Towards a taxonomy for quality control in Environmental Sciences

Jordan Maduro
jordan.maduro@student.uva.nl

September 15, 2018, 46 pages

Supervisor: Dr. Zhiming Zhao
Host organisation: Universiteit van Amsterdam

Universiteit van Amsterdam
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Master Software Engineering
http://www.software-engineering-amsterdam.nl

Contents

Abstract **3**

1 Introduction **4**
1.1 Motivation . . . 4
1.2 Research questions . . . 4
1.3 Research method . . . 5
1.4 Thesis outline . . . 5

2 Background **6**
2.1 Data Quality . . . 6
2.2 Environmental observation data . . . 6
2.2.1 Station-based data . . . 7
2.2.2 Gridded data . . . 7
2.3 Quality Assurance and Quality Control . . . 7
2.3.1 Quality Assurance . . . 7
2.3.2 Quality Control . . . 7
2.4 ENVRI Reference Model . . . 7
2.5 Taxonomy building . . . 8

3 Research Infrastructures **9**
3.1 Introduction . . . 9
3.2 Quality Control in RI . . . 9
3.3 AnaEE . . . 10
3.4 EuroArgo . . . 11
3.5 ACTRIS . . . 12
3.6 EISCAT_3D . . . 14
3.7 ICOS . . . 15
3.8 Commonalities . . . 16
3.9 Diversities . . . 16
3.10 Summary . . . 17

4 Taxonomy **18**
4.1 A taxonomy for quality control . . . 18
4.1.1 Automation levels . . . 18
4.1.2 Benchmarks . . . 18
4.1.3 Data quality . . . 19
4.1.4 Methodologies . . . 19
4.1.5 Processes . . . 19
4.1.6 Roles . . . 20
4.1.7 Standards . . . 20
4.1.8 Temporalities . . . 21
4.1.9 Tools . . . 21
4.2 Relations . . . 22

5 Harmonized Processes **25**

5.1 Introduction . . . 25
5.2 Updated Models . . . 25
5.2.1 AnaEE . . . 25
5.2.2 EuroArgo . . . 26
5.2.3 ACTRIS . . . 26
5.2.4 EISCAT_3D . . . 28
5.2.5 ICOS . . . 28
5.3 Further analysis . . . 29
5.4 Summary . . . 29

6 State of the art survey **31**
6.1 Introduction . . . 31
6.2 Survey . . . 31
6.2.1 Methodologies . . . 31
6.2.2 Standards . . . 32
6.2.3 Benchmarks . . . 33
6.2.4 Data quality . . . 34
6.2.5 Tools . . . 35
6.3 Summary . . . 36

7 Conclusions **38**
7.1 Discussion . . . 38
7.2 Achievements . . . 38
7.3 Recommendations . . . 39
7.4 Conclusion . . . 40
7.5 Future work . . . 40

A Interview transcripts **41**
A.1 ACTRIS . . . 41
A.2 EPOS . . . 42

Bibliography **44**

Abstract

The use of environmental science data in forecasting and decision making makes producing high-quality data essential. To produce high-quality data research infrastructures develop quality control solutions ad hoc. Some methodologies, tools, and techniques apply to multiple environmental domains. However, sharing these tools is difficult because of a lack in standardization or common terminology. In this thesis, we propose a taxonomy for quality control in environmental sciences. The research was conducted through literature studies, interviews, and the development of the taxonomy. Furthermore, we surveyed the state of the art quality control methods. This taxonomy can be used as a common vocabulary and to classify quality control methodologies, tools, and techniques.

Acknowledgement: The work is partially supported by EU H2020 ENVRIPLUS (654182) and VRE4EIC (676247) projects.

Chapter 1

Introduction

With the technological advancements occurring in hardware sensor technology at an exceedingly fast pace, the number of sensors and the amount of data gathered is growing exponentially [TL13, LQLG16]. Environmental research infrastructures (RI) are collecting an ever-increasing amount of observational data. A research infrastructure provides user communities with support for many research-related activities such as data curation, discovery, analytical tools, and standard operating procedures. In recent years, the importance of producing high-quality data has become the focus of many environmental research infrastructures. The published data is used in monitoring, environmental preservation, air quality and water quality assessment, climate modeling, climate forecasting, and decision making. The variable and adverse conditions where the sensors are deployed lead to many potential data quality issues. Some examples include animals and insects misusing sensors as homes or tools [SFA+00, CRP+13]; sensors placed in ship engine rooms, exposed to high temperatures and vibration [HVM+16]; regular or extreme weather events causing corrosion to equipment [SFA+00, CRP+13]; and sensor failures due to inadequate maintenance [SFA+00, CRP+13, HVM+16]. Using erroneous data can lead to invalid conclusions or incorrect analysis. Therefore, data quality control is a crucial component of any data lifecycle.

1.1 Motivation

Researchers and practitioners have identified the need for standardization within the quality control process of environmental science research infrastructures [CBK+10, CRP+13, TL13]. Scientific data is often collected, curated, preserved, and distributed by different organizations [RPMS17]. Moreover, each organization has its quality standards and quality control process in place [TL13]. Admittedly, there is no one size fits all solution for quality control. Furthermore, the type of quality control required is mostly dependent on the data type and where the sensors are located at the time of measurement. For this reason, many organizations create a quality control process ad hoc. This ad hoc development of quality control processes leads to duplication of effort across research infrastructures, countries and domains. Without standardization of the quality control process, no two datasets can be combined in any meaningful way [TL13]. Therefore, the goal is to enable research infrastructure interoperability.

The purpose of this thesis is to study whether the quality control process can be standardized across environmental science disciplines by utilizing a quality control taxonomy.

1.2 Research questions

In this thesis we focus on the following research questions:

- RQ1. How can we recommend an effective tool for controlling data quality in environmental science data management lifecycle?
- RQ2. What are the quality factors that need to be controlled in the environmental data management lifecycle?

- RQ3. What are the quality control tools and methodologies used by the environmental community?
- RQ4. What are the criteria for choosing a QC tool and what makes it a good QC tool?

1.3 Research method

The research comprised of five phases. Below we describe the different phases and purpose of each phase in detail.

- In the first phase, we gather knowledge in the effort to answer question RQ 3 and RQ 4 by conducting a literature study on state of the art in quality control and quality control practices in environmental research infrastructures. Secondly, to identify the needs of practitioners as stated by RQ 4, we conducted interviews with domain experts within the European research infrastructure communities. Lastly, to answer RQ 2, we conduct a literature study on the concept of data quality and causes of data quality issues.
- In the second phase, we conducted a literature study related to RQ 1 and RQ 3 to identify the typical components of the quality control process.
- In the third phase, to test our hypothesis that a taxonomy is the best approach to answer RQ 1, we did a literature study on taxonomy building methodologies. Subsequently, we build a taxonomy to classify the quality control process.
- In the fourth phase, we validate the taxonomy through utilization by classifying the current state of the art.
- In the last phase, we formulate a recommendation for quality control in research infrastructures based on the results of answering the research questions.

1.4 Thesis outline

In this chapter, we describe the research motivation and research questions. In chapter 2, we describe the relevant background topics related to this work. In chapter 3, we enumerate some European research infrastructure and compare the commonalities and diversities. In chapter 4, we describe the taxonomy of quality control in environmental research infrastructures. In chapter 5, we examine the previously created models using the taxonomy. In chapter 6, we apply the taxonomy and present a survey of the state of the art in quality control. In chapter 7, we conclude this thesis with a discussion, recommendations, achievements, and future work.

Chapter 2

Background

This chapter presents some background for the research presented in this thesis. It describes data quality and data quality problems, a description of observational data types, the distinction between quality assurance and quality control, a brief introduction to the ENVRI reference model, and taxonomy building.

2.1 Data Quality

Many problems with data quality can occur during the data management lifecycle. A typical problem is a sensor producing poor quality data during the data acquisition phase. There are numerous reasons why a sensor would produce poor quality data or fail. Notably, failure to do adequate maintenance, operating in adverse environments, loss of electricity, and malicious human activity [CRP+13, AJM+15, VBG+17]. Furthermore, even when properly collected, the data can still be corrupted during transmission because of adverse environmental conditions, use of an inadequate power supply, electromagnetic interference, and network congestion [CRP+13, AJM+15, VBG+17].

The data produced by environmental science sensor networks can be classified as big data. Demchenko et al. [DGdLM13] defined big data as data having the 5V properties: Volume, Variety, Velocity, Value, and Veracity. Indeed, the data produced in many of the environmental science disciplines exhibit these properties.

Data Quality is a subjective measure that is determined by the data consumer. Generally speaking, many data providers employ a measure of quality assurance and control to provide data of a specific quality level. However, data quality can have a different meaning for different data consumers depending on the context in which the data will be used. Wang et al. [WS96] describe this concept as "fit for use". Moreover, the environmental science community has adopted a similar view of data quality.

Many studies have established that data quality is a multidimensional concept [PLW02]. Many schemes have been proposed consisting of a set of dimensions which represent the different characteristics of data. However, there is no consensus on the set of dimensions which define data quality, or the exact definition of each dimension [BCFM09]. Nevertheless, there are commonalities between the sets of dimensions proposed in the literature. These include accuracy, completeness, consistency, and timeliness [BCFM09]. These dimensions often represent the "What" in measuring data quality.

In summary, data quality can be affected by many reasons. Environmental science data can be classified as big data and the challenges that come with it. Data quality is depended on the context and purpose of use. Finally, data quality is defined as a multidimensional concept.

2.2 Environmental observation data

The environmental sciences capture data about the environment using various methods. These include Argo Floats, floating and moored Buoys, Seismographs, Cabled observatories, Gliders, High-Frequency Radars, Lidars, Thermal dissipation sap flux sensors and many more [TBH+15, RHH+15,

DAB+15, AJM+15, HPA+16, OHO16, VBG+17]. The datasets produced in environmental sciences have a distinct characteristic that they have to adhere to physical laws and properties [KL15]. The observation datasets can usually be categorized into Station-based and Gridded data.

2.2.1 Station-based data

Sensor measurements are generally obtained from irregularly and non-uniformly spaced stations. These type of measurements are the most direct and therefore the least error-prone data sources available [KL15]. However, in situ sensors are limited to their spatial and temporal coverage and generally suffer from other physical limitations which introduce errors and uncertainty in these datasets.

2.2.2 Gridded data

Grid-based datasets are produced by using some interpolation, aggregation and sampling techniques to provide easily accessible datasets at a fixed spatial grid, with particular spatial resolution and available at regular time intervals [HHT+08, KL15].

2.3 Quality Assurance and Quality Control

In literature quality assurance and quality control are often mentioned in the same sentence. In this thesis, we define quality assurance and quality control as two distinct processes.

2.3.1 Quality Assurance

Quality Assurance is the process of implementing preventive measures during data acquisition to make sure the data produced is of high quality. These measures include training of personnel, implementing sensor maintenance schedules, keeping documentation, producing metadata, implementing sensor redundancy, and the proper calibration of sensors [IK95, CRP+13, AJM+15].

2.3.2 Quality Control

On the other hand, Quality Control is the process or set of steps taken to check if the quality of the data produced is of adequate quality [CRP+13]. A quality control process identifies erroneous data and annotates them with the appropriate data qualifier [Cum11, CRP+13, AJM+15, HRJM15, No16]. It usually consists of a series of checks that determine if the data conforms to a particular property. These tests can be either an automated system or performed manually by a principal investigator. Quality control ensures the consistency of the data and the proper functioning of the sensors. The results of quality control can feed back into quality assurance by detecting sensor failure early on, which can allow network operators to take appropriate actions [AJM+15].

2.4 ENVRI Reference Model

The creation of research infrastructures is a daunting task. Moreover, the complexity of research infrastructures requires a long learning process. The ENVRI [1] project identified various issues recurring in the development of environmental science research infrastructures. These problems include [idlHH]:

- Duplication of effort by tackling similar problems.
- A lack of standards which impede development and harmonization
- A significant amount of models which require a deep understanding of different research infrastructures.

[1] http://www.envriplus.eu/

The ENVRI reference model (ENVRI-RM) was created to solve these problems [ZMG+15]. The ENVRI-RM is based on the Open Distributed Processing (ODP) model and consists of five viewpoints. The viewpoints include Scientific, Computational, Information, Technological, and Engineering. The ENVRI-RM provides a way to model and communicate on an abstract level about environmental research infrastructures. Furthermore, a goal of the ENVRI-RM is to promote the interoperability of the various existing and new environmental research infrastructures. Even though the ENVRI-RM provides a general taxonomy to describe research infrastructures, it does not provide detailed terminology specifically for the quality control processes. This detailed terminology for the quality control processes is a focus of this thesis.

2.5 Taxonomy building

A taxonomy is a method of classification based on empirical similarities or shared lineage [Bai94]. Taxonomies in the classical sense are used in biology to classify living things. In Software Engineering it can be used to classify concepts within the Software Engineering knowledge areas [UBBM17]. Researchers have used taxonomies as a way to:

- Reduce the complexity and identification of similarities, differences, and interrelationships among objects [Bai94, JBV09, NVM13, UBBM17].
- Have shared vocabulary within a domain, which eases knowledge sharing [JBV09].
- Identify gaps in a knowledge field [JBV09, NVM13] .

The right structure depends on the knowledge and maturity of the domain of classification [Kwa99]. The structures include:

Hierarchy The hierarchy is the classic structure and most used by the SE literature [UBBM17]. The primary focus of this structure is the "is-a" relationship. The hierarchy always has one root that branches out into branches and leaves. In this structure, a branch always inherits from the parent. In other words, a child is a subclass of the parent. Hierarchies should be used when the domain of classification is well known.

Tree The tree structure shares the same tree-like structure as the hierarchy structure. However, the tree structure does not share the "is-a" relationship. A tree structure can be used when there are specific rules on the distinction between classes without the inheritance rule.

Paradigm The paradigm structure is used to represent the intersection of two attributes at a time. This structure can be used for discovery of gaps, similarities, differences, and relationships. Paradigms should be used when the domain of classification is well known.

Faceted analysis The faceted analysis structure is the second most used structure in SE literature [UBBM17]. A faceted analysis is used to classify entities based on usually unrelated dimensions. Each dimension, facet, in a faceted analysis should contain at least two or more classes. Each facet can have a distinct structure to represent its classes. This structure is best used when the entities do not fit in a hierarchy, or the domain of classification is either not mature or not well known.

There are three methods used to validate a taxonomy [UBBM17]. These include:

Orthogonality demonstration By demonstrating the orthogonality of dimensions and categories.

Benchmarking By comparing the proposed taxonomy to similar taxonomies.

