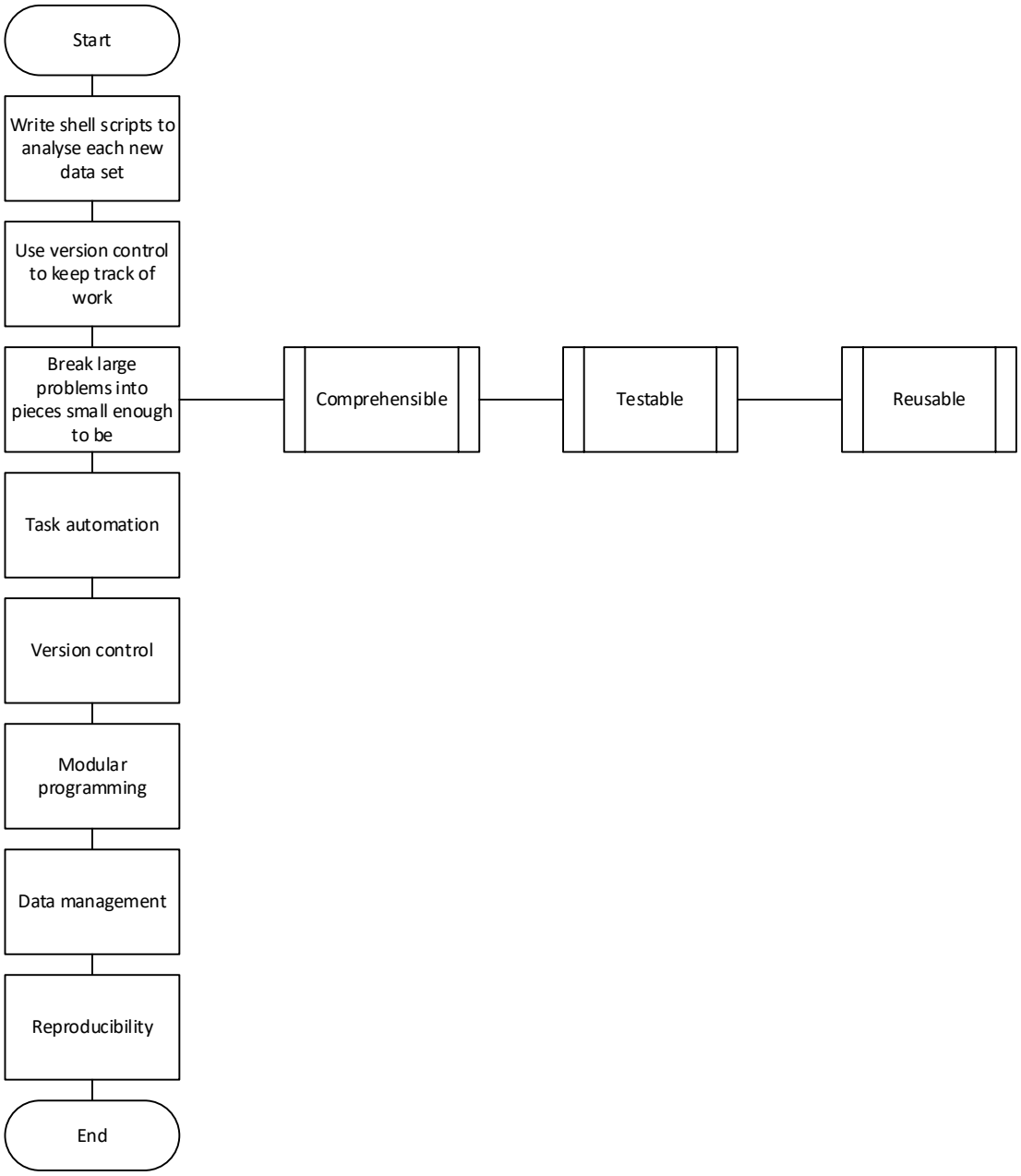
P9 - Versioning of data



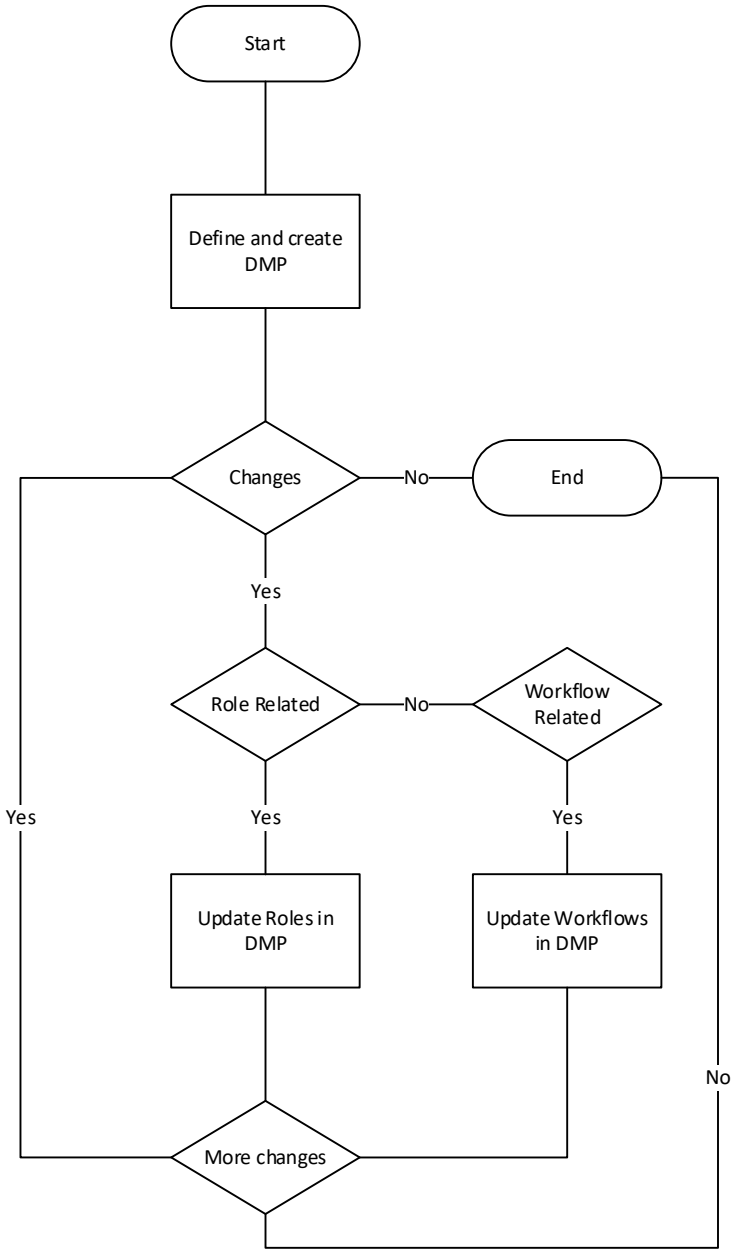
R1 - Data ownership and recognition



