A Framework to assess Data Quality in university web portals

Master's Thesis Report, November 15, 2010

Arul Mary Michel

A Framework to assess Data Quality in university web portals

MASTER THESIS

submitted in partial fulfilment of

the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

TRACK INFORMATION ARCHITECTURE

by

Arul Mary Michel

born in Pondicherry, India

Web Information Systems Group
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, The Netherlands
http://eemcs.tudelft.nl

A Framework to assess Data Quality in university web portals

Author: Arul Mary Michel
Student id: 1535854
Email: a.m.michel@student.tudelft.nl

Abstract

Data quality is often described as "data that is fit for use by data consumers". The development of the internet and the various web technologies has increased the expectations of good data quality level among the data consumers. Web portals are meant to attract a wide variety of users and serve as an important means to access data in this information era. There is a need to assess the quality of these data, and so a framework to assess the data quality is wanted. This thesis proposes a framework to assess data quality in university web portals. The proposed framework captures essential DQ dimensions in four categories namely *Intrinsic*, *Representational*, *Contextual, and Accessibility*. In order to develop this framework, DQ dimensions were derived from literature. Next to that special attention was given to data consumers (users of university web portals), this led to the expansion and modification of the dimensions to emphasize the growing importance of the user's perspective.

The set of DQ dimensions has been tested with users using two pre-validations before the final validation of the framework. The proposed framework containing 35 dimensions is validated with user groups of two different universities: TU-Delft, The Netherlands and Pondicherry University, India. The validation showed that all DQ dimensions in the framework were valuable and important from the data consumer's view of perspective. As an evaluation approach, this framework has been compared with an existing portal data quality model and found this new framework has important DQ dimensions in the context of a UWP. As a consequence, the proposed framework emphasizes the user's perspective on data quality.

Keywords: Data Quality, DQ dimensions, Data Quality framework, data consumer expectations, UWP (University web portal), UWP properties

Graduation Committee:

Prof. dr. ir. G.J. Houben, Faculty EEMCS (WIS), TU Delft
Dr. K.van der Meer, Faculty EEMCS, TU Delft
Ir. H.J.A.M. Geers, Faculty EEMCS, TU Delft

Acknowledgement

It gives me a great pleasure in acknowledging my thesis supervisor Dr. Kees van der Meer for his valuable support and guidance throughout my thesis work. His continued support led me to the right way. Besides my supervisor, I would like to thank Professor Geert Jan Houben, Coordinator of Information Architecture for his comments and suggestions on my thesis work. I would also like to express my gratitude to Professor Hans Geers, coordinator of Computer Science Masters program for his advice and support from the beginning of my study at TU Delft. In addition I would also like to thank Professor Jan Hidders for his valuable suggestions.

My sincere thanks also goes to Qi Gao PhD student of EEMCS faculty, for spending his valuable time at the beginning of my thesis work. I also thank my fellow students Michel van Tol, Eric Feliksik, Jenna Willis, and Tsigereda for their comments and suggestion on my thesis work.

I would like to take this opportunity to thank my loving husband Mr. Jegarajan for his encouragement throughout this thesis work. He gave me the confidence to begin my Master’s Program in Information Architecture at TU Delft and has made available his support in a number of ways. Without his support, this work would not have been possible.

I would also like to thank Mr. Arokia Paul Raj, Lecturer in Computer Science, Pondicherry University, for spending his valuable time with the survey administration. Special thanks to my dear friend Mrs. Aruna Kirthika, Lecturer in Computer Science for her words of encouragement and suggestions towards this thesis work. I wish to thank my parents and brothers for their encouragement and prayers.

Last but not the least I would like to thank The Almighty for his blessings and guidance that helped me to complete this work.

Contents

List of Figures

List of Tables

.

Chapter 1

Introduction

Over the past few years the amount of information available to the user has rapidly increased in size through the development of the internet and web-based technologies. Web portals are meant to attract a larger variety of users and serve as an important means to access information from the World Wide Web. Web portals contain all kinds of information and services in a unique environment to provide users with simple, quick and secure access to relevant organizational and personal data. In addition, the content of the portals may vary from one organization to another, although in general the main aim of all the portals is to create a working environment for users to find the information they need at unique point of service. The providers of web portals have different views, ideas and knowledge level in usage purposes. Consequently the information received may be erroneous, preconceived and discrepant [25], leading to data quality issues. So users need to be aware of DQ in the web portal to carry out the tasks at hand.

Data quality (DQ) or Information quality (IQ) is often described as “data / information that is fit for use” i.e., the ability of a collection of data to meet user requirements [6]. As the literature shows these two terms have been used interchangeably in various contexts. The consequences of poor quality of data are often experienced in everyday life, but often without making the necessary connections to their causes [4]. In this context, the concept of DQ is important in web portals to provide more attention to the quality of data from user’s point of view in addition to information access, as poor data quality negatively impacts on user satisfaction. In the context of a UWP (university web portal), DQ is important to facilitate a wide variety of users. For instance, lecture information of a particular course could be viewed by a diverse group of users: own students, students from other faculties, students from other universities, international students, course managers, etc. The available information about this lecture must be interpretable, relevant, complete, and timely available to all these users to maintain a good level of data quality.

As a result, DQ has to be assessed often in any web portal to achieve high quality of data. The objective of the assessment is to identify the quality of data in a web portal. So, one needs a framework capturing essential DQ aspects to assess data quality from the data consumers perspective. To be specific, assessing data quality from the data consumer perspective is the most needed DQ aspect in this information age, since they are users who use the data, so the data should fit for their requirement. DQ assessments are mainly carried through DQ frameworks consisting of categories and dimensions to identify the DQ problems. Choosing dimensions to measure the level of quality of data is the starting point of DQ-related activity. With this approach several DQ models and frameworks have been developed to assess the data quality in different domains but only less works address the quality of data in the context of web portals.

In particular, university web portals play an important role in educational field to fulfil the needs of the students and researchers by combining data from different sources, which are necessary to make decisions in the required field of study or research. The data regarding educational purposes may be easy to find through university web portals, however finding relevant and high quality of data is far more difficult. So university web portals must design and implement methodologies, frameworks and tools to assess and improve data quality in order to satisfy the needs of their users.

1.1 Problem Definition

Data quality is a multidimensional concept and it can be evaluated through different attributes or dimensions. In order to assess the quality of data in web portals, we need to identify the required DQ dimensions other than the conventional DQ dimensions. This is because the dimensions of DQ can vary depending on the context in which the data is to be used. There are several DQ models and frameworks developed on some specific domains such as web information systems, data warehousing, data integration and data mining with a few DQ dimensions. Only a few proposals of data quality utilize all possible DQ dimensions to assess the data quality. However, DQ in the context of web portals is limited and no work addresses the DQ in UWP domain in detail. It is worth to note here that ISO/IEC 9126 standards are available to define the quality of a software product but there is no such standard for DQ that provides necessary DQ dimensions to assess data quality in this context.

In particular, to assess the DQ in university web portals we need to determine the DQ dimensions from data consumer perspective which is highly important to improve the data quality of the web portal and to satisfy user’s needs. On the other hand, there is no data quality model or framework that considers the user’s requirements in a UWP domain. Hence there is a need to design a DQ framework from the perspective of data consumers of university web portals.

1.2 Research goal and questions

The research goal of this thesis work is **to develop a framework to assess data quality in university web portals,** focussed on the perspective of the data consumer or user. This framework will be validated by users of two different universities: TU Delft, The Netherlands and Pondicherry University, India respectively to perceive how far the proposed framework is suited among the university web portal users.

In order to achieve the main research goal, the following research questions have been formulated:

1. What is data quality (DQ)?
2. What is a Web portal and its characteristics?
3. What DQ attributes have been developed in a Web context?
4. What DQ attributes need to be addressed to assess data quality in university web portals (UWP)?
5. What can be recommended to improve the DQ in university web portals?

1.3 Delineation

This master thesis is my graduation project and because there is a limited timeframe for this project, not all issues concerning the data quality assessment can be addressed. Therefore I propose the following delineation:

- As stated in the research goal the main idea of this work is to develop a framework determining essential DQ dimensions required to assess data quality in university web portals and not an evaluation of UWP data quality based on this framework
- Because of time constraints and degree of importance to certain data consumer perspective, this framework will be validated only with M.Sc., and PhD students.

1.4 Scientific and Social relevance

The proposed framework can act as a starting point in assessing the DQ of university web portals. The proposed framework could help the web-based portal application designers and developers to know the needs of UWP users and thereby make changes and improve the portal to attain high data quality levels. Furthermore based on scientific technologies and advancements, tools can be developed to automate the DQ assessment process based on the dimensions proposed in this framework. Though it is a complex task in the DQ research studies, methods or algorithms can be introduced to measure the level of data accurately for each dimension.

On the other hand there are social benefits linked to this field of DQ research in UWP domain as data quality is always related with the perception of data consumers or users. Users can experience a high quality of data, if university web portals make use of this framework to assess the quality of data provided and thereby improve the area that needs attention. Moreover the results of framework validation at two different universities will be an added advantage to realize the importance of DQ dimensions proposed in the framework.

1.5 Thesis structure

This report is structured as follows:

Chapter 1 provided an introduction to this thesis with problem definition, research goal and questions and a short discussion about scientific and social relevance of this thesis.

Chapter 2 discusses the main literature contributions about DQ in order to provide an overview. Several data quality models, frameworks and the research areas of DQ are discussed in brief in this chapter

Chapter 3 explains how the proposed data quality assessment framework is developed. This chapter discusses three different phases used to develop the framework: DQ dimensions that have been collected from the literature, DQ dimensions from the data consumer perspective, and the selection of appropriate DQ dimensions for UWP based on data consumer expectations and UWP properties. These three phases are discussed in detail to show how the framework is constructed.

Chapter 4 presents the validation of the proposed framework and its results. This chapter explains how the framework is validated by means of a survey questionnaire and the value of the resulting DQ dimensions in the framework from the perspective of UWP users using statistical analyses.

Chapter 5 discusses the evaluation of the proposed framework with an existing web portal DQ model. This chapter compares each category of DQ in the proposed framework with an existing model.

Chapter 6 provides conclusions for this thesis work and provides possible future work in this context.

Chapter 2

Literature review

This chapter presents a review of the literature related to data quality. The aim of this chapter is to introduce the relevant perspectives that make DQ an important subject of research for many years. More specifically, the following topics are discussed:

- Definition of Data Quality
- Data Quality roles
- Data Quality approaches
- Research issues in Data Quality
- Data Quality in web data contexts
- Conventional characterizations of Data Quality
- Proposals of Data Quality dimensions
- Data Quality in web portal domains

2.1 Definition of Data Quality

Data quality mainly electronic data emerged as an academic research subject in the early 90's. As electronic data is so widely spread, the quality of such data becomes more important in this information era. Data represent real world objects, in a format that can be stored, retrieved, and elaborated by a software procedure, and communicated through a network [4]. Data and information are often used interchangeably in the literature. In practice, researchers intuitively differentiate between information and data, and describe information as data that have been processed and ordered. Data Quality (DQ) can be defined as "fitness for use", a relative term depending of the user's needs and requirements [6]. It refers to the degree to which data satisfy user requirements or are suitable for a specific process.

Based on this concept, the researchers have developed several data quality models to assess data quality. Generally DQ is being described or characterized by set of attributes or dimensions, which describe certain properties delivered to users. For any data quality related work, a set of data quality dimensions or attributes are examined to improve the quality of that work, where each dimension describes a specific data quality aspect. To be precise, data quality is a multidimensional concept and it can be evaluated or assessed through different attributes. The DQ literature provides several research works on data quality with multiple dimensions on different contexts. DQ has been addressed first in the context of information systems [6], and it has been extended to various contexts like web information systems, cooperative information systems, data warehouses, data mining, and data integrations systems.

2.2 Data Quality roles

We can identify three important roles in the context of data quality [13] [26]: data producers, data custodians and data consumers. ***Data producers*** are people, groups or sources who generate the data. ***Data custodians*** are people who manage resources for future use. ***Data consumers*** are users who access and use data.

Each of these roles is related with a process or task. Data producers are related with the data production process, data custodians with data storage, maintenance, security, and give access to the data and data consumers with data utilization process. Most of the DQ works assessed quality from the perspective of the data consumer. However in certain application domains, these roles may not be different. For instance, in web-based information systems, data producers and data custodians are often discussed as the same entity [27]. In this context, identifying precise data quality aspects applicable to these roles is a challenging task. The analysis of several DQ researches shows a conformance that data quality is determined mainly from the perspective of data consumers.

2.3 Data Quality approaches

Data quality is a multidimensional concept as the level of DQ is measurable by considering a set of dimensions or attributes applicable to a certain domain. Wang and Strong [6] have identified three different approaches to evaluate the DQ dimensions: intuitive, theoretical and empirical.

Intuitive approach

In the intuitive approach, the DQ dimensions are selected based on the researcher's experience and knowledge about what attributes are important in a specific context. Most of the DQ studies proposed in the DQ literature are based on this approach. The advantage of using this approach is that enables each study to select the dimensions that are relevant to the particular goals of the context.

Theoretical approach

The theoretical approach focuses on how data may become deficient during the data manufacturing process [6]. The advantage of this approach is the potential to provide a complete set of DQ dimensions. This approach assumes that an information system is a representation of a real world system as perceived by the users (Figure 1). The DQ dimensions are derived on the basis of possible inconsistencies between the user's view of the real-world system as inferred from information system and the view that can be obtained by directly observing the real-world system.

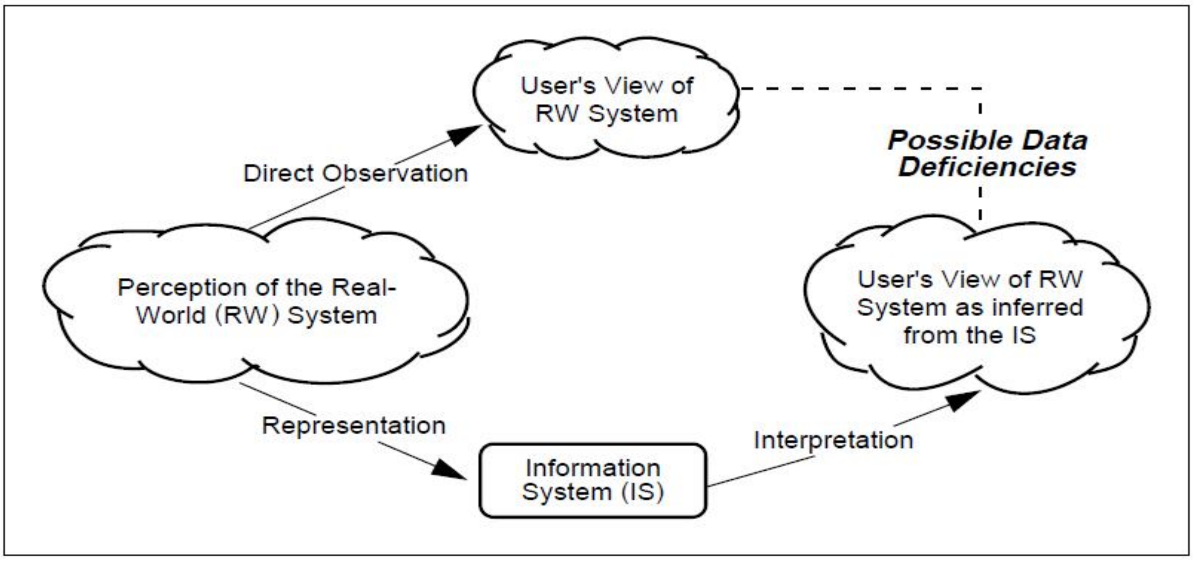

User's View of RW System
Possible Data Deficiencies
Direct Observation
Perception of the Real-World (RW) System
User's View of RW System as inferred from the IS
Representation
Interpretation
Information System (IS)

Figure 1: Possible data deficiencies in the data quality model [27]

Empirical approach

Both the theoretical and intuitive approach fail to capture the voice of the data consumer or user. The empirical approach has been used to analyze data collected from data consumers, to determine the dimensions they really use to assess whether the data are fit for use in their contexts. The empirical method is a general term for any research method that draws conclusions from observable evidence [14].

Wang and Strong [6] used this empirical approach to define the DQ dimensions, by capturing the voice of the data consumers. This approach has been carried out in two different phases. The first phase is to identify potential DQ dimensions from the perception of data consumer or dimensions that are relevant to the user by means of a survey. As a result of several DQ dimensions from the user, the first phase identified 20 DQ dimensions that were important to data consumers. Then in the second phase, these 20 dimensions have been reduced to 15 dimensions and sorted into different categories to represent a comprehensive framework of DQ from data consumers' perspectives.

2.4 Research issues in Data Quality

This section gives a short description about research issues that have been used in many academic researches on DQ. Achieving data quality is a complex, multidisciplinary area of investigation [4]. Several research topics and application domains study data quality to improve the quality of data in organizations. Mainly it has been viewed in the perspective of data consumer or user to satisfy their needs and requirements. Figure 2 shows the research issues and application domains discussed in DQ literature

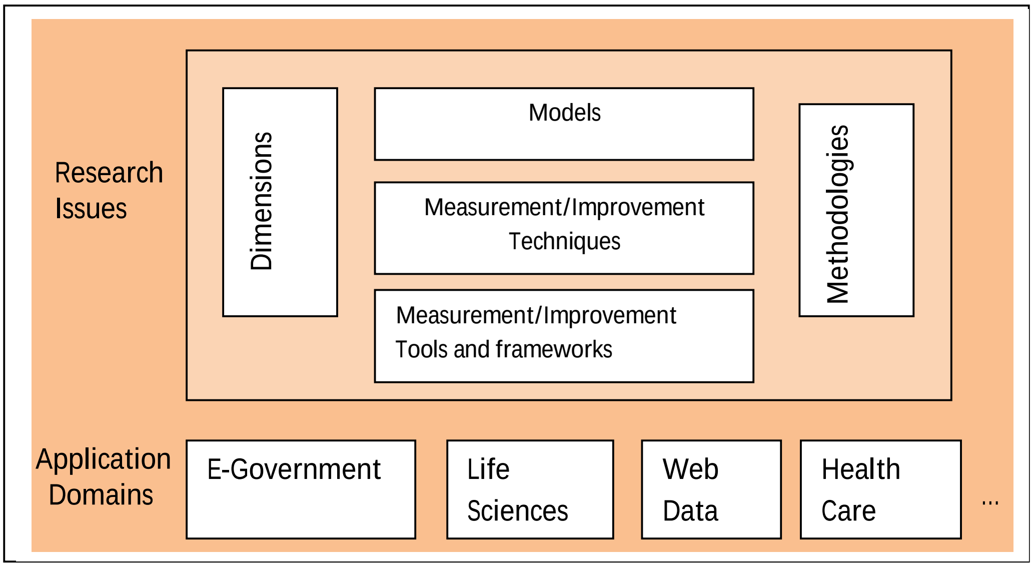


Figure 2: Main issues in data quality [4]

The most common research issues discussed are models, techniques, tools, frameworks and two 'vertical' areas: dimensions and methodologies that cross the first three (figure 2)areas. Based on these models and frameworks only the quality of data is assessed in many domains.

Models are mainly used in databases to represent data and data schemas. Models are also used in the area of information systems to represent business processes of organizations in terms of sub processes, their inputs and outputs, casual relationship between them, and functional/non-functional requirements related to processes [4]. ***Techniques*** refer to algorithms, heuristics, knowledge-based procedures and learning processes that help to identify and solve a DQ related problem. ***Methodologies*** provide guidelines to choose appropriate techniques or tools as the effective way of DQ measurement and improvement processes. ***Tools and frameworks:*** a tool is a software procedure designed, automated and provided with an interface to evaluate the DQ activities. A framework consists of a suite of coherent tools for a domain or task field.

Furthermore, the data quality is getting more attention in diverse application domains in the recent years: E-Government, Life Sciences, Web data, and Health care. However here we discuss the DQ in web data contexts. The following section will analyze the various web data contexts where DQ is essential in this information era:

2.4.1 Data Quality in web data contexts

Research on data quality has been started in the area of information systems in the early 90's and it has been extended to several other areas in the web data contexts: web-based information systems, data integration systems, collaborative information systems, data mining, data warehouses and recently in web portals. This section will briefly discuss about several contexts where quality of data is considered important to data consumers

- **Total Data Quality management (TDQM)**

In the early 1996's the Total Data Quality Management (TDQM) program at the Massachusetts Institute of Technology pioneered DQ as a research area. The goal of TDQM methodology is to provide users with a high quality of data by considering data as an information product. This methodology laid as a foundation for DQ research and attracted many researchers to research in this emerging field. The TDQM methodology consists of four phases such as Define, Measure, Analyze, and Improve [13]. The phases are iteratively executed, thus constituting a cycle. This cycle of defining, measuring, analyzing and continuously improving information quality is important to achieve a high quality of data. The definition phase of TDQM cycle identifies essential DQ dimensions and related requirements. Measurement phase produces data quality metrics for DQ dimensions to measure the quality of data appropriately. The analysis phase interprets the measurement results. The analysis phase identifies the dimensions that need improvement and the causes of potential DQ in the context. The improvement phase consists of DQ improvement techniques that allows actions to be taken for the identified data quality problems

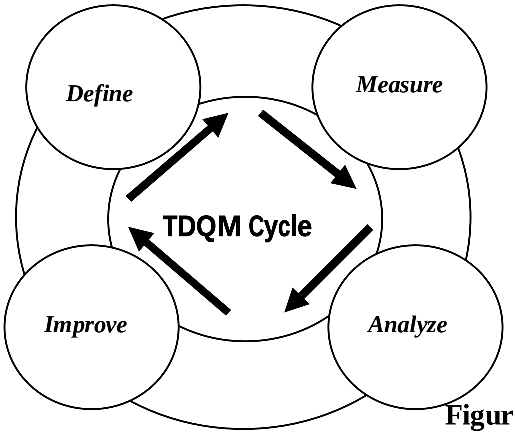


Figure 3: The TDQM cycle [13**]**

- **Web information Systems**

Web information systems (WIS) are characterized by the presentation to a wide audience of a large amount of data, the quality of which can be very heterogeneous [15]. One of the main goals of WIS is to provide information with a good quality level to the data consumers. Specifically, the evolution in time of quality information in WIS's is particularly important to satisfy the data consumers. It should provide information in a short time after it is available from information sources. Research in web information system takes this time aspect of DQ into account and focus only on a "core" set of dimensions. The DQ dimensions that are considered in this domain include *expiration, completeness, source reliability, and correctness.* But still this set of dimensions could be extended to study whether other dimensions proposed in the DQ literature are applicable to this context.

- **Co-operative information systems**

A cooperative information system (CIS) is a large scale information system that interconnects various systems of different and autonomous organizations, geographically distributed and sharing common objectives [16]. Data are fundamental to any cooperative organization, to be shared with other organization in order to fulfill the service requests from citizens. Hence the quality of data has to assured by each organization before being shared with others. If organizations exchange data without knowing their actual quality, then there is a possibility of spreading low quality of data all over the CIS [16]. And a typical aspect of CIS is high data replication by having different copies of the same data stored by different organizations. To improve quality of data in all organizations, DQ aspects like accuracy, completeness, currency and internal consistency have been proposed by researchers. However this set of dimension need not yet cover all data quality aspects in CIS.

- **Data integration systems (DIS)**

Systems need to integrate data coming from multiple and heterogeneous data sources and provide users with a uniform access to data. These systems are called data integration systems [17]. Ensuring data quality in data integration systems is particularly difficult due to the integration of data coming from multiple sources that have different schemas, representations, and administrations. So, importance has been given to DQ dimensions like *data completeness, data uniqueness, data consistency, data freshness, data accuracy, completeness, understandability, minimality and expressiveness* to evaluate the quality. In DIS, DQ dimensions studied are more as compared to co-operative information system. This is because CIS are more concerned about data values and they do not deal with aspects concerning logical schema and data formats whereas in DIS importance is given to both data values as well as data representations.

- **Data warehouses**

Ensuring a high level of data quality is one of the crucial issues in data warehousing. A data warehouse system supports information process by creating an integrated database of historical data. Generally it integrates data from multiple, incompatible systems into one consolidated database [33]. Thus the central module of any data warehouse system is the data warehouse data base, which is a single, complete, and consistent store of data that has been obtained from a variety of sources.

2.5 Conventional Data Quality dimensions

This section provides dimensions that are conventional to the DQ research. The most frequently mentioned data quality dimensions in the literature are *accuracy, completeness, timeliness and consistency*, with various definitions by different researches. Nevertheless, there exist several works that deal with additional DQ aspects, in particular in the context of management information systems. For instance in Wang and strong proposal [6] we can find different categories of DQ dimensions. Despite of several DQ dimensions, many researches address only the above mentioned conventional aspects in a broader perspective. This section provides a short description about those conventional dimensions:

- **Accuracy**

There is no standard definition for accuracy in the literature. In [2] accuracy is the defined as the extent to which the data are correct, reliable and certified free of error. It can also be defined as the closeness between a value v and a value v', considered as the correct representation of the real – life phenomenon that v aims to represent [4]

- **Completeness**

Completeness can be generally defined as "the extent to which data are of sufficient breadth, depth and scope for the task at hand" [2]

- **Timeliness**

Timeliness expresses how current data are for the task at hand. It is motivated by the fact that it is possible to have current data that are actually ineffective because they are late for a specific usage [4]. For instance, the timetable for university courses can be current with most recent data, but it cannot be timely if it is available only after the start of the classes. It can also be defined as the extent to which data is sufficiently up-to-date for the task at hand [6]. In the proposal of Bovee et al [18] timeliness is referred as datedness related with age and volatility. Age or currency is a measure of how old the information is, based on how long ago it was recorded. Volatility is a measure of information instability-the frequency of change of the value for an entity attribute.

- **Consistency**

Consistency is defined as the degree to which data managed in a system satisfy specified constraints or business rules [19]. Though consistency has been addressed in data base systems as an integrity constraint, it is considered as an important DQ problem. The above dimensions could be described with a simple record named Student, with fields or attributes such as Student ID, Name, Sex, Department and Email

Student ID	Name	Sex	Department	Email
345496840	John	Male	Computer Science	NULL
317890231	Eric	Male	Computer Science	paul@yahoo.com
567888723	Nancy	Female	Mathematics	nancy@yahoo.com
567903466	Remi	Male	Mathematics	NULL

Table 1: Example record

If the name of a student is John, then according to the definition of accuracy the value v’ = John is correct, while the value v=Jhn is incorrect, because Jhn is not a correct value according to a dictionary of English names, this is a case of low accuracy [4].

For completeness, we can consider the email field: The presence of a null value has a general meaning that, the value exists in the real world but for some reason it is not available [4]. Here in this example, the null value can have two reasons: (i) the particular student has no email address, and therefore the field is represented as null. Thus it cannot be viewed as incompleteness case (ii) the particular student has an email address but it not available or stored in the database. This second reason shows this as a case of low degree of completeness

For consistency, we can consider the values of two fields such as Students ID and Department in the above table. For instance if there exists a constraint that the student ID of computer science department should start with 34…, then the records related to computer science department may be low consistency case as student id of one record begins with 31…which shows that the student Eric may belong to another department and not computer science.

2.6 Proposals of Data Quality dimensions

In the literature there are proposals that address the classification of different DQ dimensions as opposed to conventional DQ aspects [25]. Although there are many DQ proposals, the most widely approached proposals are discussed below. These proposals provide a comprehensive list of DQ dimensions to understand the DQ aspects in a brief manner and to assess the data quality in several contexts. The dimensions in these frameworks were studied to capture a basic set of DQ dimensions.

1. Naumann [1]

In the proposal of Naumann, DQ dimensions are defined specific to integrated web information systems where quality is conceived as an aggregated value of multiple DQ criteria. The author introduces four different categories of DQ dimensions that play an important role in integrated web information system. Those four categories are as follows:

- Content related
- Technical
- Intellectual
- Instantiation-related criteria

1. *Content-related* dimensions consider the actual data that has to be retrieved. The dimensions discussed under this category include accuracy, completeness, customer support, documentation, interpretability, relevancy, value-added.

2. *Technica*l dimensions describe the hardware and software aspects used to maintain the data in an information system and the network connections between the user, mediator and sources. The dimensions addressed under this category include availability, latency, price, Quality of service, response time, security, and timeliness.

3. *Intellectual* criteria address the subjective aspects of the data sources such as believability, objectivity, reputation.

4. *Instantiation-related* criteria concern the presentation of the data. Dimensions represented under this category are amount of data, representational conciseness, representational consistency, understandability, verifiability.

Several research studies about DQ show different classifications of dimensions based on the perception of authors or the domain on which it is generated. There are several factors for the selection of DQ criteria for a particular domain or context. Below are the short descriptions of such factors proposed by Naumann [1]

- **Application**

Out of several criteria, some criteria will be more important for a particular application domain than other criteria to attain high quality of data. For instance Naumann describes two information systems that show only the dimensions that are most important to achieve good quality of data. Those two domains are; *Search Engines* and *Stock Information Systems*. Let us consider one domain here. The dimensions proposed for search engines include accuracy, timeliness, availability, completeness, latency, redundancy and response time. However in this domain, most of the dimensions are not related to the data instead to technical criteria and also few important DQ dimensions are missing such as precision, reputation and relevancy.

