Tania Islam

**Data Quality Maturity Model for Citizen Science Applications**

Examiner: Prof. Ajantha Dahanayake, Professor, PhD

Jiri Muisto, MSc
ABSTRACT

Lappeenranta University of Technology
School of Engineering and science
Master’s Degree Program in Software Engineering and Digital Transformation

Tania Islam

Data Quality Maturity Model for Citizen Science Applications

Master’s Thesis

Year: 2019

Examiner: Prof. Ajantha Dahanayake, Professor, PhD
Jiri Muisto, MSc

Keywords: Data quality, Data quality characteristics, Data quality in database, Data quality Framework, Framework for citizen science, Data Quality Maturity Model, Citizen science.

Advances in scientific projects and researches are a great initiative for real-life applications. Citizen science is the particularly new domain of science that has already proved that it can also be as helpful as classical science. There are several challenges of citizen science, but one major challenge constantly facing is the quality of collected data. There are several techniques to measure data quality as well as to check the characteristics of data quality.

This research has produced a data quality maturity model for citizen science applications. Several data quality characteristics are used particularly to citizen science applications during the development of the data quality maturity model for citizen science applications. This model can function as a tool to gauge the functional and quality requirements of a citizen science application.
ACKNOWLEDGEMENT

I would like to express warm gratitude to my parents, and it is undeniable that without their endless support it was impossible for me to come this far. Along with them, my elder sister and all other members of my family was highly supportive to me during this study.

Other than family, I must mention the most honoured person who have been of a great source of knowledge and inspiration to me throughout this whole period of my study and research. I believe my words are not enough to express my gratitude for my thesis supervisor Ajantha Dhanayake. She always has been an excellent source of knowledge, motivation and direction. Beside her My second examiner Jiri Muisto also supports me throughout this thesis and always provide me all the information I need. I would also like to thank Department of Software Engineering and Digital Transformation for giving such a nice study environment and support throughout my studies.

I would like to convey my thanks message to all who guide, support, encourage and advised me throughout my stay here in Finland during my course work and in my thesis time.

Author

Tania Islam

26.05.2019
# TABLE OF CONTENTS

## 1 INTRODUCTION

- 1.1 Research Problem ........................................................................................................ 7
- 1.2 Research objectives and questions .................................................................................... 8
- 1.3 Research tasks and directions .......................................................................................... 8
- 1.4 Research Methodology .................................................................................................... 9
- 1.5 Structure of the thesis .................................................................................................... 9

## 2 STATE-OF-THE-ART

- 2.1 About Data quality models and frameworks ................................................................. 11
- 2.2 Literature Review ........................................................................................................ 11
  - 2.2.1 Literature Review Process ...................................................................................... 12
- 2.3 Citizen Science ............................................................................................................. 13
  - 2.3.1 Citizen science definition ....................................................................................... 13
  - 2.3.2 ISO quality Standard ............................................................................................. 14
- 2.4 Data quality characteristics for citizen science applications ........................................ 15
- 2.5 Data Quality (DQ) in databases .................................................................................... 16
- 2.6 Data quality frameworks .............................................................................................. 18
  - 2.6.1 ISO data quality framework .................................................................................. 19
- 2.7 Data Quality and Citizen Science Application ............................................................. 20
  - 2.7.1 A general framework ............................................................................................. 21
  - 2.7.2 Data validation ....................................................................................................... 22
  - 2.7.3 User acceptance factors in CS applications ................................................................ 22
- 2.8 Capability Maturity Model (CMM) ............................................................................... 23

## 3 DATA QUALITY PROBLEMS AND CHARACTERISTICS IN CS APPLICATIONS

- 3.1 Data quality problems ................................................................................................... 30
LIST OF SYMBOLS AND ABBREVIATIONS

CS- Citizen Science
DQ- Data Quality
CMM- Capability Maturity Model
DQMM- Data Quality Maturity Model
IS- Information System
CL- Capability Level
RQ- Research Question
Chapter 1: Introduction

During past decades, public participations of science is increasing day by day and citizen science has been a central part in this[1]. "Citizen Science" (CS) is the collaboration where normal public are interested to giving their effort on a certain topic and researcher can find answer to those scientific questions. As a great amount of people are involving themselves with citizen science projects. There are variety of examples in citizen science applications and projects such as scientific research, environmental education and technology, community service, and ecological problem solving are some examples [2]. A good citizen science project is designed after a lot of thoughts. Citizen science is an application what allows the researchers to make report, collect data and give opportunity to citizens to directly participate to expand their research on science and management, also includes public participation with volunteer monitoring based on geographical information [2], [3].

There are lot of aspects related to citizen science applications and one of the widely discussed aspects is data quality measurement [4]. Open collaboration tackles this data quality in many ways. The quality of the information varies since it is maintained with unknown volunteers with varying levels of knowledge and skills [4].

Data validation is one of the important factors for recognition and checking the usefulness and limits the value of data collected [4]. Several studies have examined data quality in citizen science projects by way of finding out predictors of participant success. Accuracy p inside these programs tend to vary, and results are not often made accessible to the larger citizen science community. Standardizing monitoring protocols, designed by the way of authorities and field-tested with citizen scientists working below practical conditions, can enhance data quality and analyses[4].

1.1 Research Problem

This research addresses the challenges of the data quality as well as the development of a data quality maturity model for understanding the quality characteristics levels of citizen science applications. General purpose of the citizen science application is to acquire and store data that matches with citizen science applications[2], [3].The research explores the ISO data quality characteristics [5] such as reliability, consistency, availability, accessibility, accuracy, validity, legitimacy, timeliness and so on. In addition, the best possible scenarios to collect data. The main goal of this research project is to develop a model based on this data quality characteristics to measure the data quality maturity of citizen science applications.

There is another purpose of this research, that is once we have a data quality maturity model it will facilitate the ability to measure the maturity level of the citizen science application, which is lacking in the literature. This measurement of data quality level in a citizen science application will help to progress the quality of the citizen science application.
1.2 Research objectives and questions

The main goal of this thesis is to study data quality characteristics and its significance on the citizen science databases. It will help future researchers to use these findings in their data quality characteristics verifications and for applying the maturity model based on data quality characteristics.

Based on these deliberations the following research questions have been prepared:

<table>
<thead>
<tr>
<th>Research Questions (RQ)</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>• RQ1: What are the Data quality problems in Citizen Science application?</td>
<td>Identify and describe the problem of citizen science applications based on some level.</td>
</tr>
<tr>
<td>• What are the data quality characteristics specific to citizen science applications?</td>
<td>Identify quality characteristics in the CS application.</td>
</tr>
<tr>
<td>• How to develop a data quality maturity model for citizen science applications?</td>
<td>Developing maturity model related to the data quality characteristics in citizen science applications. Experiment with CS applications to validate the data quality maturity model</td>
</tr>
<tr>
<td>• What are the benefits, outcomes and challenges of data quality Maturity model?</td>
<td>Identify and describe the main benefits, outcomes and challenges of DQMM (Data quality Maturity Model) for citizen science applications</td>
</tr>
</tbody>
</table>

Table 1: Research question

1.3 Research tasks and directions

1. Describe the citizen science application.

2. Describe data quality characteristics of citizen science applications.

3. Collect relevant articles about citizen science, data quality, data quality characteristics, data quality framework, framework for citizen science applications and mostly data quality maturity model for citizen science applications.
4. Collect literature on capability maturity models in the area of Information Systems (IS)
5. Develop a data quality maturity model for citizen science applications using the knowledge of capability maturity models
6. Apply the data quality maturity model to citizen science applications
7. Assess the add value of the data quality maturity model for citizen science applications.