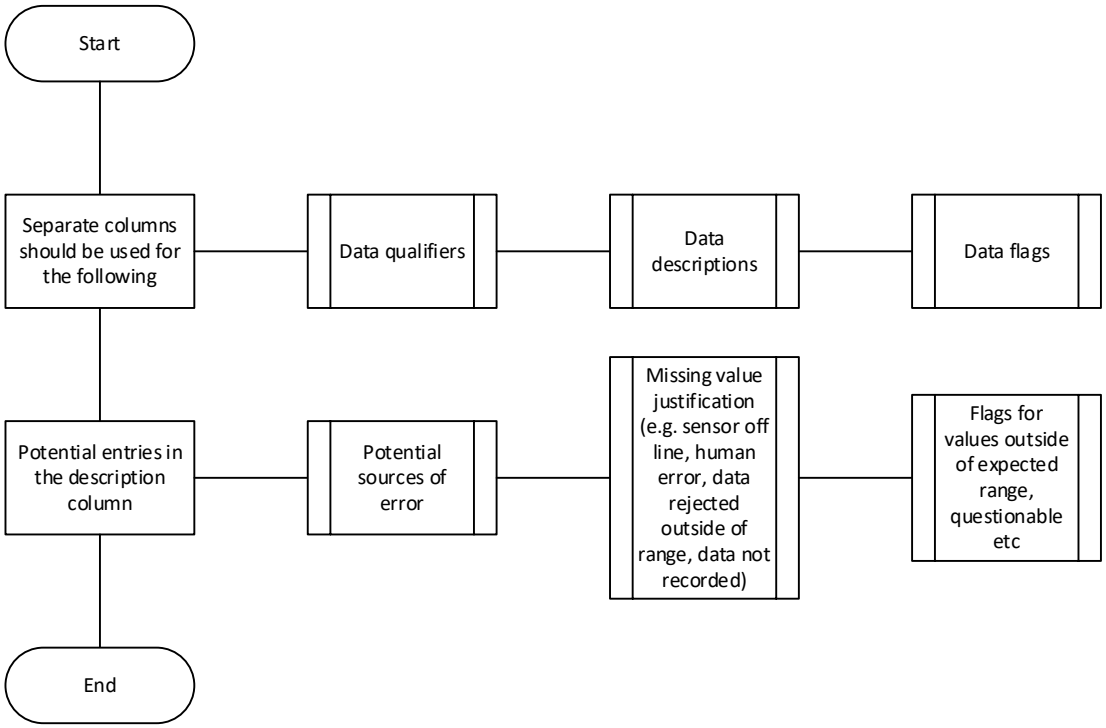
R2 - Repeatable and testable software processes to transform data