Search Engine

Accuracy	**Quality of the result ordering**
Timeliness	Update frequency of the search engine
Availability	Percentage of time the search engine is “up”
Completeness	Percentage of the Web that the search engine has indexed
Latency	Time until first web link reaches user
Redundancy	Number of redundant links in the search result
Response time	Time until the complete response reaches the user

Table 2: Data Quality dimensions proposed by Naumann for a search engine

- **Users**

User’s satisfaction is very important in any information system, so the objective of carrying out IQ- related activities in any domain is to satisfy the user. Hence it is always important to let users to participate in the selection of criteria that are used to predict the satisfaction of user.

- **Provider**

A provider is the one who makes the choice of selecting criteria to implement a certain course of action in any integrated system. For instance a price criterion cannot be used if the provider decides to offer a free service using only data sources free of charge.

- **Assessment**

Any information system should use only those DQ criteria that can be assessed accurately. If not, it makes the assessment inaccurate and also the criterion is useless.

2. Wang and Strong [6]

In the proposal of Wang and Strong, DQ dimensions have been selected by interviewing data consumers. Nearly 179 data quality dimensions have been collected from the user's point of view by means of surveys. Out of those, the authors selected 15 different dimensions. Further, these dimensions are grouped under four different categories such as Intrinsic, Accessibility, Contextual, and Representational.

DQ category	DQ dimensions
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access security
Contextual DQ	Relevancy, Value-added, Timeliness, Completeness, Amount of data
Representational DQ	Interpretability, Ease of understanding, Concise representation, Consistent representation

Table 3: Data Quality dimensions proposed by Wang and Strong

- *Intrinsic Data Quality* consists of dimensions that evaluate the quality that the data has by itself as a representations of a real world (RS).

- *Contextual Data Quality* considers DQ dimensions that evaluate quality within the context in which they are involved. For instance, the dimensions addressed in this category include relevance, timeliness, and completeness which are necessary for the completion of the task.

- *Representational Data quality* category considers DQ dimensions related to the quality of data representations i.e., the system must provide information to the users that is interpretable, easy to understand, and it should be represented concisely and consistently.

- *Accessibility Data Quality* category considers DQ dimensions related to the accessibility of the data and the level of security.

Though the proposal of Wang and Strong is an empirical approach, we cannot expect more consistency or correctness of the results gathered through user survey. Moreover the dimensions proposed may not be applicable to all contexts, on the other hand a certain domain may need, more dimensions to assess the DQ than those proposed here. However this proposal was the foundation for many DQ research studies to develop different DQ models and framework in different contexts. In comparing Naumann [1] with Wang and Strong [6] we can identify dimensions that are common to both proposal but with different categories, for instance representational DQ of Wang and Strong has been renamed as Instantiation-related criteria in Naumann with few additional dimensions such as *verifiability* and *amount of data.* In the same way the intrinsic DQ has been renamed as Intellectual criteria with same dimensions except *accuracy,* where it is included in content-related criteria of Naumann.

3. Bovee et al [18]

This proposal considers DQ dimensions by taking the view of data consumers and developed a conceptual model consisting of 4 attributes. The model has been adapted from Wang and Strong [6] and the attributes have been compared with this model to assess DQ with all its essential dimensions or attributes that determine the quality, in any domain. It consists of four attributes, namely:

Accessibility: To get information which we might find useful

***Interpretability*:** To understand the information and find meaning from it.

Relevance: To find it applicable to the domain and the context of interest

Integrity: To believe it free from defects

The last attribute *Integrity* is further classified into four sub attributes: *Accuracy, Completeness, Consistency and Existence* where existence is found absent in many studies. Although this model does not utilize all the quality dimensions that were identified in the literature, the model emphasizes some essential DQ dimensions to assess data quality.

4. AIMQ: PSP/IQ model [21]

This proposal developed a methodology for information quality assessment and improvement called as AIM quality. The main objective of this methodology is to assist organizations in achieving high quality of information, so the authors designed a model called PSP/IQ model of what DQ means to information consumers and an assessment instrument for measuring the DQ among the users of different organizations. The model has four quadrants, where the information is considered to be a product or a service, and on whether the improvements can be assessed towards a formal specification or customer expectations. The dimension in the model has been classified into four categories: sound, dependable, useful and usable information.

	Conforms to specifications	**Meets or exceeds consumer expectations**
Product Quality	Sound information IQ dimensions • Free-of-error • Concise representation • Completeness • Consistent representation	Useful information IQ dimensions • Appropriate amount • Relevancy • Understandability • Interpretability • Objectivity
Service Quality	Dependable information IQ dimensions • Timeliness • Security	Usable information IQ dimensions • Believability • Accessibility • Ease of operation • Reputation

Table 4: PSP/IQ model classification

The model measured the dimensions in four quadrants by means of a questionnaire and testing has been carried out at five organizations. Based on the reliability statistics and gap analysis techniques the DQ of the organizations were assessed at the dimensional level. Thus this model serves as a best practice to study about data quality and by analyzing these data quality classifications it is possible to identify basic DQ dimensions and methods followed to measure those dimensions. This model is the extension of Wang and Strong where the DQ dimensions were empirically derived that are important to information consumers, forming a foundation for DQ assessment in organizations.

5. Lee et al [20]

In the proposal of Lee et al, the authors consider a number of DQ dimensions that can be used to assess the DQ. Although this proposal does not contain any model or framework, the dimensions discussed here are shown to be of particular interest and importance to many organizations. In addition to the dimensions, the authors also define metrics to measure those dimensions, however the metrics will not be discussed here. The proposed dimensions are as follows:

- **Free of error**

This dimension has been used to check whether the data is correct. Though the term accuracy has been referred to the correctness of the data in many DQ works, here the authors have named the dimension as free of error instead of accuracy.

- **Completeness**

The completeness of the data has been described with three different perspectives: schema completeness, column completeness, population completeness. *Schema completeness* refers to the degree to which the entities and the attributes are not missing from the schema. *Columns completeness* refers to the missing value in a column of a table. *Population completeness* refers to the degree to which member of the population that should be present are not present. For instance, if a column should contain at least one occurrence of all 50 states, but the column contains only 43 states, then the population is incomplete.

- **Consistency**

The consistency of the data is also viewed in the proposal with difference perspectives such as consistency of redundant data in one table or in multiple tables, consistency between two related elements. For instance the name of a particular city and its postal code should be consistent with one another. And the third type of consistency focuses on consistency of format for the same data element used in different tables. But these consistencies may vary depending on the context or requirements of any organizations.

- **Believability**

Believability is described as the extent to which the data is regarded as true and credible. The authors describe the dimension with multiple variables such as the individual assessment of the credibility of the source of the data, the perceived timeliness of the data or assessment against a common standard.

- **Appropriate amount of data**

This dimension is taken into account as one of the important dimensions, refers to the degree to which the amount of data should be neither too little nor too much.

- **Timeliness**

Timeliness is the extent to which the data is up-to-date with respect to the task for which it is used. Mostly timeliness is related with volatility and currency of the data, where volatility refers to the length of the time the data remains valid and currency refers to the age of the data. In this proposal also the authors relate timeliness dimension with currency and volatility attributes to measure timeliness on various data depending on the context

- **Accessibility**

The accessibility dimension reflects the ease of attainability of the data. That is the extent to which the data is easily accessible for the required tasks.

Though the proposal does not cover a wide range of DQ dimensions, these are the commonly used dimensions in several organizations to assess their level of DQ. It would be more efficient to consider dimensions according to the perception of users in the organization. This would help to achieve a precise set of DQ dimensions as data consumers or users are the main source of organizations.

2.8 Data Quality in web portal domains

In recent years, several research works have been developed on the Web Data Quality in different domains as we discussed before. But there are far less data quality proposals in the area of Web Portals. A Web portal can be defined as a Web site that aggregates an array of content and provides a variety of services including search engines, directories, news, e-mail and chat rooms [7]. In simple terms portals provide effective and convenient access to information resources through a single gateway or aggregates information from diverse sources and make that information available to various users.

Thus today’s portals are increasingly complicated applications designed to provide users access to all information sources. At the same time, portals need to supply data with a high quality level which satisfies its users since portals are mainly differentiated based on their content and intended users. To give an overview of portal types, several categories of the web portals [7] are explained below and to understand why data quality is highly needed in these domains.

Enterprise information portal
These portals are designed basically for an organization where there are many roles and users exist. The term enterprise can be used interchangeably with corporate portal where both reflect the same idea. Enterprises portals focus on providing information to its users and allow them to access the available resources on a regularly updated manner within the enterprise. It may include facilities such as: categorization of relevant enterprise information, search engine, organizational news, and access to email, access to common software applications, document management

systems, links to internal resources and popular external web resources, and also the ability to personalize the pages.

Vertical portals

These are web portals which focus on specific industries. Vertical portals aim to aggregate information, tools and services relevant to particular groups or online trade communities of closely related industries to facilitate the exchange of goods and services.

Horizontal portals

Horizontal portals are portals that are used by a broad base of user population covering a wide range of topics and interests. It is focused on general users and tries to present something to everybody surfing the internet and allows them to spend more time on the provided information. Examples of this type of portal include msn (www.msn.com), yahoo (www.yahoo.com) where it shows up several categories of information to meet the extensive needs of the diverse users.

Community portals

Community portals are created for community groups based on a wide variety of interests. Through this portal, members of the community share their ideas or interest depending on their orientation. Examples of this type of portal include GreyPath (www.greypath.com), iVillage (www.ivillage.co.uk/).

E-marketplace portals

These portals are extended enterprise portals to support business-to-business (B2B), business-to-customer processes (B2C) such as ordering, tendering and supply of products. It facilitates the sharing of information to external business partners or groups of an individual company, customers and suppliers. Examples of this type of portal include the bookseller Amazon.com (www.amazon.com).

Though there are different kinds of portals, each portal is unique depending on the context on which it was used but the aim of many web portals is to select, organize and distribute content in order to satisfy its users/customers. On the other hand the data provided by those must be of high quality to its users to carry out their tasks and decision making. Thus the aspect of data quality needs to be significant in the area of web portals in all organizations. However, the assessment of DQ is a difficult task for the web portal designers or developers as they need to ensure whether the portal provides data at a level of quality which satisfies the needs and requirement of its users [28]. An understanding of user motivation is essential as it is a foundation for conceptual ideas of context data quality evaluation [24]. Recently research on web portal data quality is gaining more attention to assist users with a high level of data quality [12][28].

University web portals (UWP)

University web portals can be categorized under enterprise information portals, providing information to different kinds of its users such as students, staffs and outside visitors. Almost every university has a web portal that provides certain services for individuals depending on the needs of that user. Furthermore the content of the portal can influence students when choosing which university they will apply to.

Every year lots of students' visit UWP's looking for information. For instance, searching for course information, course enrollment, faculty specific information, university facilities, lecture times, class schedules etc. It is extremely important to search all information at one point from a UWP. At the same time the information presented should be of high quality and the content must be easily understandable. Thus DQ is essential in a UWP to provide users with the information they need to carry out their tasks.

Consequently, the data producers or data custodians of a web portal must be aware of the needs of their users and thereby be enabled to improve the quality of data. However the literature lacks standard methodologies or frameworks to assess DQ in web portals like ISO/IEC 9126 which was developed for software product quality. In particular, we were not able to identify previous studies that address DQ, specifically in university web portals apart from the portal data quality model (PDQM) [28], where the model implements a tool to evaluate the representational DQ category in UWP domain. Thus there is a need for more DQ research in the context of web portals, especially to the domains where high quality level of data is needed.

Chapter 3

Framework methodology

This chapter describes how the proposed DQ framework for UWP is designed. The main objective of this work is to develop a framework that is applicable to assess DQ in university web portals. To develop this framework first we need to consider an important aspect of DQ i.e., DQ dimensions, which is a cumbersome task for any data quality assessment procedure. Although data quality literature provide several classifications of dimensions, no work addresses DQ in UWP domain and also most research studies failed to capture the DQ dimensions from a user perception. The proposed framework is designed based on the following 5 phases

(i) Harvest dimensions from widely accepted models or frameworks of data quality and consider DQ dimensions discussed by different authors in different contexts of web like management information systems, web information systems, co-operative information systems, data integrations systems, web portals, and data warehousing, etc.

(ii) Gather dimensions from the data consumers (user of UWP) to observe their expectations of data quality in the UWP context by means of a survey.

(iii) Merge the DQ dimensions from phase 1 and 2 and select the appropriate DQ dimensions required to assess data quality in UWP's, by relating data consumer expectations on DQ with UWP properties based on the method of a generic portal data quality model.

(iv) From the result of phase 3 design the framework and validate it with users of two different universities to assess the importance of the selected DQ dimensions among UWP users.

(v) Evaluate the framework with an existing portal data quality model to show how far the new framework is valuable when compared to other model or framework of DQ.

The first three phases will be discussed in brief in the following sections and the last two phases will be discussed in Chapter 4 and Chapter 5 respectively.

3.1 Data quality dimensions from literature

In the first phase, DQ dimensions were identified first from widely accepted frameworks such as Wang and Strong [6], Naumann [1], Bovee [18], Lee et al [20], AIM: PSP/IQ model [23] and then DQ dimensions that were discussed in several academic studies in different contexts of web. These specific frameworks and models provide thorough classification of DQ dimensions other than conventional aspects discussed in the previous chapter. By analyzing these important classifications, it is possible to identify the basic set of DQ dimensions for DQ assessment task in various contexts of web. So these proposals were studied first to capture the concepts of DQ dimensions. As mentioned before, the most significant research works studied for this collection of DQ dimensions are provided below (Table 5).

A few of these proposals [22] [15] [5] were not discussed in chapter 2, but they were reviewed for the selection of DQ dimensions and added here to show that it is worth to study these proposals to get an understanding of DQ dimensions. For instance, the DQ dimensions proposed by Wand and Wang [5], provide a possible set of dimensions in both intrinsic and extrinsic categories, which was later extended by Wang and Strong [6] and accepted as a conceptual framework in the DQ literature. The proposals [22] discuss the same set of DQ dimensions as [1] discussed in chapter 2, but was classified under different categories. According to [4], only four DQ dimensions were proposed in the context of web information systems; however this proposal was taken into account to verify the DQ dimensions. Consequently, all these DQ research works helped to capture the relevant and important DQ dimensions and proceed to further phases.

Author	Area of research	Model/framework
1. Naumann and Rolker [22]	Data integration	3 categories and 22 dimensions
2. Naumann [1]	Integrated web information systems	4 categories and 21 dimensions
3. Bovee [18]	Conceptual model for IQ	4 dimensions
4. Lee et al [20]	Conceptual DQ model	7 dimensions
5. AIM:PSP/IQ model [23]	Conceptual IQ model for organizations	4 categories and 15 dimensions
6. Wand and Wang [5]	Information system	2 categories and 26 dimensions
7. Redman [27]	Databases	3 categories and 25 DQ dimensions
8.Pernici & Scannapieco [15]	Web information systems	4 DQ dimensions
9. Wang and Strong [6]	Information systems	4 categories and 15 dimensions

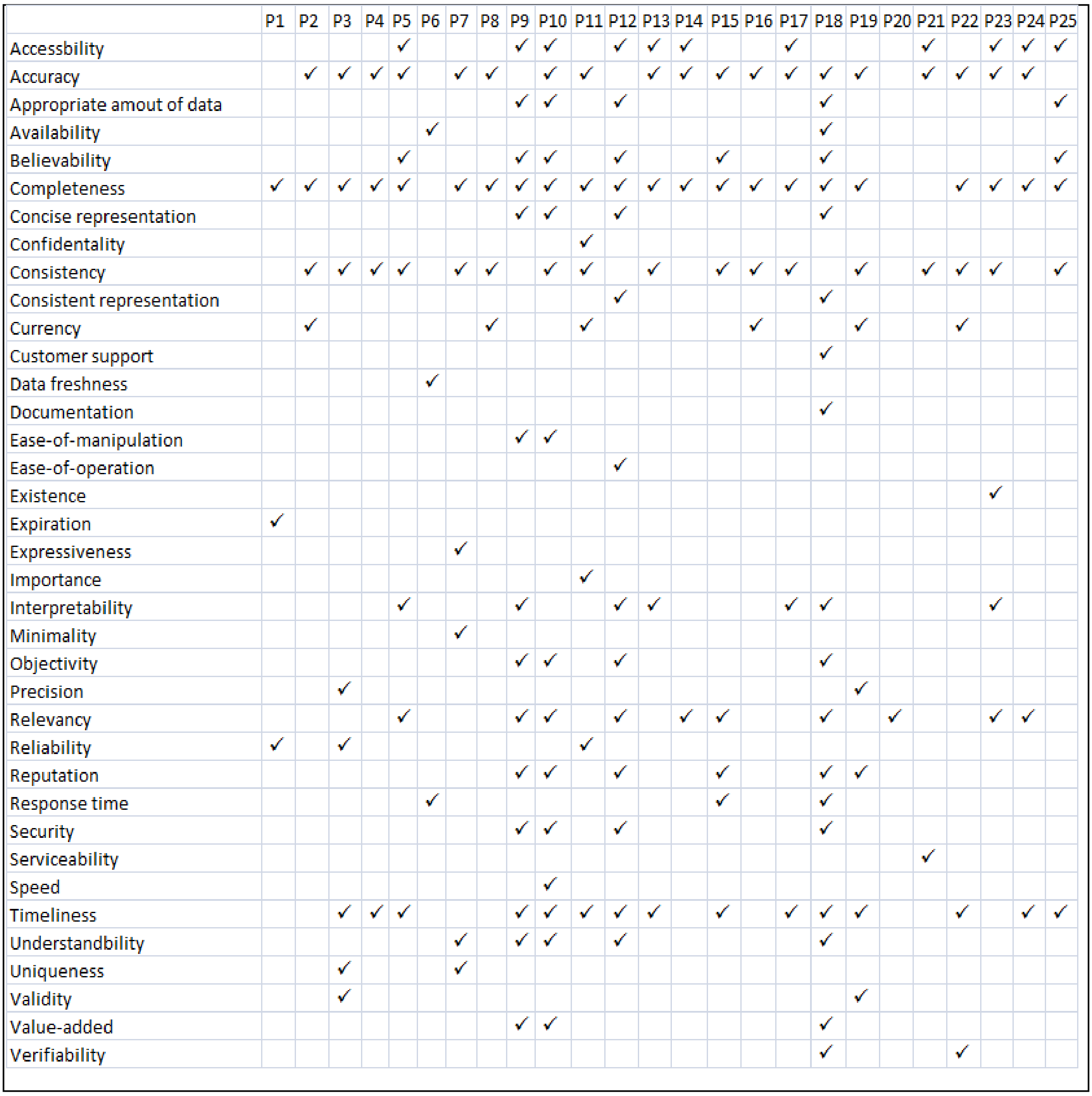
Table 5: Research works on data quality.

The main reason for this collection of DQ dimensions from different contexts is due to the absence of DQ proposals specific to university web portals and a few proposals have been proposed for domain independent web portals data quality. Due to these disadvantages, DQ dimensions have been collected among different web context. As a result of this process of dimension identification, 47 DQ dimensions were obtained from literature (Appendix A). When we compare the different proposals of DQ dimensions, we can identify two types of corrections:

(i) Same name for DQ dimensions with different meanings.

(ii) Different names for DQ dimensions with similar meaning.

These two corrections have been tackled through the widely suitable and frequently referenced definitions in the literature. Based on these two aspects, some DQ dimensions have been removed after careful analysis on their similarities and it has been summarized to 37 dimensions. Those DQ dimensions are presented in the table below (Table 6) that was proposed in various research studies. From the table we can get an overview of available DQ dimensions in the literature to assess the data quality and to observe their presence in different research contexts. The tick mark in the table shows whether that dimension has been included in that study. The dimensions *accuracy, completeness, consistency, and timeliness* have been used in most of the data quality researches. These are the conventional data quality characteristics in several contexts.



	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25
Accessbility					✓				✓	✓		✓	✓	✓			✓				✓		✓	✓	✓
Accuracy		✓	✓	✓	✓		✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	
Appropriate amount of data									✓	✓		✓						✓							✓
Availability						✓												✓							
Believability					✓				✓	✓		✓			✓			✓							✓
Completeness	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓
Concise representation									✓	✓		✓						✓							
Confidentality											✓														
Consistency		✓	✓	✓	✓		✓	✓		✓	✓		✓		✓	✓	✓		✓		✓	✓	✓		✓
Consistent representation												✓						✓							
Currency		✓						✓			✓					✓			✓			✓			
Customer support																		✓							
Data freshness						✓																			
Documentation																		✓							
Ease-of-manipulation									✓	✓															
Ease-of-operation												✓													
Existence																							✓		
Expiration	✓																								
Expressiveness							✓																		
Importance											✓														
Interpretability					✓				✓			✓	✓				✓	✓					✓		
Minimality							✓																		
Objectivity									✓	✓		✓						✓							
Precision			✓																✓						
Relevancy					✓				✓	✓		✓		✓	✓			✓		✓			✓	✓	
Reliability	✓		✓								✓														
Reputation									✓	✓		✓			✓			✓	✓						
Response time						✓									✓			✓							
Security									✓	✓		✓						✓							
Serviceability																					✓				
Speed										✓															
Timeliness			✓	✓	✓				✓	✓	✓	✓	✓		✓		✓	✓	✓			✓		✓	✓
Understandbility							✓		✓	✓		✓						✓							
Uniqueness			✓				✓																		
Validity			✓																✓						
Value-added									✓	✓								✓							
Verifiability																		✓				✓			

Table 6: DQ dimensions from DQ literature

The resulted table was obtained after the elimination of few dimensions based on the two corrections that were mentioned above. For instance, dimensions *accuracy*, *free of error* and *correctness* have the same meaning but they differ in names in different contexts.

3.2 Data Quality dimensions from data consumer perspective

The data consumer or user plays an important role in the data quality management process, where they access and use data for their individual tasks. Data consumers' assessments of DQ are increasingly important because consumers now have more choices and control over their computing environment and the data they use [3]. So, as a next phase it has been decided to gather the DQ dimensions from data consumers (here UWP users), i.e., the dimensions they consider important for assessing data quality in UWP's from their perspective. This will show their expectations on data quality items, based on this we can make sure that the DQ dimensions collected from the literature is expanded with the expectations of the user. In addition it may help to identify dimensions other than those listed above. As we discussed in previous chapter, many researches fail to capture the voice of data consumer. It would be really helpful to know about their expectations when developing a framework or tool to assess DQ, so that a high level of data quality could be achieved in any organization. With this approach the first survey has been conducted with users of TU Delft and Pondicherry University, India.

3.2.1 First Survey

An empirical approach to DQ analyzes data gathered from data consumers to determine the characteristics they use to assess whether data are fit for use in their tasks [6]. The purpose of this survey is to capture the DQ expectations of university portal users. An application specific selection of criteria would help us to identify qualitatively good data [1]. So, UWP users as data consumers have been asked to produce a list of possible dimensions that comes to their mind when they think about data quality in UWP according to Wang and Strong approach of DQ survey.

- **User group**

Users of two universities were selected for this survey such as Pondicherry University, India and Delft University of Technology, The Netherlands. For the users of Pondicherry University, the questionnaire was sent to 100 students via email explaining the nature of the survey with an html link to the survey designed with a Google document. The targeted groups were M.Sc., PhD students and professionals from different departments such as Computer science, Computer applications, and Bioinformatics. For the users of Delft University, the survey was conducted directly with 29 students from different department of EEMCS faculty: Electrical engineering, Information Architecture, Computer Science, Computer Engineering and Microelectronics. However the majority of the respondents are users of TU Delft as compared to 23 respondents from Pondicherry University.

- **Survey Instrument**

The survey has been designed with Google documents. First some general information about the users have been asked, like their category of study such as M.Sc., or PhD student, course program, sex, and name of institution, country. Following that data quality aspects have been asked with the three main questions. The first question was about DQ dimensions that they consider important when they think about DQ other than timeliness, accessibility, accuracy. In the second question, they were asked to provide any additional dimensions by listing out few of the dimensions gathered from the literature, as cues [6]. The third question asked their opinion about the assessment of DQ in UWP is highly needed or not, with two options. The questionnaire can be found in Appendix B

- **Result**

This survey produced a comprehensive list of 81 DQ dimensions, as shown in Table 7.

Access speed	Amount of data	Applicable	Attractive
Author of information	Availability	Better service	Clarity
Clear	Clearness	Comparative	Compherensenability
Comphrensive	Completeness	Concise format	Consistent of data
Continuity	Correctness	Cost effective	Cost of the data collection
Costs of the data	Creative	Credibility	Data clarity
Delivery	Depedency of data	Detail of the data	Ease of access
Ease of understanding	Easy to find	Easy to operate	Easy to read
Easy to search	Easy to understand	Easy to use	Efficiency
Findable	Flexible	Format	Identifiable
Importance	Indexable	Indexed searchability	Information importance
Informative	Innovative	Integrity	Knowledgeable
Legibility	Minimal data	No lack of data	Non-repetition
Ordered	Organized	Presentation format	Provable
Readability	Recent/old data	Relevance	Relevant information
Reliability	Representable	Searchable	Secure
Security	Service improvement	Serviceability	Simple
Simplicity	Source of the data	Structure layout	Truly
Trustworthy	Understandability	Units of measurement	Unneeded data
Uptodate information	Validity	Well organized	Well documented
Well presented			

Table 7: Data quality dimensions from data consumer perspective in first stage survey

The above list shows different DQ dimensions from user's perspectives. A same dimension or an attribute has been stated differently by different users according to their level of understanding. Hence the dimension with similar meanings has to be grouped together under a common DQ dimension name from the literature. After several runs of grouping tasks the following DQ dimensions were identified. The following attributes reflect the DQ expectations of the data consumers (UWP users)

1. Readability (Easy to read, readable)
2. Relevancy (applicable, Relevant)
3. Searchability (Easy to search, searchable, identifiable, indexed Searchability, easy to find)
4. Organized (Ordered, well-organized, arranged, well-presented, structured layout, presentation format)
5. Cost effective (Cost of the data, cost of data collection)
6. Data Clarity (Clear, clarity, clearness, correctness)

7. Provenance (Source of the data, where is the data from, author of data)
8. Accessibility (Ease of access, easy to use, access speed)
9. Believability (trustworthy, truly, credibility)
10. Uniqueness (Non repetition, minimality)
11. Documentation (Well-documentation)
12. Continuity (Flow of data)
13. Concise representation (concise, format of the data, representable)
14. Serviceability (Better service, improved service)
15. Reliability (Reliable, source reliable)
16. Appropriate amount of data (Amount of data, non-bulk data)
17. Security (Secure)
18. Attractiveness
19. Novelty (Innovative, creative, knowledgeable)
20. Simplicity (Simple)
21. Efficiency
22. Validity
23. Consistent representation (Consistent of data, same format of data)
24. Currency (Up-to-date, how old the data)
25. Completeness (Comphrensibility, complete data)
26. Understandability (Ease of understanding, Easy to understand)
27. Importance

Out of these 27 attributes, a few dimensions were found to be new when compared to the collection of DQ dimensions from phase 1 on different contexts of the web. These dimensions were not entirely new according to the DQ literature, but they were found absent in several DQ works dedicated to different application domains. Thus these data dimensions were decided to consider it along with the dimensions collected from the literature in phase 1 and also it reflects the expectations of university portal users. They are as follows

Readability	Provenance	Efficiency
Organization	Flexibility	Attractiveness
Cost-effective	Continuity	Novelty
Data clarity	Searchability	Simplicity

Table 8: Selected dimensions from data consumers

As a result, 49 dimensions were acquired (37 dimensions from literature and 12 from the user survey). Considering so many dimensions will not be efficient for any DQ evaluation purposes. Hence it has been decided to compare the DQ dimensions obtained both from the literature and the selected dimensions from the data consumers via the first survey. Thus 49 dimensions have been examined again for similarities in their names and meanings to avoid conflicts. This process of examining is repeated until a final DQ dimension set is reached.

3.3 DQ dimensions for UWP domain

In the third phase, DQ dimensions that are applicable to assess DQ in the context of UWP's were identified. This is one of the criteria to be considered while selecting DQ dimensions [1]. To check this appropriateness of DQ attributes, a portal data quality model [28] has been used, a model developed specifically for web portals. In addition to the review of DQ dimensions, this model focuses on two perspectives, to check the appropriateness of reviewed DQ dimensions in the context of web portals: the data consumer expectation of DQ and basic web portal functionalities.

However this data quality model proposal is generic and not specific to any web portal domain, so those two perspectives in this model have been adapted and extended in the context of UWP properties to analyze how far the collected DQ dimensions are applicable to a UWP. Moreover the portal functionalities of this model has been renamed as properties since functionality describes how a web portal functions or what does it performs in order to meet the needs of the user where as properties are certain features provided by a portal.

Based on the data consumer expectations and UWP properties, the DQ dimensions applicable to UWP were identified. The basic idea behind this relation is that the users of a UWP can assess the data quality when they use the different services or properties provided by the web portal [28] [29]. The following section will first give a short description about the UWP properties and data consumer expectation and then relationship between these two aspects will be described

3.3.1 University web portal properties

Users of a UWP come to know about the quality of data based on their properties provided in the portal. Thus if we know the basic properties of a UWP, then it would be helpful to identify which aspects of DQ are important for users on assessing DQ. Properties of a UWP may vary between different universities and there is no standard proposal for such properties. The web portal properties provided by the portal data quality model [28] are rather general descriptions of properties. Instead specific properties need to be determined and sorted under these general properties. The focus of this work is on the UWP domain, hence basic web portal properties applicable to universities had been chosen from literature [7] [9] [10] [11] and also by studying different university portals online. These properties are as follows:

- Data points and Integration
- Single sign on
- Personalization and Customization
- Collaboration and Communication
- User interface design
- Search and index
- Schedules and events
- Help/Support
- Registrations
- Notification and News
- Security
- Content management

3.3.2 Data consumer expectations

The data consumer expectation of DQ is another important aspect in selecting dimensions appropriate for a UWP domain. Using these data consumer expectations and portals properties we can determine what the data consumer expects from the portal and to choose DQ dimensions accordingly. The resulting dimensions will be used to achieve a high quality of data in a UWP. There are six different data consumer expectations as suggested by Redman [8], discussed below:

- **Content**

The web portal should contain explicit description to its users about the areas that the web portal content covers, its authors, original sources used etc. The provided content should satisfy the user with all needed data relevant for their tasks. In addition it should also describe how the content of the web portal should be used by its users to achieve effective usage of the web portal.