Utility demonstration By using the proposed taxonomy to classify subject matter examples.

Chapter 3

Research Infrastructures

In this chapter, we enumerate a number of European research infrastructures and describe their quality control processes.

3.1 Introduction

A research infrastructure (RI) provides user communities with support for many research related activities. These activities range from data curation, discovery and access services, analytical tools and common operating procedures. The Directorate-General for Research and Innovation defines a research infrastructure as [fRI10]:

> "research infrastructure" means facilities, resources and related services that are used by the scientific community to conduct top-level research in their respective fields and covers major scientific equipment or sets of instruments; knowledge-based resources such as collections, archives or structures for scientific information; enabling Information and Communications Technology-based infrastructures such as Grid, computing, software and communication, or any other entity of a unique nature essential to achieve excellence in research. Such infrastructures may be "single-sited" or "distributed" (an organised network of resources)

In this thesis we focus on environmental research infrastructures. Some examples include ACTRIS, AnaEE, EISCAT 3D, ICOS, and EuroArgo.

3.2 Quality Control in RI

Researchers in environmental sciences have commonly described quality control as a multiphase process. Each phase consists of multiple tests that identify potentially erroneous data and annotate them with data qualifiers. However, defining a typical quality control process is not a trivial problem. The distinct characteristics of each data type require separate quality control procedures. Besides the intrinsic characteristics of the data type, the procedures also depend on extrinsic factors such as:

- The different sensor makes and models which require distinct calibration based on the measurement location.
- The environment in which the sensors are operating.
- The available resource to perform the quality control procedures.

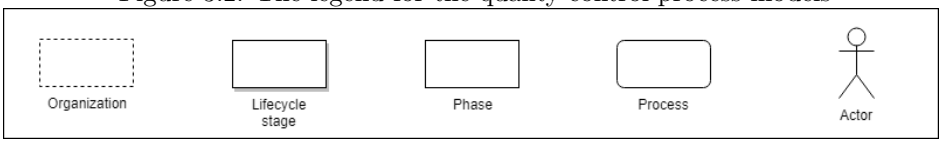
A quality control process in an RI is often across multiple phases during the data management lifecycle. Moreover, the exact phase where quality control starts and where quality control ends depends on the resources, scope, and data type of the RI. In this work, we adopted the data lifecycle model defined by the ENVRI-RM. The ENVRI-RM data lifecycle is depicted in figure 3.1

Figure 3.1: The ENVRI Reference Model data lifecycle model

To better understand the commonalities within the quality control process we examined the quality control practices described in the literature. Furthermore, we examined some representative European research infrastructures and described their quality control practices. The processes described below are based on the analysis of RI documents, published literature, and interviews with RI representatives. Moreover, we cover the quality control processes up to the publication of the primary data product.

In figure 3.2, we show a legend of the different elements used in the models below.

Figure 3.2: The legend for the quality control process models



3.3 AnaEE

The AnaEE[1] European environmental research infrastructure is a recent project that just entered the implementation phase in 2017. This project is designed to answer relevant scientific questions regarding the impact of environmental pressures on the functioning of ecosystems. The data is produced by four types of platforms. These are Open-air ecosystems, Enclosed ecosystems, Analytical, and Modeling.

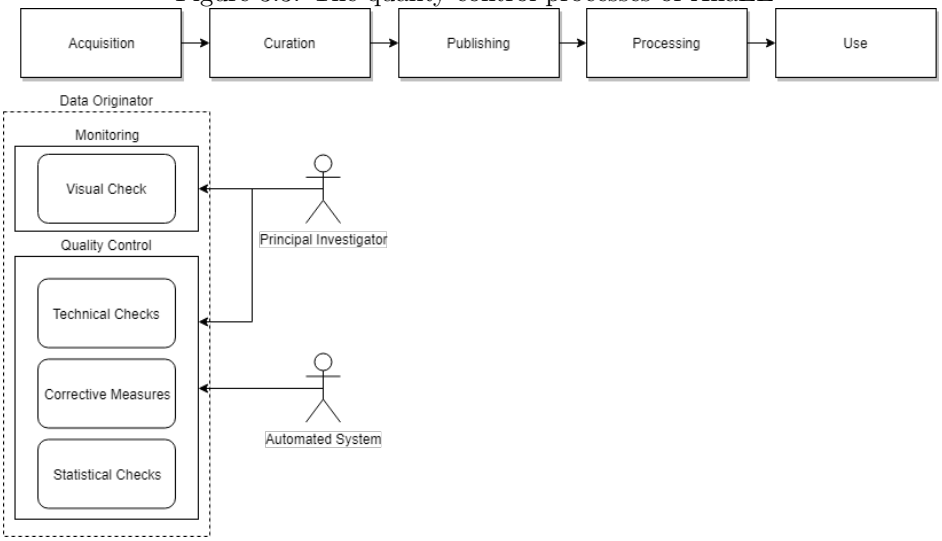
Because of the numerous possible sensors used to record the measurements, AnaEE currently leaves the quality control up to the platform performing the experiments.

An example of quality control procedures at the data originator can include a mix of automated and manual quality control. An example quality control process is described below and the procedures are shown in figure 3.3.

- A real-time visual monitoring by technicians or a scientist running the experiments.
- A program created specifically for a sensor that takes the raw data and applies quality control tests.

[1] https://www.anaee.com/

Figure 3.3: The quality control processes of AnaEE



3.4 EuroArgo

The EuroArgo[2] research infrastructure was established in 2008 and is based on the Argo[3] framework. The focus of this research infrastructure is ocean observations. Currently, it consists of 12 European member states and has a network of about 800 active free-drifting profiling floats.

Floats are instrument platforms that are deployed by one of the member organizations on behalf of the EuroArgo project. A float collects profile data by descending up to 2000m underwater and collecting measurements at predetermined intervals. After about ten days, the float will emerge and transmit the data one of the data assembly centers (DAC).

The quality control process is performed by the DAC where the data is received. The method currently consists of two phases real-time and delayed mode. The process is described below, and the procedures are shown in figure 3.4.

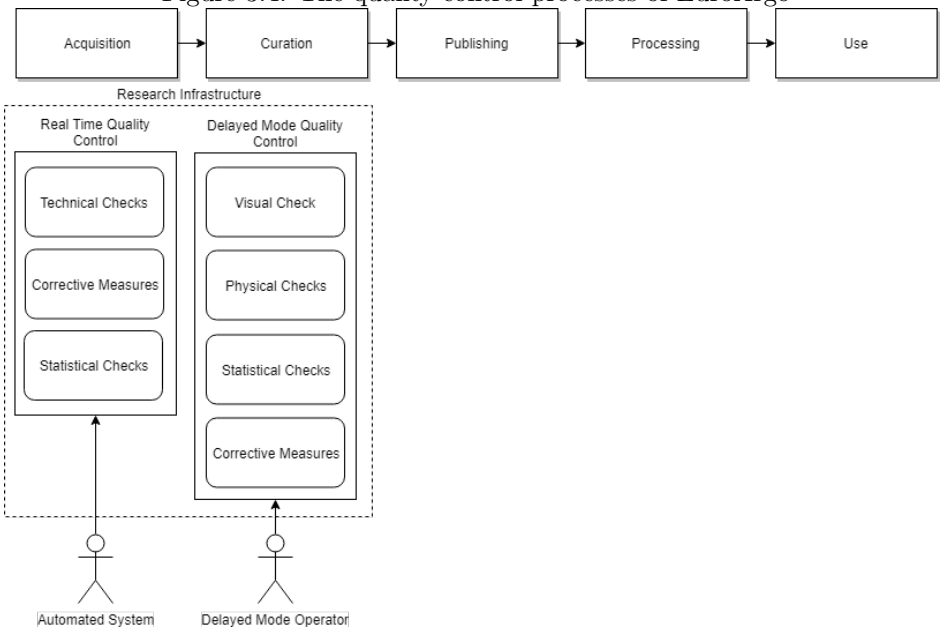
- The first is the automatic quality control based on the QARTOD[4] methodology. This methodology recommends real-time quality control checks that can be used shortly after the data is collected. About 19 tests are applied in total. These tests are automated and limited to the technical and statistical properties of the measurement.

- Second is the delayed mode quality control. This quality control process is a manual process that is done at least six months after the initial measurement is recorded. The procedures are documented in the Delayed Mode Quality Control manual and is performed by the Principal Investigator or Delayed Mode Operator. The delayed mode quality control detects sensor drift and includes a deep dive into the physical feasibility of the measurements.

The real-time quality controlled data is available after 24-48 hours. The data consumers are informed that this data has received the minimum quality control checks agreed upon in QARTOD and is provided without human intervention. The higher level of quality control data is available after the delayed mode quality control which takes a minimum time of 6 months.

[2]https://www.euro-argo.eu/
[3]http://www.argo.ucsd.edu/
[4]https://ioos.noaa.gov/project/qartod/

Figure 3.4: The quality control processes of EuroArgo



3.5 ACTRIS

The ACTRIS[5] is a European environmental research infrastructure started in 2015. ACTRIS studies about 135 different Atmospheric variables. Furthermore, the network consists of about 65 active sites that collect the data. Within the project, there are three thematic data centers. These thematic data centers are

- In-situ Aerosol and Trace Gases curated by EBAS[6].
- Aerosol remote sensing curated by EARLINET[7]
- Cloud remote sensing curated by CloudNet[8]

Each data center is responsible for the data curation of a data type. The different data types have their unique properties concerning volume and velocity which require different approaches. For this reason, a data center defines the standard operating procedures and the quality control procedures. The different quality control procedures are described below and the procedures are shown in figures 3.5, 3.6, and 3.7.

- In-situ Aerosol and Trace Gases
  - The first quality control is done by the data originators. In this phase, the data originators use their methods to flag the data.
  - The second quality control happens during submission using the EBAS-submit-tool and consists of many automated checks which determine the technical and statistical properties of the data. Furthermore, the system checks the metadata data and file format. Finally, the data is manually inspected by the thematic portal.
  - The third quality control consists of physical checks and comparison with measurements in the archive.

[5] https://www.actris.eu/
[6] http://ebas.nilu.no/
[7] https://www.earlinet.org/index.php?id=earlinet_homepage
[8] http://www.cloud-net.org/

- Aerosol remote sensing
  - The first phase is quality control done by the data originators using their methods. Additionally, they can make use of the Single Calculus Chain software provided by EARLINET.
  - The second phase is during submission of the data. Automatic quality control procedures are applied to the data. These checks assess the technical and statistical properties of the data.
  - The third phase occurs every three months when more extensive quality control checks are applied to the data. These checks assess the physical properties of the data.
- Cloud remote sensing
  - The cloud quality control procedures are fully automated. However, manual review by an expert panel occurs during the calibration of the data before final publication.

The three data types have NRT data dissemination. The data is made available to the community and public with a statement on the validity for use.

Figure 3.5: The quality control processes of EBAS

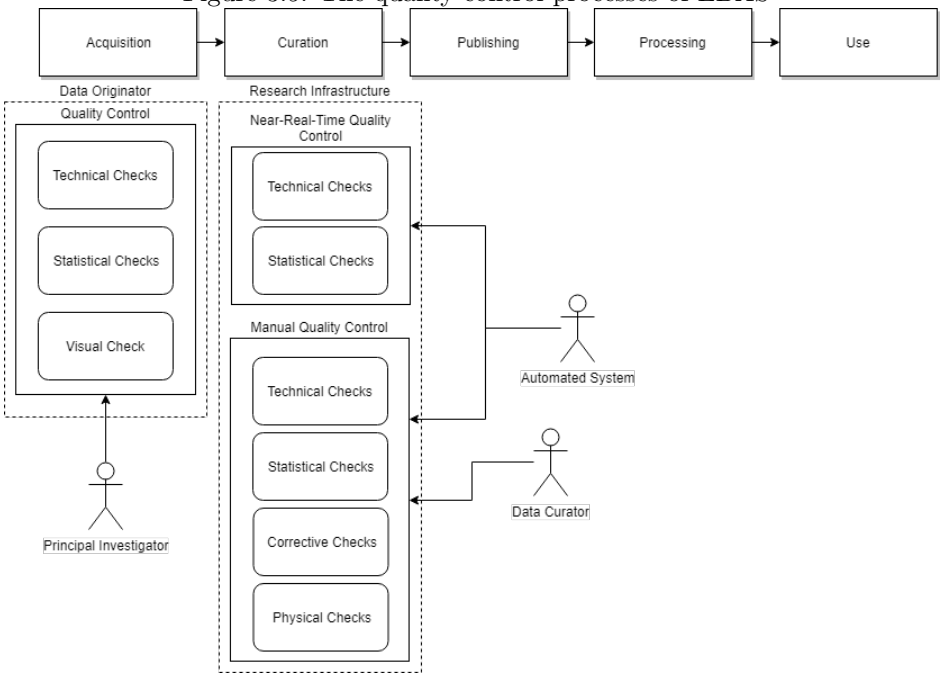


Figure 3.6: The quality control processes of EARLINET

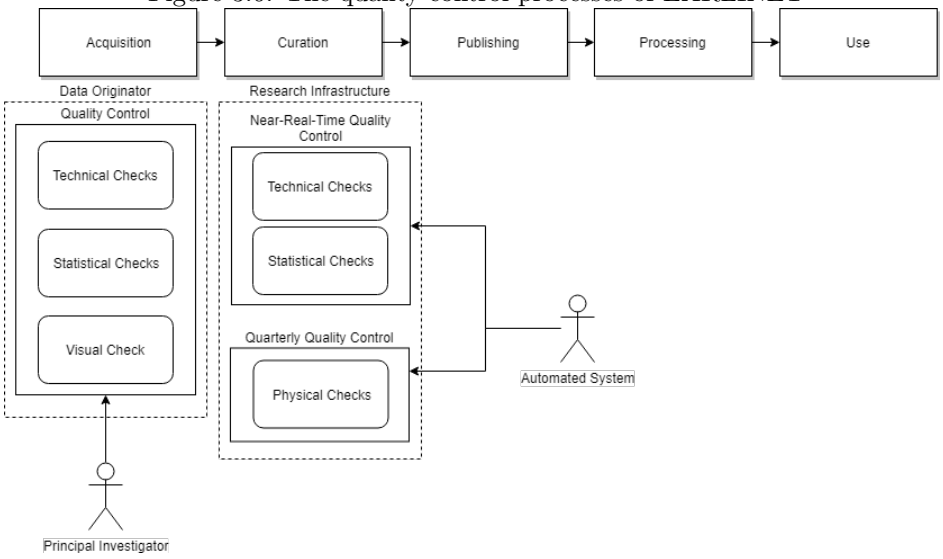
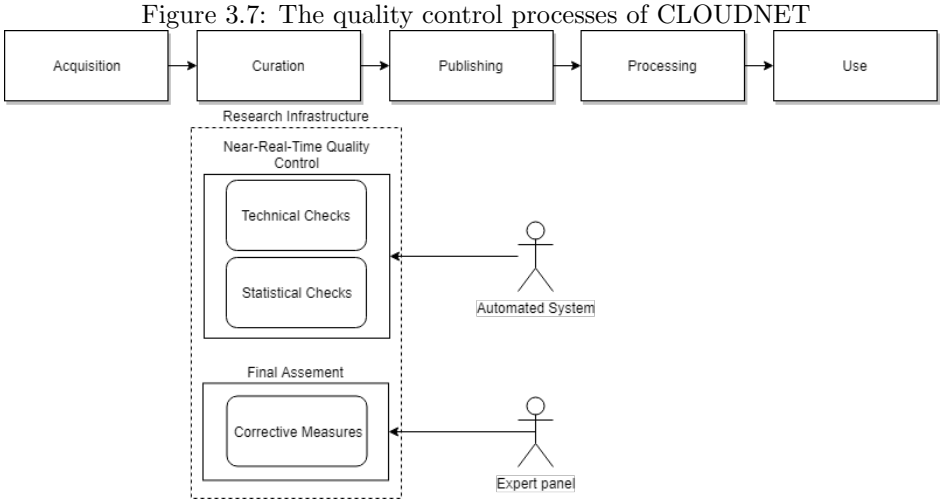


Figure 3.7: The quality control processes of CLOUDNET



3.6 EISCAT_3D

The EISCAT_3D[9] is a European environmental research infrastructure, and its initial concept was conceived in 2005. The project focuses on the Atmospheric domain and consist of 3 stations located in Norway, Sweden, and Finland. The RI is aimed at investigating how the Earth's atmosphere is coupled to space using the incoherent scatter technique. The systems used to record the data are multiphase array radar system located at each of the stations. The multiphase arrays consist of about 10.000 simple antennas.