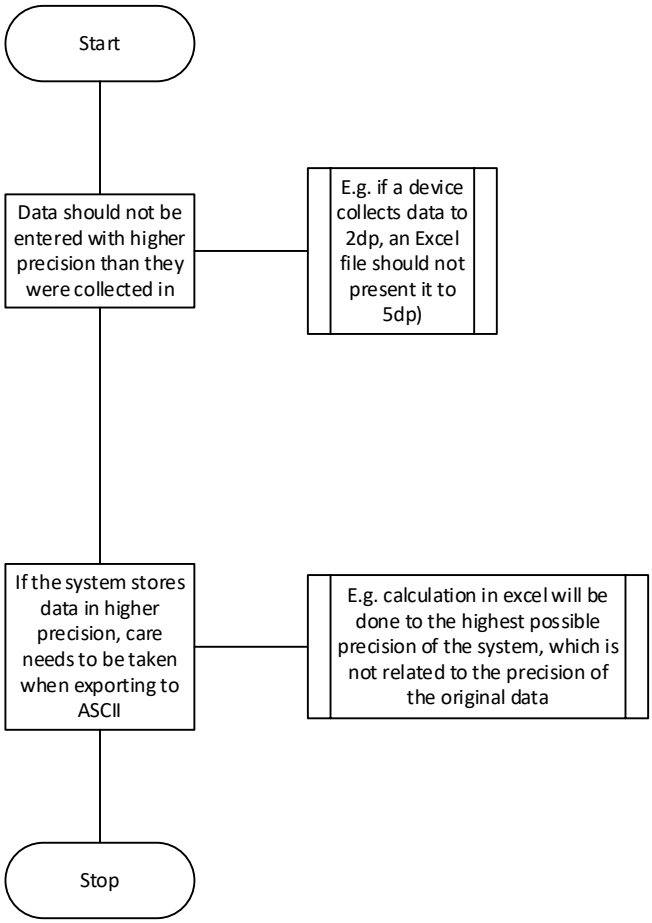
R3 - Refer back to RDM plan



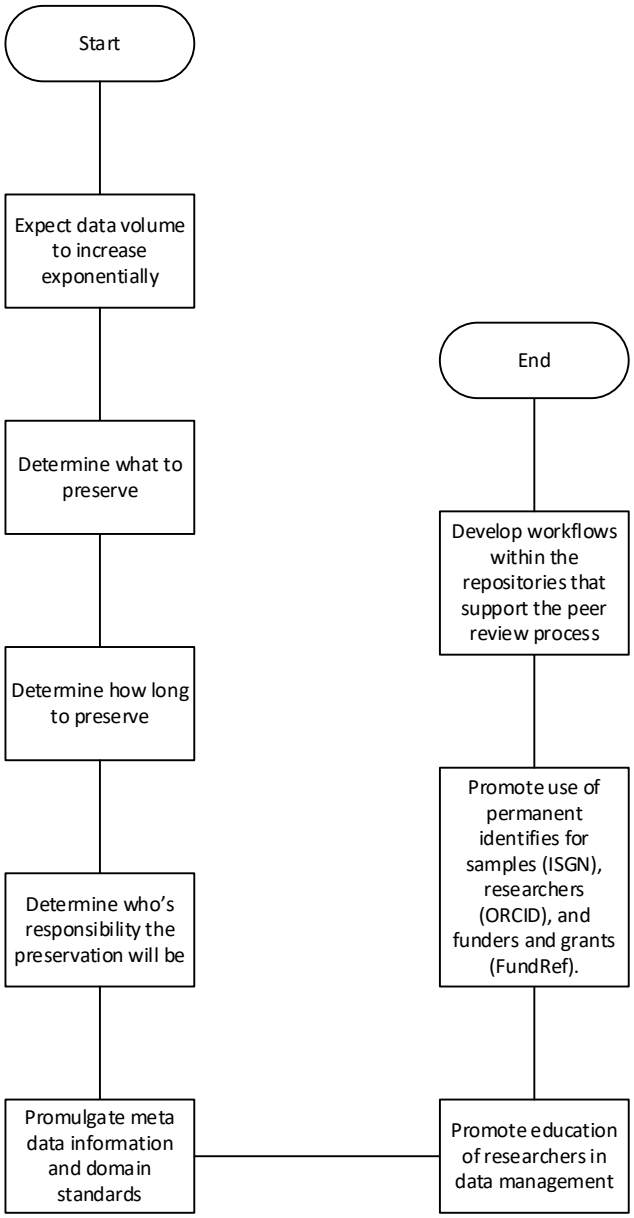
S1 - Avoid adding data descriptions on data sheets