1.4 Research Methodology
This research comprises of two main steps: a systematic literature study and a qualitative empirical research.
The literature study covers following key terms:
• Citizen Science
• Data quality characteristics
• Data quality in Databases
• Data quality Framework
• Data quality Framework and citizen science applications
• Capability Maturity Model (CMM)
Development of the data quality maturity model for citizen science applications.
The empirical study is based on two major prerequisites:
1. Achieving data quality characteristics for the data quality maturity model. Several published articles are used for this motive in the thesis.
2. Evaluating number of citizen science applications using the developed data quality maturity model

1.5 Structure of the thesis
This thesis is structured into seven different chapters.
Chapter 1 is Introduction which will provide the Problem of the research formulation that covers the research area. Also, it covers the goal of the research work.
Chapter 2 is titled as State of the Art which mainly explains literature related to this research work of Citizen science applications and data quality characteristics. Moreover, the author explains what the maturity models in CS and framework of CS are found in research papers.
Chapter 3 covers the problem and citizen science characteristics in details. Data quality characteristics problem and relation to the attributes of data quality framework.

Chapter 4 expands with checking the quality attributes for the data quality maturity models with CS application projects.

Chapter 5 describes the designed data maturity model DQMM

Chapter 6 covers the discussion and findings

Chapter 7 gives conclusions of the research including future research undertakings.
Chapter 2: State-of-the-art

2.1 About Data quality models and frameworks

The strength of computers stems from their ability to signify our physical actuality as a digital world and their potential to follow guidelines with which to manipulate that world. Ideas, images, and information that can be translated into bits of records and processed by using computers to create apps, animations, or different sorts of software program applications. The variety of directions that a computer can observe makes it an engine of innovation that is restrained only by means of our imagination. Remarkably, computers can even comply with instructions about directions in the shape of programming languages [7].

Technological advancements have improved the ways we implement successful applications. Citizen Science applications have also accommodated modern capabilities into its functionality. Most of the citizen science programs [8] have developed online records entry types with automated error checking capabilities. These types flag suspected records to permit further specialist investigation prior to their integration into an extensively used datasets. Similarly, smartphone applications [8] have been developed that allow computerized entry of area coordinates for example, related with a species sighting. These sightings should consist of a photo voucher or an identification tag for a specimen voucher. Although the capacity of these tools to enhance statistics satisfactorily has no longer been wholly tested, it is probable that they could improve data quality among all data collectors [8].

This chapter covers a literature review of related works and data quality related topics useful for the improvement as well as development of a data quality maturity model for CS applications.

2.2 Literature Review:

Following scientific databases have been used for reviewing relevant articles: ACM, IEEE, and Springer, science direct, web of science.

The keywords, used in the literature study research are ‘‘Citizen science applications – definitions’’, ‘‘data quality characteristics’’, ‘‘data quality in databases ‘‘, ‘‘ Data quality frameworks – ISO data quality framework ‘‘, ‘‘ Data quality & citizen science applications’, ‘‘ Data quality frameworks for citizen science applications’’, ‘‘ Capability Maturity Model(CMM)’’.

<table>
<thead>
<tr>
<th>Database name</th>
<th>CS definition</th>
<th>Data quality characteristics</th>
<th>Data quality in database</th>
<th>Data quality framework</th>
<th>Data quality framework and CS application</th>
<th>Capability Maturity Model (CMM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE</td>
<td>170</td>
<td>684</td>
<td>404</td>
<td>288</td>
<td>8</td>
<td>45</td>
</tr>
<tr>
<td>ACM</td>
<td>43</td>
<td>231</td>
<td>871</td>
<td>987</td>
<td>102</td>
<td>139</td>
</tr>
<tr>
<td>Database</td>
<td>102</td>
<td>190</td>
<td>304</td>
<td>865</td>
<td>111</td>
<td>67</td>
</tr>
<tr>
<td>---------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td>Springer</td>
<td>682</td>
<td>990</td>
<td>771</td>
<td>408</td>
<td>203</td>
<td>206</td>
</tr>
<tr>
<td>Science direct</td>
<td>542</td>
<td>810</td>
<td>506</td>
<td>202</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 2: Scientific databases and their representation of knowledge

### 2.2.1 Literature Review Process

Articles have been searched with publication years between 2011 and 2019 mostly. The process starts from a general phrase and then results are limited down by adding definite phrases to achieve exact results. After that it is possible to look for relevant articles. Table 2 shows information about the number of articles on each key phrase in 5 different scientific databases and sources. However, not every article in these databases contains the required information. Figure 1 summarizes the way how suitable articles are collected during this study to make a reasonably comprehensive study of the relevant topics related to the research question and research topic.

![Figure 1: The process carried out for collecting relevant scientific articles for the literature study](image-url)
2.3 Citizen science

Now-a-days the more logical apparatuses and machines you have, the more arrangements of information might be gathered. In any case, they have few issues. A few issues simply don't have legitimate or rich calculations to understand them, while it might set aside opportunity to discover an answer for other people. In any case, there is still expectation – since innovations and Internet are developing, it brings increasingly more reasons for critical thinking and citizen science is one of them. It is generally a new area of science that enables individuals to take an interest in a logical procedure by providing information. For this situation, natives are forwarder who are mentioning objective facts, gathering information, making estimations lastly translating information with no logical learning. Thus, citizen can enhance science by mentioning such objective facts around the globe in a way that would have been impractical before [9].

2.3.1 Citizen science definition

Many researchers, writers, and scientists have different opinion about Citizen Science (CS). Citizen science has been described and plotted by many types of researchers. Below table 3 presents the related definition and description of Citizen science from literature.

<table>
<thead>
<tr>
<th>Research</th>
<th>Definition and description</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Science of Citizen Science: Theories, Methodologies and Platforms[6]</td>
<td>“Citizen science involves members of the public as collaborators in scientific inquiry, creating opportunities for new or improved research”</td>
</tr>
<tr>
<td>Automated Data Verification in a Large-scale Citizen Science Project: a Case Study[7]</td>
<td>“Citizen science enlists the help of volunteers from the general public (citizen scientists) in scientific research”</td>
</tr>
<tr>
<td>Citizen science or scientific citizenship?[8]</td>
<td>“Citizen science has become an umbrella term that applies to a wide range of activities that involve the public in science”</td>
</tr>
<tr>
<td>Citizen Science[9]</td>
<td>“Citizen scientists contribute wide-ranging research that spans scales from observations of individual organisms (or even genes) to ecosystem-level assessments and analyses of images of landscapes taken by satellites”</td>
</tr>
<tr>
<td>User experience of digital technologies in citizen science[10]</td>
<td>“Citizen science involves the collaboration or partnership between professional scientists and amateurs, volunteers, and even scientists outside their prescribed role, who jointly take part in scientific endeavours”</td>
</tr>
</tbody>
</table>

Table 3: Different definitions of citizen science
Above descriptions and definitions are identified from different research papers in Table 2, it is describing that citizen science means creating and making opportunities for general citizens. Citizen science is applied in various sector with the focus of engaging people. This background has been used to calibrate the framework of data quality characteristics for different citizen science applications [11]

The definition of Citizen science makes a symbolic point of view of engaging citizen in certain scientific domain. Though some other facts lead to what is the actual impact of it and how to make this engagement process. A. Skarlatidou, M. Ponti, J. Sprinks, and C. Nold stated that CS is not only for the service but also it makes partnership between the volunteers and professionals[10]. The definition of citizen science itself point out the impact of CS in order to match people in certain activities as well as to expands value creation.

2.3.2 ISO quality Standard

ISO is an international standardization organization that has created many widely used standard for data quality. What sets ISO apart from most of the other definitions is the classification of these characteristics between inherent and system-dependent characteristics.[12]

Citizen science projects involve different types of data quality characteristics. This characteristic comes from ISO data quality standard [13]. The Data Quality model speaks to the grounds where the framework for evaluating the nature based on the items of information. In a Data Quality model, the primary Data Quality attributes must be considered while evaluating the properties of the planned information categories.

Most of the attributes of data quality characteristics are considered to gain information for making standard ISO quality. Member protection is mapped under privacy and accessibility is considered from open information point of view. Consistence, compactness and recoverability are the main attributes that are not considered. Consistence isn't considered because it is hard to see whether any standard has been pursued without seeing documentation of the task and conceivably their information demonstrate. Convey ability and recoverability are not considered because they identify with the framework, have almost no impact on others, and are hard to quantify without access to the fundamental framework [4].