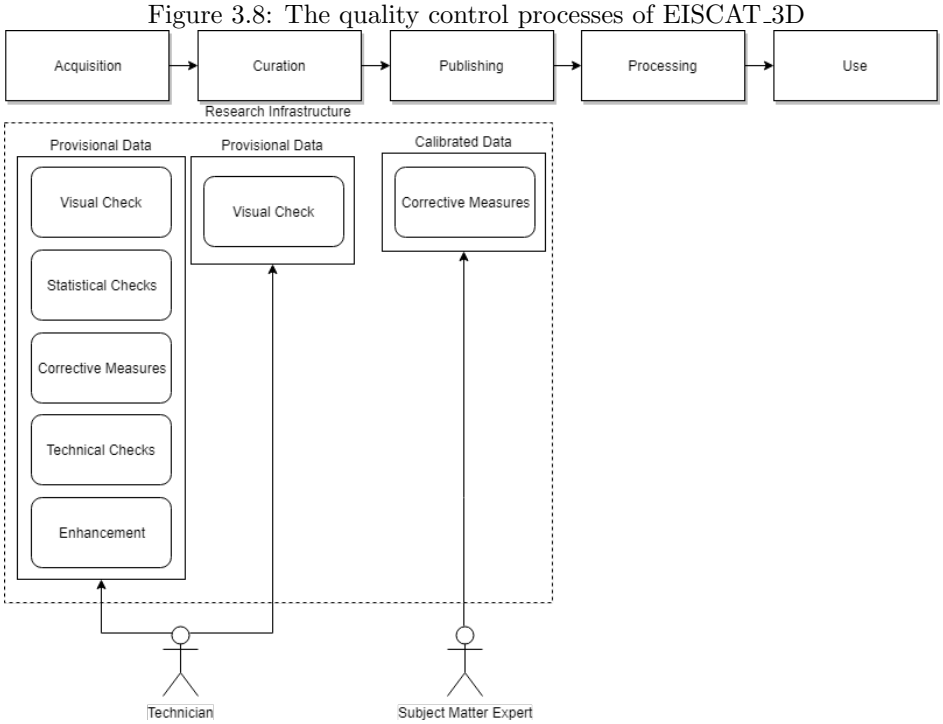
The quality control process within EISCAT_3D consists of a mix of semi-automated and manual checks. The process is described below, and the procedures are shown in figure 3.8.

- The first check is a real-time visual inspection done by the technician monitoring the system.
- The second check is a non-real-time visual check using automatically generated Error bands.
- In the last check, a domain expert calibrates the data using reference measurements.

[9] https://eiscat3d.se/

The measurement is released initially as provisional data within minutes of being recorded. The calibrated data is made available about a month after the provisional release.

Figure 3.8: The quality control processes of EISCAT_3D



3.7 ICOS

The ICOS[10] is a European environmental research infrastructure founded in 2008 to provide data on greenhouse gas concentrations. ICOS RI has more than 100 measurement stations in 12 European countries. The data collected by the RI is estimated to be more than 25TB per year. Moreover, the data is collected within three domains of environmental sciences. These domains comprise Atmospheric, Oceans, and Ecosystems. Each domain has its thematic data center that coordinates the activities within the data lifecycle.

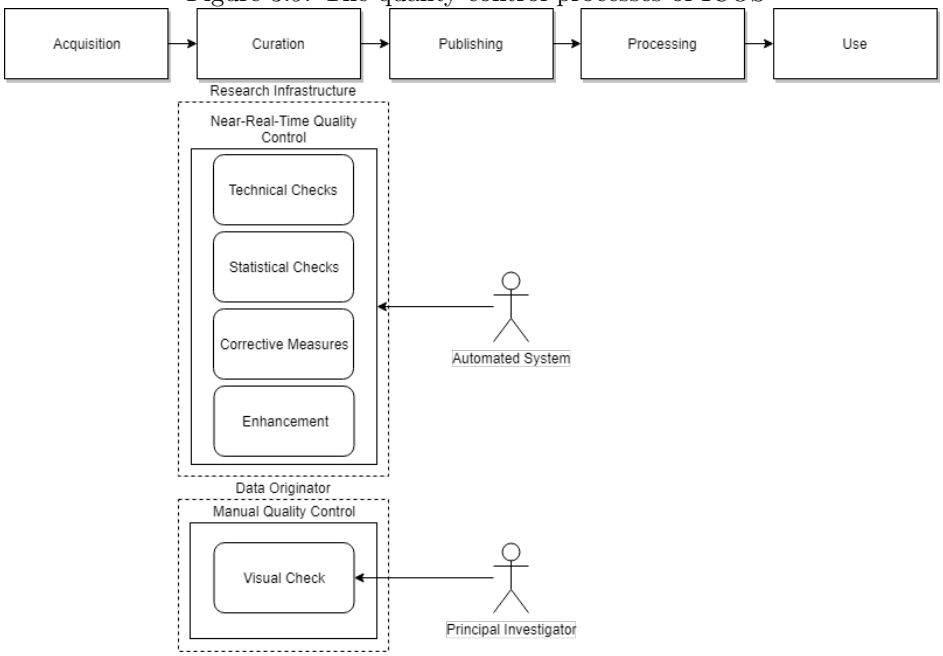
The thematic portals have developed specific quality control procedures for their respective domain. Nonetheless, the quality control procedures can be roughly divided into two types namely Near Real Time (NRT), and regular quality controlled data. The processes are described below, and the procedures are shown in figure 3.9.

- Firstly, automated tests determine the technical and statistical properties of the data. Secondly, processing, calibration, and other corrective measures are applied. Thirdly, depending on the domain, input from experts and station personnel are included. This process is called Near Real Time Quality Control.
- The normal quality control begins with the output from the automated tests performed in the NRT QC. The Principle Investigator investigates any suspicious flags and does a visual inspection of the data.

The NRT quality controlled data availability depends on the thematic portal. The Atmospheric center publishes NRT data within 38 to 48 hours of recording the measurements. The Ecosystem center takes about 10 to 30 days. Finally, the Ocean center takes between one and two months to publish its data. The higher level of quality control data is available after the regular quality control which takes a few months.

[10] https://www.icos-ri.eu/

Figure 3.9: The quality control processes of ICOS



3.8 Commonalities

The quality control processes have some common components which appear in most RIs. Many RIs use simple tests in the initial stages of quality control to catch glaring faults in the data. Identifying gross errors through the use of technical and statistical checks can reduce the noise and the amount of data that requires human inspection. Another commonality is the use of automation of different tests to improve the consistency and speed. However, the actual level of automation varies significantly between RIs. Most RIs release a Near Real Time version of their data. The dissemination of this data always occurs after quality control and usually does not include manual quality control. Nevertheless, the amount of quality control depends on the RI. An essential part of quality control within RIs is the inclusion of manual quality control. While the automated test can catch the gross errors, they can miss subtle mistakes which require more scientific insight which is much harder to include in a computerized system.

3.9 Diversities

In the above paragraph, we described the commonalities, but there are also differences in the quality control processes. The time required to release a quality controlled product depends mainly on the velocity and volume of the data type. Certain data types generate a few gigabytes per year while others can create petabytes. The amount of data can be a defining factor in the amount of automated and human quality control that is applied. RIs with larger data types tend to have more automated checks and limited human inspection. Furthermore, the number of sensors which require individual calibrations can limit the amount of quality control by the RI. Another difference among RIs is which organization does the quality control. The quality control can be done either by the data originator, RI or a combination of the two. Moreover, the position in the data lifecycle where the quality control is applied can also vary depending on the RI. While some RIs provide tools and services to aid in the quality control process, many data originators use their own means and methods to do quality control.

3.10 Summary

We examined the different RIs to understand better the quality control processes approaches and the challenges they face. As described above, RIs have some similarities and differences. In general, most RIs incorporate automation to improve the consistency and speed of their workflows. This automation varies from RI to RI and sometimes even within the same RI. For example, ACTRIS has a different workflow for each of its data types. This heterogeneity within the same RI makes it harder to combine different datasets from the same RI. Moreover, some RI develop tools which are used by their community. However, this is not always the case, which leads data originators to develop ad-hoc tools. In essence, reinventing the wheel to solve a problem which may have been solved by other data originators or other RI within the same domain. A barrier in communication is the terminology used to describe elements in the quality control process of each domain. An example of this is the name delayed-mode quality control, human intervention, human quality control, and manual quality control. All four terms refer to inspection of the data by a person in non-real-time. This difference in naming the same procedure can cause difficulty in sharing tools and workflows which might otherwise be usable by other domains.

In this work, we focus on the terminology aspect of these problems. In the further chapters, we develop and use a taxonomy to classify the components of quality control and provide a common vocabulary.

Chapter 4

Taxonomy

In this chapter, we describe the concepts and relations of the taxonomy for quality control.

4.1 A taxonomy for quality control

The taxonomy described below is intended to classify the different elements of a quality control process. We used concepts encountered in literature and our analysis of the RIs. The figure 4.1 shows a graphical representation of the taxonomy and figure 4.2 shows the tools branch.

4.1.1 Automation levels

Automation level The degree to which actions can be performed automatically.

This concept is used to indicate to what degree processes, tools or methods can function or be performed automatically. Automation is an essential aspect of quality control. Therefore, an RI will attempt to automate as much of the process as possible. However, this is not always feasible, and the level of automation of the same process may differ between RIs.

Automated The action is performed without external actors and fully automatic.

The automated level refers to tools or methods that do not require an external actor to initiate, operate or process the results. This concept is the goal of most RI because it can improve the speed and consistency of the quality control process.

Manual The action needs a human actor to perform.

While total automation might be a goal, it is not always feasible. Almost all quality control processes have manual elements. A method can be manual because it is not possible to automate them. However, particular manual methods produce better results.

Semi-automated The action is performed automatically but requires a human actor to initiate or process the results.

The automated level refers to tools or methods that require an external actor to initiate or process the results. This concept is used when total automation is not desirable or possible.

4.1.2 Benchmarks

Benchmark The comparison and evaluation of the quality produced by a quality control methodology.

A benchmark is used to determine the quality being produced by a given quality control methodology. Based on the results improvements can be made to the quality control process. Furthermore, a benchmark can be used by an RI internally or by a community within the same domain. A benchmark can be used to select the best performing quality control practices.

4.1.3 Data quality

Data quality The measure of the excellence of the data.

The abstract concept of quality refers to the measure of excellence. The measure itself varies between RIs and data consumers. In general, quality is the concept that we try to determine or improve using quality control.

Quality dimension A characteristic that can be measured to indicate quality.

The quality dimensions are universal in the sense that they can be applied to any domain. Examples of dimensions include accuracy, completeness, consistency, and timeliness.

Quality level The specification of how much quality control and processing was applied to the data.

Much like data quality, a data quality level is RI dependant. The number of levels available and the amount of quality control varies. In general, the lowest quality levels indicate zero or minimal quality control. Likewise, the top quality levels indicate the highest quality data that is produced by a specific RI.

Quality metric A method to measure the value of a quality dimension.

A data quality metric can be used to detect problems in the data. Some metrics are universal while others are specific to a study, data type or domain. A quality metric is explicitly or implicitly connected to a quality dimension. Furthermore, a quality dimension can have one or more associated quality metric. RIs or data consumers can define metrics to determine the quality of data.

Quality issue A problem that can occur in the data.

A data quality issue can be the result of any number of circumstances. Some circumstances that can cause quality issues include instrumental malfunction or miscalibration. Quality issues usually occur during the data acquisition phase. A common problem is an outlier which can appear in any data type.

4.1.4 Methodologies

Methodology A combination of techniques, methods, and tools used by RIs to quality control data.

A methodology refers to the combination of techniques, methods, and tools used during the quality control process. A standard usually establishes a methodology to achieve a specific quality goal. However, a methodology also refers to ad-hoc approaches used by RIs. When it is believed to produce data of higher quality it can be elevated to a standard.

4.1.5 Processes

Process A series of actions or steps used or required to achieve a specific goal.

The concept of process refers to one or more actions used to achieve a specific goal. Many RIs establish ad-hoc procedures for quality control. A process can consist of one step or many. Furthermore, most processes need an actor to perform the steps necessary. A quality control process itself usually comprises multiple processes.

Atomic process A single step or action required to achieve a specific goal.

An atomic process consists of one step or action. Examples of such activities include range checking, calibration, and outlier detection. The quality control in RI rarely share the same multipart processes but many times share the same atomic processes.

Multipart process A series of steps or actions required to achieve a specific goal.