- **Privacy**

Data consumers expect that the web portal should explicitly state and follow their privacy policies concerning web portal users.

- **Improvement**

Data consumers expect the service improvement of web portal to have high quality of data. Comments and suggestions on different aspects of the web portal should be received, and these comments have to be considered in a responsible manner. It should also indicate the results of any improvement measures to provide high quality of service

- **Commitment**

Data consumers of the web portal should be able to receive answers, for any questions regarding the proper use of the data or the content of the portal. The responsible data manager of the web portal should provide help facilities to guide users in their tasks.

- **Presentation**

The content of the web portal should be well-presented to its users with an easy understandable format and the choice of language has to be clear with proper definitions wherever necessary and if there are any difficult technical terms used thorough out the portal, then they have to be defined properly to increase the interpretability of the users

- **Quality of values**

The provided data values should be accurate and updated with current information. In addition it also has to be relevant to the context with data formats that properly convey the data and are easy to read, allowing users to understand the data without ambiguity.

However these expectations were also experienced in the first survey results of data quality among the UWP users, where users are more concerned about the content, presentation and improvement aspects.

3.3.3 Relationship between expectations and properties

Based on the two above mentioned aspects (data consumer expectations and UWP properties) the following matrix was formed (Figure 5). The matrix is represented with UWP properties on one side and data consumer expectations of DQ on the other side. The relationship between these two aspects is marked with an “X” symbol showing the possible expectations of the data consumer on each of the university web portal properties. For instance (Figure 4), the expectations applicable to search and index property can be *content, commitment, improvement, presentation, and quality of values*, where a university web portal user needs real content to be displayed while giving a specific search query (*content*), in case of any errors in searching they should be able to ask question and obtain answers (*commitment*), users should be able to give their suggestions and comments on data they search through the UWP (*improvement*), the answer for a given search query should be presented in a clear understandable manner *(presentation)*, and the answer displayed should be accurate, complete and up-to-date for a given search query (*Quality of values*)

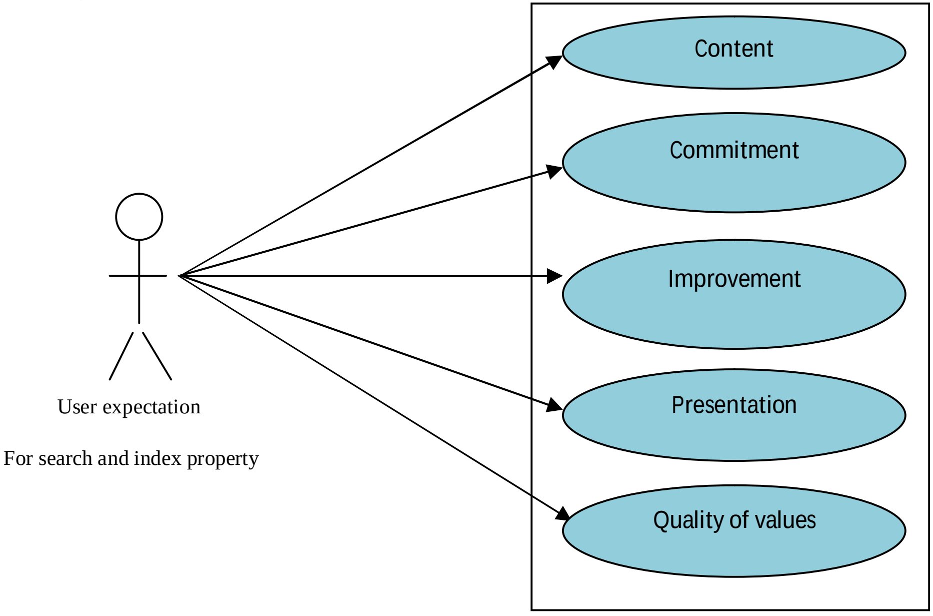


Figure 4: Expectations applicable to search and index property

University web portal properties

Category of Data Consumer expectations	Collaboration and communication	Content Management	Data Points and Integration	Help/Support	Notifications and News	Personalization and Customization	Registrations	Schedules and Events	Search and Index	Single sign-on	Security	User interface design
Content		X	X		X		X	X	X			X
Commitment	X	X	X	X					X	X		X
Improvement		X	X		X		X		X			X
Presentation		X	X	X	X			X	X		X	X
Privacy	X	X				X				X	X	
Quality of Values		X	X			X	X	X	X	X	X	X

Figure 5: Matrix for the UWP dimensions classification

- **Collaboration and Communication:** The possible expectations assigned to this property are: *Commitment* (user should expect questions regarding the use of collaboration and communication tools or subjects related to it and should be answered for the same), and *Privacy* (user should expect privacy policies to be mentioned for all users those who participate in the collaboration and communication process).

- **Content Management:** The possible expectations assigned to this property are: *Content* (Users expect descriptions about the areas covered by the web portal to see that the required data is provided for the tasks at hand), *Commitment* (user should be able to get answers about proper use of data or update schedules etc.),*Improvement* (user suggestions, comments of contents and its management should be received and any improving result should be reported), *Presentation* (the content should be easy to interpret in form of its languages, units, and data formats), *Privacy* (privacy policies for the users to get access to information sources has to be provided), and *Quality of values* (users expect the content should be correct, relevant and up-to-date).

- **Data Points and Integration:** The possible expectations assigned to this property are: *Content* (information resources or application integrated to the real content of a portal should satisfy the user in the form of appropriateness, clear and original sources, etc.), *Commitment* (users should be able to get answer for questions raised in the area of the resources or applications integrated to the portal), *Improvement* (results on any improvements about the data use, or integration resources), *Presentation* (languages, definitions used should be easy to interpret the data), and *Quality of values* (the presented information resources should be correct, up-to-date and reliable).

- **Help/Support:** The possible expectations assigned to this property are: *Commitment* (User should be able to get answers about different aspects in the web portal regarding proper use or meaning of data, update schedules etc.), and *Presentation* (Help text manuals should be easy to understand in terms of clarity, format, language etc.).

- **Notification and News:** The possible expectations assigned to this property are: *Content* (Content regarding any notification and news published should be relevant and clear), *Improvement* (Opinions / Comments of published notification or news should be easy to report by the user and any improving results should be noted), and *Presentation* (Presentation of notification and news should be clear and easily interpretable in a form of language accepted by the user).

- **Personalization and Customization:** The possible expectations assigned to this property are: *Privacy* (The user should expect privacy over the personalized data, in the UWP), and Quality *of values* (Data about the user such as personal information, study progress should be correct and up-to-date).

- **Registrations:** The possible expectations assigned to this property are: *Content* (Users should be able to find the data as correct, relevant and comprehensive to several form of registrations and presence of description of how the data should be used for the task at hand), *Improvements*

(Suggestions about the registration content or processes should be received) , and *Quality of values* (Data related to registration aspects in UWP should be correct, relevant, and up-to-date to the needs of the user).

- **Schedules and Events**: The possible expectations assigned to this property are: *Conten*t (Data regarding schedules such as class or exam schedules should be clear and easily understandable), *Presentation* (The presentation format, language for providing the schedules or events should satisfy the intended use), and *Quality of values* (Schedules and events should be correct, complete and up-do date).

- **Search and index:** The possible expectations assigned to this property are: *Content* (Results for the search query should be appropriate for the intended use and complete), *Commitment* (Errors encountered at the time of searching should be able to be reported by the user*),* *Improvement* (users should be able to give their suggestions and comments on data to a given search query), *Presentation(*the answer for a given search query should be presented in a clear understandable manner), and *Quality of values* (search results should be accurate, complete and up-to-date for a given search query).

- **Single sign-on**: The possible expectations assigned to this property are: *Commitment* (Users should be able to get answers for errors or problems encountered during single-on), and *Privacy* (User should expect privacy to information access with single-sign on).

- **Security**: The possible expectations assigned to this property are: *Presentation* (Data about the security aspects should be well-formatted with a widely accepted language), *Privacy* (Privacy policy regarding the level of access to various information resources should be stated).

- **User interface design**: The possible expectations assigned to this property are: *Content* (The design of the user interface should very well portray different aspects about the content of the UWP), *Commitment* (Users should expect answers raised for the questions about), *Improvement* (Users should be able to convey their opinion on interface design of the portal), *Presentation* (Presentation formats, language and other aspects related to presentation should be given more important to the interface design, to make it appropriate for intended use by its users), and *Quality of values* (The data that makes up the user interface design should be correct, relevant, and complete).

3.3.4 Assignment of DQ dimensions applicable to UWP

The DQ dimensions gathered from both DQ literature and the first user survey has been compared, analyzed and reduced to 39 DQ dimensions. Based on the relationship (properties, expectation) in the above matrix, those 39 DQ dimensions have been analyzed for appropriateness in the UWP context. This has been done by assigning appropriate DQ dimensions with data consumer expectations for each portal property. As we stated before, users of a UWP can assess the data quality of the web portal based on the provided properties. With this approach table 9 shows the expectations for several UWP properties and DQ dimensions applicable to this relationship:

Data Consumer Expectations	**University web portal properties**	**DQ dimensions assigned for expectations and properties relation**
1.Content	• Content Management • Data Integration • Notification and News • Registrations • Schedules and events • Search and Index • User Interface Design	Concise representation, Searchability, Organized, Currency, Availability, Provenance Ease-of-operation, Interpretability, Relevancy, Completeness, Timeliness, Reputation, Expiration, Accuracy, Appropriate amount of data, Readability, Understand ability, cost-effectiveness, Value-added
2.Commitment	• Collaboration and Communication • Content Management • Data Integration • Help/Support • Search and index • Single sign-on • User interface design	Accessibility, Reliability, Timeliness, Availability Security, User support, Understandability, Organized, Novelty
3. Improvement	• Content Management • Data Integration • Notification and News • Registrations • Search and Index • User Interface design	Accessibility, Relevancy, Reliability, Completeness Documentation, Searchability, User support, Serviceability, Availability
4. Presentation	• Content Management • Data integration • Help and Support • Notification and News • Schedules and events • Search and Index • Security • User Interface Design	Concise representation, Consistent representation, Accessibility, Appropriate amount of data, Completeness, Ease-of-operation, Understandability, Attractiveness, Uniqueness, Readability, Data Clarity, Availability, Flexibility, Organization, User Support
5. Privacy	• Collaboration and Communication • Content Management • Personalization and Customization • Single-sign on • Security	Security, Response time, Understandability, Confidentiality, Expiration, Objectivity
6. Quality of values	• Content Management • Data Integration • Personalization and Customization • Registrations • Schedules and events, • Search and Index • Single sign-on • Security • User Interface Design	Accuracy, Availability, Provenance, Relevancy, Searchability, Accessibility, Believability, Understandability, Concise representation, Objectivity, Completeness, Uniqueness, Value-added, Response time, Verifiability, Data Clarity, Documentation

Table 9: Assignment of DQ dimensions applicable to UWP with data consumer expectations and UWP properties

3.3.5 Validation of DQ dimensions

The DQ dimensions chosen for a UWP in the above mentioned phase was validated with five users (three master students and two computer science lecturers) based on their experience as web portal users. Instructions were provided to users for this testing of DQ dimensions, by asking them to remove the DQ dimension that does not suit under a category or to add dimensions that need importance in a category. Definitions for each DQ dimensions were also provided along with the instructions to the users, in order to have clear results. Based on the feedback collected from the user, DQ dimensions were added, deleted and revised in each category after which 37 dimensions remained.

The validation was done in order to make sure whether the chosen dimensions are applicable from the perspective of UWP users, as we can identify the required dimensions by relating to the properties of the portal used by the data consumers. Furthermore, it was carried out to realize the dimensions that are not applicable to the context. Although this validation was based on the aspects of existing portal data quality model, the results showed required DQ dimensions to assess the data quality in a UWP domain. Figure 6 shows the resulting DQ dimensions applicable to UWP

Accessibility
Accuracy
Appropriate amount of data
Attractiveness
Availability
Believability
Completeness
Concise representation
Confidentiality
Consistent representation
Cost-effectiveness
Currency
Customer Support
Data clarity
Documentation
Ease of operation
Expiration
Flexibility
Interpretability
Novelty
Objectivity
Organization
Provenance
Quality of Service
Readability
Relevancy
Reliability
Reputation
Response time
Searchability
Security
Timeliness
Understandability
Uniqueness
Validity
Value added
Verifiability

Figure 6: Final set of DQ dimensions for the framework

3.4 The Proposed DQ framework

Thus the proposed framework is developed to assess DQ in university web portals. The DQ dimensions obtained so far has been classified into four categories in the framework. This classification of DQ dimensions is based on Wang and Strong DQ framework [6], where there are 15 DQ dimensions from the perception of data consumers. This framework has been chosen as it is considered as the most significant and popular classification of DQ dimensions in the data quality literature. Based on the definitions [6] of each category, the DQ dimensions in the new framework are categorized with four categories: *Intrinsic, Representational, Contextual, and Accessibility*.

The *Intrinsic* DQ category consists of dimensions that evaluate the quality that the data has by itself. The dimensions *completeness* and *timeliness* were given importance to be categorized under intrinsic category since it is necessary to have all data about the university in a UWP and also should be timely available to its users. The *Representational* DQ category considers dimensions that the systems must provide to its users that are interpretable, understandable, readable, and it should also be represented concisely and consistently. The *Contextual* DQ category considers dimensions that evaluate quality within the context of the data used by the data consumers. This is the only category in the framework that has more essential DQ dimensions. The *Accessibility* DQ category refers DQ dimensions related to the accessibility of the data and the level of security.

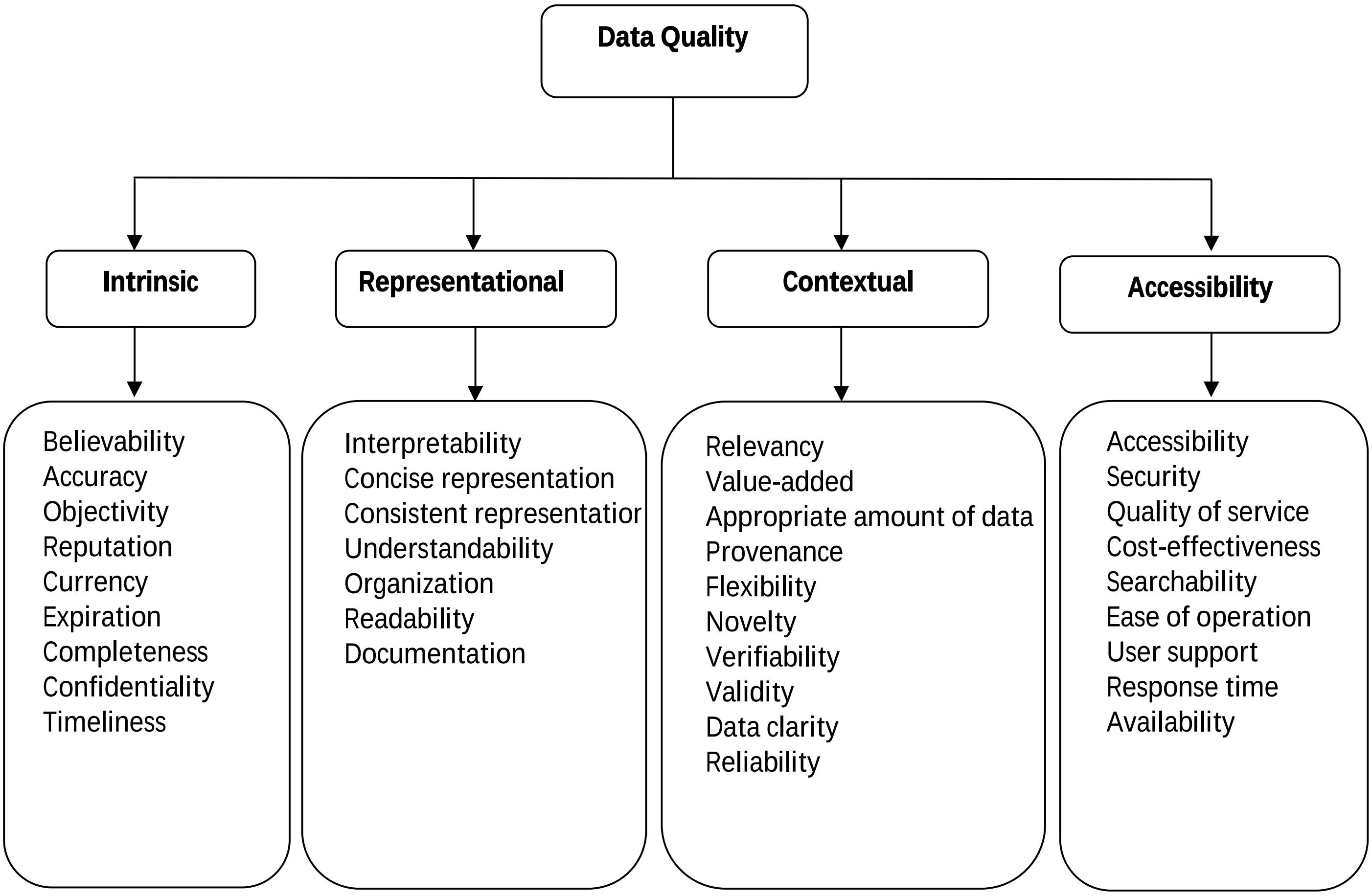


Figure 7: The proposed DQ framework

This framework was tested by means of a survey with users of two different universities to assess how far the framework is wanted among the users of university web portals. User groups of two universities have been selected and asked to rate the importance of each dimensions in the framework, which will be discussed brief in the next chapter. The results of this validation showed that all the selected dimensions are highly essential among the users of a UWP.

Furthermore the proposed framework is evaluated with an existing portal data quality model, a generic model for web portals. Though this model is not entirely specific to the UWP domain, it provides a comprehensive set of dimensions, which was also taken into account in this new data quality framework with additional DQ dimensions which will be discussed in chapter 5.

Chapter 4

Framework validation

This chapter describes how the proposed framework is validated based on the results of the survey conducted at two different universities (Technical University of Delft, The Netherlands, Pondicherry University, India,). The objective of this survey was to gather user opinion about the importance of each of the DQ dimensions in the proposed framework. The following sections of this chapter will explain about the framework validation in detail.

4.1 Selection of participants

To validate the new framework a sample of 150 users were chosen for this survey from Pondicherry University, and 60 users from TU Delft University. The sample consisted of users from different departments, however the main contributors for this survey were Computer Science students of both universities. The selected samples were M.Sc., PhD students and professionals from both universities with experience as portal users. Users of the UWP will be more concerned about the quality of data to carry out their individual tasks. So, the data consumers selected to validate the framework are users of UWP's.

4.2 Validation Instrument

The validation instrument here is the survey questionnaire, which was designed to collect the importance of DQ dimensions under each category in the proposed framework. The survey questionnaire was developed with 37 questions; one for each DQ dimension. Each of the questions stated the DQ aspect in a clear way. The questionnaire asked the respondents to rate the importance of each DQ dimension on a Likert scale of 1-5 where 1 is 'not important' and 5 is 'highly important'. For the users of Pondicherry University, the questionnaire was sent via online with the help of Google document application, designed for preparing questionnaires. For the users of TU Delft, the questionnaires were distributed directly to the users, in a printed format and the purpose of the survey and time period to answer the questionnaire were explained to the users.

Before conducting the full study using the survey questionnaire, a pilot study was conducted to validate the questionnaire as such in order to get reliable responses. Ten users from both the universities were selected to participate in this pilot study. 6 of these 10 users were participated in this pilot study: 3 M.Sc., students from Computer Science department, 2 PhD students from Management Studies and 1 software professional. The pilot study helped to identify the feasibility of the questionnaire and the following comments were reported in order to improve the design of the survey questionnaire.

Findings in pilot study

- Time taken to complete the whole questionnaire were reported as the maximum
- Different method of scaling were suggested based on the statement of each DQ aspects in the questionnaire and also due to the inconsistency felt between the statements and the scaling
- A Few statements were considered as difficult to understand and unclear
- A Few statements were identified as irrelevant to the context of DQ aspects.

As a result of these findings, the survey questionnaire was revised and changes were made, in order to help users in understanding the questions and to show their importance easily with less time.

4.3 Data Analysis

The survey has been performed with a total expected sample of 210 users. In practice, the questionnaire was answered by 116 users from Pondicherry University, as the remaining students did not participate, so the response rate was 77%. In Delft University, the same questionnaire was answered by 52 users, as the remaining users did not attend it, so the response rate was 87% which was 10% higher than PU. The participation of users from both the universities are shown below

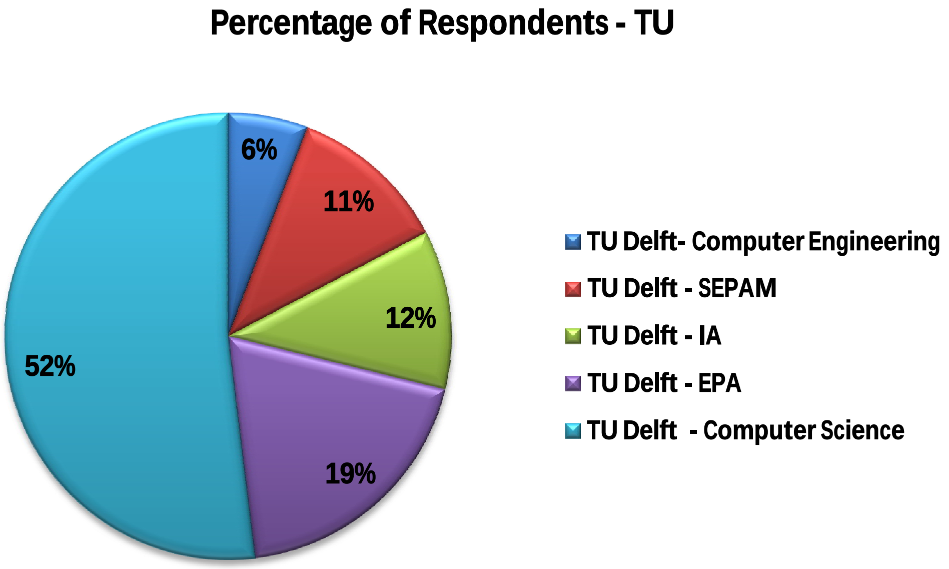


Figure 8: Percentage of respondents - TU Delft University

From TU Delft, 52 % of the respondents were from Computer Science department, 19% of the respondents were from Engineering and Policy Analysis department, 12% of the respondents were from Information Architecture department, 11% of the respondents were from SEPAM (Systems Engineering, Policy and Management) and 6 % of the respondents were from Computer Engineering department. This shows that the majority of the respondents are Computer Science students from EEMCS faculty.

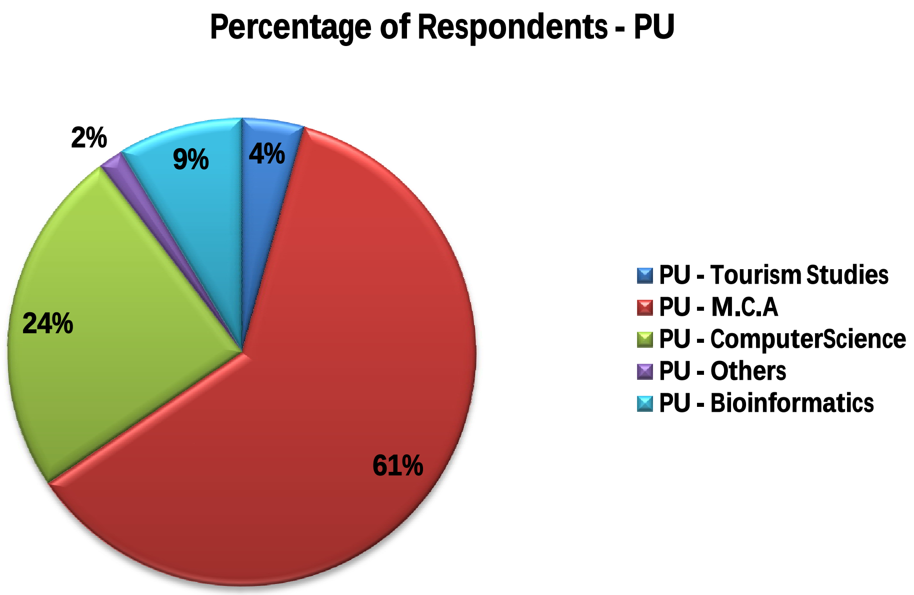


Figure 9: Percentage of respondents - Pondicherry University

From the figure 9, we can see the percentage of respondents from Pondicherry University, India: 61% of the respondents were from Master of Computer Applications department (MCA)-PU, 24 % of the respondents were from Computer Science department-PU, 9 % of the respondents were from Bioinformatics-PU, 4% of the respondents were from tourism studies-PU, and 2 % of the respondents were from other departments such as commerce and finance departments respectively. So, the majority of the respondents are MCA students from Pondicherry University. The percentage of overall user's category are shown below, where most of the respondents were M.Sc., students.

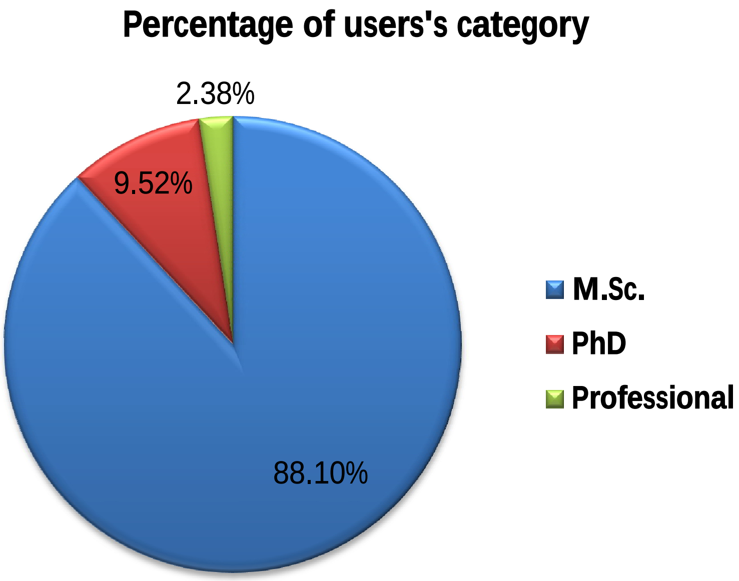


Figure 10: Percentage of user's category

Statistical analyses were performed using SPSS 17.0 for windows. The responses from the Google document application were exported to SPSS statistical package and all the responses have been screened to check the existence of any unanswered questionnaire. Responses from the both the universities were sorted according to the name of the institution and merged as a single set of data. The responses for 37 DQ dimensions of both universities were arranged and analyzed carefully, according to the number of items in each DQ category.

4.3.1 Descriptive Statistics Interpretation

In order to analyze the survey responses, first descriptive statistics of the responses were computed using SPSS package by selecting the 37 dimensions under each category in the framework. Descriptive statistics of 37 DQ dimensions are presented in Table 10 with columns N, Minimum, Maximum, Mean and Standard deviation, where N represents the No. of respondents or cases, Minimum and Maximum represents the range of the values for each DQ dimension and Standard Deviation represents a measure of the dispersion of a set of data from its mean value. Probably, the most often used descriptive statistics is the mean average value. A dimension mean was computed as the average of the responses to all of the items in the survey instrument.

Most of the 37 items had a range of 1 - 5 where “1” is not important and “5” is highly important. There are few dimensions: Interpretability, Understandability, Attractiveness, Verifiability, Data clarity, Searchability, Ease of operation, Believability, and Accuracy, where no one considered these DQ dimensions as “Not important”.

The range provided here represents the likert scale value given to the questions. From this we can also identify which items were considered as highly important or moderately important to users. The values for above described columns are showed in the descriptive statistics table below (Table 10).The individual descriptive statistics tables for both the universities are available in the Appendix E and Appendix F.