The data quality characteristics presented in ISO data quality standard are elaborated below with respect to citizen science applications. This data quality characteristics descriptions given below that are adjusted to citizen science applications situations are found in [3]
2.4 Data quality characteristics for citizen science applications [12]

**Accuracy**: Data accuracy is partition into syntactic and semantic accuracy. Syntactic accuracy means that the given data matches the syntax in the data model for example: number is number, text is text, date is date etc. Semantic accuracy means the given data’s semantics matches the semantics in the data model such as when asking a location, the location makes sense. [12]

**Completeness**: Data completeness explains how complete the submitted data is against all possible attributes of the data entity. The attributes refer to not only all mandatory but also optional fields the participant can fill, and these are put against what the participant submits. [12]

**Consistency**: Data consistency address how consistent data is across the tables and database. The data should be consistent in all of them if the same data entity is in different tables. And, the data entities within the same table should be consistent and not have varying data in same attributes. Consistency is checked from syntactic accuracy and completeness.[12]

**Credibility**: Data credibility mentions how credible a data entity is. If the data is gotten through sensors, the credibility is measured by the credibility of that sensor. The credibility is measured by the credibility of that human when data is gotten through human observation. Researchers are thought to always give credible data and only their methods are questioned but the credibility of the human is questioned when a citizen gives data.[12]

**Correctness**: Data correctness means the degree of which data attributes are of the right age. For example, the difference of observation to the difference when submitting an observation. Both times should be mentioned If the time is different.[12]

**Accessibility**: Accessibility described how accessible the data is in a specific context. Is the available data need specific types of equipment or configurations. In this research, accessibility is measured how data is accessed by participants or outsiders in the project excluding the download of data.[12]

**Confidentiality**: Confidentiality means that the data is only accessible and usable by authorized users only. In this research, privacy is considered as the major factor in confidentiality.[12]

**Efficiency**: Efficiency points to how efficiently the data can be processed and accessed, how well the system performs with the given data. [12]

**Precision**: Precision describes to how exact the attributes are within the data. The acceptable degree of precision can be defined within a project. Being able to give estimates decreases the precision of data but giving an estimate is better than not giving any information.[12]

**Traceability (provenance)**: Traceability addressed to how well the changes and access to the data can be traced. This can be called data provenance. [12]

**Understandability**: Understandability means how understandable the data is, how well can people understand and interpret the data when reading, and how well is the data represented with appropriate symbols, language, and units.[12]
Availability: Availability refers to how available the data is for authorized and unauthorized user applications. In this research, availability refers to if data can be downloaded from the project.[12]

2.5 Data Quality (DQ) in databases

Data quality analysis plays a very important role within the method of knowledge mining, aiming at guaranteeing accuracy, completeness, consistency, validity, uniqueness, timeliness, and stability etc of knowledge [11]. several researchers like better to study strategies of knowledge conversion automation and cleansing from the angle of schema and integration [11]. Researchers have done abundant work on totally different fields of knowledge quality as well as detection of duplicate records, examination of knowledge integrity, verification of knowledge validation, integration between heterogeneous knowledge sources and style of ETL tools for knowledge extraction etc[14].

Data quality is implemented in databases using several different concepts [14]. Those concepts are used so that there are specific components and subcomponents where the list can be changed after initial definition. The information from different folders can be merged into final output to guarantee data quality.

Database systems are designed to store what data should be used in analysis, and the technical implementation reflects this principle[15].

DQ will become a precondition to guarantee the information values, so that its use may be high quality and promote the company’s competitiveness. Data administration is increasingly beneficial property significant in the modern-day corporate scenarios, stimulating the introduction of a records administration maturity model[16].

The specialized literature proposes different DQ management methodologies, such as, TDQM (Total Data Quality Management)- TDQM adopts the information perspective as a product, which has a defined production cycle[18], TQIM (Total Quality Information Management)-TQIM considers that a data quality project is beyond a data improvement or cleaning process[18], DWQ (Data Warehouse Quality Methodology)- It is very important for data warehouse designers to follow a consolidated and robust conceptual design methodology, as the development of a data warehouse is a very expensive process[20], AMIQ (A methodology for information quality assessment)- AMIQ uses the generic term information, and performs qualitative evaluations using questions which apply to structured data, but may refer to any type of data, including unstructured ones[21], HDQM (Heterogeneous Data Quality Management)-HDQM can be extended to other unstructured data types (like texts, images and sounds) by encompassing dimensions, metrics and improvement techniques tailored to specific type of data[21].

Researchers identified five dimensions of data quality and seven categories of data quality assessment methods[17]. The dimensions are: Completeness, correctness, concordance,
plausibility, currency. The categories are: Gold standard, data element agreement, Element presence, Data source agreement, Distribution comparison, Validity check, Log review [17].

![Data quality dimensions and categories](image)

**Figure 2: Data quality dimensions and categories** [17]

One of the largest difficulties in conducting the overview of the result is the inconsistent terminology used to discuss records quality. We had not expected, for example, the overlap of terms between dimensions, or the fact that the language inside a single article was occasionally inconsistent. The lookup community has largely failed to advance or undertake a consistent taxonomy of data pleasant [17].

Interdisciplinary lookup addressing the challenges of information quality touches a diversity of subjects which include economics, psychology, records systems, statistics mining, database technology, and many others. While statistics systems lookup concentrates on the modelling of many facets and dimensions of facts quality, science has a center of attention on the algorithmic and statistics administration factors of analysing statistics with respect to diverse first-class criteria and maintaining or growing quality. This Special Issue on Data Quality in Databases highlights four of the present day computational and algorithmic effects [18]:

— to analyse data quality, use of data mining techniques;

— methods to implant information quality dimensions into a (sensor stream) database;
— a contemporary method to clean data within a database;
— an access to make use of data quality values by incorporating them into database queries. [18]

2.6 Data quality frameworks

Data quality may be a cross disciplinary space wherever central themes and topics are related to the methodologies. Central to many studies is the information quality. So, the size ought to characterize varied information properties like accuracy, currency, completeness. Dimensions are typically self-addressed for assessment functions, and accustomed to living the standard of information. It's long been acknowledged that information is best delineate or analysed via multiple dimensions and there are many proposals in information quality analysis literature for varied classifications and definitions of information quality dimensions [19].

According to [19, 20] a practical and an ordinary framework for DQ aware query answering ought to be decoupled from the precise definition of a DQ dimension. Querying in records integration systems with consideration of records first-rate is studied in a quantity of works [20].

Data exceptional troubles arise when dealing with multiple information sources. This increases the data cleaning needs significantly. In addition, the large dimension of records units that arrive at an uncontrolled speed generates an overhead on the cleaning processes. In the below framework, it uses a combined approach, which consists of a information satisfactory management gadget that deals with best rules for all the pre-processing things to do prior to facts analysis[21].

![Data quality measurement framework (21)](image-url)
The idea of the value of poorly structured knowledge created in immense amounts, requests a far required knowledge quality assessment framework and potential knowledge quality metrics [26]. sometimes the linking of knowledge sets is completed through a method named ETL (Extract remodel Load) with the intention to make an information deposition framework for economical data processing processes to be enforced [27].

DQ frameworks typically follow the 3 distinct steps: state reconstruction, assessing/measuring the standard then upgrading the quality [21]. this is often the overall approach that the majority methodologies take, as well as that of the Heterogeneous knowledge Quality Methodology (HDQM) [21]. HDQM focuses the nonuniformity as outlined by a mixture of structured, semi structured and unstructured knowledge sources. They analyse the standard of "Conceptual Entities" that mix quality measures from all completely different sources with reference to an entity to provide a top quality of that abstract entity. One issue with most standing frameworks is their lack of qualitative assessments[22].