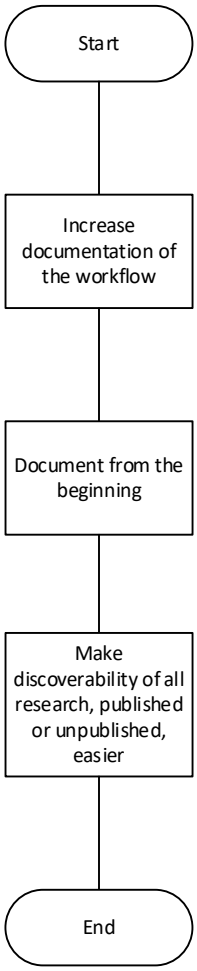
S2 - Guidelines on data precision



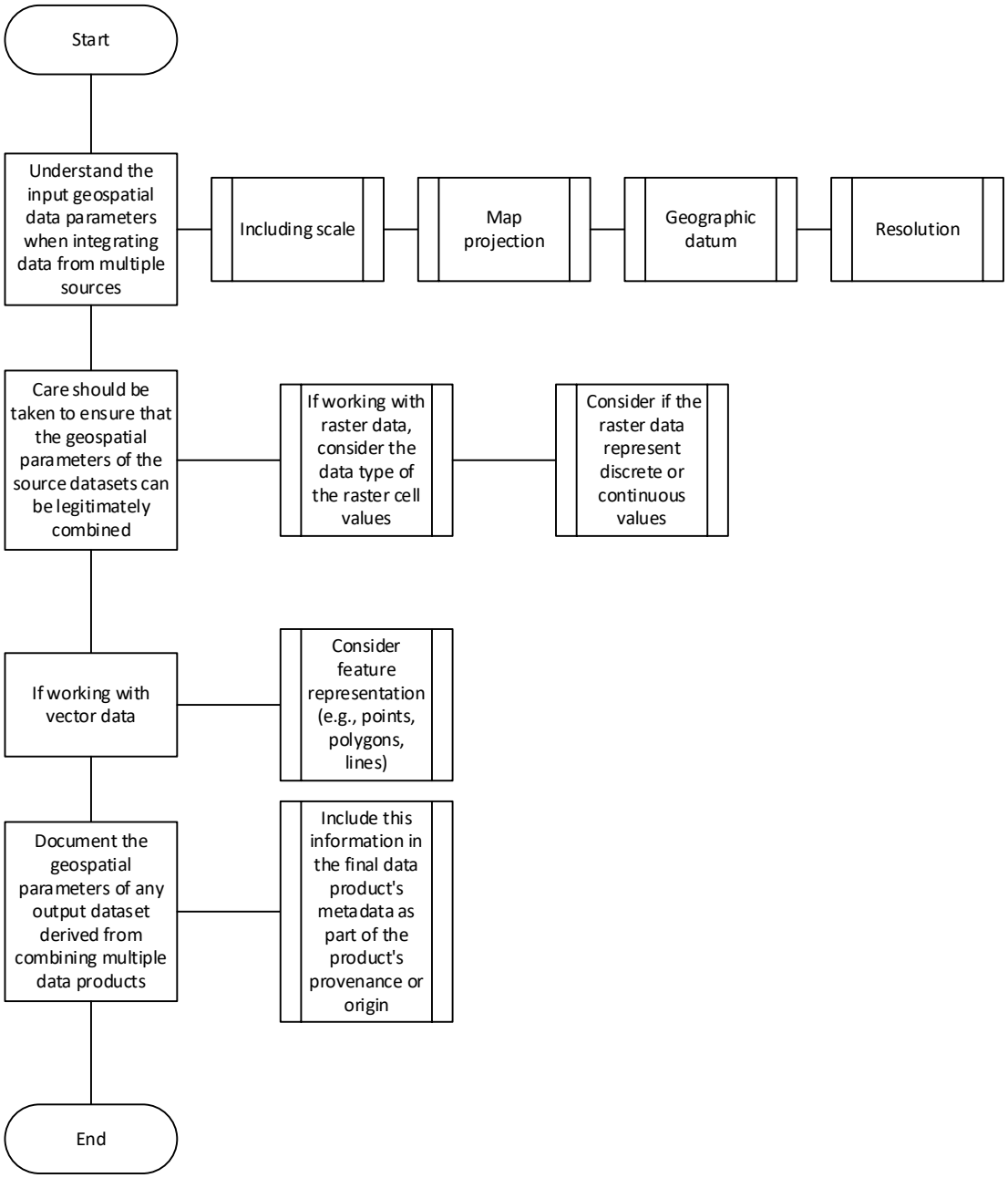
T1 - Data discovery and stewardship guidelines