The multipart process is used to describe the different phases in an RI's quality control process. An example of a multipart process is the Near Real-Time quality control. In this process, multiple steps are performed autonomously by the system to check the quality and annotate the data.

4.1.6 Roles

Role The position or function held by someone or something in the quality control process.

Within any process, there are actors required to perform actions. Within a quality control process, an actor can have different job titles, experience working with systems, and theoretical background. An actor is not limited to people. Any (semi-)autonomous system can also be considered an actor. As such, we define roles to designate the class of actor used or required to perform a particular action.

Domain expert A person with the theoretical background related to the research.

A domain expert also referred to as a subject matter expert is usually the principal investigator. This person has the theoretical knowledge and deep understanding to catch the subtle errors which occur during the experiments. Another trait of a domain expert is local knowledge. This knowledge is used to distinguish anomalies from real extreme events.

System Any non-human actor that performs actions (semi-)autonomously.

Most QC workflows include some degree of automation. Whether it be a script written ad-hoc by a scientist to clean and process her sensor data automatically during collection or a central automated system which is used to quality control the data by the whole RI during submission. These systems perform a task (semi-)autonomously.

Technician A person with training in the monitoring, operation, engineering, and maintenance of the processes used for quality control.

The role of a technician is given to a person that has training in the various aspects of running the quality control process. An example of a technician is a staff member at a measurement site that is monitoring the charts and flags any gross errors which could be the result of instrument malfunction.

4.1.7 Standards

Standard A method, technique or requirement that is regulated or widely accepted as a credible means to perform quality control.

In general, an RI establishes a body within the organization to develop quality control standards. These standards range from internationally accepted standards to guidelines specific to the study or sensors used. Furthermore, they usually appear in the form of a document that is published and made available to practitioners within a particular domain.

Domain standard A methodology that applies to specific domains or data types.

The domain-specific standards concentrate on a specific field. These standards focus on what is being measured and how to best deal with the unique issues that arise within a particular area. An example of these standards are manuals that provide the best type of checks that need to be applied. However, they defer the concrete implementation to practitioners.

General standard A methodology appropriate for multiple domains or data types.

The general standards focus on the general aspects of the quality control process. An example of a universal aspect of quality control within the environmental domain is the spatiotemporal aspect. A general standard would prescribe the best practices to deal with issues related to spatiotemporal data. In general, these standards do not focus on a specific measurement or study. For this reason, these standards apply to multiple fields.

Measurement standard A methodology that is specific to an experiment, instrument or study.

A measurement standard has a narrow focus on a specific study, instrument or parameter. These standards are usually RI specific. An example is the standard operating procedure document. This document lays out the steps, equipment, calibrations, and settings required for a specific measurement.

4.1.8 Temporalities

Temporality How an operation is performed in relation to time.

Many processes, tools, and techniques depend on the concept of temporality. RIs use this concept to distinguish the time required to processes data.

Near real-time An operation that occurs soon after the data is collected.

Near real-time operations can be more complicated than real-time operations because they collect more data points. Particular methods work best after enough data has been collected. The time span considered Near real-time varies significantly between RIs and datatypes.

Non-real-time An operation that occurs long after the data is collected.

Non-real-time operations are actions that occur long after the data is collected. In RIs these include the production of higher-level data products. The benefit of non-real-time is the amount of data that can be processed which allows very subtle and systematic errors to be detected.

Real-time An operation that occurs immediately after the data is collected.

Real-time operations are often associated with a single data point or dataset. Most checks in the initial stages of quality control are applied in real-time. The time span considered real-time varies depending on the RI.

4.1.9 Tools

Tool A functionality provided by a software system used to assist in or perform an action.

Tools can take many forms. They can be desktop applications, web applications, scripts or files. A software system usually implements one or more tools. Tools can have different temporalities and automation levels. These tools are used during the various stages of the quality control process. An RI develops tools to help their community improve the quality control efforts. Nevertheless, specific tools can be shared between RI or domains.

Data analytics tool This category classifies tools which have the primary purpose of giving the user some insight into the data.

Analysis tool This class of tools provides insight into the characteristics of the data.

Collection analysis tool A collection analysis tool compares the measurement to a large collection of data points.

Measurement analysis tool A measurement analysis tool looks at the characteristics of the given data point.

Stream analysis tool A stream analysis tool compares a data point to adjacent data points.

Visualization tool This class of tools provides a graphical representation of the data.

Graph visualization tool These tools show the data in a graph.

Map visualization tool These tools plot spatial data using any map.

Plot visualization tool These tools plot data points on a diagram.

Data contextualization tool This category classifies tools which have the primary purpose of giving context to the data and circumstances around the stewardship of the data.

Annotation tool This class of tools helps to add more information about the data.

Flagging tool These items can add flags to each data point.

Metadata tool This class of tools gives information about the circumstances in which the data was collected.

Metadata authoring tool These tools help create or modify metadata.

Metadata conformance tool These tools check the metadata for conformance to a particular specification.

Provenance tool This class of tools gives information about the steps taken to produce the data.

Provenance authoring tool This item allows the documenting of the steps applied to the data.

Provenance document tool These items comprise any format to store the provenance.

Data processing tool This category classifies tools which have the primary purpose of transforming and shaping the data for the next step in the quality control process.

Interoperability tool This class of tools takes data from one environment and transfers them to another environment.

Conversion tool A conversion tool takes data from one format to another.

Export tool An export tool is used to extract data within a system to a specific type of data.

Import tool An import tool is used to load a specific type of data into a system.

Processing tool This class of tools can transform data.

Storage tool This class of tools includes items that allow the user to store data during the quality control process.

Database tool A database to store and query the data.

File Format tool A file-based system to store and query the data

Workflow tool This class of tools helps to automate different tasks and processes.

Workflow design tool This class of tools enables the creation of workflows.

Workflow execution tool This class of tools can execute an existing workflow.

4.2 Relations

We identified specific relations between the concepts of the taxonomy. These relations show how the different concepts can interact and be used together to classify elements in the quality control process. Figure 4.3 depicts a graphical representations of the relations between concepts.

- The ***Methodology*** consists of one or more ***Processes***, ***Roles***, and ***Tools***.
- A ***Process*** can be part of a ***Process***
- A ***Process*** has an ***Automation level***
- A ***Process*** has a ***Temporality***
- A ***Tool*** has an ***Automation level***
- A ***Tool*** has a ***Temporality***
- A ***Tool*** can be part of a ***Process***
- A ***Role*** can use ***Tool***
- A ***Role*** can perform a ***Process***

Figure 4.1: A graphical representation of the taxonomy for quality control

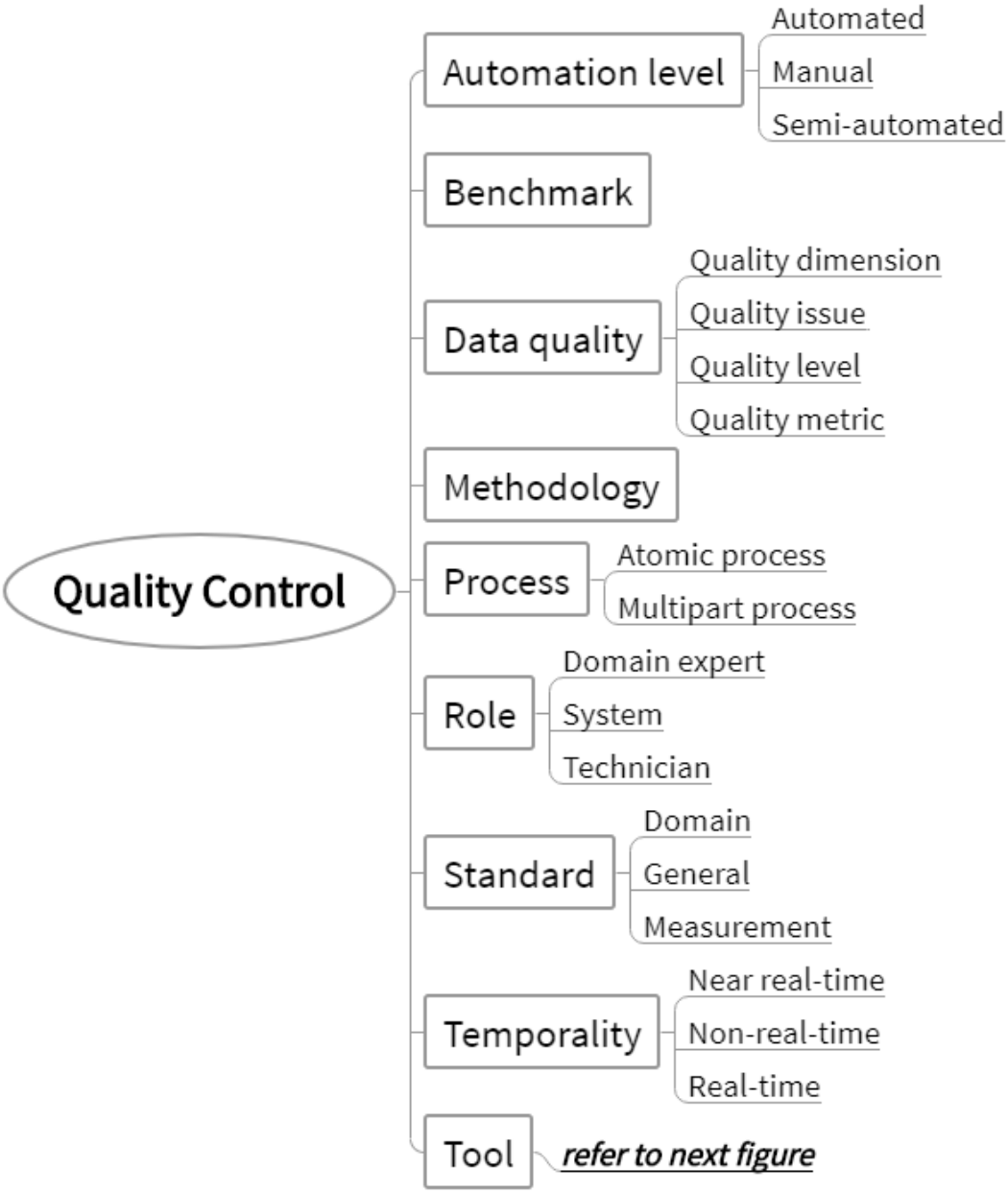


Figure 4.2: A graphical representation of the taxonomy for quality control tool branch

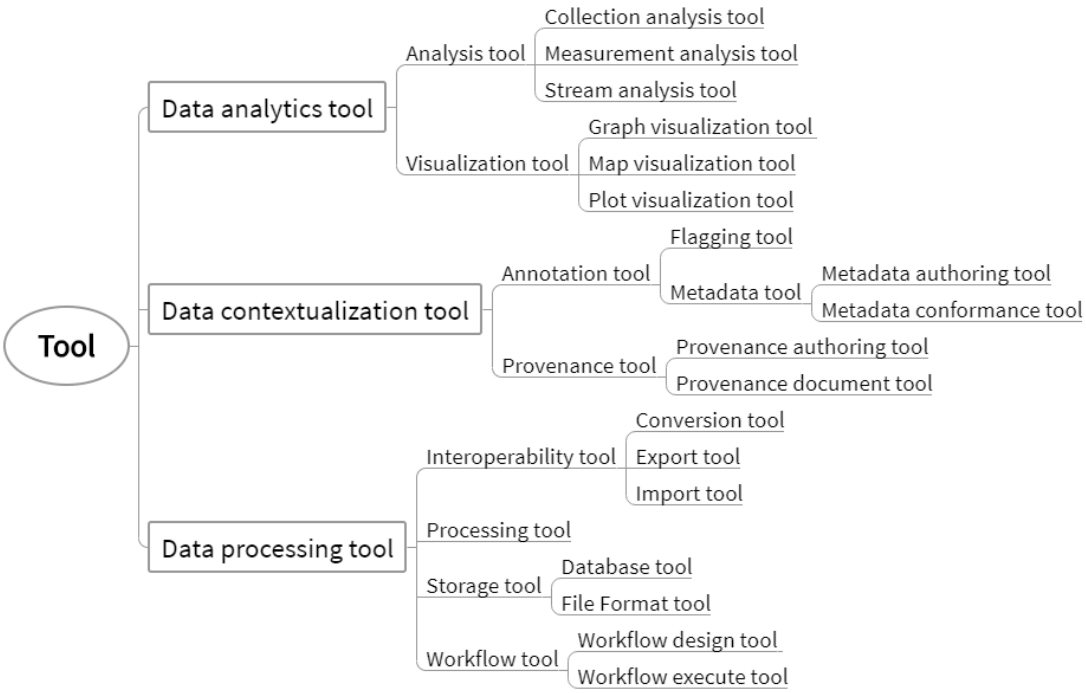
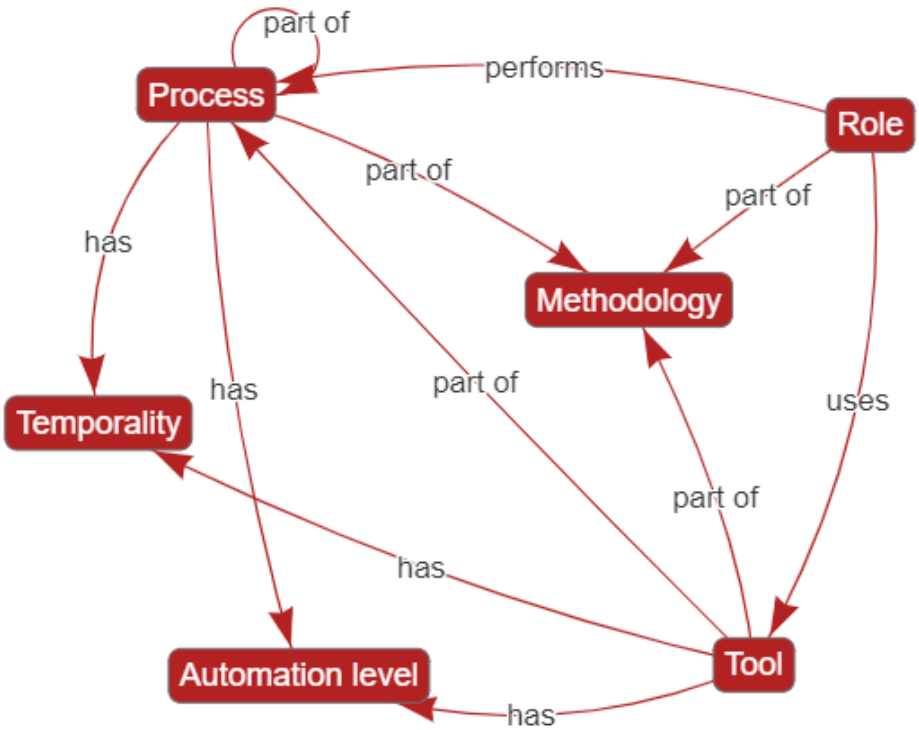


Figure 4.3: A graphical representation of the relations between concepts



Chapter 5

Harmonized Processes

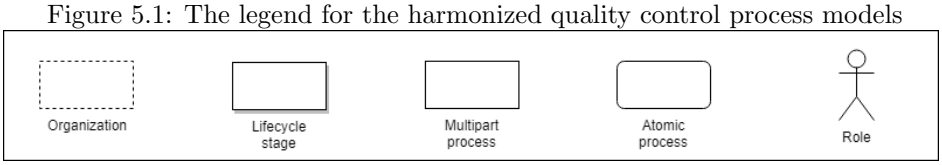
In this chapter, we revisit the quality control processes modeled in chapter 3 and attempt to harmonize them using our proposed taxonomy.