There has so much complexity to complete the data base on knowledge. The impact to complete the knowledge is so much than the impact of the data itself[20]

Improve daily operations, Make decisions and change strategy, increase financial impact, Secure data and comply for all stakeholders, Existence of source of data, recognize properties of data that have awesome weight distinguishing proof in connection to another information source, nature of data and how to proceeds to execute are the main factor [23].

2.6.1 ISO data quality framework

ISO 8000 comprise a process reference model that has namely a maturity model and established principles on it [13]

<table>
<thead>
<tr>
<th>Framework</th>
<th>Data management</th>
<th>Data quality management</th>
<th>Data governance</th>
</tr>
</thead>
<tbody>
<tr>
<td>English [24]</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>CALDEA[25]</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>IQM3[26]</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>IQM[27]</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Aiken et al.[28]</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMM[29]</td>
<td>x</td>
<td>x</td>
<td>X</td>
</tr>
<tr>
<td>LAIDQ[30]</td>
<td>x</td>
<td>x</td>
<td>X</td>
</tr>
<tr>
<td>ISO 8000/61[31]</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>DAMA[16]</td>
<td>x</td>
<td>x</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 4: Data maturity model classifications according to their scope[31].

- facilitates self-assessment;
- provides a basis to be used in method improvement and capability determination;
- considers the context during which the assessed method is implemented;
- produces a method rating;
- addresses the power of the method to realize its purpose;
- is acceptable across all application domains and sizes of organization;
- could offer an objective benchmark between organization [5].

Process assessment model is consisting of a standard that is intended to reflect the whole software engineering activities, the set of work products of a new process should already be founded inside this model, and most time, no additional work product description should be required. However, the new process may be outside the domain of this process assessment model. In this case, new work products should be identified, and their traits ought to facilitate the mapping between the new process and any specific organisation methods [13].

2.7 Data Quality and Citizen Science Application

We live in an era of the evolution of mobile devices and web technology that allow us to access data anytime and anywhere. The internet gives the citizens to engage in collaborative knowledge building process. Data quality is a key aspect of citizen science that requires further investigation. Data quality in community based monitoring programs are common ground to related work in the area of collaborator network to evaluate quality data through different measures [32].

Data quality in CS applications are multi-dimensional. Some metrics are mission dependent, such as timeliness of information for a question. Accuracy is the diploma to which data are right overall, while bias is systematic error in a dataset [33].

Data base of citizen science no longer solely an attribute of data however also a method of keeping data quality. It is regarded to be “the approaches and results of evaluating and enhancing the utility of data.” When reviewing data quality, specialists are asked to decide the chance that a given document is reliable; this is how the truth of observational information is established. we anticipate provenance and stewardship to be at least as essential for assess-ability. Indeed, the requirements of scientific understanding manufacturing routinely require such facts [10].

Data collected by subject scientists may be achieved at very little price, enabling research to assemble knowledge over longer periods of time and across broader specializations [35].
Data filters, data excellent models, estimating know-how degree are some techniques utilized in CS initiatives. Data is a fundamental challenge in any sensor network, particularly when the sensor community consists of a big quantity of volunteer observers that have differing competencies to accurately perceive project objectives. A foremost statistic best wished in extensive scale citizen science tasks is the capability to filter misidentified information. Automated statistics generation is a challenge for improving records quality in CS applications [36].

Citizen science is, at its core, a way of empowering people to help with large-scale scientific research. Perhaps the most successful citizen science challenge to date is the Christmas Bird Count (CBC), a task started over a century in the past by way of the Audubon Society, which asks volunteers at some stage in the vacations to remember the number of birds in each species they sight [34].

Quality control via expert overview implies validation, which means that a third party evaluates facts and determines whether it is acceptable. Citizen science offers possibilities for human beings with one-of-a-kind backgrounds and cultures to address society-driven question. Several researches show that by making use of splendid coaching it is possible to accumulate information equal to those gathered by way of experts.

In [12] data has been collected by going through different citizen science projects and their applications. Some of the projects have been found from different articles and most have been found through SciStarter, which describes over 1400 different project. Citizen science project platforms offer a fast and easy way to start citizen science projects. People can create new projects in under half an hour and start collecting data. Many of the presented data quality characteristics are vital to perceived data quality. According to [12] accuracy, credibility and completeness are the most easily perceived characteristics and often the baseline for perceived data quality. These characteristics are mostly tied to participants especially if the validation is done by community. The lack of quality control over these characteristics is easily perceived by participants creating a lack of trust over the collected data [12].

2.7.1 A general framework

Citizen-science programs consist of a manner for balancing inevitable conflicts between competing dreams and creating everyday opportunities and mechanisms for neighbourhood oversight. This may additionally contain inviting community leaders onto oversight or advisory committees, web hosting ordinary casual interplay between scientists and local neighbourhood individuals and/or offering advanced scientific training to a subset of neighbourhood individuals. This can be achieved by partnering with existing local organizations, as in the Celebrate Urban Birds program [35] or by leveraging existing channels, as in the White Earth [35] example, where the connection between the college and tribal leadership provided a means to guide citizen science efforts to align with and support cultural practices and community values [35]. General framework mostly depends on data validation.
2.7.2 Data validation

Data validation is one of the qualities to measure the CS applications. The semi-automated process of on line data filters and reviewing described in figure 4 helped researchers to validate reports submitted by a community of dispersed individuals in a continental-scale citizen-science venture.[36]

![Figure 4: Data validation process][36]

The validation process should work like in the figure 4 for CS applications. The data screening and validation system describe number of enormous observations and typographical error for the database that not only receive the error message also correct the wrong message from the database. This type of correction is helpful for the participants to become more knowledgeable about CS the applications [43].

Most data quality issues relate to amateurs giving data instead of professional researchers. According to a survey [12] inadequacy of participants is the main concern of many citizen science projects. The inadequacy leads to poor and non-reliable observations, which in turn leads to low quality data. Similarly, another survey [8] found out, that data quality is a big concern and common methods to increase quality are expert validation and volunteer training. The survey suggests that automated tests and user-friendly tools can help improve the quality of data [12].

2.7.3 User acceptance factors in CS applications

There are three widely acknowledged user acceptance factor in CS. Those are quality of Data, privacy and motivation [37].
Quality of data: A well known computational approach is excess where different laborers rehash the same errand. A notoriety show is another component proposed to extend the unwavering quality of citizen-collected information.

Privacy: Tangible information is our advanced impression, inserting points of interest of way of life. There’s an inherent conflict between information sharing and security. Area information may be a principal portion of the information being collected in citizen science applications, which when coupled with date and time information raises several extra concerns around security.

Motivation: Pulling in and holding volunteers is vital for planning technology-in pondered science ventures. In this way, motivational variables need to be tended to when building an application. Analysts have found that inherent and collective inspirations like individual interface, delight, taking after social standards, or affirmation are overwhelming variables.[37]

<table>
<thead>
<tr>
<th>Authoring tool</th>
<th>Use of technology</th>
<th>User acceptance factors</th>
<th>Authoring type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical</td>
<td>Various</td>
<td>No support for distribution</td>
<td>Various</td>
</tr>
<tr>
<td>Sensr</td>
<td>Locations/Photos/Text</td>
<td>Automatically launched in a web platform</td>
<td>Filtering/Validation</td>
</tr>
<tr>
<td>Ushahidi</td>
<td>Various</td>
<td>Automatically launched within an app</td>
<td>Various</td>
</tr>
<tr>
<td>Noah</td>
<td>Locations/Photos/Text</td>
<td>No support for distribution</td>
<td>Collection process</td>
</tr>
<tr>
<td>Epicollect</td>
<td>Locations/Photos/Text</td>
<td>No support for distribution</td>
<td>Not specified</td>
</tr>
</tbody>
</table>

Table 5: Some factor in user acceptance[37]

Researchers have used actual systems and different platform to support citizen science. The technology concern also measures quality. Twitter is a great platform for customize the flexibility. So, there is a need to build CS platform to maximize the capacity of mobile devices [37].