Table 10: Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
1.Interpretability	168	2	5	4.21	.649
2. Concise representation	168	1	5	4.02	.796
3.Consistent representation	168	1	5	3.76	1.018
4. Understandability	168	2	5	4.27	.740
5.Organization	168	1	5	4.05	.953
6. Attractiveness	167	2	5	3.78	1.048
7. Uniqueness	167	1	5	3.69	1.074
8. Readability	168	1	5	4.11	.841
9. Documentation	167	1	5	4.14	.835
10.Value-added	167	1	5	4.01	.780
11.Relevancy	168	1	5	4.29	.727
12. Appropriate amount of data	168	1	5	4.10	.831
13.Provenance	168	1	5	3.61	1.021
14. Flexibility	168	1	5	3.79	.972
15. Novelty	168	1	5	3.70	1.002
16. Verifiability	166	2	5	4.07	.818
17. Validity	168	1	5	3.99	.954
18. Data clarity	168	2	5	4.26	.744
19.Reliability	166	1	5	4.40	.785
20. Security	167	1	5	4.22	.899
21.Quality of service	167	1	5	4.02	.846
22. Accessibility	167	1	5	4.23	.821
23. Cost-effectiveness	165	1	5	3.99	.950
24. Searchability	167	2	5	4.40	.728
25. User Support	166	1	5	4.03	.950
26. Response time	165	1	5	4.13	.823
27. Availability	165	1	5	4.23	.746
28.Ease of operation	167	2	5	4.04	.787
29.Believability	168	2	5	4.21	.825
30.Accuracy	167	2	5	4.41	.632
31. Objectivity	165	1	5	4.02	.883
32. Reputation	168	1	5	3.90	.891
33. Currency	167	1	5	4.39	.835
34. Expiration	166	1	5	4.01	.918
35. Completeness	167	1	5	4.06	.848
36. Confidentiality	167	1	5	4.35	.776
37. Timeliness	168	1	5	4.30	.732

On the other hand, the mean value of most of the DQ dimensions was higher than 4. That is, thirty-one of the 37 items (83 % of dimensions) had a mean value equal or greater than 4. This shows that most of the items surveyed were considered to be important DQ dimensions applicable to the UWP domain. Based on these results a hypothesis has been made: DQ dimensions which had a mean of 3.5 or more would be considered as valid DQ dimension to the proposed framework. Dimensions that did not fulfill this condition would be removed from the framework. This mean value of 3.5 has been taken since it is the midpoint of the scale range 1-5 (considering the scales “moderately important”, “important”, and “highly important”)

From table 10, we can identify that mean value of most of the DQ dimensions are above 4 showing high level of importance by the users. However, there are a few dimensions with mean value of above 3.5 and below 4: *Consistent representation, Attractiveness, Uniqueness, Provenance, Flexibility, Novelty, Validity, Cost-effectiveness, and Reputation*. Out of these 9 dimensions, *Validity, Cost-effectiveness*, are considered to be highly important, since they are closer to mean value 4 with 3.99, 3.99, respectively. Though the mean value of other DQ dimensions (*Consistent representation, Attractiveness, Uniqueness, Provenance, Flexibility, Novelty and Reputation)* are above 3.5, they are likely to be moderately important. To be specific, data consumers of UWP not considered these dimensions as very essential data quality aspects, though the range is between 2 and 5. DQ dimensions with mean value of 4 and above are considered to have a high level of importance and are presented in table below (Table 11).

No	Dimension Category	DQ Dimensions	Mean value	Level of importance
1	**Intrinsic**	Believability	4.21	High
		Accuracy	4.41	High
		Objectivity	4.02	High
		Reputation	3.90	Moderate
		Currency	4.39	High
		Expiration	4.01	High
		Completeness	4.06	High
		Confidentiality	4.35	High
		Timeliness	4.30	High
2	**Representational**	Interpretability	4.21	High
		Concise representation	4.02	High
		Consistent representation	3.76	Moderate
		Understandability	4.27	High
		Organization	4.05	High
		Attractiveness	3.78	Moderate
		Uniqueness	3.69	Moderate
		Readability	4.11	High
		Documentation	4.14	High
3	**Contextual**	Value added	4.01	High
		Relevancy	4.29	High
		Appropriate amount of data	4.10	High
		Provenance	3.61	Moderate
		Flexibility	3.79	Moderate
		Novelty	3.70	Moderate
		Verifiability	4.07	High
		Validity	3.99	High
		Data clarity	4.26	High
		Reliability	4.40	High
4	**Accessibility**	Security	4.22	High
		Quality of Service	4.02	High
		Accessibility	4.23	High
		Cost-effectiveness	3.99	High
		Searchability	4.40	High
		User support	4.03	High
		Response time	4.13	High
		Availability	4.23	High
		Ease of operation	4.04	High

Table 11: DQ dimensions with level of importance

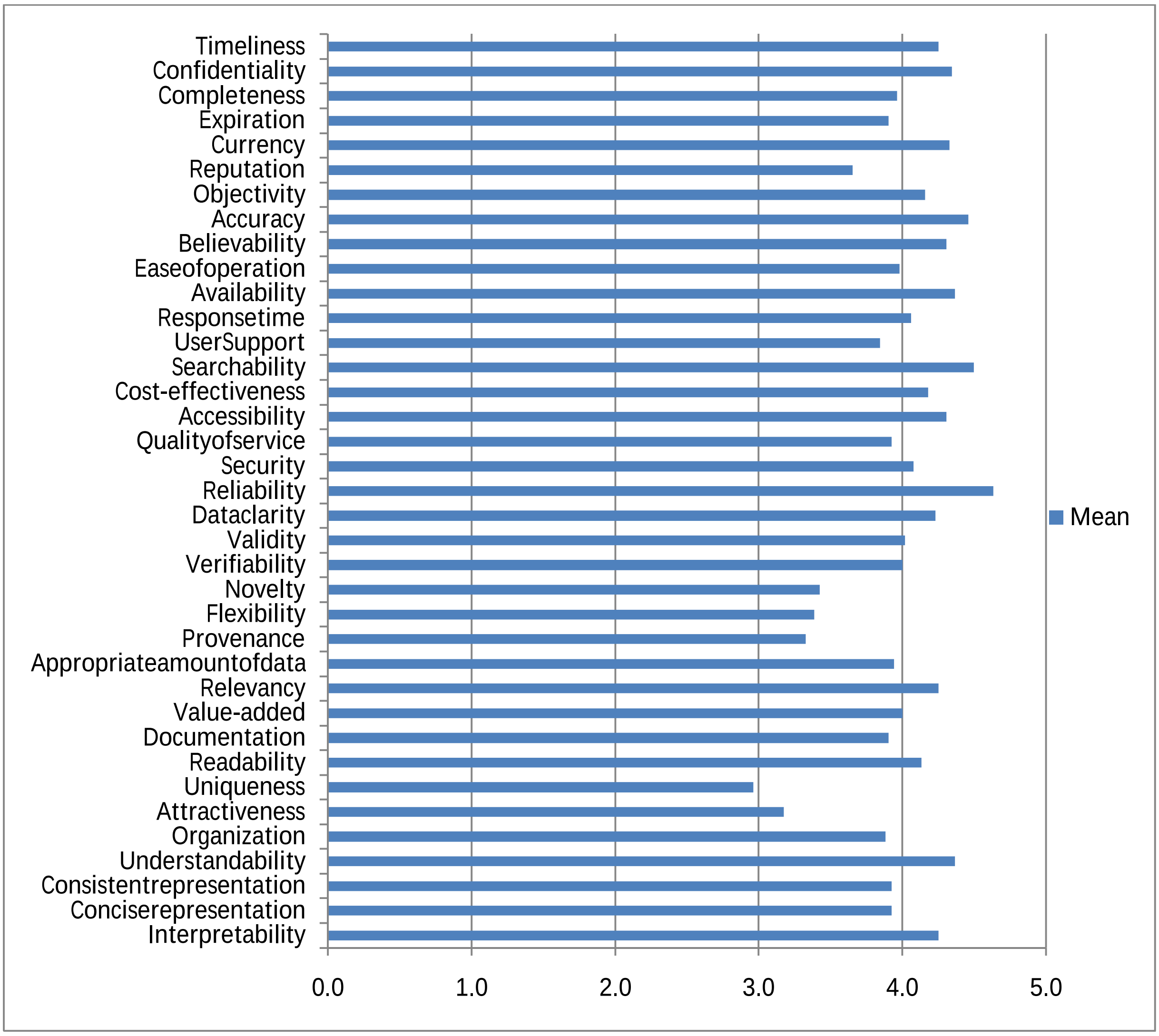


Figure 11: Importance of DQ dimensions

Descriptive analyses of the DQ dimensions are represented in a graph format to have a clear view. As the graph in Figure 11 shows, the mean value of most of the DQ dimensions are above 4, showing more importance to the UWP domain. In particular, the highly rated DQ dimensions are *Timeliness, Confidentiality, Currency, Accuracy, Searchability, Reliability, and Understandability* with a mean value close to 4.5. The rest of the items are above 4 or close to 4. On the other hand, as discussed above the graph shows a few dimensions that fall within a mean value of 3.5 to 4 which are considered as moderately important.

The level of importance has been verified with two other graphs drawn separately for both the universities (Appendix G and H). Those two graphs also reflected the same level of provided importance for each DQ dimensions with minor variations. The results of descriptive statistics analysis showed that all DQ dimensions in the proposed framework are significant to assess or evaluate the DQ of a UWP. As discussed above, a few dimensions had a mean value of 3.5 to 3.7; *Consistent representation, Provenance, Flexibility, Novelty, Uniqueness, and Attractiveness.* Though these dimensions are absent in most of the DQ literature, they are considered as moderately important by the users. This shows that besides the other important DQ dimensions, data consumers (users) of a UWP also expect moderate importance to these items in the framework.

4.3.2 Reliability Analysis Method

Though the descriptive statistics analysis showed the average score of the responses for each DQ dimension, it does not show the consistency of the responses by the users for different DQ dimensions. Thus, reliability analysis method was selected to check the consistency of the items in the survey questionnaire. To analyze the data, a statistical package named SPSS (versions 17.0) has been used as mentioned before.

A very common measure of reliability analysis in the research literature is Cronbach's alpha. Construct reliability of the DQ dimensions was tested using this Cronbach alpha. In this approach, each item is correlated with every other item in the instrument. Cronbach‘s alpha, a measure of construct reliability was computed for each DQ dimensions in the framework to assess the degree of internal consistency among a set of items (Survey questionnaire questions). Cronbach alpha values generally range between 0 and 1. When the alpha value is close to 1, then the level of internal consistency of the items in the questionnaire is high. Generally, alphas of 0.70 or above represent satisfactory reliability of the set of items measuring the dimension.

The reliability statistics table (Table 12) indicates that Cronbach's alpha for the 37- item scale is 0.866, indicating good reliability. The values which range from 0.86 to 0.85 indicate that the measures of each DQ dimension are reliable (Table 14). The statistical reliability analysis table displays information about the scale as if it were calculated without each item. This gives some information on how individual items contribute to the whole. The information that are important in this reliability analysis are given below with short descriptions:

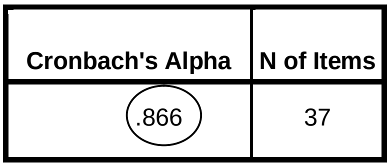
a) Overall alpha: The Cronbach alpha was constructed for the 37 items. Table 4 shows the overall alpha value as .866 which is good and indicates high internal consistency among the 37 items. This means that the respondents who selected high scores for one item also intended to select high scores for some other items. Similarly respondents who selected low scores for one item intended to select low scores for some other items. If this value is low, then the consistency between the scores from one item and the other items is poor.

b) Corrected Item-Total correlation: This column displays the correlation between one item and the sum score of the other items in the survey instrument.

c) Cronbach's alpha if item deleted: This column displays Cronbach's alpha that would result if a given item were deleted. This column of information is very important for determining which items from among a set of items contribute to the total alpha value. For instance, omitting item E*ase of operation* would result in a reliability of .859. This is a decrease from calculation with all the other items (.866). Thus we consider that this item *Ease of operation* to be useful and contribute positively to the total Cronbach's alpha value. On the other hand, if the alpha is higher than the overall score, then the item should be removed.

Table 12: Reliability statistics

(Overall Cronbach's Alpha)

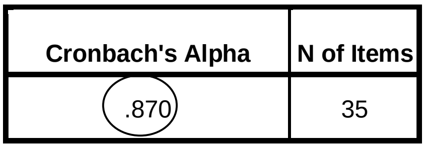


Cronbach's Alpha	N of Items
.866	37

Although the overall Cronbach value is .866, two DQ dimensions such as *Attractiveness* and *Uniqueness* does not show good consistency with the rest of the items and also their Cronbach's alpha value is .868 and .867 respectively (Table 14) which is greater than the overall Cronbach value (Table 12). Though they showed moderate importance in the analysis of descriptive statistics, the same result does not end up with reliability analysis which is important to measure the reliability of each dimension. If the alpha value of any item in the scale is higher than the overall Cronbach's alpha value then the item should be removed and the analysis should be re-run to get a high score for all overall reliability analysis. Thus the two DQ dimensions have been removed and the analysis was repeated. The following table shows the new overall Cronbach's value realized after re-run of the analysis.

Table13: Reliability Statistics

(Overall Cronbach's Alpha – new)



Cronbach's Alpha	N of Items
.870	35

This shows that there is an increase in the overall Cronbach alpha .866 to .870 (Table 13) after the removal of those two dimensions indicating that the measures of each dimension are reliable. The new statistical result of overall reliability analysis is found in Appendix D.

Consequence: The number of dimensions in the framework decreases from 37 to 35 using this method.

Table 14: Statistical Results for overall Reliability Analysis of proposed framework

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1.Interpretability	146.65	169.517	.182	.866
2. Concise representation	146.83	167.938	.217	.866
3.Consistent representation	147.13	165.673	.236	.866
4. Understandability	146.55	166.047	.354	.863
5.Organization	146.83	162.426	.394	.862
6. Attractiveness	147.06	167.386	.164	.868
7. Uniqueness	147.13	165.432	.235	.867
8. Readability	146.72	163.653	.392	.862
9. Documentation	146.68	162.488	.471	.861
10.Value-added	146.85	165.298	.353	.863
11.Relevancy	146.54	168.049	.258	.865
12. Appropriate amount of data	146.73	163.647	.424	.862
13.Provenance	147.32	162.071	.380	.863
14. Flexibility	147.02	164.583	.309	.864
15. Novelty	147.17	160.663	.451	.861
16. Verifiability	146.77	164.136	.397	.862
17. Validity	146.85	160.797	.470	.860
18. Data clarity	146.59	164.513	.428	.862
19.Reliability	146.45	164.839	.361	.863
20. Security	146.64	166.071	.257	.865
21.Quality of service	146.83	161.576	.487	.860
22. Accessibility	146.65	164.874	.348	.863
23. Cost-effectiveness	146.86	164.068	.327	.864
24. Searchability	146.41	165.223	.412	.862
25. User Support	146.81	160.842	.488	.860
26. Response time	146.72	163.116	.426	.862
27. Availability	146.63	167.254	.270	.865
28.Ease of operation	146.82	161.196	.551	.859
29.Believability	146.65	164.566	.371	.863
30.Accuracy	146.44	166.678	.355	.863
31. Objectivity	146.87	165.217	.303	.864
32. Reputation	146.95	161.172	.473	.860
33. Currency	146.46	163.713	.391	.862
34. Expiration	146.83	161.616	.453	.861
35. Completeness	146.76	164.788	.349	.863
36. Confidentiality	146.51	165.527	.333	.864
37. Timeliness	146.52	165.808	.351	.863

Representational DQ:

The Cronbach's alpha values of items under representational category shows good internal consistency range between .865 and .870, where .870 is the overall alpha value (Table 15). The items *Interpretability, Concise representation, Consistent representation* have alpha values equal to Cronbach's alpha value (.870) showing better internal consistency of responses. However, their item-total correlation are low, particularly item interpretability has a very weak correlation with .188. The rest of the items such as *Understandability, Organization, Readability and Documentation* have a lower Cronbach's alpha than the overall alpha. In addition, the corrected item-total correlation scores are also good. This shows these items have a better consistency of responses and correlation between each item and the rest of items in the scale. To be specific, item *Documentation* has a good correlation (.449) with rest of the items and it has a low Cronbach's value showing good consistency of responses. Thus all the dimensions of representational DQ category are reliable to assess DQ in university web portals.

	DQ dimensions	**Scale Mean if Item Deleted**	**Scale Variance if Item deleted**	**Corrected Item-Total Correlation**	**Cronbach's Alpha if Item Deleted**
Representational DQ	1. Interpretability	139.17	156.999	.188	.869
	2. Concise representation	139.35	155.584	.217	.869
	3. Consistent representation	139.65	153.009	.251	.870
	4. Understandability	139.07	153.700	.358	.867
	5. Organization	139.34	150.642	.377	.866
	6. Readability	139.23	151.536	.388	.866
	7. Documentation	139.19	150.774	.449	.865

Table 15: Reliability analysis results for representational category

Contextual DQ

All the items of contextual category show a very good consistency of responses with alpha values lower than overall Cronbach's alpha. The Cronbach's alphas of this category range from .864 to .867 (Table 16). The corrected item-total correlation of all the items are >.3 except the item *Relevancy.*

The item *Relevancy* has a low correlation score of .263 with Cronbach's alpha of .865 which is less than the overall alpha. Thus the item has a good internal consistency despite of low correlation score between items. In a nutshell, all items have a good and positive internal consistency of responses and the correlation between an item and the rest of the scales are also satisfactory except relevancy. Thus all items of Contextual category are considered as important to be included in the framework, in order to assess the DQ of a UWP. Thought the items *Provenance, Flexibility, Novelty* showed a moderate importance by users in descriptive statistics analysis method, here the reliability analysis showed that there is a better consistency of responses among these items. Thus they can be considered as important dimensions for assessing data quality in UWP's.

	DQ dimensions	Scale Mean if Item Deleted	Scale Variance if Item deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Contextual DQ	8. Value-added	139.36	152.997	.356	.867
	9. Relevancy	139.05	155.608	.263	.868
	10. Appropriate amount of data	139.25	151.328	.430	.865
	11.Provenance	139.83	149.925	.380	.866
	12. Flexibility	139.53	153.016	.280	.869
	13. Novelty	139.68	149.186	.425	.865
	14. Verifiability	139.29	152.179	.384	.866
	15. Validity	139.37	148.704	.471	.864
	16. Data clarity	139.10	152.050	.442	.865
	17.Reliability	138.96	152.227	.380	.866

Table 16: Reliability analysis results for contextual category

Accessibility DQ

The items of accessibility category show a good consistency of responses with Cronbach's alpha range from .862 to .869 which are lower than overall alpha value (Table 17). The corresponding item-total correlations indicate that there is a positive correlation between the scores on the one item and the combined scores of the other items in the scale. Except the items *Security* and *Availability*, the rest of the items has a score >.3, which is considered as a better correlation between items.

A rule of thumb is that the corrected item-total correlation should preferably be .30 or higher. Otherwise there is a possibility of weak correlation between an item and sum of the other items in a scale. In this accessibility category, the items Security and Availability has a low correlation score of .250 and .286 respectively. This indicates that there is no significant correlation between these items and sum of rest of the items in the scale. However there is a strong and positive internal consistency of responses with Cronbach's alpha range of .868 and .869 respectively. Thus when compared to the representational category, items of accessibility category shows a good internal consistency of the responses, influencing these items to be included in the framework to assess the data quality efficiently.

	DQ dimensions	Scale Mean if Item Deleted	Scale Variance if Item deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Accessibility DQ	18. Security	139.15	153.943	.250	.869
	19. Quality of Service	139.35	149.369	.492	.864
	20. Accessibility	139.16	152.270	.366	.866
	21. Cost-effectiveness	139.37	151.390	.347	.867
	22. Searchability	138.92	152.732	.427	.865
	23. User support	139.32	149.132	.470	.864
	24. Response time	139.23	151.106	.418	.865
	25. Availability	139.15	154.636	.286	.868
	26. Ease of operation	139.33	148.975	.558	.862

Table 17: Reliability analysis results for accessibility category

Intrinsic DQ

In this category, all items show a strong and positive consistency of responses with a Cronbach's alpha range from .864 to .867 (Table 18). These are quite less than the overall Cronbach's alpha value, indicating high level of internal consistency of the responses. Their corrected item-total correlation are also good with a range >.3, showing good correlation between the scores on one item and the combined scores of other items in the scale. Thus all items of intrinsic category have more importance to be included in the proposed DQ framework. Among the four categories in the framework, intrinsic category is the only category indicating a strong and positive consistency of responses on items without any weak or low correlation.

	DQ dimensions	Scale Mean if Item Deleted	Scale Variance if Item deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Intrinsic DQ	27.Believability	139.16	152.014	.387	.866
	28. Accuracy	138.95	154.219	.365	.867
	29. Objectivity	139.38	152.425	.328	.867
	30. Reputation	139.47	149.338	.460	.864
	31. Currency	138.97	151.422	.395	.866
	32. Expiration	139.35	149.530	.452	.864
	33. Completeness	139.27	152.307	.361	.866
	34.Confidentiality	139.03	153.073	.343	.867
	35. Timeliness	139.03	153.589	.348	.867

Table 18: Reliability analysis results for intrinsic category

The results of reliability analysis showed that the DQ dimensions in the proposed framework are valid indicating good consistency of responses and correlation between the items. Few dimensions under representational category indicate inconsistent responses. Particularly the item *Attractiveness* had a Cronbach's alpha value greater than the overall alpha value. Furthermore, the corrected-item correlation also showed a weak correlation. This highlights that the item *Attractiveness* has a low consistency compared to the other items in the representational category. All the other categories in the proposed framework produced consistent results and it's therefore reliable. Thus the proposed framework contains 35 important DQ dimensions necessary to assess the DQ in a UWP.

Chapter 5

Evaluation of the framework

This chapter introduces an approach for the evaluation of the proposed DQ assessment framework for a UWP with an existing DQ model designed for generic web portals. The validation results of the new DQ framework showed that all DQ dimensions proposed in the framework are highly important from the perspective of the data consumers (users of University web portals). This framework has been developed to help the web portals designers or developers to assess and improve DQ in UWP. Data quality (DQ) or Information quality (IQ) is vital in a UWP. Yet, despite a decade of research in DQ, web portal lacks comprehensive methodologies for DQ assessment and improvement. Till now, there is no standard DQ model that is applicable to all domains, hence DQ works have been carried out to fulfil the needs of each domain.

This new framework has been realized after studying significant DQ models, frameworks [1] [6][18][20][23], and fields that conducted DQ research with specific dimensions. Especially empirical analysis has been carried out to capture the dimensions that are important to data consumers (users of University web portals) as well as to validate the dimensions proposed in the framework. In such a way, the new framework has been designed based on the expectation and perception of UWP users.

Whenever a framework is designed, we often experience a question like, how can we determine whether the designed framework is suitable to the context? On the other hand, there are no methods available to determine which framework would be the best in the given context. Though it is critical to answer this question, it has been decided to compare this new framework with any related framework in the literature, since there are no specific methodologies for framework evaluation. However only few researches have been carried out in the context of web portals, particularly no work addresses the DQ in a UWP. Similar studies have been proposed in a broader perspective, though not specific to universities. One among those studies is portal data quality model (PDQM), a generic DQ proposal made in the context of web portals [28]. A short description about PDQM is provided below:

5.1 Model for evaluation

The model acts as a starting point for DQ research in the domain of web portals. It consists of two main phases. The first phase of this model provides a set of DQ attributes (here dimensions were mentioned as attributes) to assess the DQ in web portals [28]. To create this set of attributes, the model made an extensive study on several DQ research works from literature. With the help of existing DQ attributes from the literature the model filtered out attributes relevant for web portals using the functionalities of web portals.

Though this model is not specific to any web portal domain, it can be tailored to include and modify the DQ dimensions according to the domain of interest. Hence in the new framework, the dimensions applicable to a UWP were chosen based on two significant aspects of this model such as data consumer expectations and basic web portal functionalities which we discussed in Chapter 3. The other part of this web portal data quality model, is to provide a tool to assist the developer and administrators to assess and maintain the DQ of web portals. This tool is yet to be completed to evaluate all categories of DQ dimensions, at this moment the tool implements the DQ evaluation only for representational category in the university web portal domain. Despite the fact of evaluating DQ dimensions under representational category in a UWP domain, the model does not provide any detailed validation methods to verify whether all DQ dimensions are highly acceptable in the context of UWP. The reason is that dimensions proposed for one domain may not be applicable to all domains of web portal. However the PDQM tool part will not be considered for this thesis work.

5.2 Comparison of the New Framework and PDQM

When we evaluate this new data quality framework with PDQM, we can identify newly added DQ dimensions in each category. These dimensions were not included in PDQM, but were found relevant and useful in the context of a UWP. The importance ratings provided for those dimensions by the users of university web portals are high except the dimension *provenance*, however it is considered as moderately important. The newly added dimensions in the proposed framework are as follows:

- ✓ Verifiability
- ✓ Data clarity
- ✓ Readability
- ✓ Quality of Service
- ✓ Cost-effectiveness
- ✓ Searchability
- ✓ Confidentiality
- ✓ Provenance

In the new framework, the dimensions were classified based on a widely accepted and suitable DQ conceptual model that was followed till now in almost all DQ works. Each category in the new framework has been evaluated with the category of PDQM, where the *accessibility* DQ category was called as operational. The following graphs were also created for the existing model along with the new framework based on their validation results. However the dimensions were clearly tested and validated in the new framework as compared to PDQM through reliability analysis, which is the most important analysis to be done for any survey analysis to check the reliability of the responses among the data consumers.

The existing portal data quality model lacks this essential validation. The following section compares the DQ dimension in both newly proposed framework and PDQM. Dimensions such as *Verifiability, Data clarity, Readability, Quality of service, Cost-effectiveness, Searchability, Confidentiality,* and *Provenance* do not exist in PDQM, however in the UWP domain these dimensions were considered as important DQ aspects to assess the data quality.

DQ category	New Framework	PDQM
Intrinsic	Believability Accuracy Objectivity Reputation Currency Expiration Completeness Confidentiality Timeliness	Accuracy Duplicates Believability Objectivity Reputation Traceability Expiration Currency
Contextual	Value-added Relevance Appropriate amount of data Flexibility Provenance Novelty Verifiability Validity Data Clarity Reliability	Validity Value-added Relevance Completeness Flexibility Novelty Reliability Completeness Applicability Specialization
Representational	Interpretability Concise representation Consistent representation Understandability Organization Readability Documentation	Concise representation Consistent representation Attractiveness Understandability Interpretability Appropriate amount of data Documentation
Accessibility	Security Accessibility Quality of Service Cost- effectiveness Searchability User support Response time Availability Ease of operation	
Operational		Availability Accessibility Security Response time Interactivity Ease of operation User support

Table 19: Comparison of DQ dimensions in the proposed framework and the portal data quality model

5.2.1 Intrinsic DQ

The dimensions introduced in this DQ category are ***completeness, confidentiality*** *and* ***timeliness*** (Figure 12). In PDQM, the ***completeness*** and ***timeliness*** dimensions were classified under contextual category, whereas in the new framework these two dimensions were found relevant to be under intrinsic category. As per the definition of intrinsic DQ, these two dimensions should be inherent to data provided by a UWP. For instance, we can relate a low completeness and timeliness value for the course schedules in a university, if such a schedule becomes available after the commencement of classes with incomplete data.

Confidentiality is another dimension which should be inherent to keep the personal data of students and staffs always confidential. Due to the privacy aspects, users of a UWP might intend to keep their information confidential. PDQM lacks this dimension in the intrinsic category (Figure 13). In the following graph of intrinsic DQ dimensions (Figure 12), we can see that the importance rating value given by users are ≥ 4. This shows that these dimensions in the new framework are necessarily important and consistent to assess the intrinsic DQ in university web portals compared to PDQM. The vertical axis of the graph indicates the importance ratings of each dimension in the framework and not the DQ assessment of any web portals. The horizontal axis represents the dimensions proposed in the new framework with different classifications. The coloring schema such as orange color bar in the graph indicates the dimension that is found new in the proposed framework and the blue color indicates dimensions that also exist in PDQM. The same case is applied to graph of PDQM where we can see two dimensions such as ***traceability*** and ***duplicates*** which were found to be absent in new framework.