5.1 Introduction

In the sections below we describe how the taxonomy can be applied to the models previously created in chapter 3, to harmonize them. The goal of this harmonization is to make it easier to compare and reason about the different models.

In order to understand the updated models, we present an updated legend that incorporates concepts from our taxonomy. The legend replaces the elements Phases and Process with Multipart processes and Atomic processes respectively. Furthermore, it replaces Actor with Role.

Figure 5.1: The legend for the harmonized quality control process models

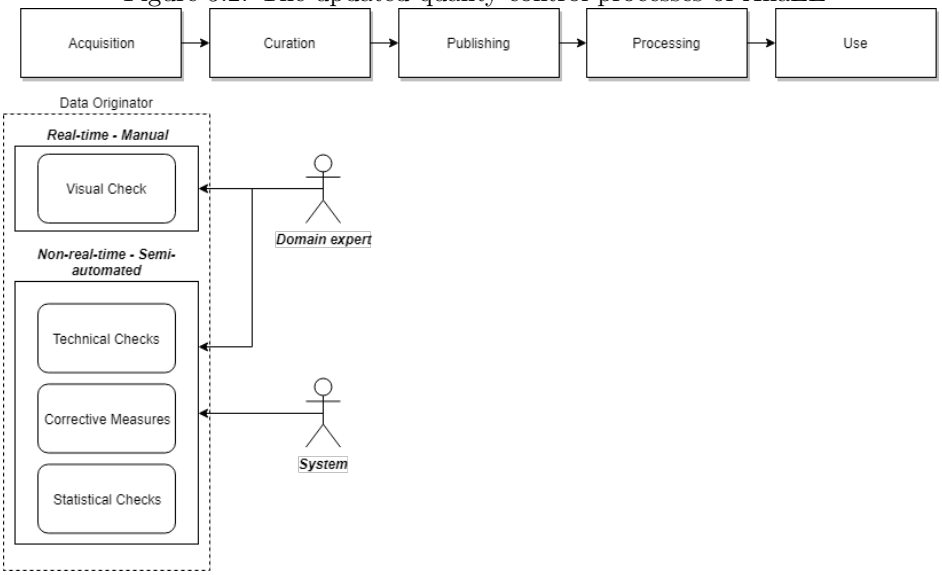


5.2 Updated Models

5.2.1 AnaEE

In the updated model of the AnaEE quality control process, we replaced Principal investigator and Automated System with Domain expert and System, respectively. Furthermore, we changed to names of the Multipart processes to reflect their Temporality and Automation level. See figure 3.3 for the original model and 5.2 for the updated model.

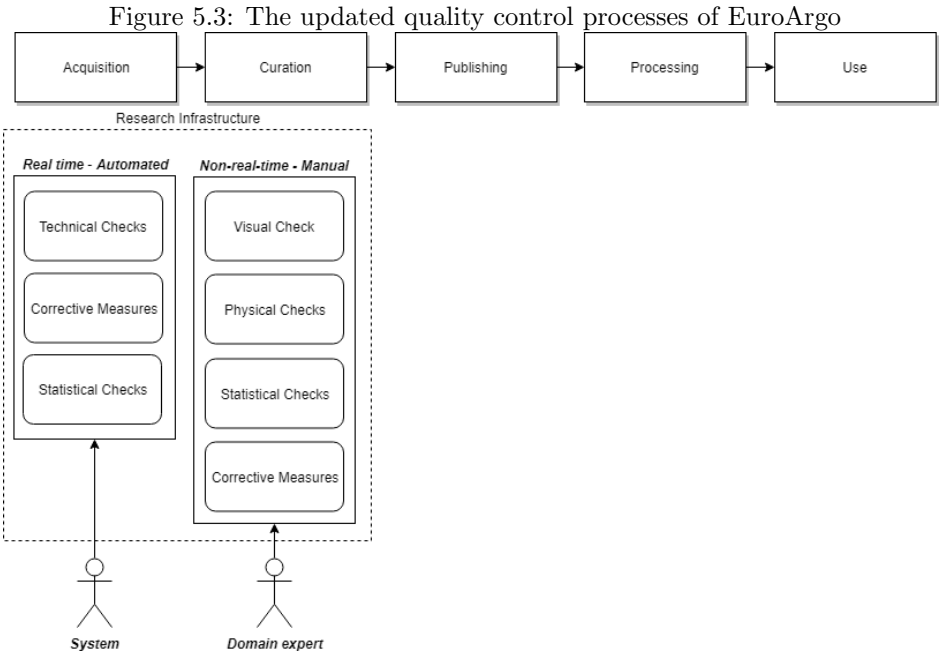
Figure 5.2: The updated quality control processes of AnaEE



5.2.2 EuroArgo

In the updated model of the EuroArgo quality control process, we replaced Delayed Mode Operator and Automated System with Domain expert and System, respectively. Furthermore, we changed to names of the Multipart processes to reflect their Temporality and Automation level. See figure 3.4 for the original model and 5.3 for the updated model.

Figure 5.3: The updated quality control processes of EuroArgo



5.2.3 ACTRIS

In the updated models of the ACTRIS quality control process, we replaced Expert panel, Principal investigator, and Data Curator with Domain expert. Moreover, The Automated System was replaced by the System role. Finally, we changed to names of the Multipart processes to reflect their Tempo-

rality and Automation level. See figures 5.4, 5.5, and 5.6 for the original models and 5.4, 5.5, and 5.6 for updated models.

Figure 5.4: The updated quality control processes of EBAS

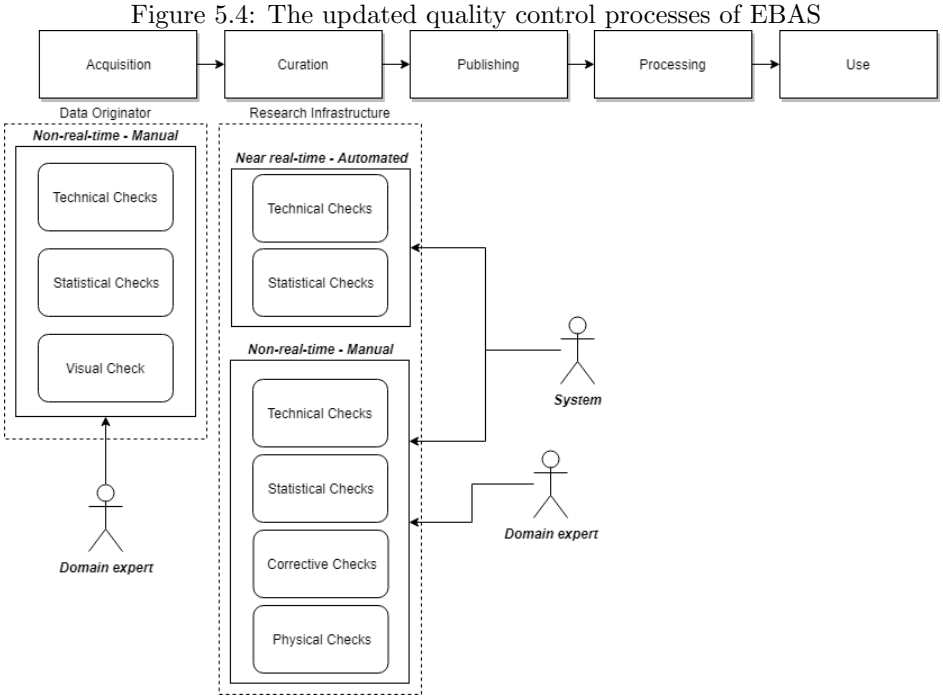


Figure 5.5: The updated quality control processes of EARLINET

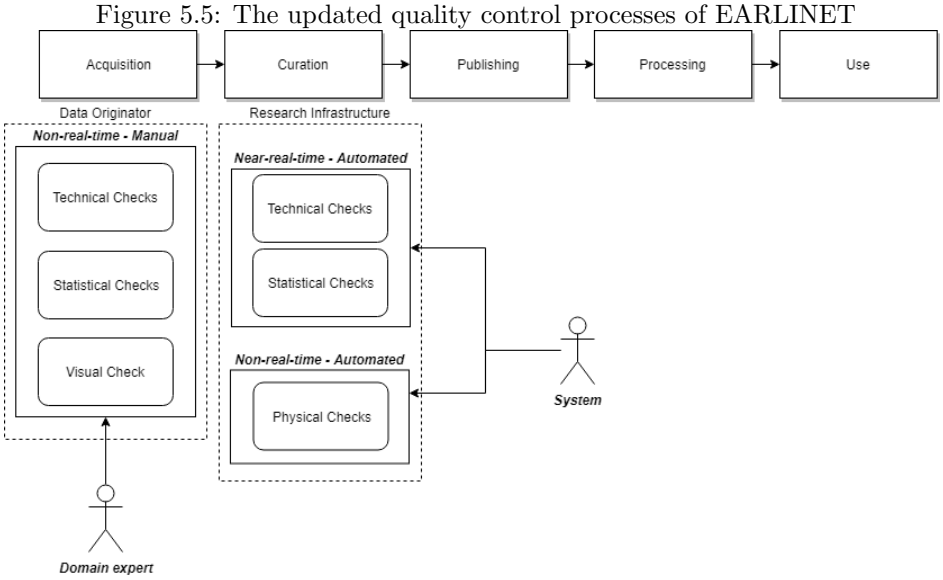
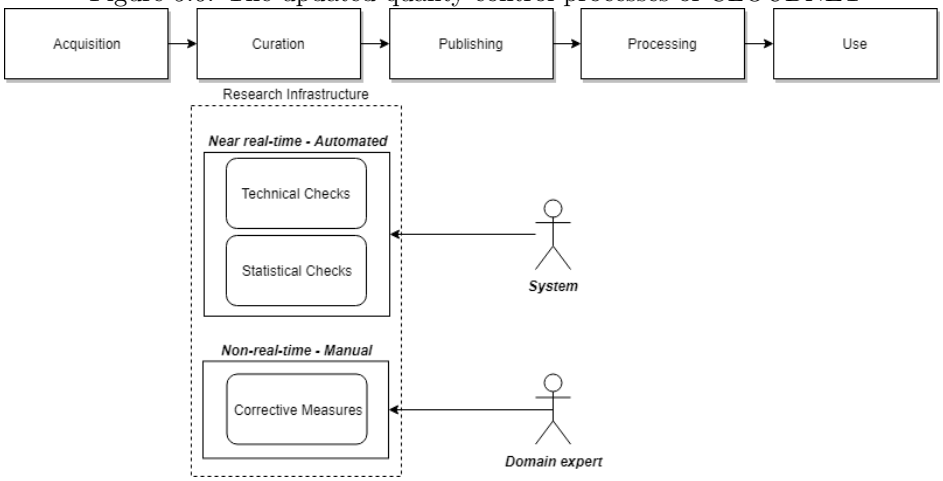


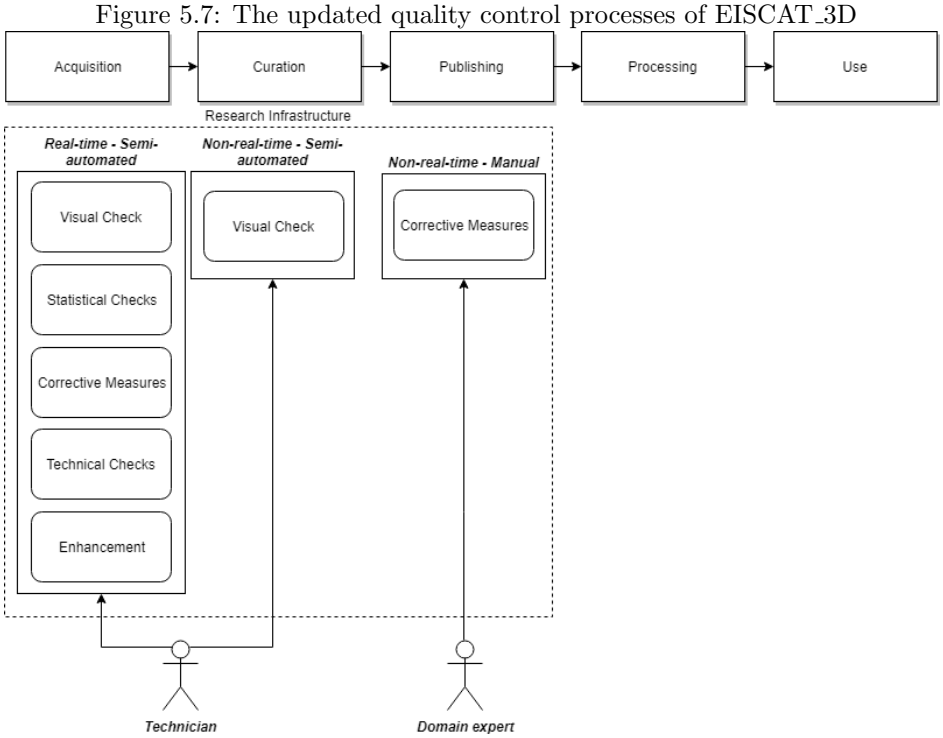
Figure 5.6: The updated quality control processes of CLOUDNET



5.2.4 EISCAT_3D

In the updated model of the EISCAT_3D quality control process, we replaced Subject Matter Expert with Domain expert. Furthermore, we changed to names of the Multipart processes to reflect their Temporality and Automation level. See figure 3.8 for the original model and 5.7 for the updated model.