2.8 Capability Maturity Model (CMM)

Early research of data quality cantered on data status quality and data benefit quality. Afterward, the significantly moved from data status quality to information structure quality. More as of late, data quality administration considerations have included data administration as well as data service, value, and structure. Underneath figure shows that data quality variables can be partitioned into three spaces: data value components (information esteem and information benefit), data structure variables, and data service quality variables.[38]
Data quality management system step by step move into Capability Maturity Model (CMM). The CMM is based on the idea of continues improvement and aid organizations in prioritizing its improvement efforts. The CMM proposes five different maturity levels, in which achieving each level of maturity establishes a different component in a software process, resulting in an increase in the process capability of an organization. Each maturity level forms a foundation for the next. The five maturity levels are described below [46].

1. At level 1: Initial, an organization regularly does not give a steady environment for creating and keeping up program. These sorts of firms have challenges with commitment of staff, and this could result in emergency. In an emergency, ventures regularly forsake arranged methods. Center is given to people, not organizations[39].

2. At level 2: Repeatable, arrangements for administration and strategies to actualize those approaches are built up. Unused ventures are based on encounters with comparable ventures. Extend benchmarks are characterized and the organization guarantees that they are loyally taken after. Level 2 organizations are restrained since venture arranging and following are steady and prior triumphs can be rehashed[39].

3. At level 3: Defined, a commonplace prepares for creating and keeping up computer program over the organization is archived, counting the software-engineering and administration forms. A characterized handling contains a coherent, coordinates set of well-defined programs designing and administration forms, both are steady and repeatable. This capability is based on a common, organization-wide understanding of exercises, parts, and obligations [40]

4. At level 4: Managed, an organization sets quantitative quality objectives for both items and forms with well-defined and steady estimations. An organization-wide handling database is utilized to gather and analyze the information accessible from a project’s characterized forms. The dangers included in moving up the learning curve and carefully overseeing. When limits are surpassed, supervisors take activity to rectify the circumstance[39].
5. At level 5: Optimizing, the complete organization is centered on nonstop enhancement. The organization has the capacity to identify shortcomings and reinforce the method proactively, with the objective of avoiding problems. At level 5, squander is unsatisfactory; organized endeavours to expel squander result in changing the framework by changing the common causes of wastefulness. Decreasing squander happens at all development levels, but it is the center of level 5. Change happens both by incremental headways within the existing and by developments in innovations and strategies [40].

![Image: Five levels of software process maturity](image)

**Figure 6: Five levels of software process maturity[40]**

The maturity model provides a good overview of each maturity level. Firm can use this model to think about their current or desirable state with regards to data quality management. It is important to note that every level set up the formation for the next level. The contribution of data quality management for below model is that it characterizes some aspects of data quality within. In addition, each maturity level gives a certain amount of attention to data quality[39].
Data quality maturity model is patterned after the Capability Maturity Model (CMM) developed by means of the Software Engineering Institute at Carnegie Mellon University. Capability maturity models are management equipment that signify tiers of organizational refinement in addressing design, implementation, manufacturing, trouble resolution, and so on. These kinds of models have been applied to many application domains, which include software program development, programmer development, and mission administration domains. This data excellent maturity model defines five levels of maturity, ranging from an preliminary stage where practices and policies are ad hoc, to the highest in which tactics and practices lead to continuous measurement, improvement, and optimization[41].

Below table describe the mapping of maturity stages in accordance to expectations and data quality dimensions.[41]

<table>
<thead>
<tr>
<th>Maturity level 1: Person dependent and basic.</th>
<th>Maturity level 2: Policies, standards and procedures.</th>
<th>Maturity level 3: Data model optimizations.</th>
<th>Maturity level 4: Managed and standardized.</th>
<th>Maturity level 5: Continues improvement.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.) Focused on individuals, therefore tasks are person dependent.</td>
<td>1.) Policies, standards and procedures are defined or updated.</td>
<td>1.) Changes in the data structure of the firm are reflected to the data model.</td>
<td>1.) The data within the firm is standardized. This ensures sharing and reusing.</td>
<td>1.) The firm is continuously improving.</td>
</tr>
<tr>
<td>2.) Knowledge is lost when individual is not available.</td>
<td>2.) Earlier success can be repeated, because these are defined accordingly.</td>
<td>2.) The firm educates her employees so that new knowledge and skills can be acquired.</td>
<td>2.) The firm sets quantitative goals for both products and processes with well-defined measurements.</td>
<td>2.) The firm can identify their strengths and weaknesses.</td>
</tr>
<tr>
<td>3.) Systems in the firm are not maintained regularly.</td>
<td></td>
<td></td>
<td></td>
<td>3.) The main focus of the firm is to reduce waste.</td>
</tr>
</tbody>
</table>

Table 6: Maturity model for determining the level of data quality management [44]
<table>
<thead>
<tr>
<th>Level</th>
<th>Characterization(expectations)</th>
<th>Characterization(dimensions)</th>
</tr>
</thead>
</table>
| Initial | • Data quality activity is reactive  
• No capability for identifying data quality expectations  
• No data quality expectations have been documented | • No recognition of ability to measure data quality  
• Data quality issues not connected in any way  
• Data quality issues are not characterized within any kind of management taxonomy |
| Repeated | • Limited anticipation of certain data issues  
• Expectations associated with intrinsic dimensions of data quality (see chapter 8) associated with data values can be articulated  
• Simple errors are identified and reported | • Recognition of common dimensions for measuring quality of data values  
• Capability to measure conformance with data quality rules associated with data values |
| Defined | • Dimensions of data quality are identified and documented  
• Expectations associated with dimensions of data quality associated with data values, formats, and semantics can be articulated using data quality rules  
• Capability for validation of data using defined data quality rules  
• Methods for assessing business impact explored | • Expectations associated with dimensions of data quality associated with data values, formats, and semantics can be articulated  
• Capability for validation of data values, models, and exchanges using defined data quality rules  
• Basic reporting for simple data quality measurements |
| Managed | • Data validity is inspected and monitored in process  
• Business impact analysis of data flaws is common  
• Results of impact analysis factored into prioritization of managing expectation conformance  
• Data quality assessments of data sets performed on cyclic schedule | • Dimensions of data quality mapped to a business impact taxonomy  
• Composite metric scores reported  
• Data stewards notified of emerging data flaws |
Maturity models replicate the extent to which key approaches or things to do are defined, managed, and accomplished efficiently and produce dependable results. They generally describe the characteristics of a pastime at a range of exceptional tiers of performance [50]. Approaches to decide manner or functionality maturity are increasingly applied to various factors of product development, both as an assessment instrument and as phase of an improvement framework [45, 46].

While making the design of maturity model, the design, quality, aspects & products must be considered. On a coarse level, it is advocated to structure maturity fashions hierarchically into more than one layers. On a certain level, it outlines a meta-model together with factors such as competence objects, maturity levels, criteria, and methods for statistics collection and analysis. It perceive the following components: levels, descriptors, descriptions for every level, dimensions, system areas, things to do for every process area, and a description of each exercise as carried out at a certain maturity level [43]. According to [43] the maturity model design layout should be as in the below table:

| Optimized | • Data quality benchmarks defined  
• Observance of data quality expectations tied to individual performance targets  
• Industry proficiency levels are used for anticipating and setting improvement goals  
• Controls for data validation integrated into business processes | • Data quality service level agreements defined  
• Data quality service level agreements observed  
• Newly researched dimensions enable the integration of proactive methods for ensuring the quality of data as part of the system development life cycle |