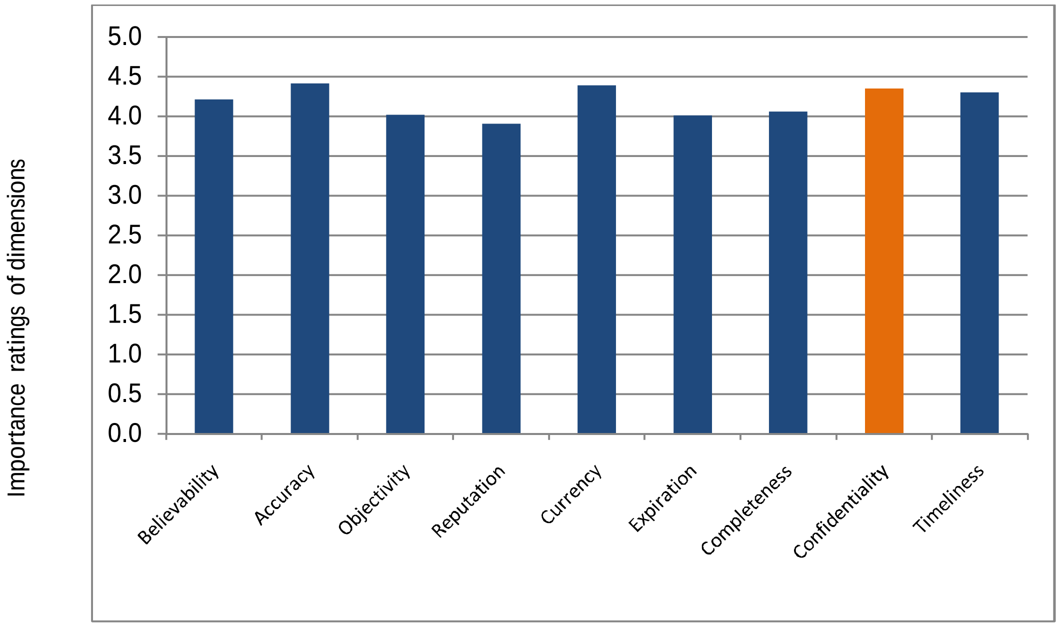


Figure1 2: Intrinsic DQ dimensions - New Framewrok

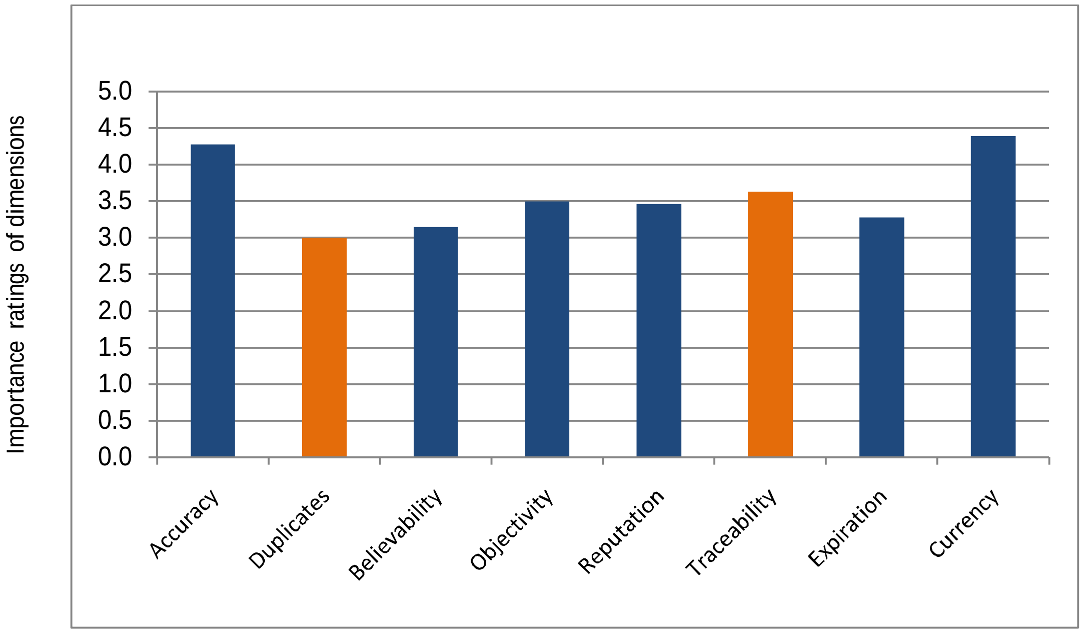


Figure 13: Intrinsic DQ dimensions **-** Portal Data Quality Model

Figure 13 shows the valuations of intrinsic DQ dimensions in PDQM. Though this validation is generic, we can see that most of the DQ dimensions were considered as moderately important except ***currency*** and ***accuracy.*** However in PDQM the dimensions that are above 3 (importance rating value) were considered as important to assess data quality in web portals.

5.2.2 Contextual DQ

The dimensions introduced in contextual DQ category are ***provenance, verifiability*** *and* ***data clarity*** (Figure 14). These were not included in PDQM, but when we want to assess DQ in UWP, these DQ dimensions should be taken into consideration. Since this new framework has been analyzed in the context of UWP, the dimensions were found applicable and useful to improve the contextual DQ. For instance, ***provenance*** is one of the DQ dimensions gaining increasing importance in information systems [4]. A UWP should provide the sources of available information such as where does the data come from or who is author of the information, when is created, etc., for its users to acquire more knowledge in the relevant areas or subjects. This would help the users to interpret the data semantics more accurately, and to resolve conflicts among the data retrieved from different sources. In addition, the dimension ***appropriate amount of data*** has been included in representational category of PDQM, whereas in the new framework it is reasonable to mention it under contextual category, as the amount of data depends on the context. According to the context the amount of data would be modified to satisfy the specific data consumers.

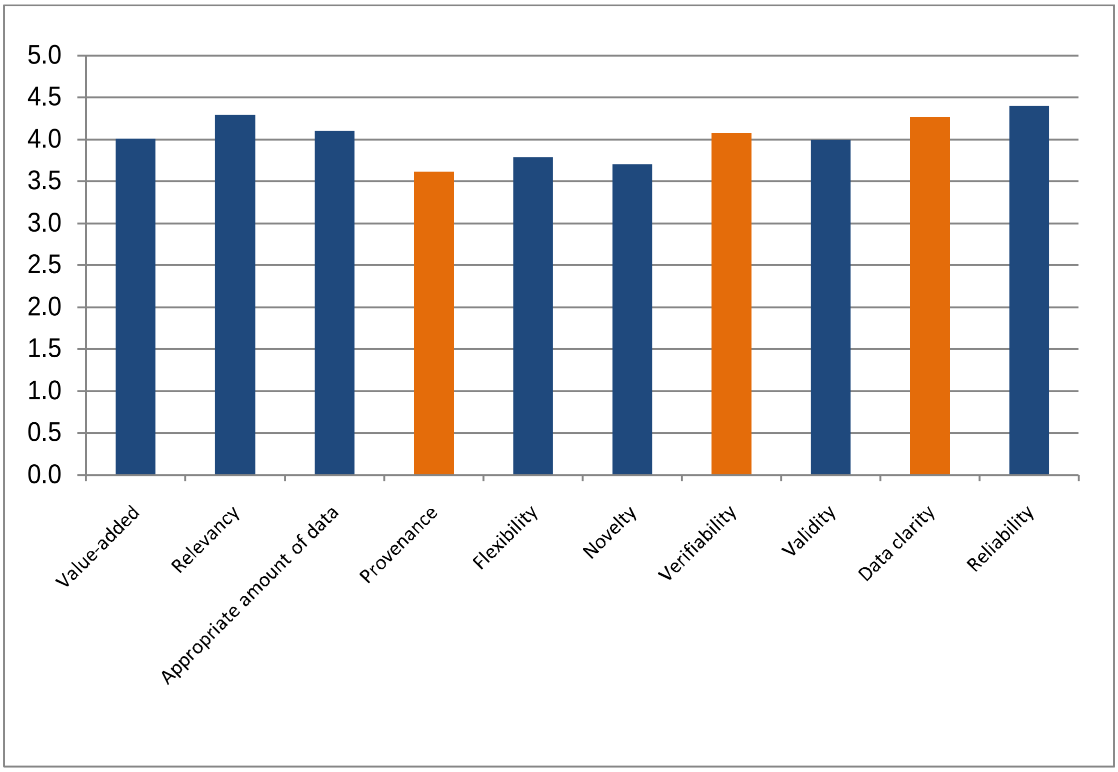


Figure 14: Contextual DQ dimensions – New Framework

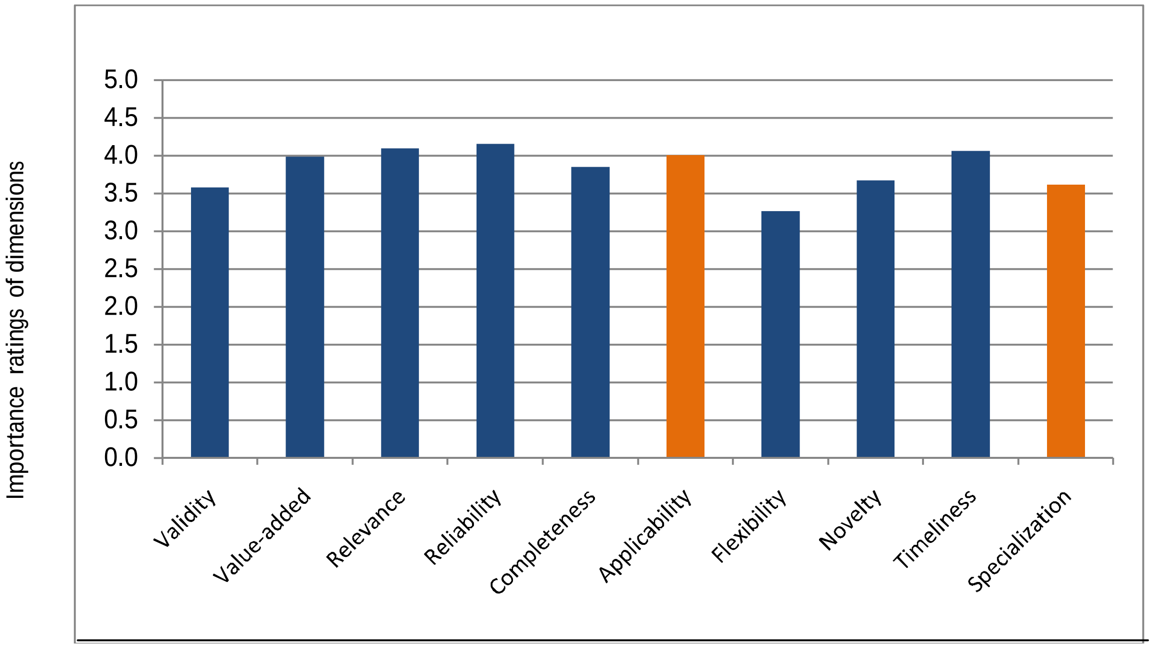


Figure1 5: Contextual DQ dimensions – Portal Data Quality Model

From the graph of new framework contextual DQ dimensions (Figure 14) we can see that this is the only category with a higher number of dimensions. Most dimensions have equal importance ratings among users of UWP with importance rating value greater ≥ 4, showing these are significant in assessing the contextual DQ. On the other hand, the dimensions ***provenance, flexibility and novelty*** were accepted as moderately important DQ dimensions, the same has been reflected in PDQM, where ***flexibility*** was even lower than new framework (Figure 15). This shows that users were not much concerned about these DQ dimensions; however in new framework these dimensions were resulted from the first survey carried out to capture the expectation of data consumers among users of two university web portals.

5.2.3 Representational DQ

Most of the dimensions in this category are same as in PDQM, however the dimension ***readability*** is highly important to reflect the meaning of real world information in a clear way. Like the other web portals, UWP should not be concerned only with the bulk of information which is the nature of all upcoming web portals. Instead it should give attention to the readability concept to make the data readable or easy to read and understand by means of consistent layout, common font size, color, line spacing between the lines etc. This was lacking in PDQM and therefore we can say PDQM is an incomplete representational DQ category. Hence importance has been expanded in the new framework with the *readability* dimension.

In addition, almost all DQ dimensions of this category have equal importance ratings among users of a UWP with importance rating value greater ≥ 4, except *consistent representation* dimension (Figure 16). Comparing this new framework with generic portal data quality model, we can see that most of the dimensions have moderate importance (Figure 17). This is caused by differences in quality requirements, i.e. the dimensions interested or essential for some domain of web portals may not be same for the other domains.

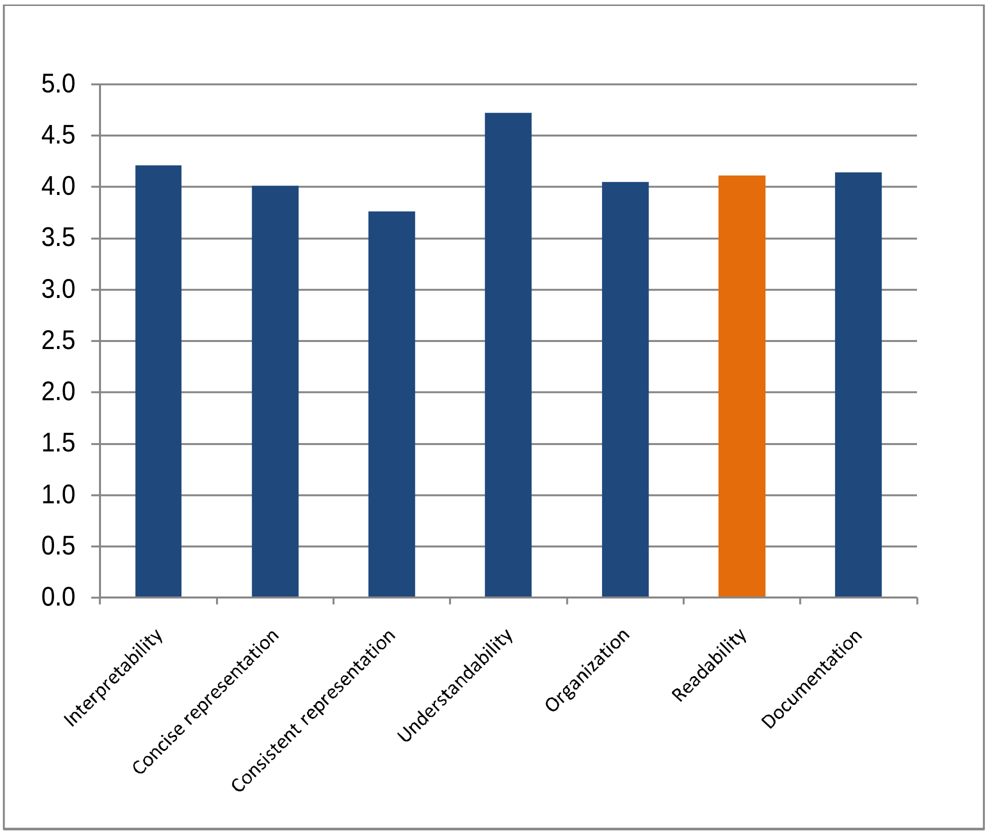


Figure 16: Representational DQ dimensions – New Framework

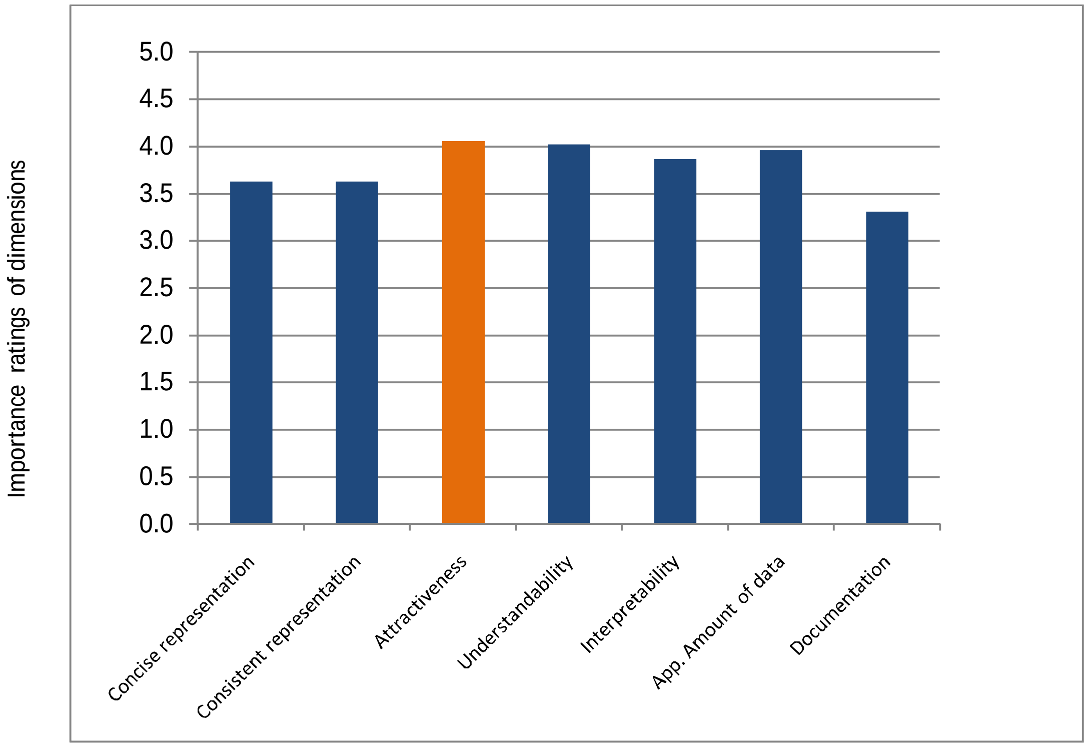


Figure 17: Representational DQ dimensions – Portal Data Quality Model

5.2.4 Accessibility Category DQ

In the new framework, three different DQ dimensions have been introduced in the accessibility category. They are: ***Quality of service, Searchability***, and ***Cost-effectiveness*** (Figure 18). Compared to PDQM, the accessibility category of the new framework has more valuable dimensions to assess the accessibility of data. For instance, *cost-effectiveness* is a highly important DQ dimension applicable for a UWP, because almost every UWP has integrated library access, where this dimension is highly wanted by students in accessing the paid articles and e-journals via university authorization. In PDQM this accessibility category is renamed as operational (Figure 19). In the graph of accessibility DQ dimensions-new framework, we can see that all dimensions are with importance rating value greater ≥ 4, showing these dimensions are necessary to assess the accessibility of UWP (Figure 18). No attribute was given moderate importance under the accessibility category in the new framework.

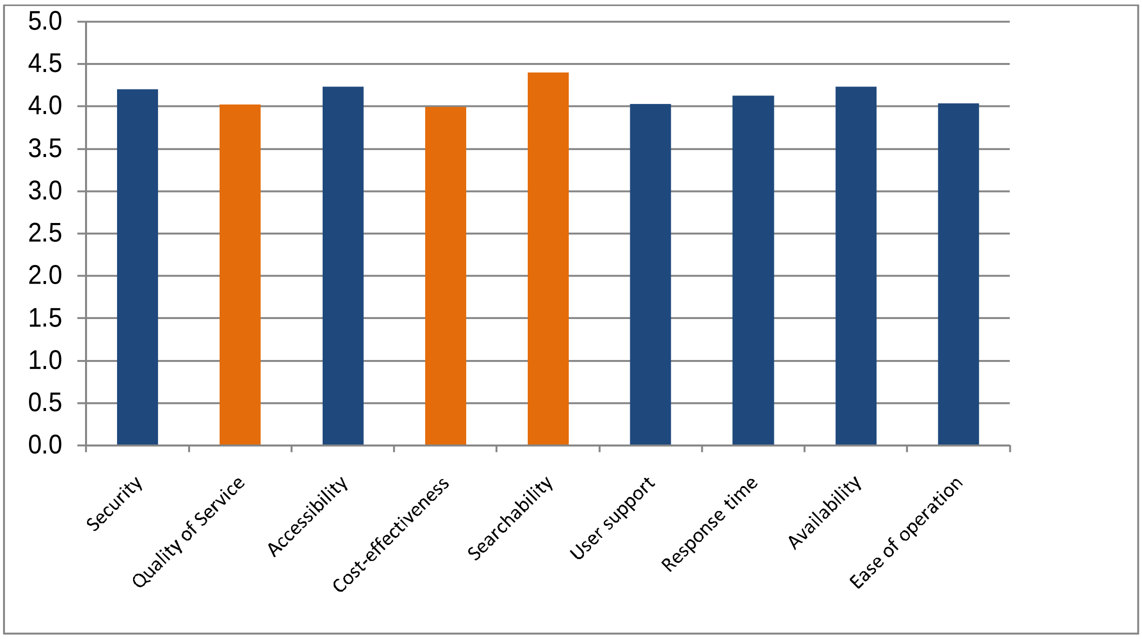


Figure 18: Accessibility DQ dimensions – New Framework

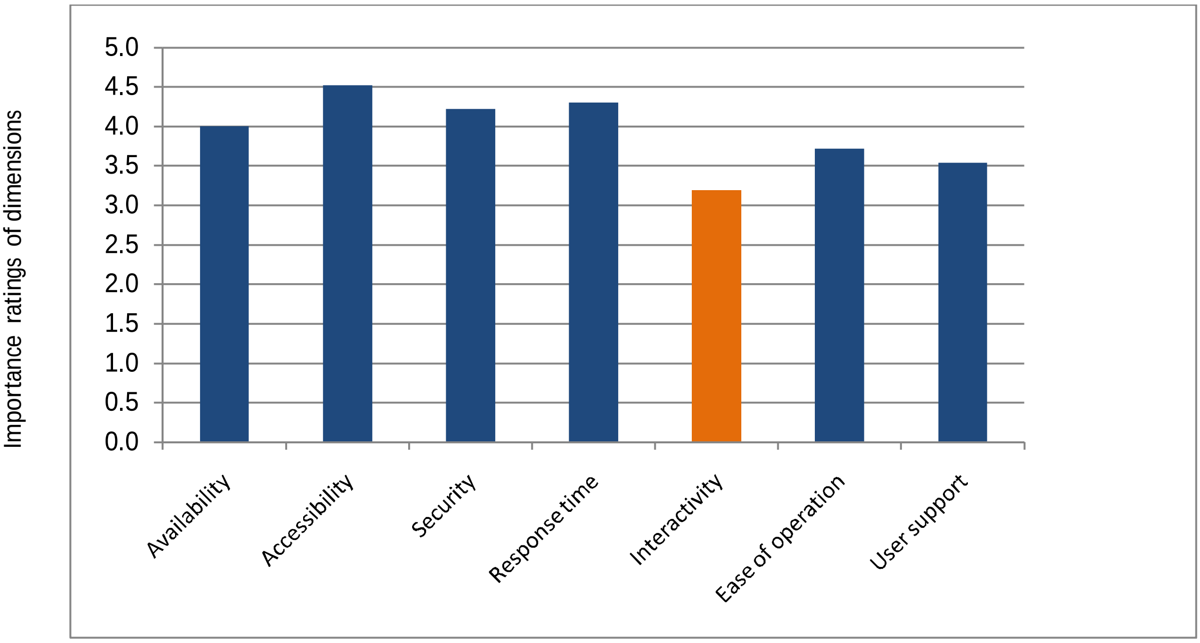


Figure 19: Operational DQ dimensions – Portal Data Quality Model

Apart from this classification of dimensions in the new framework, some dimensions of PDQM do not exist in the new framework. They are: ***duplicates, traceability, applicability, specialization, attractiveness** and **interactivity** (*orange colored bars in all Portal data quality model graphs***)**. In the new framework ***applicability*** has been excluded for the reason that it is more or less related with a relevancy DQ dimension and the dimensions ***specialization, traceability, duplicates*** and ***interactivity*** have been excluded since they were not considered to be applicable for evaluating the DQ of university web portals, moreover they were not widely discussed in DQ literature. However the dimension ***attractiveness*** chosen for representational category first, has been removed later since there was no reliability with the user's responses in the process of validation.

Chapter 6

Conclusion and Future work

The objective of this thesis was to develop a framework to assess DQ in university web portals. In order to develop this framework several DQ models and frameworks in different web contexts have been studied, to identify the essential DQ dimensions in the literature, the task which is essential in any DQ related work. It is not adequate to evaluate DQ by considering only the DQ dimensions in the literature. Indeed, it is essential to take into account the data consumer expectations or requirements, as DQ is the ability of a data collection to meet user requirements. In the context of web portals we have a variety of users who access the information and have different expectations about the quality of the data according to their needs. So, in addition to DQ dimensions from the literature, this work also considered the expectations of UWP users by means of a short survey conducted among users groups of two different universities: TU Delft, The Netherlands and Pondicherry University, India. The DQ dimensions gathered from the literature and the user survey has been compared and analyzed until a final set of DQ dimensions were obtained.

From this set of dimensions, DQ dimensions appropriate for UWP have been identified based on a process that considers the two important aspects: basic UWP properties and data consumer expectations on DQ. Currently there is no standard for assessing DQ in web portals, particularly no work has been dedicated in the context of the UWP domain. Hence a generic portal data quality model has been used for this process of obtaining DQ dimensions applicable for a UWP context. A first pre-validation has been carried out to validate the chosen DQ dimensions, which helped to discover the dimensions that are not applicable to the context. As a result, a framework has been designed and realized with essential DQ dimensions classified into four categories: *Intrinsic, Contextual, Representational, and Accessibility* to assess DQ in university web portals.

The proposed framework has been validated at two different universities: TU Delft -The Netherlands, Pondicherry University- India. The selected users groups for this survey: M.Sc., PhD students and professionals. To address this framework validation, a survey questionnaire has been prepared with 35 DQ dimensions in the framework to rate the importance of those DQ dimensions from the perspective of UWP users. Before sending the questionnaire to the user groups, a second pre-validation was conducted to validate the survey questionnaire with selected group of users to improve the formation of the questionnaire.

The results of the framework validation showed that most of the DQ dimensions have high level of importance and also there was a good consistency of responses among the users for most of the DQ dimensions. This result was observed through two methods: Descriptive statistics interpretation and reliability analysis method. The first method showed the level of importance of the DQ dimensions and the second method obtained the reliability of responses among DQ items in the questionnaire. Based on this validation, the proposed framework has been realized with 35 essential DQ dimensions to assess the DQ in UWP's.

Furthermore, as a process of evaluation this new framework has been evaluated with the generic portal data quality model (PDQM) to realize how far this framework is effective from the perception of similar existing model. Though the existing model produces a comprehensive set of dimensions for web portals in the DQ literature, it lacks essential DQ dimensions to assess DQ in a UWP domain. This new framework encapsulates all the essential DQ dimensions in the context of a UWP from the perspective of the users. In a nutshell, the new framework was found valuable and applicable in the context of UWP with ratings as important from the data consumers (for this case: the users of university web portals).

As a future work, this DQ framework could be improved by giving importance to specific properties of the university web portals. The properties mentioned in this thesis are the basic UWP properties, hence more specific properties could be taken into account and classified to identify significant DQ dimensions to assess DQ in a UWP domain. This could reduce the number of DQ dimensions and make the DQ assessment efficient. This framework would help the developer and administrator of the UWP to find the needs of their users and thereby make change and improve the portal to attain high quality of data. It would be interesting to create a tool to assess these DQ dimensions in the framework and exhibit the level of data quality of a certain UWP.

References

[1] Naumann, F (2002): Quality –Driven Query Answering for Integrated Information Systems, Springer Publications

[2] L. Pipino, Y. Lee, and R. Wang Data Quality Assessment (2002): Communications of the ACM, Vol. 45, pp: 211-218

[3] D. Strong, Y. Lee and R. Wang, Data Quality in context (1997): Communications of the ACM, Vol.40, pp: 103-110

[4] C. Batini, M. Scannapieco (2006): Data quality – Concepts, Methodologies and Techniques, Springer publications

[5] Y. Wand and R.Y. Wang (1996): Anchoring data quality dimensions in ontological foundations, Communication of the ACM, 39(11), pp: 86-95

[6] R.Y. Wang and D.M. Strong, Beyond Accuracy (1996): What Data Quality Means to Data Consumers, Journal of Management Information Systems 12, pp: 5-34

[7] A.Tatnall (2005): Web Portals – The New Gateways to Internet Information and Services, Idea group publications

[8] T.C. Redman (2001): Data Quality-Field Guide, Digital press Publications

[9] Dias (2001): Corporate Portals: a literature review of a new concept in information management, International Journal of Information Management, pp: 269-287

[10] Collins, H (2001): Corporate Portals: Revolutionizing information access to increase productivity and drive the bottom line, AMACOM publications

[11] Collins, H (2003): Enterprise knowledge portals: Next-Generation portal solutions for dynamic information access, better decision making, and Maximum Results, AMACOM publications

[12] C.A. Calero, C. Caballero, I. Piattini, M., (2007): Towards a Data Quality Model for Web Portals, Springer-Verlag Berlin Heidelberg

[13] R.Y.Wang, M. Ziad, Y.W. Lee., (2001): Data Quality, Kluwer Academic publishers

[14] S.E. Madnick, R.Y.Wang, Y. W. Lee., H. Zhu (2009): Overview and Framework for Data and Information Quality research, ACM Journal of Data and Information Quality, pp: 2.1-2.22

[15] B. Pernici and M. Scannapieco (2002): Data Quality in web information systems, Springer-Verlag Berlin Heidelberg

[16] M. Mecella, M. Scannapieco, A. Virgillito, R. Baldone, T. Catarci, C. Batini (2002) Managing Data Quality in Cooperative web information systems, Springer-Verlag Berlin Heidelberg

[17] J. Akoka, L. Berti-Equille, O. Boucelam, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoue-thion, Z. Kedad, S. Nugier, V. Peralta, S. Sisaid-Cherfi: A Framework for Quality evaluation in data integration systems, ICEIS 2007 - Proceedings of the 9th International Conference on Enterprise Information Systems

[18] Bovee M., Srivastava R. R., Mak B.R (2001): A Conceptual Framework and Belief Function Approach to Assessing Overall Information Quality. In proceedings of the 6th International Conference on Information Quality

[19] K. Sattler : Data Quality Dimensions (2009), Encyclopedia of Database systems, Springer Science+Business Media

[20] Yang W. Lee Leo L. Pipino James D. Funk Richard Y. Wang: Journey to Data Quality, 2006 Massachusetts Institute of Technology

[21] Caro, A et al (2006): A first approach to a Data Quality Model for Web Portals. Springer – Verlag Berlin Heidelberg.

[22] Naumann, F., & Rolker, C. (2000). Assessment methods for information quality criteria. In 5th International Conference on Information Quality, pp: 148-162

[23] Y.W.Lee, D.M.Strong, B.K.Khan, R.Y.Wang: AIMQ- A methodology for information quality assessment (2002), Information and Management, pp: 133-146

[24] Dudar, Z. and Medovoy, A. (2009): The Data Quality Estimation for the Information Web Resources. Proceedings of the 10th international conference of CAD systems in microelectronics, IEEE Xplore publications pp: 405-406

[25] Knight, S et al (2005): Developing a Framework for Assessing Information Quality on the World Wide Web. Journal of Informing Science, pp: 159-172

[26] M.Gertz, M.T. Ozsu, G.Saake, Kai-Uwe Sattler (2004): Data Quality on the web. SIGMOD Record, Vol.33, pp: 127-132

[27] T.C. Redman. Data Quality for the Information Age, Artech House, 1996.