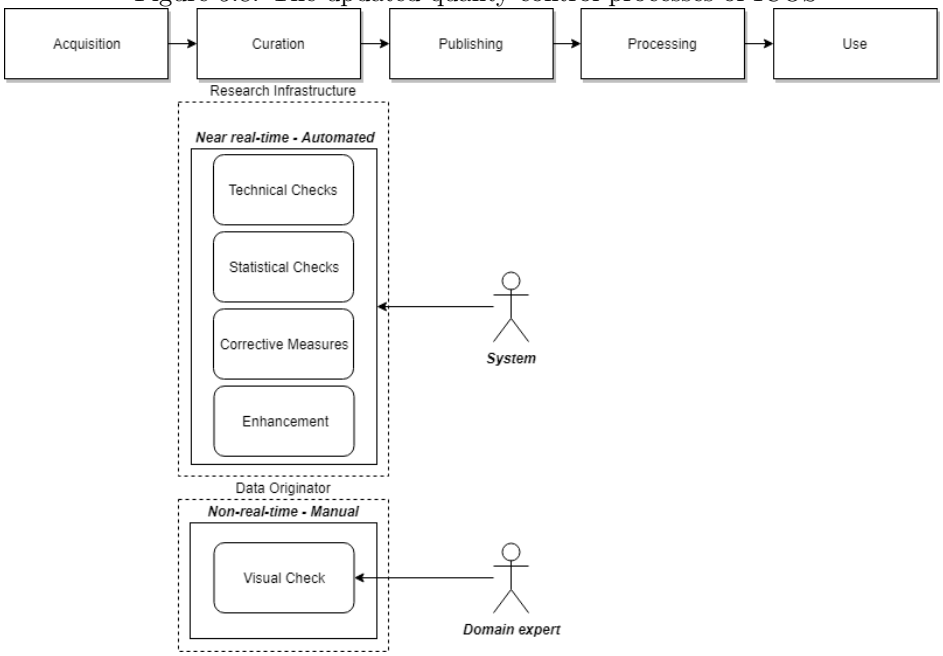
Figure 5.7: The updated quality control processes of EISCAT_3D



5.2.5 ICOS

In the updated model of the ICOS quality control process, we replaced Principal investigator and Automated System to Domain expert and System, respectively. Furthermore, we changed to names of the Multipart processes to reflect their Temporality and Automation level. See figure 3.9 for the original model and 5.8 for the updated model.

Figure 5.8: The updated quality control processes of ICOS

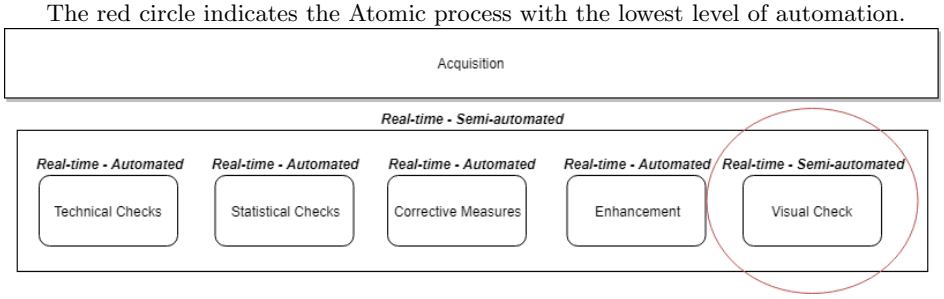


5.3 Further analysis

For further analysis, we use EISCAT_3D as an example. We can describe the Real-time Semi-automated Multipart process, which is implemented in the Acquisition phase, using our taxonomy to see what the Automation level and Temporalities are of its Atomic processes. In figure 5.9, we can see that for the most part, that the Atomic processes are Real-time and Automated. However, the Visual check is Semi-automated which makes the whole Multipart process classified as Semi-automated. Indeed, the Multipart process is classified by its lowest level of Automation and the slowest Temporality of its atomic processes.

Figure 5.9: The analysis of the Real-time Automated Multipart process in EISCAT3d

The red circle indicates the Atomic process with the lowest level of automation.



5.4 Summary

Using the taxonomy concepts, we see that the models become much more convenient to compare. The roles have been consolidated, and the updated models now provide more information on the Temporality and Automation level of a Multipart process. In literature, researchers have identified automation level as an essential method to improve the consistency and the speed of checking the quality of data [CRP+13, TL13, HVM+16]. By automating the processes where possible, the subjectivity is lowered or removed altogether. In the updated models, we can see which Atomic processes

determine the automation level and temporalities of the Multipart process. The terminology provided by the taxonomy supports the cross-domain analysis of processes. The harmonized models eliminate domain-specific terms which allow for a straightforward comparison of the different RI quality control methods.

Chapter 6

State of the art survey

In this chapter, we survey methodologies, standards, benchmarks, data quality, and tools in the environmental sciences.

6.1 Introduction

The state of the art survey is essential because it provides an overview of quality control research efforts and practices in the environmental sciences. This survey is structured using the concepts in the top level taxonomy. We focus on methodologies, standards, benchmarks, data quality, and tools because they are necessary for the recommendations to RIs. Additionally, the application of the taxonomy helps to identify weaknesses in the taxonomy structure.

6.2 Survey

6.2.1 Methodologies

Researchers in environmental sciences have identified quality control as a Multipart process. Each phase consists of multiple tests that identify potentially erroneous data and annotate them with flags.

Shafer et al. [SFA+00] describe a five-part quality control approach consisting of a mix of automated and manual quality control procedures.

- The first part consists of an automated sensibility check on each observation and automated procedure that confirms there aren't any gaps in the data.
- The second part consists of real-time monitoring of the data by staff, technicians or data manager to make sure that the data collected and the system performance is acceptable.
- The third part consists of a series of automated quality control checks that evaluates if the data contains subtle errors.
- The fourth part is performed by the data manager that goes over the results of part three and assess if any action needs to be taken.
- The fifth part consists of a monthly visual assessment of the data. This step detects sensor drifts and bias.

Cummings et al. [Cum11] describe a four-part process to automatically quality control oceanographic data.

- In the first part of quality control, they apply sensibility checks on each observation. If the observation fails any one of these checks, then the observation is removed from the dataset.

- The second part consists of complex quality control procedures where the observation is subjected to a series of tests. The cumulative score of these tests determine if the observation will be annotated with either accept, reject, or schedule for manual non-real-time assessment.
- An analysis of the system itself is performed in the third part. This stage will detect marginally acceptable data that passed the earlier stages of the quality control.
- The fourth part is designed to minimize the impact of assimilating observation on the model forecast error.

Taylor et al. [**TL13**] describe a three-part quality control approach consisting of automated and manual procedures and internal and external audits.

- The first part consists of automated quality control checks designed to annotate suspicious data.
- The second part consists of manual visual inspection to classify flagged data from the previous stage as either poor quality or high quality.
- The last part is a mixture of internal audits by the organization and external audits by the user community.

Abeysirigunawardena et al. [**AJM+15**] describe a three-part quality control process using a mix of automated and manual quality control.

- The first part consists of real-time automated quality control checks before the data is ingested into the database. These tests are designed to detect sensor malfunctions and erroneous data at a regional level.
- The second part consists of near real-time or automated non-real-time testing. These tests are applied to consecutive data points to determine the validity of the value.
- The last part consists of a manual review by a domain expert. These tests are designed to detect any data that has been mistakenly flagged as bad by the automated quality control procedures.

These methodologies provide insights into the quality control methods in different environmental domains outside of the RIs examined in chapter 3.

6.2.2 Standards

Domain standard

Quality Control of Biogeochemical Measurements The [JHR+18] is a domain standard for the quality control of biogeochemical data. The goal of this standard is to harmonize the quality control and quality assurance procedures related to 8 biogeochemical parameters. It focuses on the parameter and is sensor independent. The standard describes the use of quality flags various tests and techniques used to check the data depending on the parameter.

SeaDataNet standards for quality control The [Sea10] is a domain standard for the quality control of oceanographic data. This standard is based upon multiple other standards within the oceanographic domain. It focuses on the type of quality control test that can be performed regardless of the instrument or parameter. It describes a series of automated checks that focus on the technical aspects of the data. Furthermore, it also describes some scientific tests to further quality control data.

Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data The [KGS+13] is a domain standard for marking or flagging data. The quality flags are intended for use in the oceanography and marine meteorology domain. It describes a flagging scheme consisting of five flag values. QARTOD has adopted this standard as their recommended flagging scheme.

General standard

Best practices for sensor networks and sensor data management The [ESI] is a general standard for the establishing of sensor networks and sensor data management. It is available from the ESIP EnviroSensing Cluster community wiki. It contains a chapter on the best practices related to sensor data quality which introduces the concepts of QA and QC on data streams. Furthermore, it explains other ideas related to data quality qualifiers and data levels.

ISO 9001:2015 The ISO 9001:2015 [ISO15b] is a general standard for quality management systems. It provides a set of principles and framework that applies to any organization. This systematic approach to quality management has been applied by more than 1 million organizations worldwide. An important principle of the standard is customer focus. A customer expects a certain level of quality to be provided by a product or service which in turn an organization has to deliver. Essentially, ISO 9001 sets some organizational and process requirements that an organization has to adhere to in regards to quality management. Moreover, organizations meeting those requirements are eligible for certification. Some organizations use this standard as a base for their quality management standards.

ISO 19157:2013 The ISO 19157:2013 [ISO15a] is a general standard for data quality in geographic information. It provides a set of principles and framework that applies to any organization using geographic data. The standard defines some quality dimensions which describe the spatial data quality. These quality dimensions include Completeness, Logical consistency, Positional accuracy, Thematic accuracy, Temporal quality, and Usability. Furthermore, it provides data quality metrics used to describe each of the quality dimensions for a given dataset.

Measurement standard

Real-Time Quality Control of In-situ Temperature and Salinity Data The [dat16] is a measurement standard that focuses on the real-time quality control of temperature and salinity data. It describes the flag scheme and a series of test to detect errors during in situ observations. Furthermore, it describes steps required for the quality assurance of the measurements such as sensor calibrations, sensor comparison, and a deployment checklist.

We make a distinction in our taxonomy between the different standard types. Using this distinction makes it easier to identify where there is a need for standards.

6.2.3 Benchmarks

GODAE Ocean Data Quality Control Intercomparison Project The [CBK+10] is a benchmark. A method employed by this benchmark is the comparison of the results of automated quality control processes to the outcomes of manual non-real-time quality control of oceanographic data. The project examines the practices of five oceanographic data centers. This project aims to identify the most effective quality control processes for oceanographic data. Additionally, the results can be used by the data center to evaluate their quality control processes.

Argo real-time quality control intercomparison In [WSH15], the authors describe the results of a benchmark of Argo float real-time quality control processes. The benchmark utilized Argo float data from 2007-2011 inclusive. The analysis included the data from four data centers. The results concluded that the real-time quality control processes were consistent in detecting bad data. However, the differences in the number of good profiles incorrectly flagged as bad were notable among the different systems, in this case, the Australian Bureau of Meteorology's real-time processes providing the best results.

AutoQC The AutoQC [IQu15] is a benchmark created by the IQuOD (International Quality Controlled Dataset) initiative. The project intends to implement automated quality control checks used by practitioners and determine the best combination of test to quality control a data set of known

quality. The aim is to find the combination of tests that can detect the incorrect data while not rejecting a lot of the correct data. Moreover, AutoQC can be used to benchmark novel approaches. The tests are written in Python and are made available under an opensource license.

Currently, the benchmark node does not contain any child nodes. Although, examining the different methods of benchmarking further could identify subclasses in this branch.

6.2.4 Data quality

Quality dimensions

Batini et al. [BBL+17] describes the use of seven quality dimensions to describe the quality of remote sensing data. These dimensions include Accuracy, Completeness, Redundancy, Readability, Accessibility, Consistency, and Trust of Sources. The authors explain that ideally, one should strive all of the quality dimensions. However, the cost and benefit will have to be taken into account.

Quality issues

Pastorello et al. [PAP+14] observed error patterns in time series data. The authors describe many common issues and their sources related to the quality control of carbon, water, energy fluxes, and micro-meteorological data. They propose 18 observational patterns grouped by seven classes to classify these data issues. The authors argue the use of multiple data types identified a broader range of data quality issues.

Quality levels

The quality levels [1][2][3] in most RI vary significantly depending on the data type and purpose. In all cases, there is a level 0. The level 0 indicates that there was no processing or quality control applied to the data set. This data is the raw output of the instrument. Further, the level 1 data would contain a single parameter with some (near) real-time automated quality control. The level 2 data would have received extensive manual and automated quality control. Finally, the level 3 data would have undergone an extensive scientific manual non-real-time quality control.

Quality metrics

Roarty et al. [RSK+12] propose an automated quality control process for High-Frequency radar measurements. In the paper, they describe the use of four metrics to evaluate the quality of the data. These quality metrics include average radial bearing, spectra merged count, radial count, and data latency. Furthermore, they describe a metric for the temporal and spatial coverage of a High-Frequency radar network.

Ringler et al. [RHH+15] describe the use of 18 quality control metrics related to the quality of seismic data. These metrics give the data consumers the ability to assess the quality of data for their purpose. The metrics include Availability, Gap count, Timing quality, Mass position, NLNM deviation, Dead channel, Station deviation, Event compare synthetic, Difference, Coherence, Event compare strong motion, Days since last calibration, Calibration error estimates, Data-earth tide synthetic, Relative orientation, P-wave particle motion, Event SNR, and Clip detection. The authors implemented 11 of the 18 metrics in their software system that can calculate them automatically.

The different data quality aspects surveyed help answer the questions related to the what(dimensions), why(issues), how(metrics) and when(levels) of quality control.

[1] https://otc.icos-cp.eu/data-levels-quality-access
[2] https://www.actris.eu/About/ACTRIS/ACTRISglossary.aspx
[3] https://eiscat3d.se/sites/default/files/EISCAT_3D_Sixth_User_Meeting/21_Enell.pdf

6.2.5 Tools

Ocean Data View

In 2002, Schlitzer [Sch02] introduced Ocean Data View. A freeware standalone desktop application developed in C for Windows and Linux operating systems created for use with oceanographic data. Because of its underlying generic principles, it can be applied to multiple domains. The primary focus of ODV is the data analytics. With five different modes, ODV helps the user explore the data visually. Moreover, it contains a dedicated view for quality control. Besides data analysis ODV also contains data contextualization tools which include metadata and flagging. These are used to annotate the data during the quality control. Finally, it provides numerous data processing tools to import, convert and export the data. ODV is very prominent in the oceanographic domain with a reported 50.000 users worldwide.