Table 7: Mapping of maturity level in data quality expectations and dimensions [41]
<table>
<thead>
<tr>
<th>Group</th>
<th>Design Principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 BASIC</td>
<td>a) Application domain and prerequisites for applicability</td>
</tr>
<tr>
<td></td>
<td>b) Purpose of use</td>
</tr>
<tr>
<td></td>
<td>c) Target group</td>
</tr>
<tr>
<td></td>
<td>d) Class of entities under investigation</td>
</tr>
<tr>
<td></td>
<td>e) Differentiation from related maturity models</td>
</tr>
<tr>
<td></td>
<td>f) Design process and extent of empirical validation</td>
</tr>
<tr>
<td>1.2</td>
<td>Definition of central constructs related to maturity and maturation</td>
</tr>
<tr>
<td></td>
<td>a) Maturity and dimensions of maturity</td>
</tr>
<tr>
<td></td>
<td>b) Maturity levels and maturation paths</td>
</tr>
<tr>
<td></td>
<td>c) Available levels of granularity of maturation</td>
</tr>
<tr>
<td></td>
<td>d) Underpinning theoretical foundations with respect to evolution and change</td>
</tr>
<tr>
<td>1.3</td>
<td>Definition of central constructs related to the application domain</td>
</tr>
<tr>
<td>1.4</td>
<td>Target group-oriented documentation</td>
</tr>
<tr>
<td>2.1 DESCRIPTIVE</td>
<td>Intersubjectively verifiable criteria for each maturity level and level of granularity</td>
</tr>
<tr>
<td>2.2</td>
<td>Target group-oriented assessment methodology</td>
</tr>
<tr>
<td></td>
<td>a) Procedure model</td>
</tr>
<tr>
<td></td>
<td>b) Advice on the assessment of criteria</td>
</tr>
<tr>
<td></td>
<td>c) Advice on the adaptation and configuration of criteria</td>
</tr>
<tr>
<td></td>
<td>d) Expert knowledge from previous application</td>
</tr>
<tr>
<td>3.1</td>
<td>Improvement measures for each maturity level and level of granularity</td>
</tr>
<tr>
<td>3.2</td>
<td>Decision calculus for selecting improvement measures</td>
</tr>
<tr>
<td></td>
<td>a) Clarification of relevant objectives</td>
</tr>
<tr>
<td></td>
<td>b) Clarification of relevant factors of influence</td>
</tr>
<tr>
<td></td>
<td>c) Distinction between an external reporting and an internal improvement perspective</td>
</tr>
<tr>
<td>3.3</td>
<td>Target group-oriented decision methodology</td>
</tr>
<tr>
<td></td>
<td>a) Procedure model</td>
</tr>
<tr>
<td></td>
<td>b) Advice on the assessment of variables</td>
</tr>
<tr>
<td></td>
<td>c) Advice on the concretization and adoption of the improvement measures</td>
</tr>
<tr>
<td></td>
<td>d) Advice on the adaptation and configuration of the decision calculus</td>
</tr>
<tr>
<td></td>
<td>e) Expert knowledge from previous application</td>
</tr>
</tbody>
</table>

Table 8: A framework of general design principles for maturity models [47]
Chapter 3: Data quality problems and characteristics in CS application

3.1 Data quality problems

In recent years, citizen science projects have grown dramatically. They combine Web-based social networks with community-based data systems to harness collective knowledge and observe a specific scientific problem[44]. However, there are some inherent weaknesses. The restricted coaching, information and experience of contributors and their relative knowledge cause poor quality, deceptive or maybe malicious information being submitted[45].

The unavailability of a ‘scientific method’[46] and the use of non-standardized and low quality design methods of data collection[47] often lead to incomplete or inaccurate data. Also, the lack of commitment from volunteers in collecting discipline information [45][48] can cause gaps within the information across time. afterward, these problems have caused several scientific communities to question the seriousness of such undertakings [53].

3.1.1 Issues defined

There are significant number of issues found for the lack of data quality. [44]:

- shortage of validation and consistency confirmation;
- absence of a defined and extensible data model and comparable schema;
- shortage of data quality assessment measures;
- shortage of automated metadata/data;
- shortage of user verification and automatic attribution of data to individuals
- shortage of assessment to volunteers on their data;
- shortage of graphing, trend analysis and visualization tools that enable errors or removed data to be easily detected and corrected[44].

3.2 Data analysis for evaluation

The systematic assortment and analysis of knowledge required to assess the “strengths and weaknesses of programs, policies, personnel, products, and organizations” to boost their overall effectiveness. A comparison of learning gains for people through citizen-science participation is critically necessary for understanding whether or not the program is meeting its instructional and volunteer engagement goals[49].

Evaluating the impacts of citizen-science projects on gaining knowledge at the level of the individual, program, or community can also ultimately increase the possibilities of assignment success and contribute to the society. Although researchers have begun to discover the
integration of results at more than one scales, it is argued that this must be at the forefront of future citizen-science initiatives.[49]

For the further scientific use of the data, the data quality is very important precondition and it should be measured. We followed the data quality framework suggested by[50] which is assured of four data quality attributes: 1) intrinsic data quality which means the credibility or accuracy of the data, 2) contextual data quality which means how consistent, timely and complete the data is, 3) representational data quality which means how explicable and easy to use the data is and 4) accessibility which means how easy the data is to access and use.[51]

A difficult element of this problem is that researchers for example working within the equal organic or ecological disciplines do not always agree upon taxonomic keys. In fact, many researchers develop their personal key editions to guide their personal endeavours. Furthermore, keys are normally written for expert users, and are often complex, quite variable and difficult to translate into a form that will be suitable for use in a socio-computational system.[51]

3.3 Data quality integrity

Problems with the integrity of information may undermine the validity of a citizen science project. Although difficulties with information can manifest in any type of research, citizen science tasks may have extra issues with data due to the fact citizens probably have no longer had coaching in scientific information management or research integrity, and consequently might also not apprehend how to collect, record, or manage information properly. They may want to make systematic blunders that adversely have an impact on the validity of the data.[52]

There are quite a few strategies scientists can use to address this issue. Before data collection begins, scientists can furnish citizens with appropriate education on how to make observations, use scientific instruments, and document records and manage research records. They can also ask residents questions about how they are collecting, recording, and managing information to make possible that they are following correct guidelines, and they can ask residents for additional documentation to help their data, if needed. When making record is completed, scientists can evaluate the data once more to make sure that it meets scientific standards. They may additionally need to discard or correct data that they accept.[52]

3.4 Mechanism for data quality

Professional scientists are getting involved with members of the public in order to complete scientific research for citizen science because it is a form of research collaboration [53]. Data validation process is one of the biggest mechanisms for checking data quality. Most initiatives appoint a couple of mechanisms to make certain data quality and the responses validation techniques with some of them designed into information entry systems [59]:

31
• Instruments label annually at a significant Observatory
• Evaluating observer dependability with analysis program
• Participants known and qualified
• Participants send samples for identification
• Know credibility of contributors.
• Measurements
• Done by employees of the project
• Test in any respect coaching categories
• Scientific QA/QC in any respect levels; peer-review
• Results are reviewed by neutral committee and project groups
• Manual filtering of bizarre reports
• Piloting efforts to possess knowledgeable participants and ground-truth participant submissions
• On-line information entry subject to see and knowledgeable review and analysis

3.5 Data quality attributes

Data quality is a multi-dimensional construct consisting of a variety of attributes itself and it is necessary to any science project. Data quality, as a facet effect of crowdsourced scientific efforts, requires extra study.

Data quality attributes are: 1) intrinsic data quality which means the credibleness or accuracy of the data, 2) contextual data quality which means how consistent, timely and complete the data is, 3) representational data quality which means how explicable and easy to use the data is and 4) accessibility which means how easy the data is to access and use. Evaluating only the accuracy of data gathered by a social-computational machine is not adequate; additionally the methods that system format can affect the contextual, representational, and accessibility qualities of the data [54].