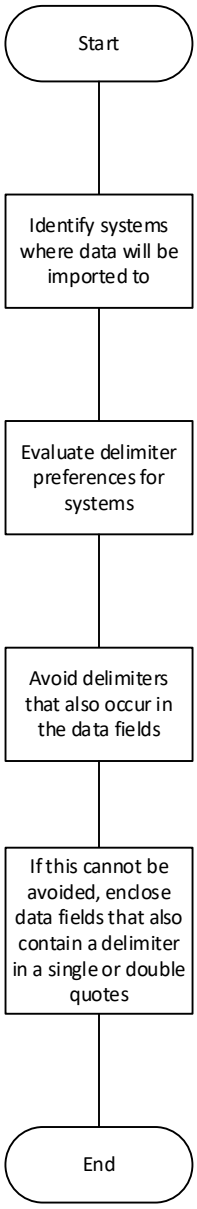
T2 - Guidelines to make data reproducible



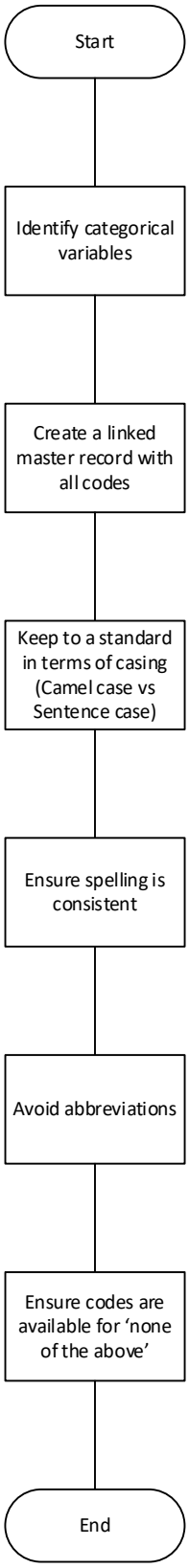
U1 - Parameter guidelines for geospatial data



U2 - Guidelines for field delimiters



U3 - Standardise on codes



ADDENDUM E: LETTER FROM LANGUAGE EDITOR

TANYA-LEE STEWART LANGUAGE EDITING SERVICES

TR Stewart – BA (General with English and Linguistics)

7 Siesta
8 Kamp Street
Potchefstroom, 2531

Tel: 0845567745
tanyaleestewart01@gmail.com

19 November 2017

To whom it may concern

DECLARATION OF LANGUAGE EDITING

Re: RESEARCH DATA MANAGEMENT FRAMEWORK FOR A SOUTH AFRICAN UNIVERSITY-BASED RESEARCH ENTITY

This serves to confirm that I undertook the language editing for the above-mentioned document on behalf of TV Bester, NWU student number 10728929.

All language errors identified were corrected electronically and marked with the 'track changes' function. The implementation of suggestions was left up to the author.

Should you have any queries please contact me on 084 556 7745.

Yours sincerely

TR Stewart
Member: South African Translators' Institute
SATI registration no: 1003470

ADDENDUM F: TURNITIN REPORT

turnitin

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	**PETRA BESTER**
Assignment title:	**TV Bester second review**
Submission title:	**TV_Bester_mini-dissertation_20No..**
File name:	**TV_Bester_mini-dissertation_20No..**
File size:	**4.18M**
Page count:	**177**
Word count:	**27,776**
Character count:	**219,967**
Submission date:	**20-Nov-2017 05:16AM (UTC+0200)**
Submission ID:	**882696958**


Research data management framework for a South African university-based research entity

TV Bester

orcid.org/0000-0003-2194-322X

Graduation March 2017