Quince

In 2016, Hellstrom et al. [HVM+16] described the quality control process used for the Near Real Time data flow in the ICOS RI. The three thematic portals have distinct processes. Notably, the Ocean Thematic portal was in the process of developing a comprehensive quality control system. The system would take the raw streaming data or non-real-time data and perform automatic quality control checks. Furthermore, it would provide tools to generate plots for monitoring and to allow users to perform manual non-real-time quality control. The benefits of such a system would be to automate and standardize the quality control applied within the ocean portal.

Single Calculus Chain

In 2015, DAmico et al. [DAB+15] introduced the Single Calculus Chain (SCC). The SCC is a web-based server application created for the automated quality control of Lidar data in EARLINET. It provides data handling and data analysis tools. It is composed of three modules. The first module pre-processes and applies instrument calibration and corrective measures to the data. The second module takes the input of the first module and employs optical processing algorithms. The last module coordinates the processes between the previous two modules. The SCC is part of the quality control process of the ACTRIS RI. It is intended to automate the quality control process of lidar data completely. Because of the heterogeneity of the network, it is essential to have a centralized quality control system. With SCC, EARLINET can assure that the quality control is applied in a standard way. Furthermore, it allows the near real-time dissemination of data within 30 minutes of measurement.

Baseliner

In 2016, Oishia et al.[OHO16] described Baseliner, an interactive MATLAB based desktop application to quality control data from thermal dissipation probes. The open source program was designed to assist the user in applying data processing, data contextualization, and data analytics in non-real-time. The system provides a user the ability to import raw measurements and apply automated quality control process based on user-defined parameters. Furthermore, the system includes provenance tools to track the changes applied to the raw data. The Baseliner software was designed to standardize the approach for quality control and processing. Moreover, its design and open source allow the user to extend the system easily.

ODM Tools

In 2015, Horsburgh et al. [HRJM15] introduced ODM Tools, an updated version of their quality control system. ODM Tools is a cross-platform, open-source desktop application developed in Python that allows the user to perform non-real-time quality control with provenance. It provides data analytics, data contextualization, and data processing. The unique selling point of ODM tools is its ability to encode the user's actions in the graphical user interface into a Python script workflow that can later be run on the data. That script is a provenance tool which allows edits to the data to be

tracked. They argue the use of both user interface and scripting caters to all scientists regardless of the level of programming skills.

SOCIB glider Toolbox

In 2015 Troupin et al. [TBH+15] introduced the SOCIB glider toolbox. This toolbox consists of multiple MATLAB scripts that provide the user with data analysis, data context, and data handling tools. The challenges of glider data are in the numerous types of file formats which require conversion before further processing. The focus of this toolbox is to provide a user-friendly way to set up a glider data processing chain. It produces three different levels of data which can be configured by the user depending on their needs. The system can perform in real-time and non-real-time modes.

IMOS Toolbox

In 2016, Hidas et al. [HPA+16] explain the data management infrastructure for Australia's Integrated Marine Observing System (IMOS). In this paper, they describe the use of a MATLAB program to process and quality control data. The IMOS Toolbox is a freely available open-source system that provides data analytics and data processing tools. The system takes various instrument-specific formats, it applies some automated quality control checks and allows for manual non-real-time quality control procedures. The result is a quality controlled netCDF file.

Data Quality Analyzer

In 2014, Ringler et al. [RHH+15] described the data quality analyzer (DQA). The DQA is an open source application for the quality control of seismic data. It provides the user with data analytics tools. The DQA has three main components. Firstly, a backend Java application called SEEDscan. This application calculates the data quality metrics. Secondly, is a Postgres database which stores the metrics for each data file from the different stations. Lastly, a web interface where the user can view the metrics and assess the data quality. At the time of writing the DQA implemented 11 of the 18 proposed quality metrics. The authors claim that the quality of data depends on the use which is best determined by the data consumer. The DQA makes that determination possible.

Our taxonomy can classify tools using fine-grain concepts. However, because of time constraints, we limited the classification in the above survey to the higher level concepts.

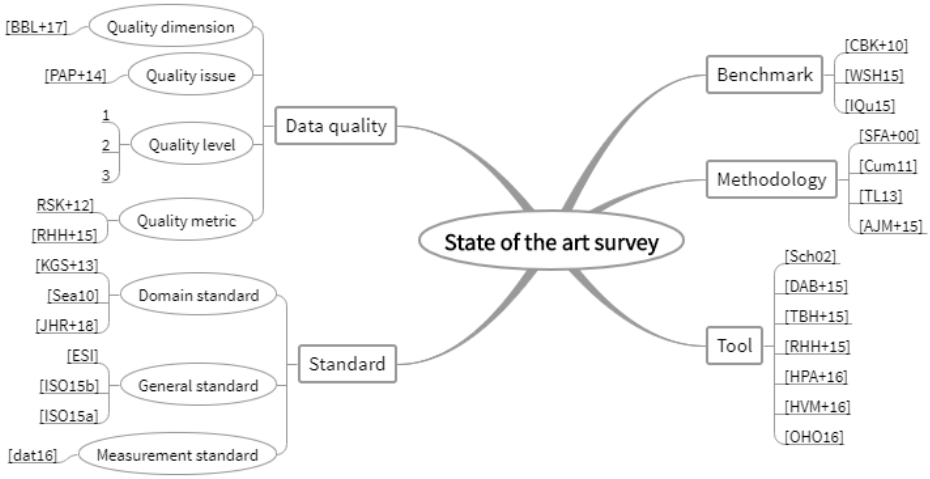
6.3 Summary

In table 6.1 and figure 6.1, we present a summary of the articles in the survey. We considered multiple items and selected at least one for each category. In some cases, we omitted articles with minimal information gain. For example, QARTOD published 11 measurement standards in total at the time of writing. However, we decided to include one example which is representative of the available manuals.

Table 6.1: Summary of state of the art survey

Classification	Citation
Benchmark	[CBK+10] [WSH15] [IQu15]
Methodology	[SFA+00] [Cum11] [TL13] [AJM+15]
Domain standard	[KGS+13] [Sea10] [JHR+18]
General standard	[ESI] [ISO15b] [ISO15a]
Measurement standard	[dat16]
Tool	[Sch02] [DAB+15] [TBH+15] [RHH+15] [HPA+16] [HVM+16] [OHO16]
Quality dimension	[BBL+17]
Quality issue	[PAP+14]
Quality level	1, 2, 3
Quality metric	[RSK+12] [RHH+15]

Figure 6.1: A tree view of state of the art survey



Chapter 7

Conclusions

In the previous chapters, we presented the results of the taxonomy building and survey. In this chapter, we conclude this thesis with a discussion of those results, answers to the research questions, recommendations, a conclusion of the work, and provide future work.

7.1 Discussion

The ENVRI-RM[ZMG+15] provides a general taxonomy for research infrastructures. However, it does not provide details of the quality control process within the data management lifecycle. It models quality control as a service of the data curation phase. Our taxonomy explores the quality control process in greater detail.

In chapter 5, we explored how the taxonomy can be applied to modeling the quality control process. We show which roles and organizations perform the quality control. Moreover, we show the composition of the different Multipart processes for each RI. Additionally, we used the ENVRI-RM data lifecycle model to illustrate where quality control processes are performed in the data management lifecycle. By providing a common vocabulary for the quality control processes, we can compare the models created across domains. Furthermore, the added insight on the automation level and the temporality of a process allows RIs to discover locations for improvements.

The tool branch focuses on atomic functionalities provided by a software system. This fine-grain classification supports the comparison of similar systems. For example, ODV and DQA both provide Visualization tools. However, ODV includes a Map visualization tool while DQA does not. Admittedly, these tools are for different domains and purposes. Another example is the comparison between ODV and ODM Tools. Both systems were initially designed for the quality control of oceanographic and hydrologic data. Both provide Data analytics tools, Data processing tools, Data contextualization tools. However, ODM Tools provides Workflow tools while ODV does not.

Because of time constraints, we did not publish the taxonomy using a formal representation. This formal representation would make it easier to release the taxonomy for use by the environmental community.

7.2 Achievements

RQ 1. How can we recommend an effective tool for controlling data quality in environmental science data management lifecycle?

We provide a useful method to recommend quality control methodologies, techniques, and tools. Initially, intended to classify tools in oceanography, we expanded the scope and were able to classify elements within multiple environmental domains successfully. However, a more systematic approach towards the selection of literature can contribute to the scientific rigor of the results. Furthermore, using techniques such as automatic keyword extraction and community consensus can reduce the subjectivity in the naming of the concepts of the taxonomy.

RQ 2. What are the quality factors that need to be controlled in the environmental data management lifecycle?

The answer to this question depends on the domain. In chapter 3, we examined the quality control processes of some representative research infrastructures. Most environmental domains have quality levels which denote that a dataset has undergone a predetermined amount of quality control. However, in seismology, quality metrics are used to determine the quality of a dataset using user defined parameters.

The nature of continuous monitoring makes the data unpredictable and difficult to quality control. Acceptable values can fall in a range which is determined by physical, temporal, and spatial factors. Furthermore, these factors are universal in the environmental sciences and can form the basis of any quality control process.

RQ 3. What are the quality control tools and methodologies used by the environmental community?

In chapter 6, we presented the results of the survey to answer this question. We surveyed the available environmental science quality control literature to determine what quality control tools and methodologies are currently available.

Initially, we focused on the oceanographic community which yielded good results. However, exploring other domains proved much harder. This difficulty can be attributed to the lack of knowledge outside of oceanography of the primary author.

Nevertheless, we managed to gather tools and methodologies from oceanography, ecosystems, seismology, and atmosphere. These items are representative of the environmental domain and were categorized together using the proposed quality control taxonomy.

RQ 4. What are the criteria for choosing a QC tool and what makes it a good QC tool?

Eight criteria emerged from the interviews (see A) with the representatives from the European research infrastructures for choosing a quality control tool. The criteria are Open-source, Level of support, Adoption rate, Maturity, Longevity, Functionality, Software Quality, and Programming language. At least, two are explicitly reinforced in the literature. In [HRJM15], Horsburgh et al. described providing functionality and developing open-source to foster adoption. Moreover, Oishia et al. [OHO16] state that the open-source design of Baseliner provides flexibility for adding more functionality. Interestingly, only one interviewee mentioned software quality as a criterion. It seems that creating quality control tools should include some requirement for the quality of the software responsible for performing quality control.

7.3 Recommendations

Based on the results of this research we can provide a recommendation for the development of quality control tools. An effective quality control strategy is the use of a centralized web application. This application can be used by the data providers to submit or manually quality control data. The application should support both the automated procedures and the manual steps required to do the quality control from the raw data. This kind of tool assures that all the data passes through the same quality control processes. This centralized tool provides the following benefits:

- Standardized methods that can be reapplied to all data when updated
- Can implement provenance mechanism in once place
- Speeds up the processing time by using the raw data
- Faster dissemination of data

Another recommendation is the use of the proposed taxonomy as a common vocabulary. By adopting a common vocabulary, research infrastructures can easily identify methodologies, tools, and tech-

niques that can be shared between them. This sharing will reduce the duplication of effort and support quality control methods by creating larger user communities.

7.4 Conclusion

In this thesis, we studied how we can recommend an effective tool for quality control. Researchers have identified the need for standardization within the quality control process of environmental science research infrastructures. There is no universal process for quality control because the checks required are dependent on the data type. As a result, many research infrastructures create quality control methodologies, tools and techniques ad hoc which leads to duplication of effort. Without standardization of the quality control process, no two datasets can be combined in any meaningful way.

We first explored the literature and conducted interviews to identify which methods are available and how to select the right one. We examined the concepts of data quality and data quality issues which is helpful in defining what the purpose is of a method. Furthermore, we discovered how and where the methods could be applied within the data management lifecycle. We proposed a taxonomy to be able to classify the various elements of the quality control process including an extensive branch that provides a fine-grained classification of tools. To validate, we examined state of the art quality control methods by applying the taxonomy.

Using the proposed taxonomy, we can classify the different elements within the environmental science quality control process. Additionally, the survey provides an overview of the available quality control methodologies, tools, and techniques.

7.5 Future work

In this thesis, we created a taxonomy with terms found in the literature. However, we did not take into account the subjectivity that would be introduced when choosing terms ad hoc. A research avenue would be to create an environmental science corpus for use with automatic keyword extraction methods. Moreover, the keywords produced by those methods can be validated by involving multiple practitioners from different environmental domains.

Because of time constraints, we did not encode the taxonomy into a formal representation. A possible next step in this research could be the encoding of the taxonomy in a formal notation such as SKOS Simple Knowledge Organization System[1] and the machine-readable Resource Description Framework (RDF)[2] format.

[1] https://www.w3.org/TR/2009/REC-skos-reference-20090818/
[2] https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

Appendix A

Interview transcripts

A.1 ACTRIS

The datasets include metadata describing the operating procedures, standards and quality assurance used by the instrument during collection. There are three types of data observed. The first is single parameter time series. These include properties of aerosol particles and trace gas in the atmosphere. Second, is cloud radar data, which is a height time series profile that shows the thickness of the clouds. The cloud radar data is the largest type of data by volume. The third one is LIDAR which produces the same kind of cloud radar data but in lower quantities. The last two data types require processing at the Radar location before transmission.

The stations apply quality control to the data before providing it to the data center. That quality control is not uniform across all stations because of resource constraints. To improve this LIDAR data is moving to a machine controlled quality control. Another initiative to improve the uniformity of quality control while still capturing the conditions present at the station is the implementation and use of a centralized business workflow tool where the stations can do the quality control.

The primary focus of the cloud radar is near-real-time dissemination as a validation dataset to do operational weather forecasting. The sheer volume of cloud radar data makes it impossible to manually quality control. The quality control is fully automated while providing adequate quality for use.

The software tools are mostly built in-house because of the specialized nature of the software and lack of comparable software tools on the market. There is a need for a man-in-the-middle workflow system which is not possible with Taverna. They are investigating the possibility of using a business workflow system.

The man-in-the-middle workflow allows a scientist to manually investigate the time series and apply special tools to quality control the data. They look at closure between parameters to detect trends.

The reason they put the effort in quality control and quality assurance is to avoid instrument drift. The aim is to detect drift in a ten-year time series. Some of the quality control checks performed are outlier checks, spike checks, threshold, closure checks and metadata consistency. A dataset is a collection of data that is internally homogeneous. The metadata can indicate a problem if the instruments suddenly changed. That would require representing the data as a new dataset. The user knows that he can use the datasets together but needs to pay attention that there might be a rupture between them.