3.5.1 Improvements of Data quality[12]

Accuracy: Data accuracy should be handled with more care. Syntactic accuracy checks are already implemented in most projects. Linguistic accuracy is harder to check when trying to give freedom to the user, but some things should be checked. First, when a user is asked to give a location such as a country and city, there should be some automated check to see if such a country or city exists. Another method is to contribute the user with a predefined set of options the user can select from and if the user cannot find what they are looking for, ask them to fill a different field. When this field is filled, it is immediately flagged for possible moderators or community to check. This method can work with cities, countries, species, numbers, observation characteristics and many others if implemented correctly. Implementing this type of semantic accuracy check would lead to higher quality data and it can be used to prioritize flagged entries over normal entries for moderators or community to check.[12]
Completeness: Completeness is a double-edged sword in citizen science applications. Scientists want as complete data as possible, but most participants are unwilling or unable to provide such complete data. Participants are often amateurs and using their own time so they may not be able to give all the information a scientist is asking for. This is the reason why most projects only have some mandatory fields and many optional for the participant to fill and then have the option to edit the submission later. If all the fields are required, there are often less than five different fields for participants to fill. Observation information such as name of observed species or phenomena, and short description of the observation, location, time of observation, and participant information are the most common mandatory fields across all projects. These be minimum required fields for the observation input. The short description can be free word, selected from a list, or measurement information. Participants should also have a way to modify their submissions as many citizen science projects already enable, but this modification option is only available for logged in users. It could be enabled for anonymous users as well if they are asked to provide contact details such as an email address. The project could send an observation ID and a authentication key via email that could be used to edit the submitted data later.\[12\]

Consistency: Data consistency is tied to the accuracy and completeness characteristics. If the other two can be resolved, the data should be consistent. \[12\]

Credibility: There are different ways to check the credibility of data, but it should be a three-staged check. First stage would be an automated validity check. Data input should be checked syntactically and semantically and then input should be checked against different categories like some projects already do. This will reduce the amount of poor data and the amount of data possible experts must go through. Depending on the number of data submissions, the input can be checked against other submissions. Second stage should be to have the data checked by other community members. If there are others disagreeing with some data, it will show up during the last stage and helps the possible experts to take a closer look on specific data. The final step is to have experts validate data. The data can be ranked from 1-10 during the first and second stage according to the tests and community opinions. The lower the rank is, the more likely the data is inaccurate or false and the higher the rank is, the more likely the data is correct. This will help the experts to screen the highly ranked data more quickly and be careful with the lower ranked data.\[12\]

Correctness: Correctness is one of the most well-handled characteristics out of all. However, it should never be an exact time that is asked from the participant. Asking a participant to provide an approximate time should be done by giving the participant the option to choose a time window, not just one specific approximation. This would help participants and raise the quality of data. When saying an observation is made approximately at 4 p.m., it can be that the participant remembers it poorly and the observation is made around 5 p.m. If the participant is asked to provide a time window, they could say it is between 3-5 p.m. and that would be more accurate as the correct time falls within a time window and is not outside the time participant submits. \[12\]
Accessibility: In some projects, participant is asked to provide data without getting anything out of it. When a birding enthusiast reports a bird sighting, he would probably want to know where other birds can be found around his area of operation. If the data is not accessible in any way, the bird enthusiast might lose interest in the whole project. Therefore, there should always be some way to access the data and most preferably, a visual way and not just a table of raw data. A visual output is more meaningful and easier to understand than just the collection of data. This visual output can be a map, analysis charts or something else if applicable. Even in projects, where participants are asked to analyse pictures, there can be a visual output or chart of the current state and results. Accessibility can also provide easier error checking. If there are data points that are outside of an expected range, they are easier to notice when the data is put on top of a map or in a chart. [12]

Confidentiality (privacy): Privacy is lacking in half of the projects and it has the easiest improvement out of all the data quality characteristics. Instead of showing participants’ real name, show a username if a login is required for the system. Otherwise, just writing “human observation” or something else is enough. It does not mean that the data the scientists have should not include names, it means that the data that is shown to anyone outside do not need to know the actual name of the participant. Even in the case of reusing data, it is enough if the data is validated and there is a username pointing to the correct participant or the validation process can later be asked from the data provider. Having a person’s real name on a website with a handful of different locations creates privacy issues and concerns. This solution will lessen the privacy issues but not remove them completely as location information can still be harmful even as anonymous data. People do have different perceptions of privacy and all projects should have an option to change privacy settings such as showing a username instead of a real name. Additionally, if a user changes their settings after submitting data with the old settings, this change should be reflected in the submitted data. People often change privacy settings when they notice something is off and want to protect their privacy but if the change in settings does not change the situation in any way, people may lose their trust on the system itself. [12]

Efficiency: It is difficult to provide exact improvement suggestions without knowing the whole structure but in general, improving the handling of data or the server where the project is should improve efficiency. Handling of data is the only thing that affect the efficiency and the handling is related to many things such as underlying database, database management system and so on. Server itself does not improve efficiency per se, it mainly handles the data faster even if efficiency is poor. [12]

Precision: Precision is another double-edged sword. Most participants are not able to provide extremely precise data like Latin name for an animal and that can be handled during validation process. There are ways to improve the precision of data for example giving predefined options or fields to the participant or filtering wrong information out of the data. The latter can be done by asking different attributes from the participant and then giving possible options to the user to choose from (participant saw something small, black and flying, was it a crow or raven perhaps?). However, this can create time consuming tasks for the participants and it needs to be carefully designed and implemented so the participants are not demotivated. [12]
Traceability (provenance): Provenance is something that should be implemented with every citizen science project. Having other users edit data and not knowing who did what creates issues, especially if the data can be edited after being validated and this new change does not trigger a new validation. There can also be users edit correct data into wrong data and this decreases accuracy. Having no way to know who did it and what is the previous information creates problems not only to the community but to the scientists as well. Implementing full provenance can be difficult as every data related task needs to be recorded. Additionally, the provenance should be implemented in a transparent way so that the participants can also see provenance data on some level at least. This transparency not only increases the credibility of the data but can also reduce some misbehavior of participants. For example, if one participant edits correct data into wrong data on purpose, this change could be seen by everyone else. If the project employs a reputation model on their community, this misdeed would reduce reputation and so the transparent provenance would act as a prevention method.[12]

Understandability: As understandability is well managed in most projects, there are not many things to do to improve it. Explaining what each data attribute means and possibly contains, helps understanding the data. Another way is to remove the redundant data when showing it to others. Some data is unnecessary for third parties and removing that from the dataset will increase the readability and understandability of the data. Third way is to have a visual representation of the data such as a map or different analyses on the project site. Most people will not bother going through a list of different observations and a visual representation is easier and faster to understand. [12]

Availability: Data availability is an important factor in modern society, where people are being pushed towards openness through open publishing, open data etc. Putting the data available is not a difficult task but there are some things to consider when data is open to everyone. First, are all the attributes necessary for outsiders? Many datasets include the real name of the participant but that is irrelevant information and should be omitted from the dataset as it is also a privacy issue if combined with location information. Second, what format should the data have? Using good standards such as JSON, CSV or XML give an edge when opening data, as others can reuse it more easily. Having the data as an excel spreadsheet is helpful if the data only needs to be read but there should still be another format available. Additionally, there should be metadata found with the downloaded data so when people reuse the data, they can show where the data comes from and what credibility it has. Without metadata, it is difficult for outsiders to point people to the source and it is difficult to prove the validity of the data they are using. When a project ends, the metadata should still point to the people responsible of the project, so they can still be used as references. [12]

Data quality has been measured via ISO/IEC 25012 trendy with the aid of dividing information first-rate into separate characteristics. Different mapping methods and programming libraries have multiplied the range of maps used on websites and in citizen science projects. Maps are an effortless way to visualize data, if applicable, and they are convenient for others to understand. This has elevated the accessibility of data, which has been noted to be one of the issues of citizen science stakeholders [67]. Having other approaches to show information than in raw form, will
increase the interest on the project. There are nonetheless methods for initiatives to improve on accessibility as most projects solely had the maps as nicely as uncooked data. Few projects had taken the more step of which includes analyses as well, which enlarge their accessibility.
Chapter 4: Data Quality Attributes and DQMM (data quality maturity Model)

4.1 Data management maturity

According to [55] The development of the data management maturity model follows the rules bestowed on top to confirm the validity for a tutorial style science analysis.