[28] Caro, A., Calero, C. and Piattini, M. (2007): A Portal Data Quality Model for Users and Developers, In *ICIQ2007, The 12th International Conference on Information Quality*

[29] Caro, A., Calero, C., Caballero, I., Piattini, M.(2008): A proposal for a set of attributes relevant for Web portal data quality. Software Quality Journal 16, 513–542 (2008)

[30] M.Helfert, Eitel von Maur (2001): A Strategy for Managing Data Quality in Data Warehouse Systems. In proceedings of the International Conference on Information Quality, 2001

References for DQ dimensions from literature

P1. B. Pernici and M. Scannapieco (2002): Data Quality in web information systems, Springer-Verlag Berlin Heidelberg

P2. M. Mecella, M. Scannapieco, A. Virgillito, R. Baldone, T. Catarci, C. Batini (2002) Managing Data Quality in Cooperative web information systems, Springer-Verlag Berlin Heidelberg

P3. B. Piprani, D. Ernst (2008): A Model for Data Quality Assessment, Springer-Verlag Berlin Heidelberg

P4. M.Gertz, M.T. Ozsu, G.Saake, Kai-Uwe Sattler (2004): Data Quality on the web. SIGMOD Record, Vol.33, pp: 127-132

P5. M.Helfert, Eitel von Maur (2001): A Strategy for Managing Data Quality in Data Warehouse Systems. In proceedings of the International Conference on Information Quality, 2001

P6. V. Peralta, R. Ruggia, Z. Kedad, M. Bouzeghoub (2004): A Framework for Data Quality Evaluation in a Data Integration System, Proceedings of the 19th international conference

P7. J. Akoka, L. Berti-Equille, O. Boucelam, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoue-thion, Z. Kedad, S. Nugier, V. Peralta, S. Sisaid-Cherfi: A Framework for Quality evaluation in data integration systems, ICEIS 2007 - Proceedings of the 9th International Conference on Enterprise Information Systems

P8. M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella, R. Baldoni (2004): A platform for exchanging and improving data quality in cooperative information systems, Information systems 29, 551-582

P9. L. Pipino, Y. Lee, and R. Wang Data Quality Assessment (2002): Communications of the ACM, Vol. 45, pp:211-218,

P10. M. B. Parker, V. Moleshe, R. De la Harpe, R. G. B. Wills (2006): An evaluation of Information quality frameworks for the World Wide Web, In 8th Annual Conference on WWW Applications, pp: 1-11

P11. M. G. Fugini, M. Mecella, P. Plebani, B. Pernici, M. Scannapieco (2002): Data Quality in Cooperative Web Information Systems, Kluwer Academic Publishers

P12. Y.W.Lee, D.M.Strong, B.K.Khan, R.Y.Wang: AIMQ- A methodology for information quality assessment (2002), Information and Management, pp: 133-146

P13. C. Cappiello, C. Francalanci, B. Pernici (2004): Data quality assessment from user's perspective, IQIS - Proceedings of the international workshop on Information quality in information systems, pp: 68-73

P14. Y. J. Kim, R. Kishore, G. Lawrence Sanders (2005): Understanding Data Quality in the context of E-Business system, Communications of the ACM, Vol.48, pp: 75-81

P15. K. Sattler : Data Quality Dimensions (2009), Encyclopedia of Database systems, Springer Science+Business Media

P16. N.K. Yeganeh, S. Sadiq, K. Deng, X. Zhou (2009): Data Quality Aware Queries in Collaborative Information Systems, Proceedings of the joint international conference on Advances in Data and Web Management, pp: 39-50

P17. M. Scannapieco, T. Catarci (2002): Data Quality under the Computer Science perspective, Journal of Archivi & Computer, Vol.2, pp:1-12

P18. Naumann, F (2002): Quality –Driven Query Answering for Integrated Information Systems, Springer Publications

P19. L. Berti-Equille (2007): Measuring and Modelling Data Quality for Quality-Awareness in data mining, Springer-Verlag Berlin Heidelberg, pp: 101-126

P20. M. Rese, G. Graefe, V. Herter (2004): Relevance as an Information Quality Problem in Corporate Decision-Making Processes, Proceedings of the 16th Conference on Advanced Information Systems, pp: 25-36

P21. C. S. Carson (2000): What is Data Quality? A Distillation of Experience, Statistics Department, International Monetary Fund

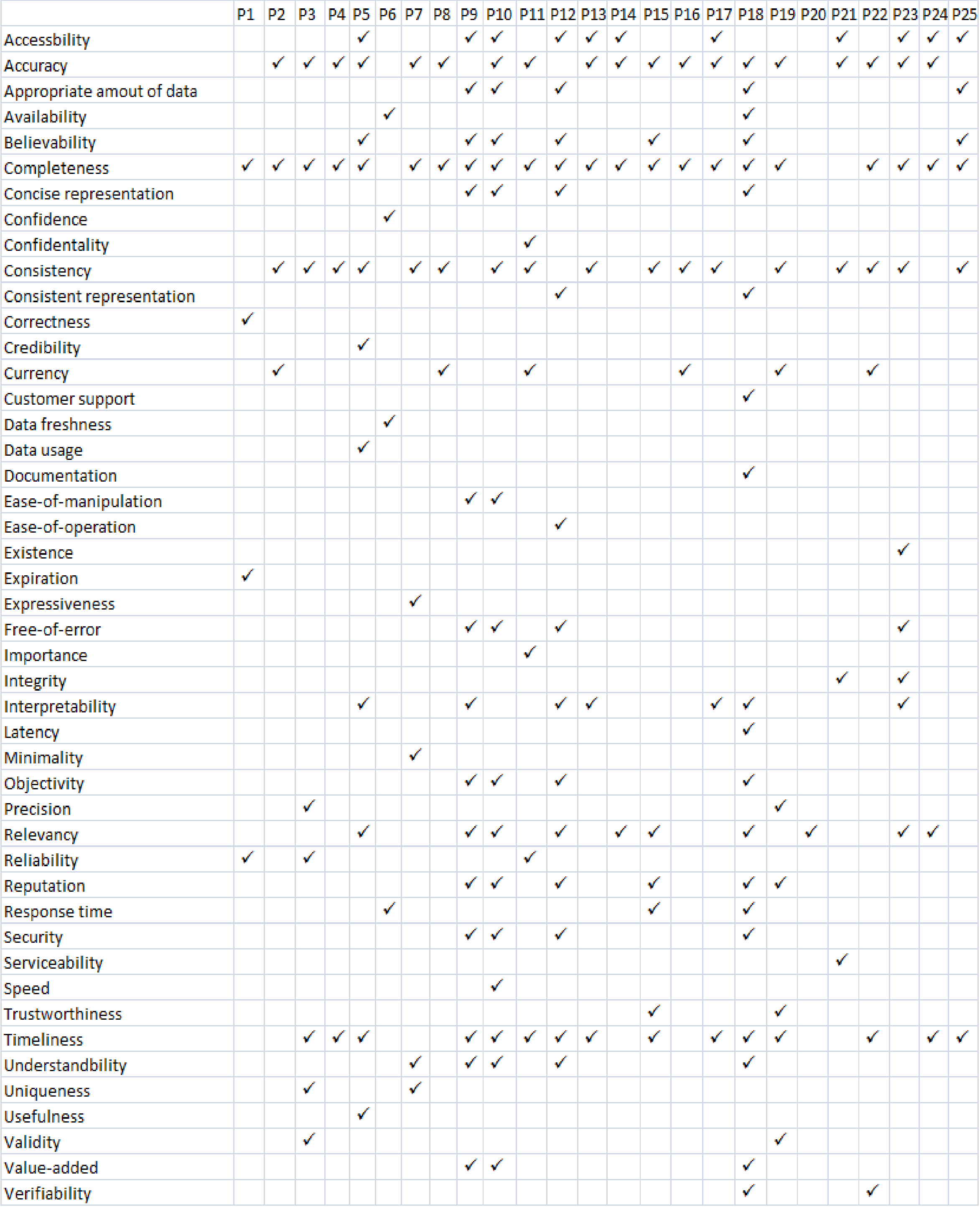
P22. M. Scannapieco, P. Missier, C. Batini (2005): Data Quality at a Glance, Datenbank-Spektrum, Vol.14

P23. Bovee M., Srivastava R. R., Mak B.R (2001): A Conceptual Framework and Belief Function Approach to Assessing Overall Information Quality. In proceedings of the 6th International Conference on Information Quality

P24. Y. W. Lee, D. M. Strong (2004): Knowing-Why About Data Processes and Data Quality, Journal of Management Information Systems, Vol. 20, pp. 13–39

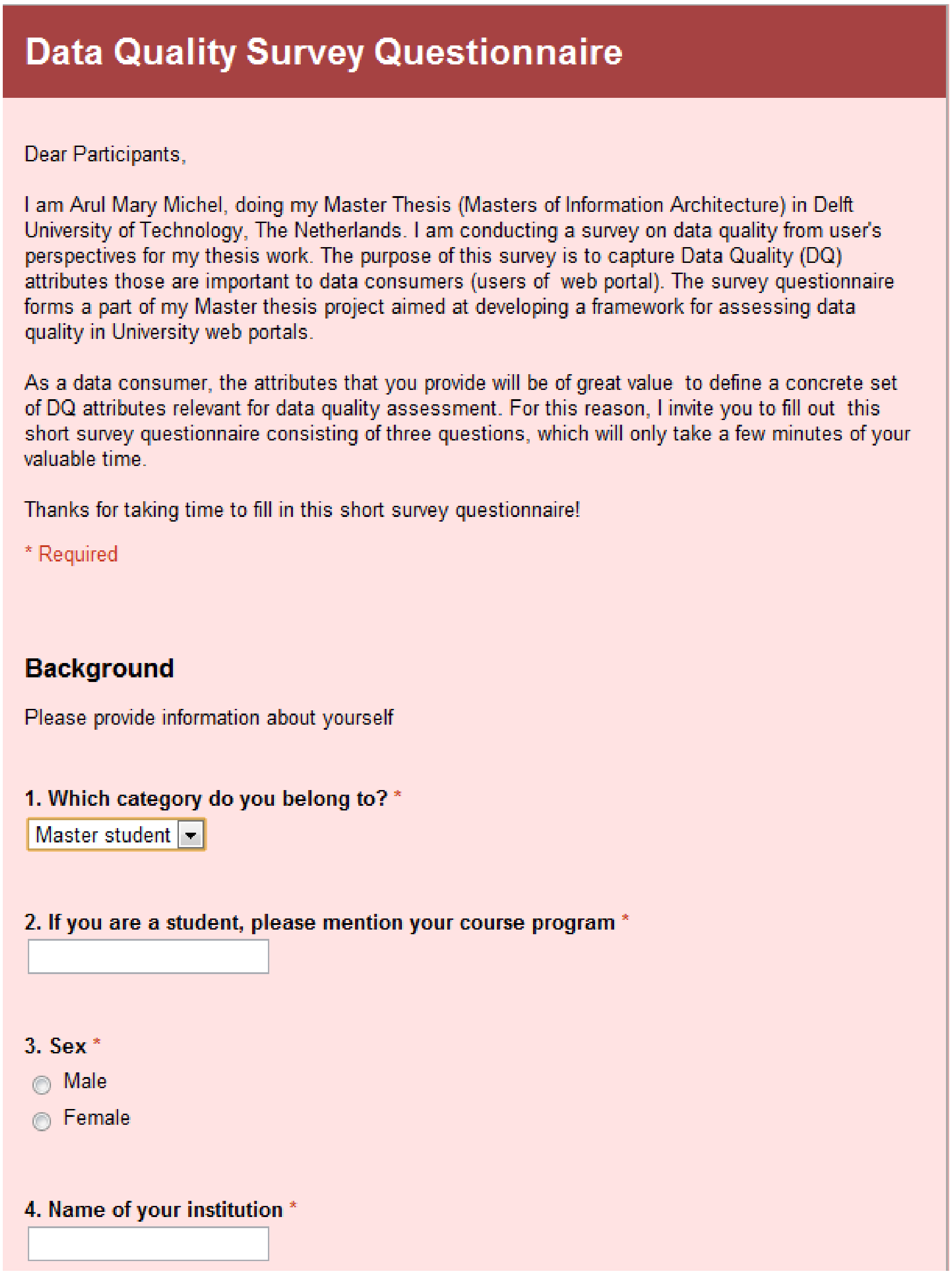
P25. Yang W. Lee Leo L. Pipino James D. Funk Richard Y. Wang: Journey to Data Quality, 2006 Massachusetts Institute of Technology

Appendix A - Data Quality dimensions from literature (47 dimensions)



	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25
Accessbility					✓				✓	✓		✓	✓	✓			✓				✓		✓	✓	✓
Accuracy		✓	✓	✓	✓		✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	
Appropriate amout of data									✓	✓		✓						✓							✓
Availability						✓												✓							
Believability					✓				✓	✓		✓			✓			✓							✓
Completeness	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓
Concise representation									✓	✓		✓						✓							
Confidence						✓																			
Confidentality											✓														
Consistency		✓	✓	✓	✓		✓	✓		✓	✓		✓		✓	✓	✓		✓		✓	✓	✓		✓
Consistent representation												✓						✓							
Correctness	✓																								
Credibility					✓																				
Currency		✓						✓			✓					✓			✓			✓			
Customer support																		✓							
Data freshness						✓																			
Data usage					✓																				
Documentation																		✓							
Ease-of-manipulation									✓	✓															
Ease-of-operation												✓													
Existence																							✓		
Expiration	✓																								
Expressiveness							✓																		
Free-of-error									✓	✓		✓											✓		
Importance											✓														
Integrity																					✓		✓		
Interpretability					✓				✓			✓	✓				✓	✓					✓		
Latency																		✓							
Minimality							✓																		
Objectivity									✓	✓		✓						✓							
Precision			✓																✓						
Relevancy					✓				✓	✓		✓		✓	✓			✓		✓			✓	✓	
Reliability	✓		✓								✓														
Reputation									✓	✓		✓			✓			✓	✓						
Response time						✓									✓			✓							
Security									✓	✓		✓						✓							
Serviceability																					✓				
Speed										✓															
Trustworthiness															✓				✓						
Timeliness			✓	✓	✓				✓	✓	✓	✓	✓		✓		✓	✓	✓			✓		✓	✓
Understandbility							✓		✓	✓		✓						✓							
Uniqueness			✓				✓																		
Usefulness					✓																				
Validity			✓																✓						
Value-added									✓	✓								✓							
Verifiability																		✓				✓			

Appendix B - First survey Questionnaire (Wang and Strong survey approach)



Data Quality Survey Questionnaire

Dear Participants,

I am Arul Mary Michel, doing my Master Thesis (Masters of Information Architecture) in Delft University of Technology, The Netherlands. I am conducting a survey on data quality from user's perspectives for my thesis work. The purpose of this survey is to capture Data Quality (DQ) attributes those are important to data consumers (users of web portal). The survey questionnaire forms a part of my Master thesis project aimed at developing a framework for assessing data quality in University web portals.

As a data consumer, the attributes that you provide will be of great value to define a concrete set of DQ attributes relevant for data quality assessment. For this reason, I invite you to fill out this short survey questionnaire consisting of three questions, which will only take a few minutes of your valuable time.

Thanks for taking time to fill in this short survey questionnaire!

* Required

Background

Please provide information about yourself

1. Which category do you belong to? *

Master student ▾

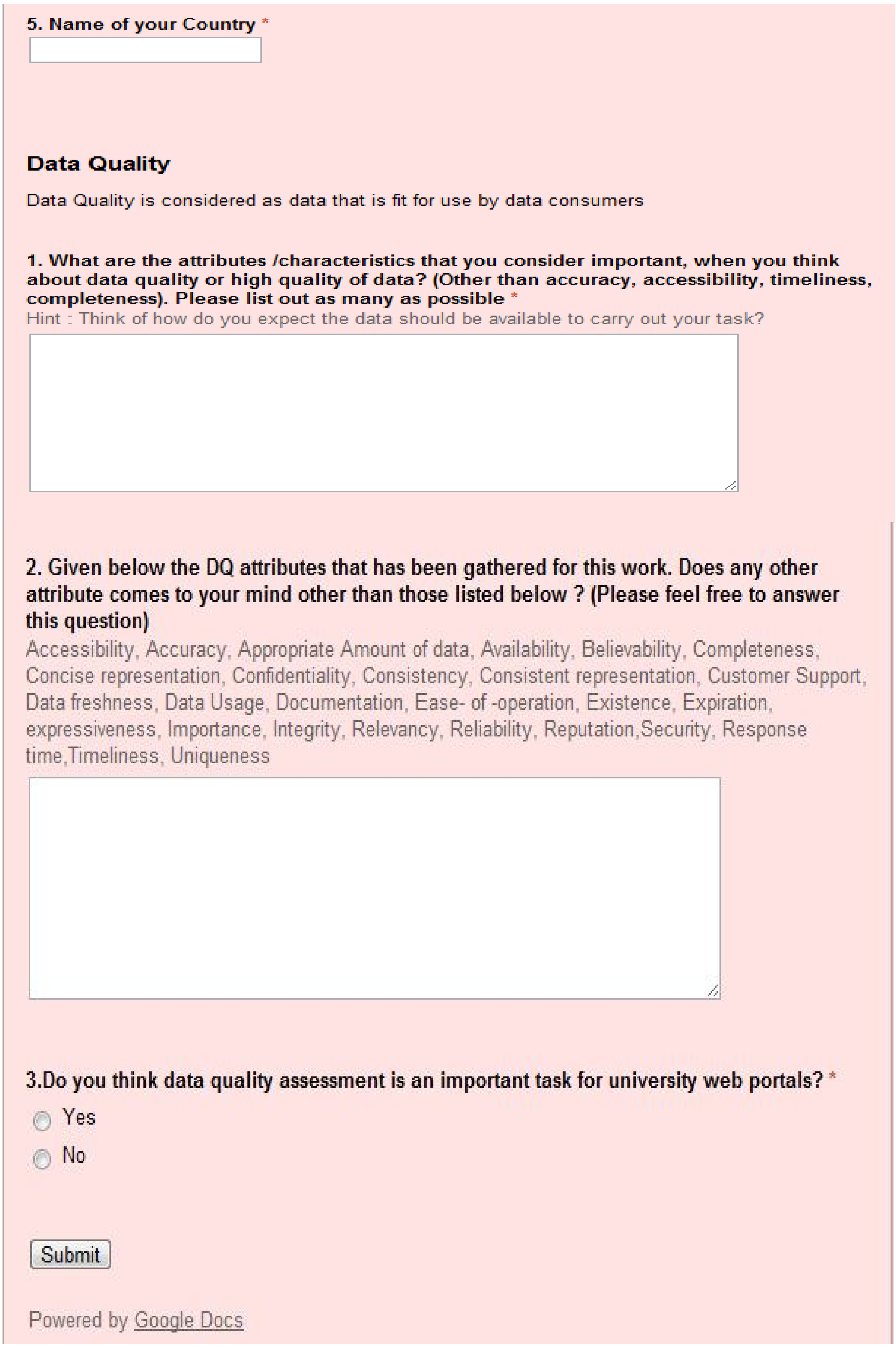
2. If you are a student, please mention your course program *

3. Sex *

○ Male

○ Female

4. Name of your institution *



5. Name of your Country *

Data Quality

Data Quality is considered as data that is fit for use by data consumers

1. What are the attributes /characteristics that you consider important, when you think about data quality or high quality of data? (Other than accuracy, accessibility, timeliness, completeness). Please list out as many as possible *

Hint : Think of how do you expect the data should be available to carry out your task?

2. Given below the DQ attributes that has been gathered for this work. Does any other attribute comes to your mind other than those listed below ? (Please feel free to answer this question)

Accessibility, Accuracy, Appropriate Amount of data, Availability, Believability, Completeness, Concise representation, Confidentiality, Consistency, Consistent representation, Customer Support, Data freshness, Data Usage, Documentation, Ease- of -operation, Existence, Expiration, expressiveness, Importance, Integrity, Relevancy, Reliability, Reputation,Security, Response time,Timeliness, Uniqueness

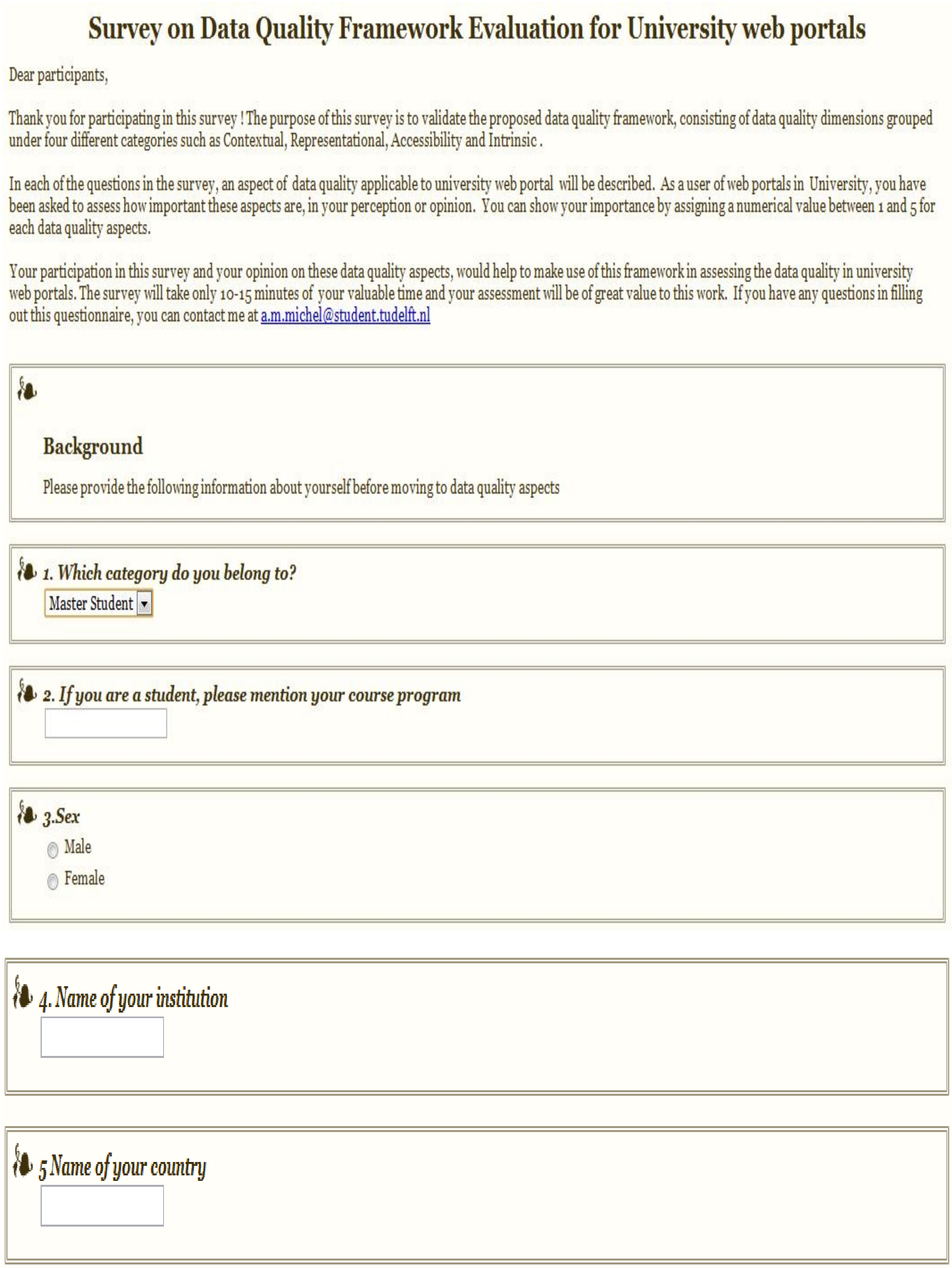
3.Do you think data quality assessment is an important task for university web portals? *

○ Yes

○ No

Powered by Google Docs

Appendix C - Second survey Questionnaire to validate the proposed framework



Survey on Data Quality Framework Evaluation for University web portals

Dear participants,

Thank you for participating in this survey ! The purpose of this survey is to validate the proposed data quality framework, consisting of data quality dimensions grouped under four different categories such as Contextual, Representational, Accessibility and Intrinsic .

In each of the questions in the survey, an aspect of data quality applicable to university web portal will be described. As a user of web portals in University, you have been asked to assess how important these aspects are, in your perception or opinion. You can show your importance by assigning a numerical value between 1 and 5 for each data quality aspects.

Your participation in this survey and your opinion on these data quality aspects, would help to make use of this framework in assessing the data quality in university web portals. The survey will take only 10-15 minutes of your valuable time and your assessment will be of great value to this work. If you have any questions in filling out this questionnaire, you can contact me at a.m.michel@student.tudelft.nl

Background

Please provide the following information about yourself before moving to data quality aspects

1. Which category do you belong to?

Master Student ▾

2. If you are a student, please mention your course program

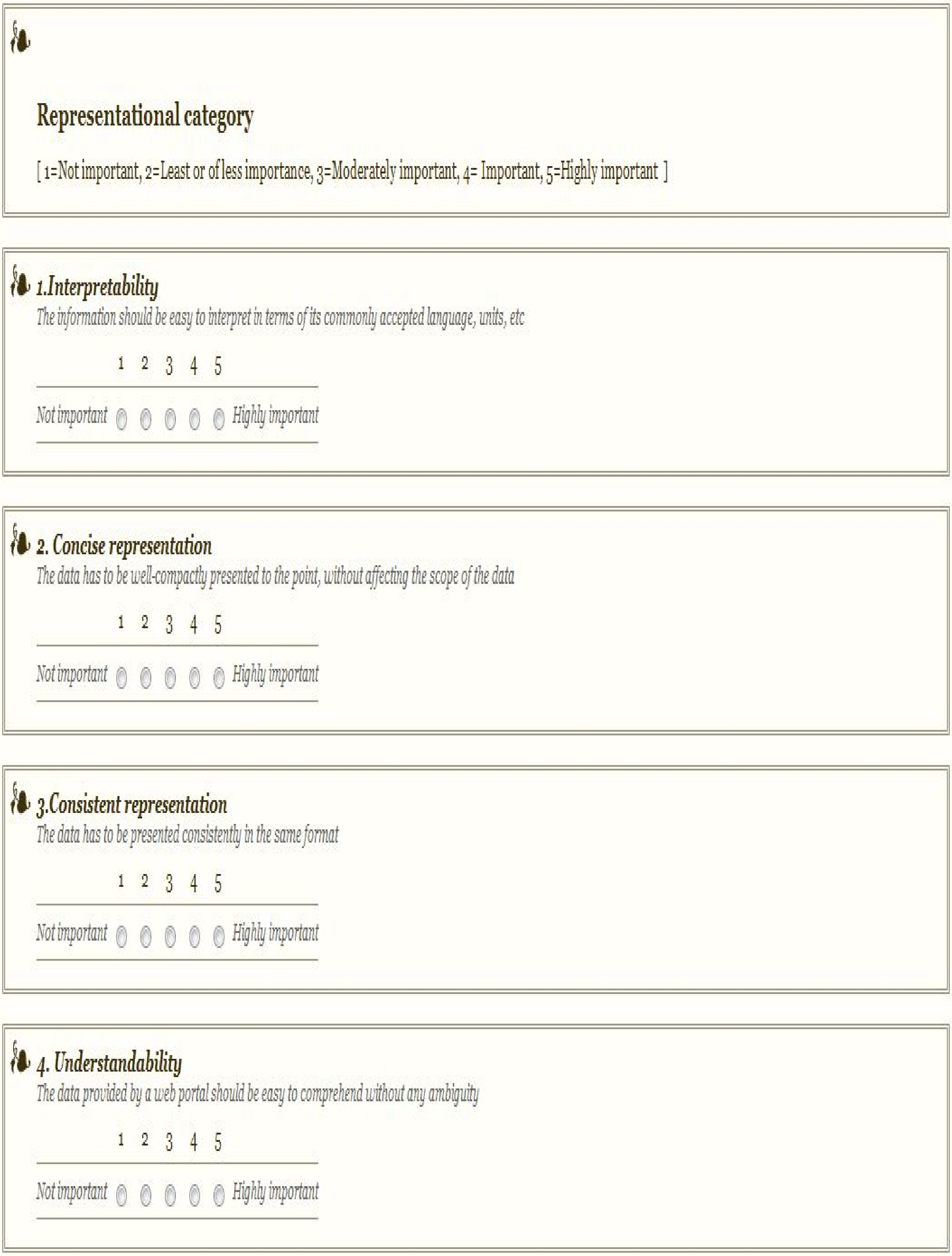
3.Sex

○ Male

○ Female

4. Name of your institution

5 Name of your country



Representational category

[1=Not important, 2=Least or of less importance, 3=Moderately important, 4= Important, 5=Highly important]

1.Interpretability

The information should be easy to interpret in terms of its commonly accepted language, units, etc

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

2. Concise representation

The data has to be well-compactly presented to the point, without affecting the scope of the data

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

3.Consistent representation

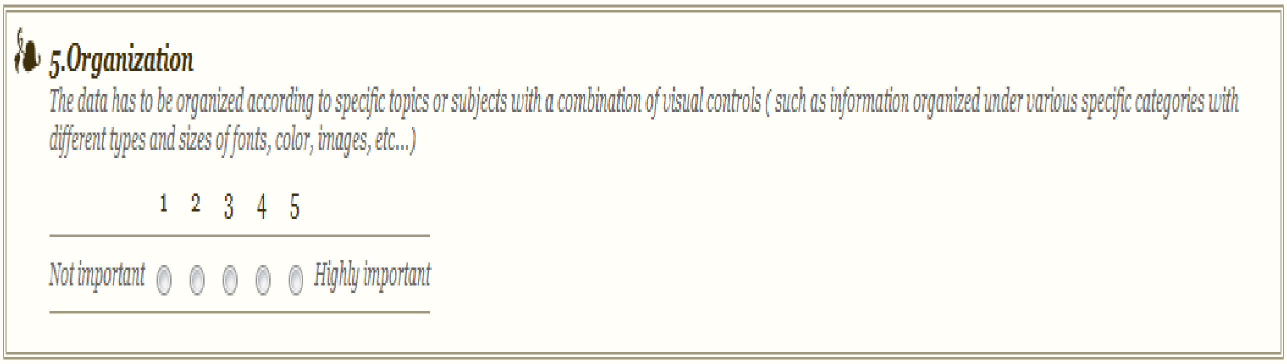
The data has to be presented consistently in the same format

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

4. Understandability

The data provided by a web portal should be easy to comprehend without any ambiguity

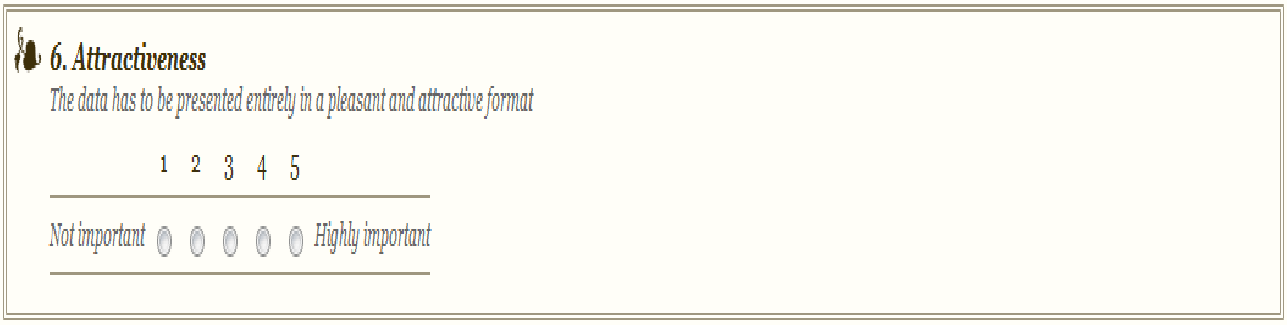
	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



5.Organization

The data has to be organized according to specific topics or subjects with a combination of visual controls (such as information organized under various specific categories with different types and sizes of fonts, color, images, etc...)