ACTRIS has different data levels. The quality checks happen at different places. The first data level is zero which is annotated raw data coming from the instrument. The level zero data is either manually or automatically quality controlled. Some outlier and threshold checks are being down manually by a human or scientist. The next data level is one which only contains the final quality controlled parameter and has the original temporal resolution. Data level two only includes hourly averages. The next quality control is during import into the archive where the data is checked if it is consistent with data already in the archive.

The quality checks have to be calibrated per station. The current issue in quality control is the lack of uniformity.

The criteria for selecting a tool is the functionality, longevity, the maturity, the level of support available, the adoption rate and the programming language. In this case, python would have the advantage.

A.2 EPOS

One area that started formalized, standardized quality control is on the basic observation (waveform data from thousands of seismographs). There is an international federation, the FDSN, The International Federation of Digital Seismograph Networks, that looks at standards and best practices. Within that federation, there are regional organizations, e.g., IRIS from us and ORFEUS in Europe. A few years ago they started with the development of quality metrics the one could apply as processing on the incoming data, to look for specific quality parameters that have to do with the usefulness of the waveform. These include items like the number of spikes, gaps, noise level and overall completeness of data sets.

These are a relatively standardized set of parameters that are added as a metadata layer on the data. These parameters can be used to query the data depending on the use-case. There is currently an interface under development to do on the fly quality processing based on user-defined quality parameters.

That makes the search and discovery process much more efficient. Before, one would have to download the data and throw two-thirds of it away for whatever reason.

Another area which includes implicit quality assessment is the packaging of earthquake hazard assessment. That is where all the knowledge about the performance of earthquakes and their effects into a package and calculate the probabilities of shaking for a given region, time and shaking strength. Multiple experts review and each opinion of every viable expert, in theory, have been accounted for.

However, this is not a real quality control because of the significant uncertainties and the widespread of potential configurations and outcomes. These are the only area with some standardization. The rest is just quality assessment by peer comments and it's an evolutionary process. There are a couple of products which are produced in a standard way for more than a decade because the method of producing them proved to be of high quality concerning usefulness and content.

Both methods are applied to the same place in the data management lifecycle. It can be used as a cross-check to evaluate if an instrument is producing bad data. Even if its bad data, data is never thrown away.

The calculation method is to capture all the branches of the logic tree in the provenance and keeping that provenance with the product so that people can evaluate and reason about the usefulness of the quality. That is an implicit quality description. That is an abstract description of how the product came together. There are standardized interfaces for particular data for a specific purpose that come in a specific configuration. However, scientists do not want to standardize software because it takes away the freedom of doing something fun with the algorithm.

The way standards are created is that there are five ways of doing something and 80% of people do it in 1 way. Then why not agree to all do it in that way. That is a kind of natural evolution, and the international body tries to look after it. The standard could be a standardized algorithm, software implementation or a specific piece of code. However, using a specific could lead to platform issues. However, having five different software that relies on proper coding of the method and the proper interfaces need to check that comes out is what should be done.

There are no guidelines on what is the workflow for seismology or the software stack. It is institution or community dependent. Though, most everybody uses Linux.

There isn't a real need for cloud. There are some needs of HPC applications, but most processing can be done on a mini cluster in the basement. The usage profile of cloud is not quite there yet.

The amount of data is overall in the petabytes, but the sum that is needed for specific scientific purposes is a couple of tens of terabytes. The scalability would be excellent.

It is essential when planning the infrastructure and budget to have a reliable cloud that will be here for the next seven years. EUDAT has come along providing software as a service that is useful while requiring with minimal configuration.

There isn't a need for using a commercial cloud, because it is just mostly for storage. Some large

agencies use the cloud for scaling up their web appearance.

In seismology, there are very rapid peaks when there is a massive earthquake many people want to access the website, and that is difficult to maintain on internal infrastructure, while the system is idling around 99.5% of the time. When you want 24/7 99.9% availability cloud becomes expensive.

The recommendation of algorithms may be trivial especially since there may be multiple algorithms that produce more or less the same result.

People might follow the recommendations because when they start from scratch, they have no clue. However, when the recommendation is questioned, they might realize its probably arbitrary.

A point of interest is the quality metrics of software programs. Where tools automatically evaluate the source reports specific quality metrics. That is from a managerial point of view interesting when selecting application critical software. Where you start out with software that has an agreed quality to it. A way to prevent software vendors from going closed source in an attempt to make more money on the code, only open-source software development is sponsored.

Opensource is not always cheaper because there is always a component of maintenance. Price is a fake issue. However, for the principles of transparency control, reproducibility and extensibility opensource is much easier.

Bibliography

[AJM+15] Dilumie Abeysirigunawardena, Marlene Jeffries, Michael G. Morley, Alice O.V. Bui, and Maia Hoeberechts. Data quality control and quality assurance practices for Ocean Networks Canada observatories. pages 1–8. IEEE, October 2015. URL: http://ieeexplore.ieee.org/document/7404600/, doi:10.23919/OCEANS.2015.7404600.

[Bai94] Kenneth D. Bailey. *Typologies and taxonomies: an introduction to classification techniques*. Number 07-102 in Quantitative applications in the social sciences. Sage Publications, Thousand Oaks, Calif, 1994.

[BBL+17] Carlo Batini, Thomas Blaschke, Stefan Lang, Florian Albrecht, HM Abdulmutalib, Á Barsi, György Szabo, and Zs Kugler. Data quality in remote sensing. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 42, 2017.

[BCFM09] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41:16:1–16:52, 2009.

[CBK+10] James Cummings, Gary Brassington, Robert Keeley, Matthew Martin, and Thierry Carval. Godae ocean data quality control intercomparison project. *Proceedings of OceanObs 09: Sustained Ocean Observations and Information for Society*, 2010.

[CRP+13] John L. Campbell, Lindsey E. Rustad, John H. Porter, Jeffrey R. Taylor, Ethan W. Dereszynski, James B. Shanley, Corinna Gries, Donald L. Henshaw, M. Esther Martín, Wade Sheldon, and Emery R. Boose. Quantity is nothing without quality : Automated qa / qc for streaming environmental sensor data. 2013.

[Cum11] James A Cummings. Ocean data quality control. In *Operational oceanography in the 21st Century*, pages 91–121. Springer, 2011.

[DAB+15] G D'Amico, A Amodeo, H Baars, I Binietoglou, V Freudenthaler, I Mattis, U Wandinger, and G Pappalardo. Earlinet single calculus chain–overview on methodology and strategy. *Atmospheric Measurement Techniques*, 8(11):4891–4916, 2015.

[dat16] Manual for real-time quality control of in-situ temperature and salinity data. 2016.

[DGdLM13] Yuri Demchenko, Paola Grosso, Cees T. A. M. de Laat, and Peter Membrey. Addressing big data issues in scientific data infrastructure. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 48–55, 2013.

[ESI] Best practices for sensor networks and sensor data management. URL: http://wiki.esipfed.org/index.php/EnviroSensing_Cluster.

[fRI10] Directorate-General for Research and Innovation. *European Commission: Legal framework for a European Research Infrastructure Consortium - ERIC Practical Guidelines*, 2010. doi:DOI:10.2777/79873.

[HHT+08] MR Haylock, N Hofstra, AMG Klein Tank, EJ Klok, PD Jones, and M New. A european daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research: Atmospheres*, 113(D20), 2008.

[HPA+16] Marton G Hidas, Roger Proctor, Natalia Atkins, Julian Atkinson, Laurent Besnard, Peter Blain, Philip Bohm, Jon Burgess, Kim Finney, Dan Fruehauf, et al. Information infrastructure for australias integrated marine observing system. *Earth Science Informatics*, 9(4):525–534, 2016.

[HRJM15] Jeffery S. Horsburgh, Stephanie L. Reeder, Amber Spackman Jones, and Jacob Meline. Open source software for visualization and quality control of continuous hydrologic and water quality sensor data. *Environmental Modelling & Software*, 70:32 – 44, 2015. URL: http://www.sciencedirect.com/science/article/pii/S1364815215001115, doi:https://doi.org/10.1016/j.envsoft.2015.04.002.

[HVM+16] Margareta Hellstrm, Alex Vermeulen, Oleg Mirzov, Simone Sabbatini, Domenico Vitale, Dario Papale, Jerome Tarniewicz, Lynn Hazan, Leonard Rivier, Steve Jones, B Pfeil, and Truls Johannessen. Near real time data processing in icos ri. 11 2016.

[idlHH] A. ieva de la Hidalga, , and A. Hardisty. How the envri reference model helps to design research infrastructures. In *ENVRIplus Newsletter No.2*.

[IK95] A Chidi Ibe and Gunnar Kullenberg. Quality assurance/quality control (qa/qc) regime in marine pollution monitoring programmes: the gipme perspective. *Marine Pollution Bulletin*, 31(4-12):209–213, 1995.

[IQu15] Autoqc. https://github.com/IQuOD/AutoQC, 2015.

[ISO15a] ISO. *ISO 19157:2013 - Geographic information – Data quality*, 2015.

[ISO15b] ISO. *ISO 9001:2015 - Quality management systems - Requirements*, 2015.

[JBV09] N. Juristo, V. R. Basili, and S. Vegas. Maturing software engineering knowledge through classifications: A case study on unit testing techniques. *IEEE Transactions on Software Engineering*, 35:551–565, 01 2009. URL: doi.ieeecomputersociety.org/10.1109/TSE.2009.13, doi:10.1109/TSE.2009.13.

[JHR+18] Pierre Jaccard, Dag Oystein Hjemann, Jani Ruohola, Anna Birgitta Ledang, Sabine Marty, Trond Kristiansen, Seppo Kaitala, and Antoine Mangin. Quality control of biogeochemical measurements. 2018.

[KGS+13] S Konovalov, Hernan Garcia, Reiner Schlitzer, Laure Devine, Cynthia Chandler, Gwenaelle Moncoiffe, T Suzuki, and A Kozyr. *Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data*. Paris, 01 2013.

[KL15] A. Karpatne and S. Liess. A guide to earth science data: Summary and research challenges. *Computing in Science & Engineering*, 17(6):14–18, Nov.-Dec. 2015. URL: doi.ieeecomputersociety.org/10.1109/MCSE.2015.127, doi:10.1109/MCSE.2015.127.

[Kwa99] Barbara H. Kwasnik. The role of classification in knowledge representation and discovery. *Library Trends*, 48, 1999.

[LQLG16] Yingjian Liu, Meng Qiu, Chao Liu, and Zhongwen Guo. Big data challenges in ocean observation: a survey. *Personal and Ubiquitous Computing*, 21:55–65, 2016.

[No16] WMO No. *100, 2016*, 3 edition, 2016.

[NVM13] Robert C Nickerson, Upkar Varshney, and Jan Muntermann. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359, 2013.

[OHO16] A Christopher Oishi, David A Hawthorne, and Ram Oren. Baseliner: an open-source, interactive tool for processing sap flux data from thermal dissipation probes. *SoftwareX*, 5:139–143, 2016.

[PAP+14] Gilberto Zonta Pastorello, Deborah A. Agarwal, Dario Papale, Taghrid Samak, Carlo Trotta, Alessio Ribeca, Cristina Poindexter, Boris Faybishenko, Dan Gunter, Rachel Hollowgrass, and Eleonora Canfora. Observational data patterns for time series data quality assessment. *2014 IEEE 10th International Conference on e-Science*, 1:271–278, 2014.

[PLW02] Leo Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Commun. ACM*, 45:211–218, 2002.

[RHH+15] AT Ringler, MT Hagerty, J Holland, A Gonzales, Lind S Gee, JD Edwards, David Wilson, and AM Baker. The data quality analyzer: A quality control program for seismic data. *Computers & Geosciences*, 76:96–111, 2015.

[RPMS17] Hampapuram Ramapriyan, Ge Peng, David Moroni, and Chung-Lin Shie. Ensuring and improving information quality for earth science data and products. *D-Lib Magazine*, 23(7/8), jul 2017. URL: https://doi.org/10.1045/july2017-ramapriyan, doi:10.1045/july2017-ramapriyan.

[RSK+12] Hugh Roarty, Michael Smith, John Kerfoot, Josh Kohut, and Scott Glenn. Automated quality control of high frequency radar data. In *Oceans, 2012*, pages 1–7. IEEE, 2012.

[Sch02] Reiner Schlitzer. Interactive analysis and visualization of geoscience data with ocean data view. *Computers & geosciences*, 28(10):1211–1218, 2002.

[Sea10] Seadatanet standards for quality control, 2010.

[SFA+00] Mark A Shafer, Christopher A Fiebrich, Derek S Arndt, Sherman E Fredrickson, and Timothy W Hughes. Quality assurance procedures in the oklahoma mesonetwork. *Journal of Atmospheric and Oceanic Technology*, 17(4):474–494, 2000.

[TBH+15] Charles Troupin, Joan Pau Beltrán, Emma Heslop, Marc Torner, Bartolomé Garau, J Allen, Simón Ruiz, and Joaquín Tintoré. A toolbox for glider data processing and management. *Methods in Oceanography*, 13:13–23, 2015.

[TL13] Jeff R Taylor and Henry L Loescher. Automated quality control methods for sensor data: a novel observatory approach. *Biogeosciences*, 10(7):4957–4971, 2013.

[UBBM17] Muhammad Usman, Ricardo Britto, Jrgen Brstler, and Emilia Mendes. Taxonomies in software engineering: A systematic mapping study and a revised taxonomy development method. *Information and Software Technology*, 85:43 – 59, 2017. URL: http://www.sciencedirect.com/science/article/pii/S0950584917300472, doi:https://doi.org/10.1016/j.infsof.2017.01.006.

[VBG+17] Michael F. Vardaro, Leila Belabbassi, Lori M. Garzio, Michael Smith, Friedrich Knuth, John Kerfbot, Sage Lichtenwalner, and Michael F. Crowley. Ooi data quality procedures and tools building on the first year of operations. *OCEANS 2017 Anchorage*, pages 1–5, 2017.

[WS96] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *J. of Management Information Systems*, 12:5–33, 1996.

[WSH15] R Wedd, Marc Stringer, and Keith Haines. Argo real-time quality control intercomparison. *Journal of Operational Oceanography*, 8(2):108–122, 2015.

[ZMG+15] Zhiming Zhao, Paul Martin, Paola Grosso, Wouter Los, Cees De Laat, Keith Jeffrey, Alex Hardisty, Alex Vermeulen, Donatella Castelli, Yannick Legré, et al. Reference model guided system design and implementation for interoperable environmental research infrastructures. In *e-Science (e-Science), 2015 IEEE 11th International Conference on*, pages 551–556. IEEE, 2015.