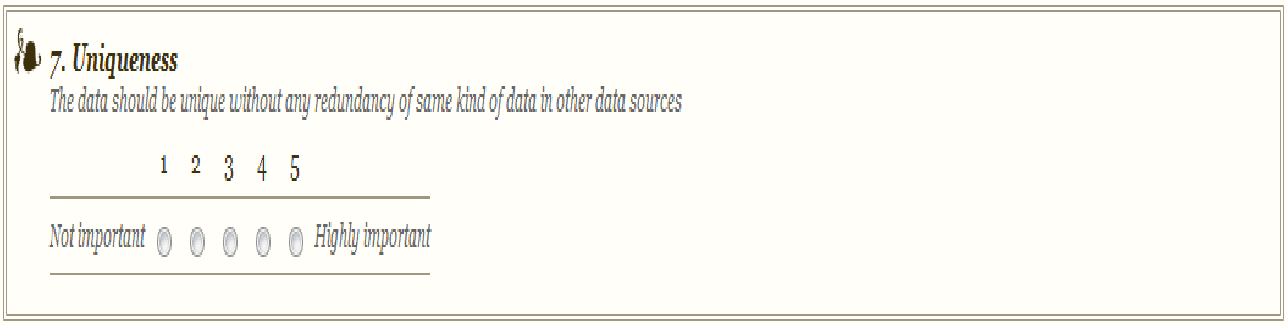
	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



6. Attractiveness

The data has to be presented entirely in a pleasant and attractive format

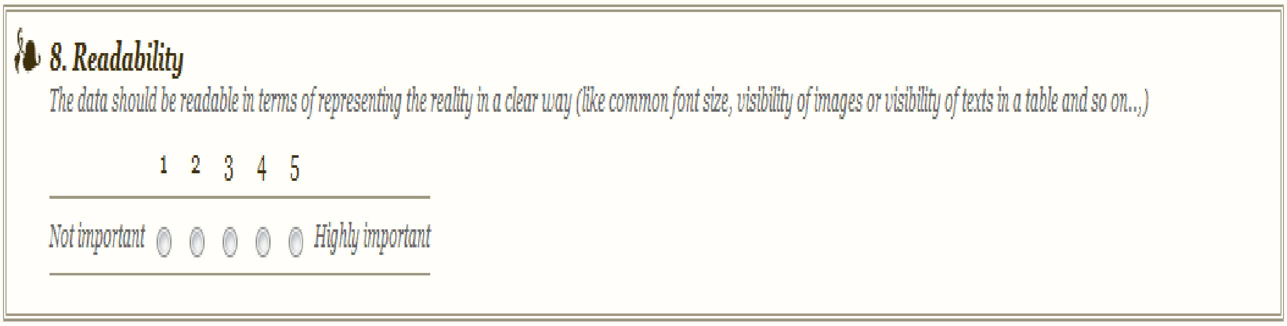
	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



7. Uniqueness

The data should be unique without any redundancy of same kind of data in other data sources

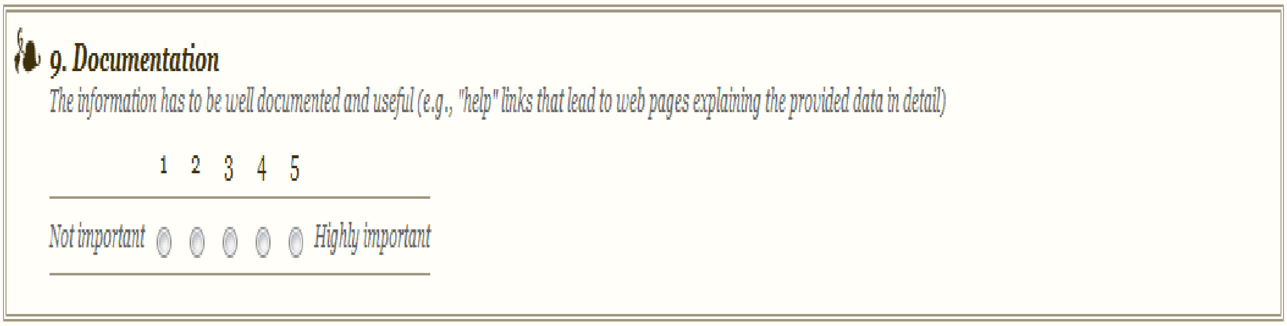
	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



8. Readability

The data should be readable in terms of representing the reality in a clear way (like common font size, visibility of images or visibility of texts in a table and so on..,)

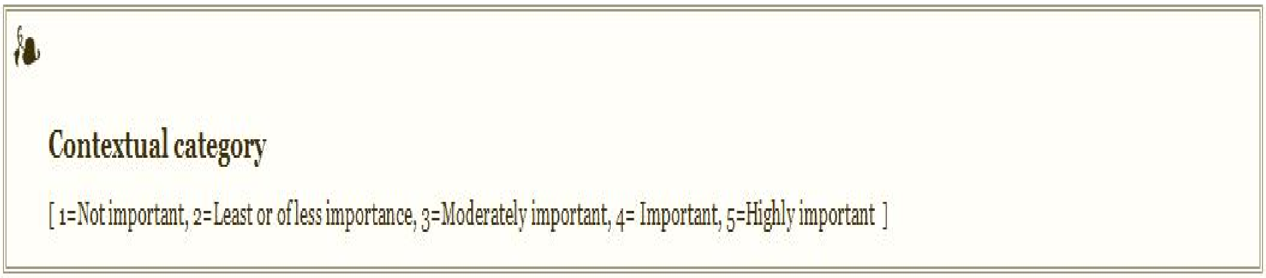
	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



9. Documentation

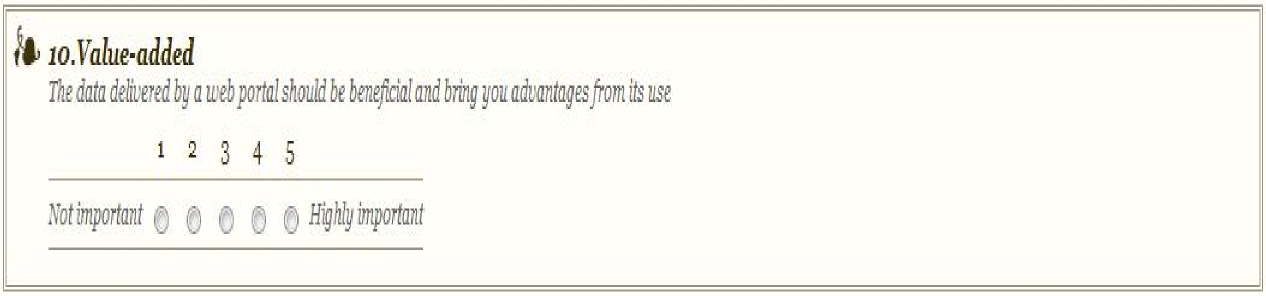
The information has to be well documented and useful (e.g., "help" links that lead to web pages explaining the provided data in detail)

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



Contextual category

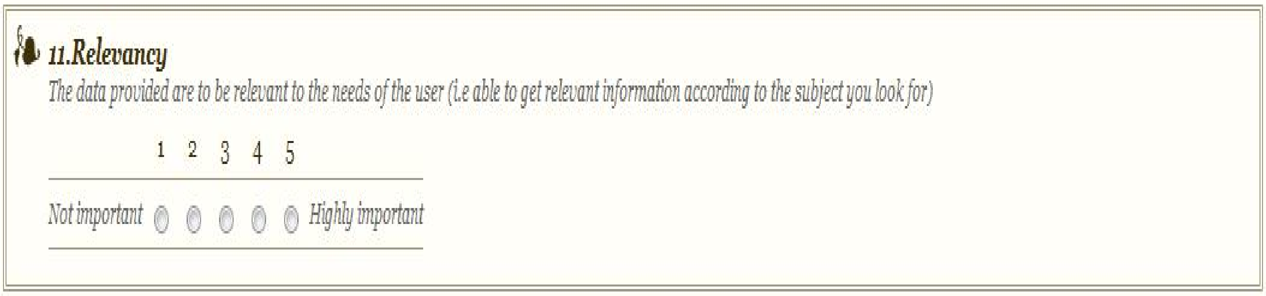
[1=Not important, 2=Least or of less importance, 3=Moderately important, 4= Important, 5=Highly important]



10.Value-added

The data delivered by a web portal should be beneficial and bring you advantages from its use

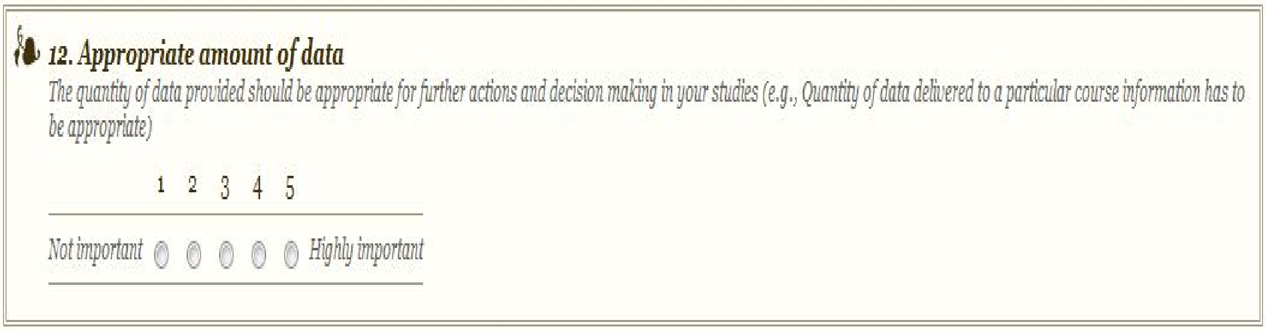
	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



11.Relevancy

The data provided are to be relevant to the needs of the user (i.e able to get relevant information according to the subject you look for)

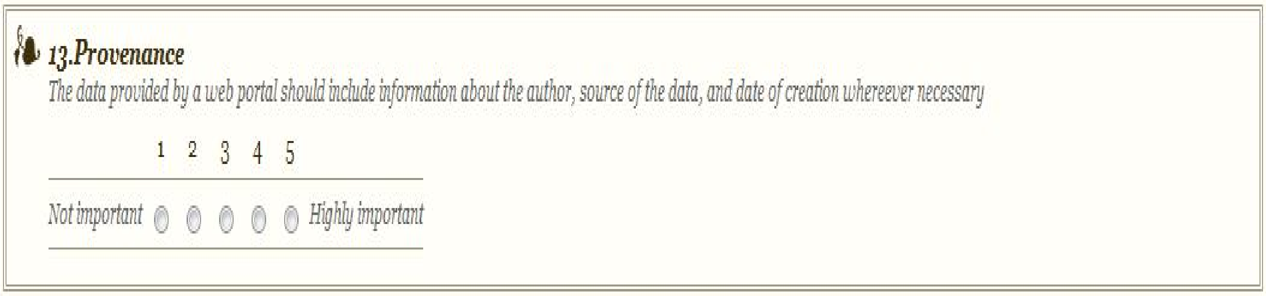
	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



12. Appropriate amount of data

The quantity of data provided should be appropriate for further actions and decision making in your studies (e.g., Quantity of data delivered to a particular course information has to be appropriate)

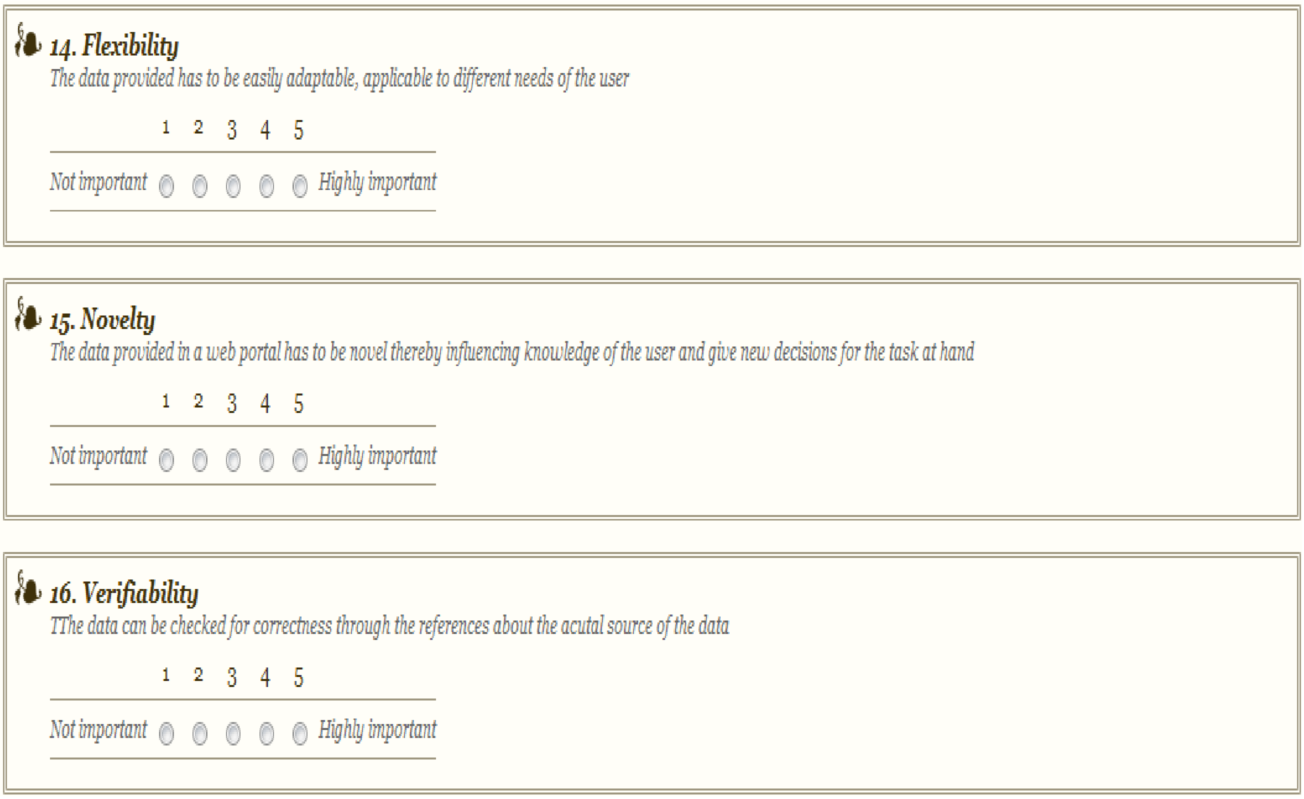
	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



13.Provenance

The data provided by a web portal should include information about the author, source of the data, and date of creation whereever necessary

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



14. Flexibility

The data provided has to be easily adaptable, applicable to different needs of the user

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

15. Novelty

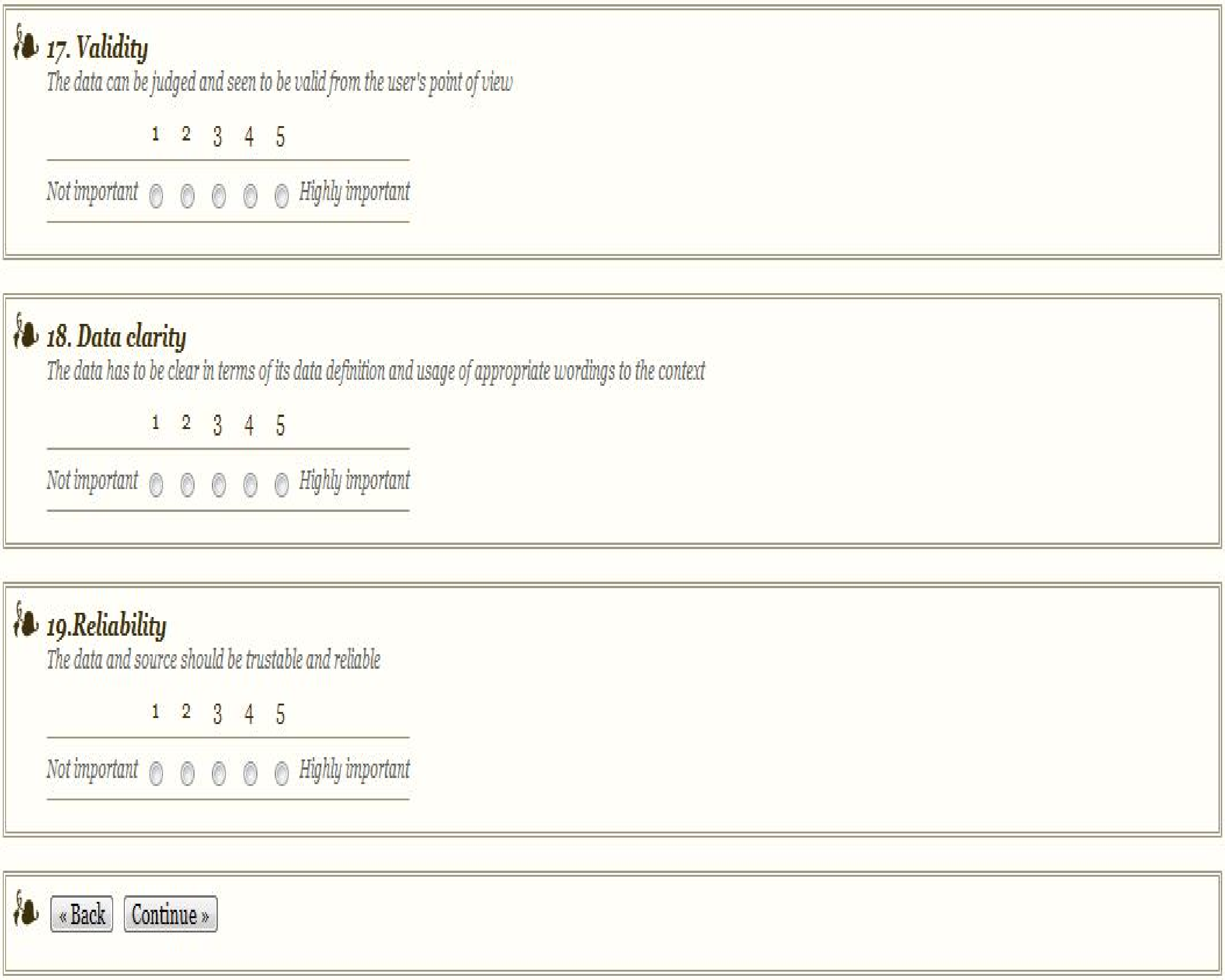
The data provided in a web portal has to be novel thereby influencing knowledge of the user and give new decisions for the task at hand

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

16. Verifiability

TThe data can be checked for correctness through the references about the acutal source of the data

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



17. Validity

The data can be judged and seen to be valid from the user's point of view

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

18. Data clarity

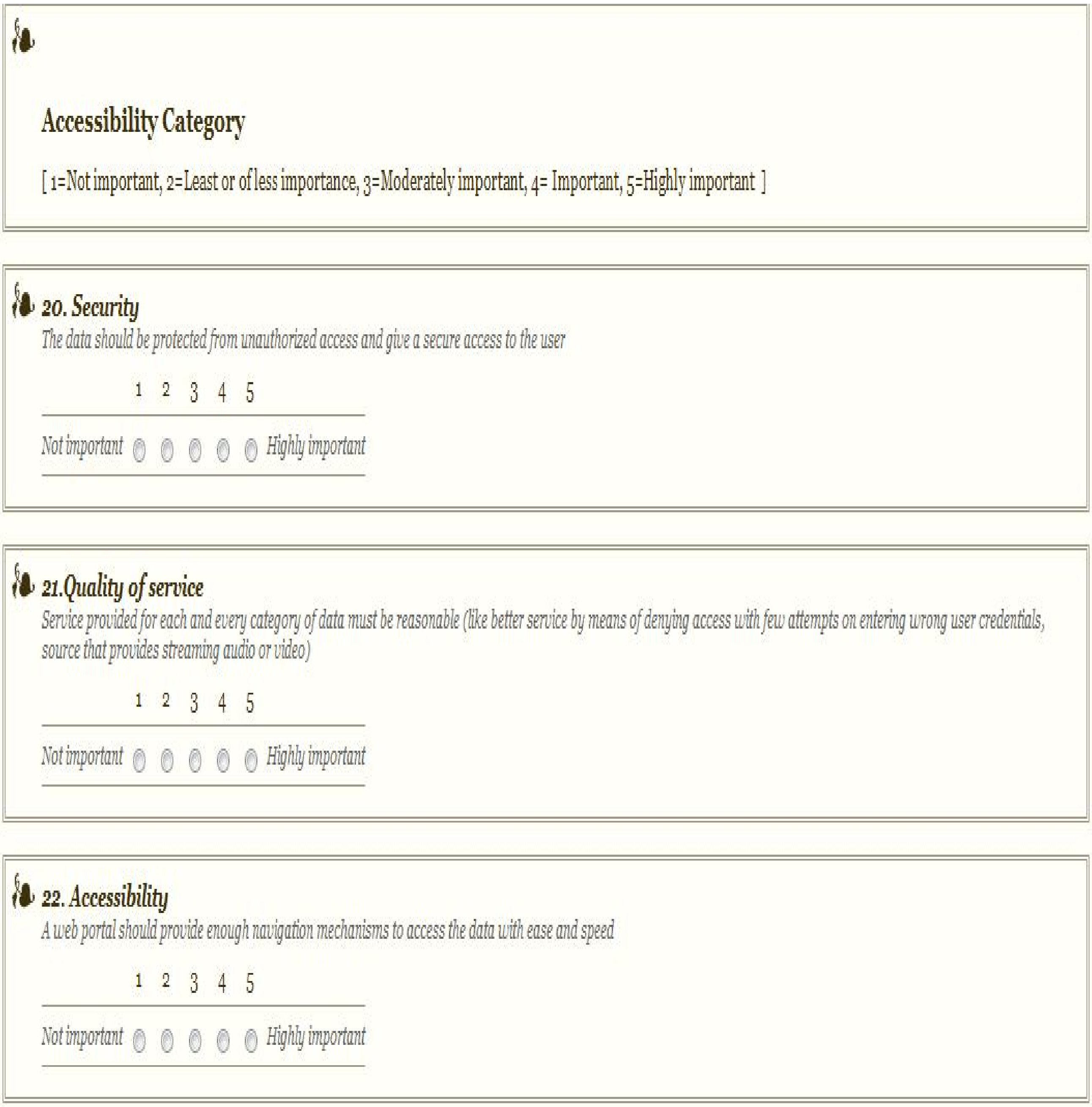
The data has to be clear in terms of its data definition and usage of appropriate wordings to the context

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

19.Reliability

The data and source should be trustable and reliable

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*



Accessibility Category

[1=Not important, 2=Least or of less importance, 3=Moderately important, 4= Important, 5=Highly important]

20. Security

The data should be protected from unauthorized access and give a secure access to the user

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

21.Quality of service

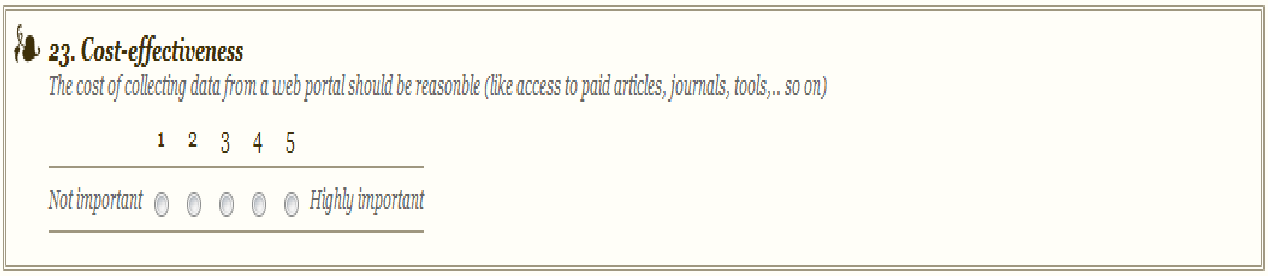
Service provided for each and every category of data must be reasonable (like better service by means of denying access with few attempts on entering wrong user credentials, source that provides streaming audio or video)

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

22. Accessibility

A web portal should provide enough navigation mechanisms to access the data with ease and speed

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

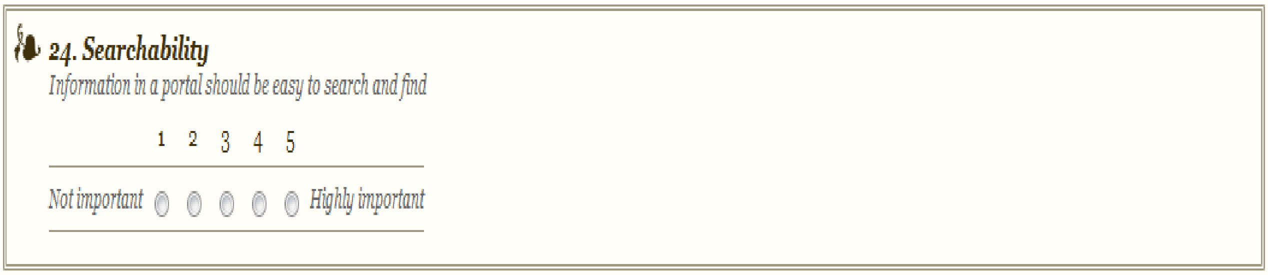


23. Cost-effectiveness

The cost of collecting data from a web portal should be reasonble (like access to paid articles, journals, tools,.. so on)

1 2 3 4 5

Not important ○ ○ ○ ○ ○ *Highly important*

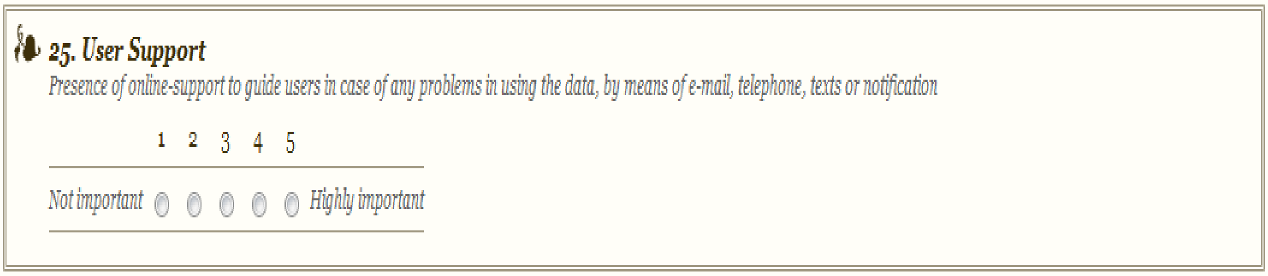


24. Searchability

Information in a portal should be easy to search and find

1 2 3 4 5

Not important ○ ○ ○ ○ ○ *Highly important*

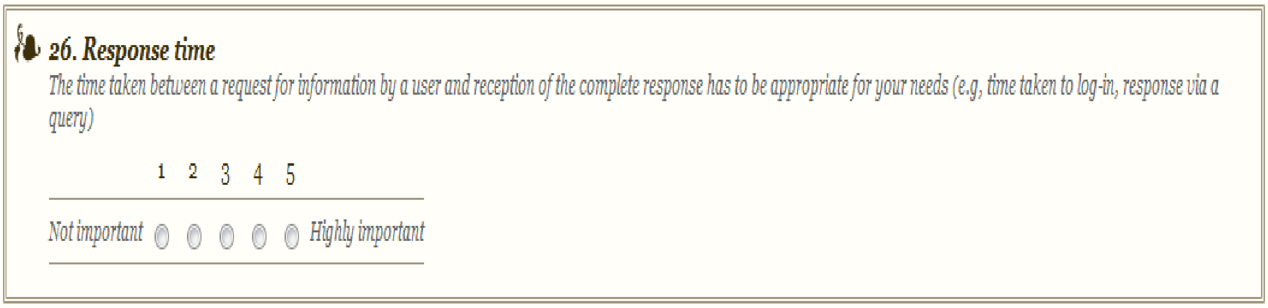


25. User Support

Presence of online-support to guide users in case of any problems in using the data, by means of e-mail, telephone, texts or notification

1 2 3 4 5

Not important ○ ○ ○ ○ ○ *Highly important*

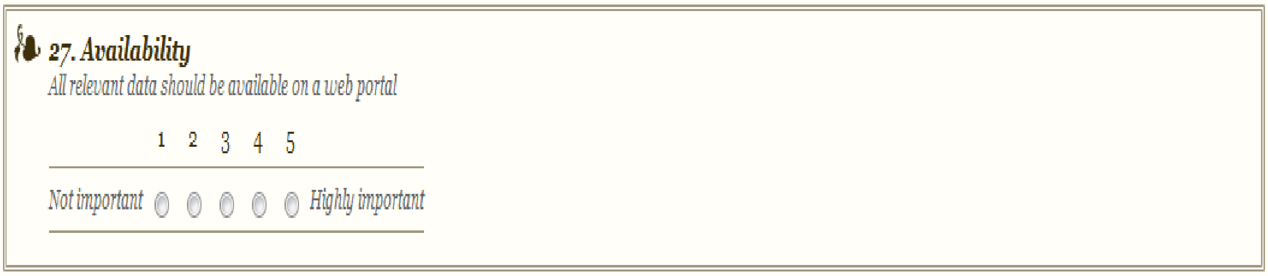


26. Response time

The time taken between a request for information by a user and reception of the complete response has to be appropriate for your needs (e.g, time taken to log-in, response via a query)

1 2 3 4 5

Not important ○ ○ ○ ○ ○ *Highly important*

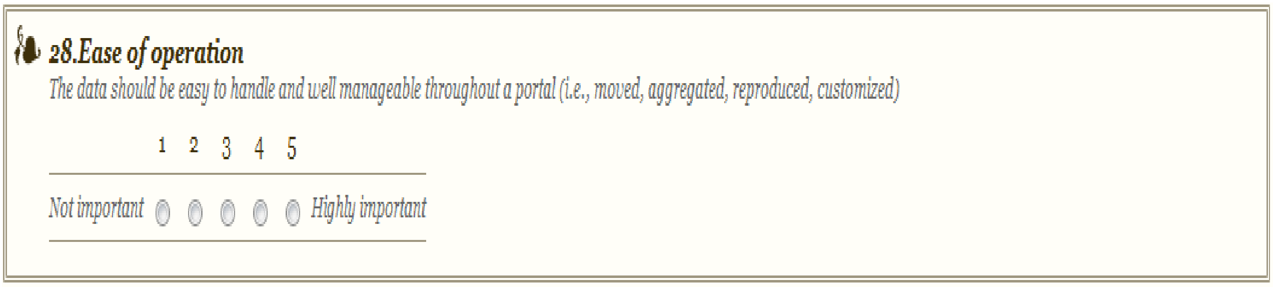


27. Availability

All relevant data should be available on a web portal

1 2 3 4 5

Not important ○ ○ ○ ○ ○ *Highly important*

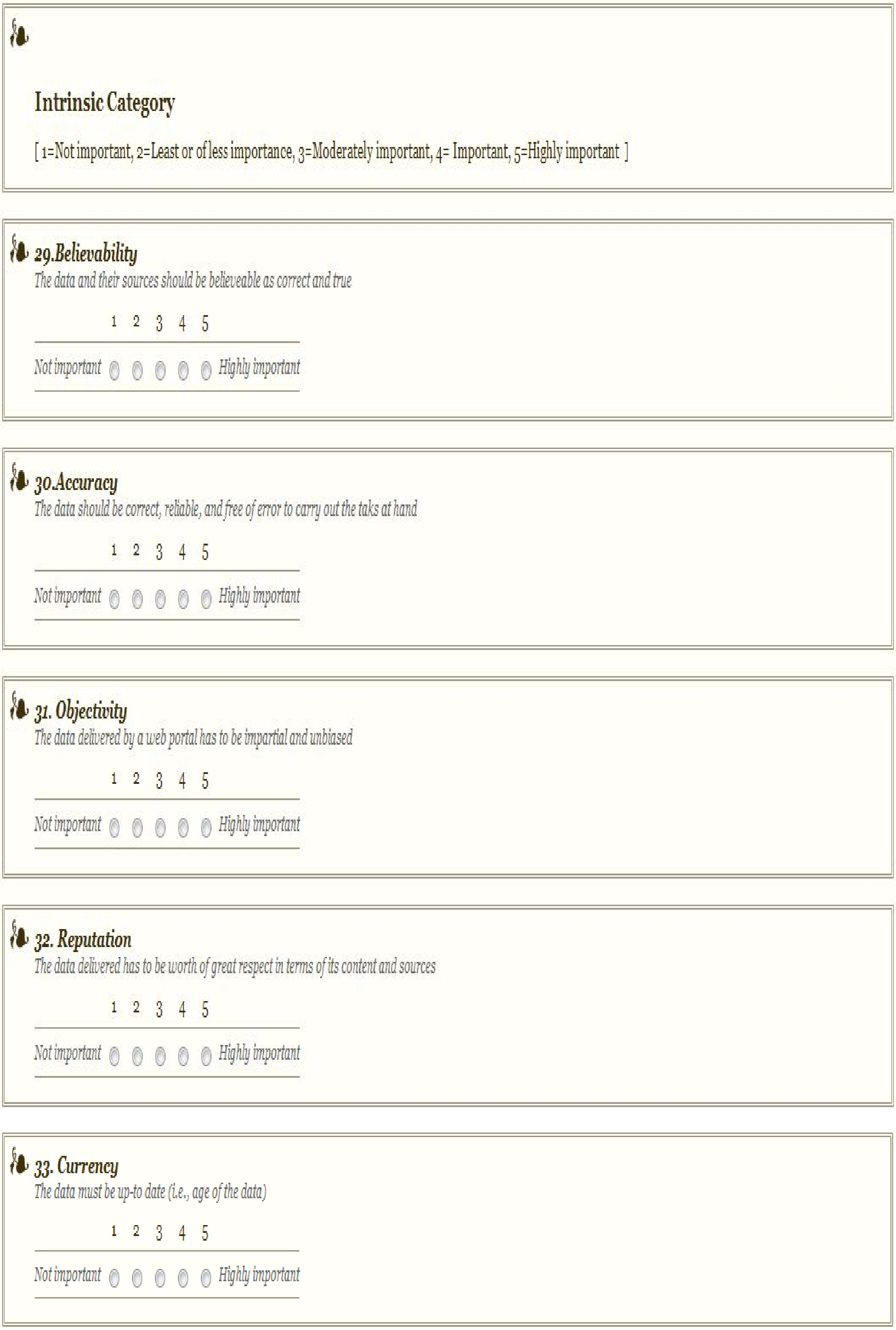


28.Ease of operation

The data should be easy to handle and well manageable throughout a portal (i.e., moved, aggregated, reproduced, customized)

1 2 3 4 5

Not important ○ ○ ○ ○ ○ *Highly important*



Intrinsic Category

[1=Not important, 2=Least or of less importance, 3=Moderately important, 4= Important, 5=Highly important]

29.Believability

The data and their sources should be believeable as correct and true

1 2 3 4 5

Not important ○ ○ ○ ○ ○ *Highly important*

30.Accuracy

The data should be correct, reliable, and free of error to carry out the taks at hand

1 2 3 4 5

Not important ○ ○ ○ ○ ○ *Highly important*

31. Objectivity

The data delivered by a web portal has to be impartial and unbiased

1 2 3 4 5

Not important ○ ○ ○ ○ ○ *Highly important*

32. Reputation

The data delivered has to be worth of great respect in terms of its content and sources

1 2 3 4 5

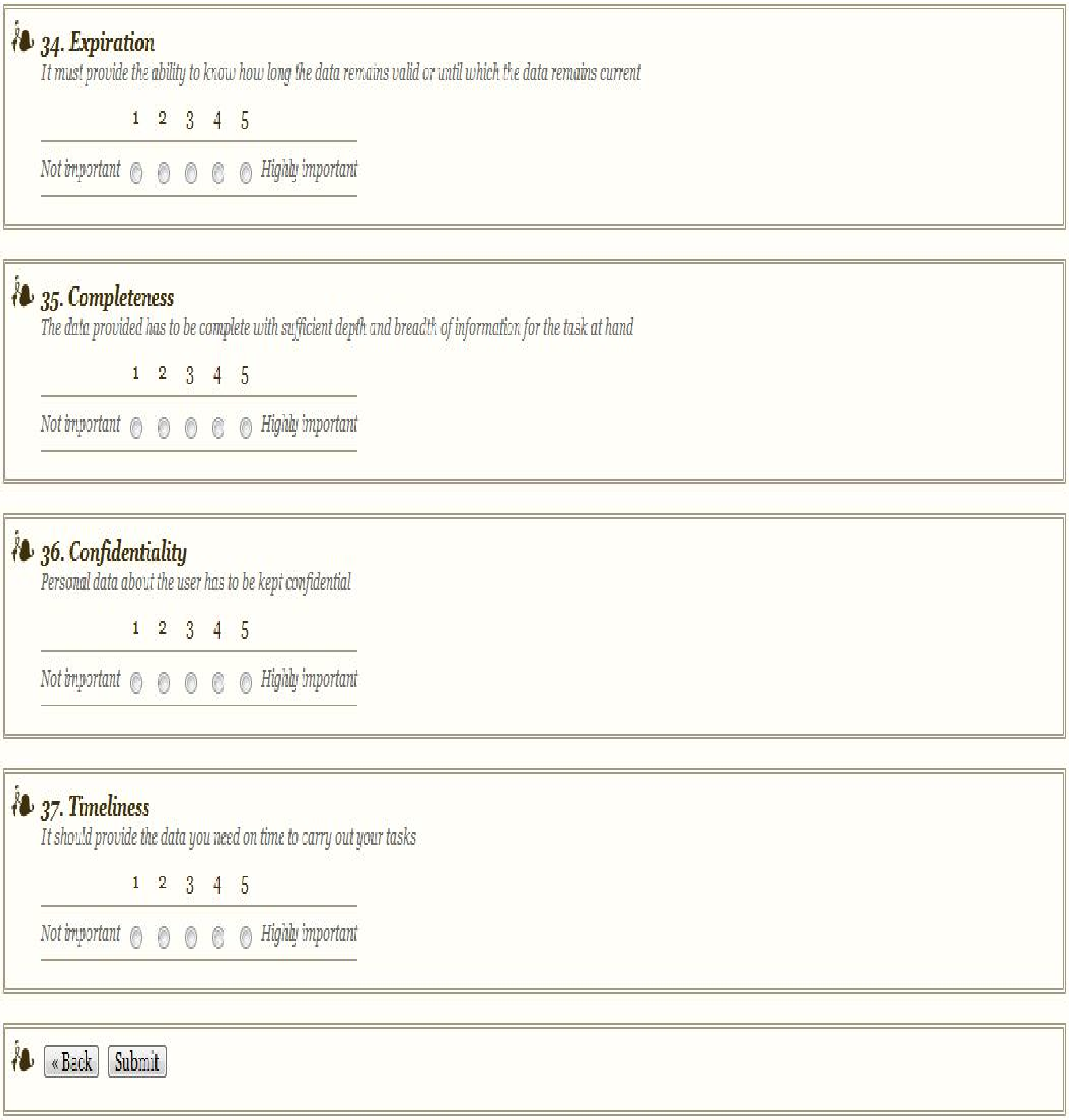
Not important ○ ○ ○ ○ ○ *Highly important*

33. Currency

The data must be up-to date (i.e., age of the data)

1 2 3 4 5

Not important ○ ○ ○ ○ ○ *Highly important*



34. Expiration

It must provide the ability to know how long the data remains valid or until which the data remains current

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

35. Completeness

The data provided has to be complete with sufficient depth and breadth of information for the task at hand

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

36. Confidentiality

Personal data about the user has to be kept confidential

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

37. Timeliness

It should provide the data you need on time to carry out your tasks

	1	2	3	4	5	
Not important	○	○	○	○	○	*Highly important*

Appendix D – Reliability Analysis

Reliability analysis statistics of the proposed framework

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1.Interpretability	139.17	156.999	.188	.869
2. Concise representation	139.35	155.584	.217	.869
3.Consistent representation	139.65	153.009	.251	.870
4. Understandability	139.07	153.700	.358	.867
5.Organization	139.34	150.642	.377	.866
6. Readability	139.23	151.536	.388	.866
7. Documentation	139.19	150.774	.449	.865
8.Value-added	139.36	152.997	.356	.867
9.Relevancy	139.05	155.608	.263	.868
10. Appropriate amount of data	139.25	151.328	.430	.865
11.Provenance	139.83	149.925	.380	.866
12. Flexibility	139.53	153.016	.280	.869
13. Novelty	139.68	149.186	.425	.865
14. Verifiability	139.29	152.179	.384	.866
15. Validity	139.37	148.704	.471	.864
16. Data clarity	139.10	152.050	.442	.865
17.Reliability	138.96	152.227	.380	.866
18. Security	139.15	153.943	.250	.869
19.Quality of service	139.35	149.369	.492	.864
20. Accessibility	139.16	152.270	.366	.866
21. Cost-effectiveness	139.37	151.390	.347	.867
22. Searchability	138.92	152.732	.427	.865
23. User Support	139.32	149.132	.470	.864
24. Response time	139.23	151.106	.418	.865
25. Availability	139.15	154.636	.286	.868
26.Ease of operation	139.33	148.975	.558	.862
27.Believability	139.16	152.014	.387	.866
28.Accuracy	138.95	154.219	.365	.867
29. Objectivity	139.38	152.425	.328	.867
30. Reputation	139.47	149.338	.460	.864
31. Currency	138.97	151.422	.395	.866
32. Expiration	139.35	149.530	.452	.864
33. Completeness	139.27	152.307	.361	.866
34. Confidentiality	139.03	153.073	.343	.867
35. Timeliness	139.03	153.589	.348	.867

Appendix E – Descriptive Statistics (Pondicherry University)

Descriptive Statistics

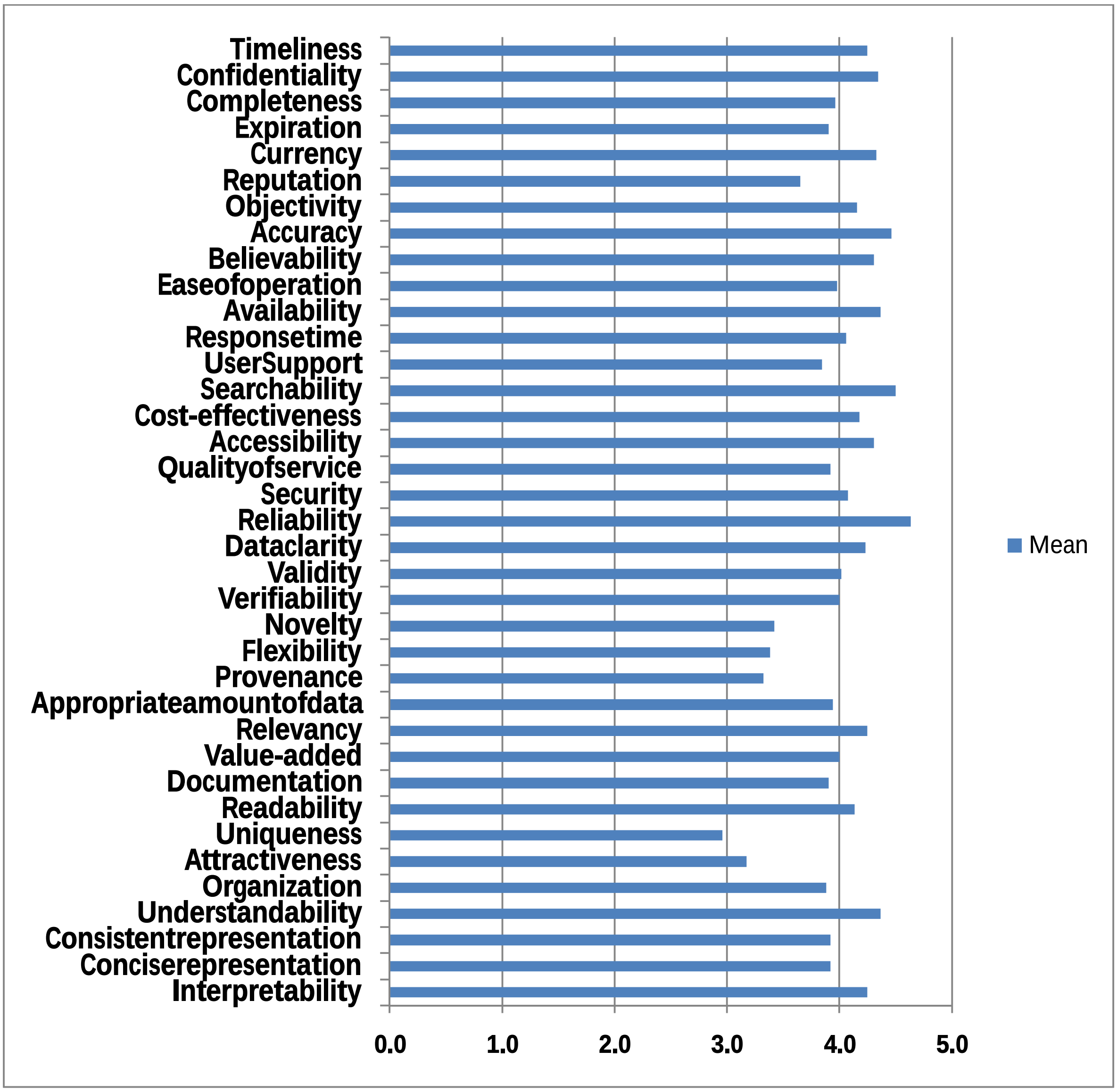
	N	Minimum	Maximum	Mean	Std. Deviation
1.Interpretability	116	1	5	4.10	.762
2. Concise representation	116	1	5	4.07	.852
3.Consistent representation	116	1	5	3.68	1.052
4. Understandability	116	2	5	4.23	.773
5.Organization	116	1	5	4.12	.961
6. Attractiveness	115	2	5	4.06	.967
7. Uniqueness	115	2	5	4.03	.863
8. Readability	116	1	5	4.09	.884
9. Documentation	115	1	5	4.24	.790
10.Value-added	115	1	5	4.01	.822
11.Relevancy	116	1	5	4.30	.794
12. Appropriate amount of data	116	1	5	4.17	.837
13.Provenance	116	1	5	3.73	.981
14. Flexibility	116	1	5	3.97	.918
15. Novelty	116	1	5	3.82	.992
16. Verifiability	114	2	5	4.10	.798
17. Validity	116	1	5	3.97	.991
18. Data clarity	116	2	5	4.28	.753
19.Reliability	114	1	5	4.29	.849
20. Security	115	2	5	4.28	.801
21.Quality of service	115	1	5	4.06	.820
22. Accessibility	115	1	5	4.20	.850
23. Cost-effectiveness	115	1	5	3.90	1.009
24. Searchability	115	2	5	4.35	.738
25. User Support	115	1	5	4.11	.925
26. Response time	114	1	5	4.17	.872
27. Availability	113	1	5	4.17	.743
28.Ease of operation	115	2	5	4.07	.769
29.Believability	116	2	5	4.16	.864
30.Accuracy	115	2	5	4.38	.670
31. Objectivity	114	1	5	3.96	.949
32. Reputation	116	1	5	4.02	.884
33. Currency	115	1	5	4.42	.838
34. Expiration	114	1	5	4.05	.891
35. Completeness	115	1	5	4.10	.872
36. Confidentiality	115	1	5	4.35	.738
37. Timeliness	116	1	5	4.33	.732

Appendix F - Descriptive Statistics (TU Delft University)

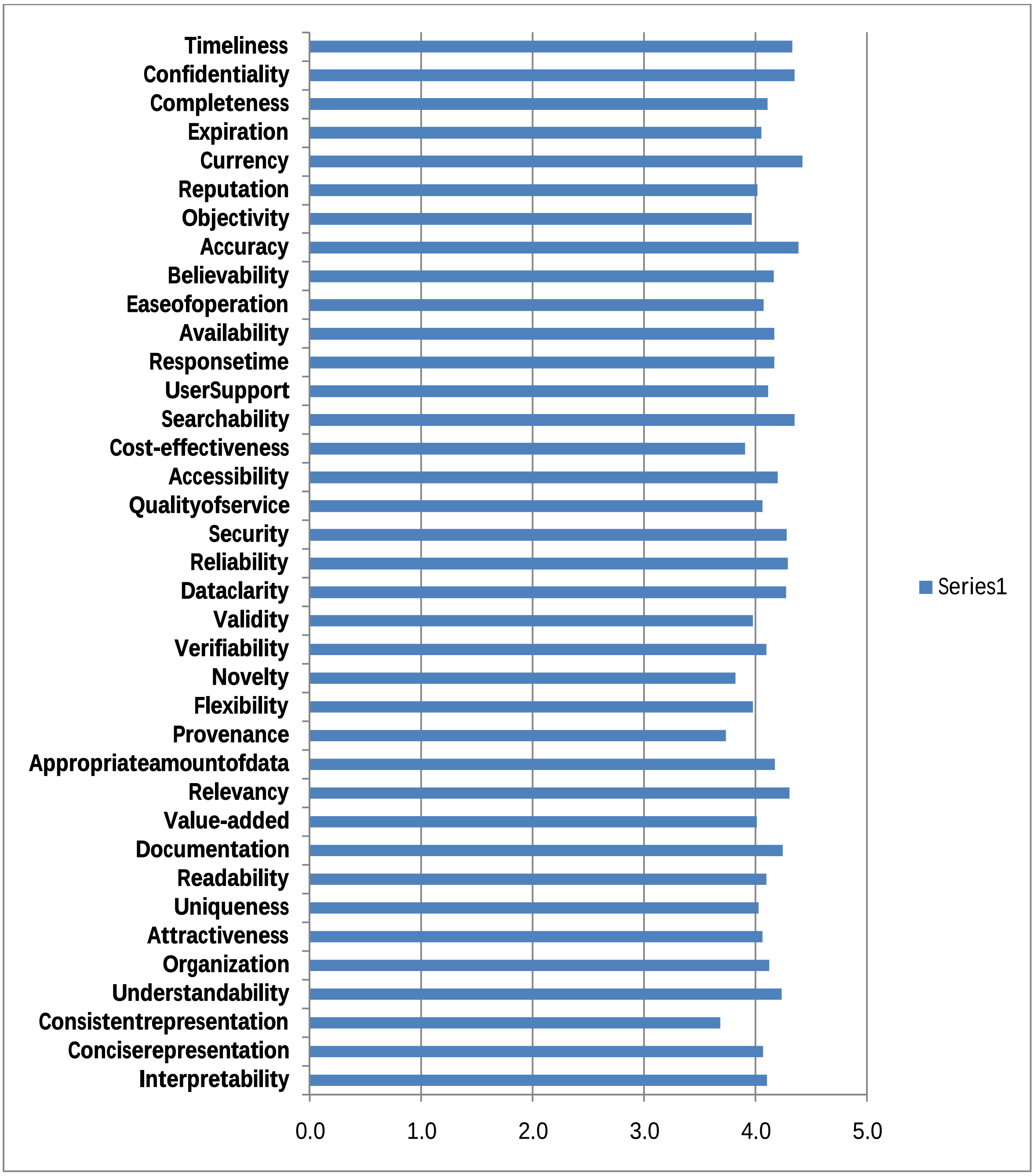
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
1.Interpretability	52	3	5	4.25	.682
2. Concise representation	52	2	5	3.92	.652
3.Consistent representation	52	2	5	3.92	.926
4. Understandability	52	3	5	4.37	.658
5.Organization	52	1	5	3.88	.922
6. Attractiveness	52	2	5	3.17	.964
7. Uniqueness	52	1	5	2.96	1.137
8. Readability	52	2	5	4.13	.742
9. Documentation	52	2	5	3.90	.891
10.Value-added	52	3	5	4.00	.686
11.Relevancy	52	3	5	4.25	.556
12. Appropriate amount of data	52	2	5	3.94	.802
13.Provenance	52	1	5	3.33	1.061
14. Flexibility	52	1	5	3.38	.973
15. Novelty	52	1	5	3.42	.977
16. Verifiability	52	2	5	4.00	.863
17. Validity	52	2	5	4.02	.874
18. Data clarity	52	3	5	4.23	.731
19.Reliability	52	3	5	4.63	.561
20. Security	52	1	5	4.08	1.082
21.Quality of service	52	2	5	3.92	.904
22. Accessibility	52	2	5	4.31	.755
23. Cost-effectiveness	50	3	5	4.18	.774
24. Searchability	52	2	5	4.50	.700
25. User Support	51	2	5	3.84	.987
26. Response time	51	3	5	4.06	.705
27. Availability	52	3	5	4.37	.742
28.Ease of operation	52	2	5	3.98	.828
29.Believability	52	2	5	4.31	.729
30.Accuracy	52	3	5	4.46	.541
31. Objectivity	51	3	5	4.16	.703
32. Reputation	52	1	5	3.65	.861
33. Currency	52	2	5	4.33	.834
34. Expiration	52	1	5	3.90	.975
35. Completeness	52	2	5	3.96	.791
36. Confidentiality	52	2	5	4.35	.861
37. Timeliness	52	2	5	4.25	.738

Appendix G - Importance of DQ dimensions (TU Delft)



Appendix H - Importance of DQ dimensions (Pondicherry University)



Appendix I- Definitions of DQ dimensions

1. Accessibility The extent to which data is available, or easily and quickly retrievable *1
2. Amount of data The extent to which the volume of data is appropriate for task at hand *1
3. Accuracy The extent to which data are correct, reliable and certified free of error *1
4. Availability The extent to which data are available to its users
5. Believability The extent to which data is regarded as true and credible *1
6. Completeness The extent to which data are of sufficient breadth and depth for the task at hand *1
7. Concise representation The extent to which data is compactly represented *1
8. Consistent representation The extent to which data is presented in the same format and compatible with previous data *1
9. Interpretability The extent to which data is in appropriate language, symbols, units, and the data definitions are clear *1
10. Objectivity The extent to which data is unbiased, unprejudiced, and Impartial *1
11. Relevancy The extent to which data is applicable and helpful for the task at hand *1
12. Reputation The extent to which data is highly regarded in terms of its source or content *1
13. Security The extent to which access to data is restricted appropriately to maintain its security
14. Timeliness The extent to which the age of the data is appropriate for the task at Hand *1
15. Understandability The extent to which data is easily comprehended *1
16. Quality of Service The extent to which a measure for the transmission and error rates of web sources is guaranteed *2
17. Currency The extent to which the web portal provides non-obsolete Data *3
18. User support The extent to which the amount and usefulness of human help via email, text or telephone *2
19. Documentation The amount and usefulness of documents with metadata *2
20. Ease of operation The extent to which data is easy to manipulate and apply to different tasks
21. Expiration The extent to which the date until which the data remain current is known *3
22. Response time Amount of time until complete response reaches the user
23. Value-added The extent to which data are beneficial and provide advantages from their use *1
24. Readability The extent to which data is readable whenever it represents the meaning of the reality represented by the schema in a clear way *4
25. Searchability The extent to which the information is easily searched and indexed

26. Provenance The extent to which the source of the data is mentioned (i.e., from where the data comes from and from whom and how) *4

27. Organization The extent to which data is organized and presented (structure layout, colour, text, font, images) *3

28. Flexibility The extent to which data are expandable, adaptable and easily applied to other needs *1

29. Novelty The extent to which data obtained influence knowledge and new decisions *3

30. Cost-effectiveness The extent to which the cost of collecting appropriate data is Reasonable *1

31. Data clarity The extent to which the data is clear without any ambiguity for decision making

32. Verifiability Degree and ease with which the data can be checked for Correctness *2

33. Validity The extent to which users can judge and comprehend data delivered by the portal *3

34. Confidentiality Reflects the degree to which the user information is kept Confidential *1

35. Reliability The extent to which the users can trust the data and their sources *3

*1Wang and Strong (1996)

*2 Naumann (2002)

*3 Caro et al (2008)

*4 Batini &Scannapieco (200)