According to [56] A new maturity model proposed in order to implement a systematic reuse approach, to improve the success of researches. The goal of the reuse capability maturity model is: 1) identifying maturity model based on maturity levels, Reuse Process Area and a set of Reuse Practices; 2) identifying capability model based on capability levels; and 3) assessment model defines a way to assess the state of a reuse practices by determining its score and relevance[56].

The sources for how to improve the data, the focus is the following[55]:

- Estimate of data quality
- Impact on business
- Perception of quality gaps
- Development

The capability level of an organization arranges a way to characterize its achievement by improving processes that interrelated to a given operation area and goes through 0 to 5: Incomplete level; Performed; Managed; Defined; Quantitatively Managed; Optimizing[56].

4.2 CMMI (Capability Maturity model Integration) components

The below figure illustrates the CMMI components. The CMMI define two types of levels: capability and maturity levels[56].

![Figure 7: CMMI Model component][56]
4.3 Improving quality through maturity model

Maturity reflects the level of organizational development which can be used to determine the capability of organizations to perform certain activities. Maturity models are an important tool to assess the quality and effectiveness of processes. The Software Engineering Institute at Carnegie Mellon University evaluating maturity and it became popular with the introduction of the CMM for software defined. Maturity models can be used to classify organizational strengths and weaknesses, and as tools for benchmarking information[57].

Maturity models have their origin within the field of total quality management. They drive strategically coupled continuous improvement so need an intensive understanding of an organisation’s current position and wherever it aims to be within the future. Commitment to the present amendment is crucial and needs the support and involvement of senior management and data quality characteristic, the maturity model within the amendment domain suggests that a lot of concepts developed to handle broader business amendment are applicable to the project management atmosphere. Some suggests that the project management framework could be a good selection for guiding the implementation of a amendment initiative in an exceedingly business[57].

Despite the acceptance of maturity models into the accepted bodies of project management, excellent knowledge, evidence as to the extent of use and impact of the models is very limited. It is recognized only three researches prior to their very own that explore the topics of use and impact. They make the point that in many cases where studies have taken place, the records are no longer available to the public[57].

Prescriptive models surpass descriptive ones when you consider that they are accurate no longer only for assessing the here-and-now (also known as the “as-is” situation) but additionally to point out the way to improve the stage maturity through enabling corporations to improve a roadmap for improvement. Organizations making use of these kinds of models to gain the capacity to measure and check their abilities at a given point in time and to have hints on improvement measures. Some of the information on data quality initiatives stand to contribute to enhance the data produced. However, it remains that none of these initiatives are in particular aimed at improving the quality (defined by its completeness, uniqueness, timeliness, validity, accuracy, and consistency) of the data generated for the monitoring. [58]

4.3.1 CMM in a project

Looking at the software engineering enterprise, the place the present maturity fashions originated, it is easy to see that there are many approaches to make the decision of any single software program problem. Software improvement efforts commonly consist of many greater variables, unknowns, and intangibles than we would consider “normal” for initiatives in many different industries. Because of this complexity, the anticipated result of a software program project can also be extra based on the “star” developer in a business enterprise than something
else. Unfortunately, megastar builders go away, and when they do or when initiatives get so massive and complicated that the developer’s influence on them is no longer dominant, the version in venture outcomes turns into brilliant and leads to inevitable frustration and disappointment. Obtaining predictable effects becomes an actual challenge. It is logical that these of us in the mission management area analyse from the efforts to enhance effectiveness in the software program industry. Applying task administration ideas in any organisation has many similarities to the complexities and intangibles of software program development. Obtaining consistent results in any undertaking surroundings involves appreciation and measuring as many variables as those that exist in the software development industry. In short, the need becomes so necessary that the agency must reply to growing commercial enterprise pressures. Many organizations will obtain substantial benefit by way of achieving the “repeatable process” level. In effect, an excellent mannequin for the size of task administration maturity creates a strategic design for moving undertaking administration forward.[59]

4.4 Maturity and data quality

Involves the things to do related to keeping and managing the organization’s data quality, in terms of open data current definitions and stage of interoperability of the organization’s data. Its CL (Capability Level) is resolved by the consecutive variables:

- Data arrangement: Use of non-proprietary arrangement that simplified interoperability.
- Complementary Data: The data are not blocked by copyright or other legal restrictions that could limit their restate.
- Primary Data: It should be primary data, avert secondary processed data.
- Data integrity: Data must be complete[60].

Data Restate

- Open data disclosure status. Its CL is resolved by means of variables:
- Open Data Developed Initiatives: Number of Open Data tasks carried out or in progress, and popularity of them.
- Number of Open Data Available: Volume (ratio) of OD facts units published by using the entity in relation to the global data furnished by means of several capabilities.
- Single Access Point: Accessibility of access indexes to data indexes.
- Data Access Measurement: Presence and the board of access pointers as well as downloading of data collections, examination of results and proposition of progress measures (selection criteria, promotion, etc.) [60].
The direct relationship between capability maturity and the expected quality of data illustrates in the below figure. Subsequently, advancing capability maturity of the entities responsible for revealing the advancement and it contributes to higher quality of information, and therefore, to better observing of the worldwide improvement agenda[58].

![Figure 8: Maturity and data quality][58]

### 4.5 DQ attributes for CS applications:

According to [12] we have number of Citizen Science project lists and then we have data quality Attributes from [54] and identifying all the attributes, by studying the DQ characteristics we can now map the DQ characteristics into the DQ attributes as follows:

<table>
<thead>
<tr>
<th>DQ Attributes</th>
<th>DQ Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contextual data quality</td>
<td>-Completeness</td>
</tr>
<tr>
<td></td>
<td>-Efficiency</td>
</tr>
<tr>
<td></td>
<td>-Traceability</td>
</tr>
<tr>
<td>Accessibility data quality</td>
<td>-Correctness</td>
</tr>
<tr>
<td></td>
<td>-Accessibility</td>
</tr>
<tr>
<td></td>
<td>-understandability</td>
</tr>
<tr>
<td>Representational data quality</td>
<td>Consistency</td>
</tr>
<tr>
<td></td>
<td>-Credibility</td>
</tr>
<tr>
<td></td>
<td>-Confidentiality</td>
</tr>
<tr>
<td>Intrinsic data quality</td>
<td>-Accuracy</td>
</tr>
<tr>
<td></td>
<td>-Precision</td>
</tr>
<tr>
<td></td>
<td>-Availability</td>
</tr>
</tbody>
</table>

*Table 9: Selection of DQ characteristics according to DQ attributes*
The identification of CS projects and data quality attributes with matching data characteristics is given below.

<table>
<thead>
<tr>
<th>Project lists</th>
<th>Contextual data quality</th>
<th>Accessibility data quality</th>
<th>Representational data quality</th>
<th>Intrinsic data quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Completeness</td>
<td>- Correctness</td>
<td>- Consistency</td>
<td>- Precision</td>
</tr>
<tr>
<td></td>
<td>- Efficiency</td>
<td>- Accessibility</td>
<td>- Credibility</td>
<td>- Availability</td>
</tr>
<tr>
<td></td>
<td>- Traceability</td>
<td>- understandability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ancient tree inventory</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Atlas of living Australia</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Bio collect (Australian citizen science apps)</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Bleach petrol</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Budbrust</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>CitSci</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CoCoRaHs</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>CosmoQuest</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>CrowdCrafng</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>EarthDrive</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>eBird</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EpiCollect5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Eterna</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>eTick</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>fold it</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Globe at Night</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>iNaturalist</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>iRecord</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>iSpot</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Laji</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Landslides @NASA</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Linking Landscapes for Massachusetts Wildlife</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>MLMP Data Portal</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MNHNLI Data portal</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Nature’s Calendar</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>NatureSpot</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Open Phylo</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Project Noah</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Riverwatch</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Shark Trust</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Social Enterprise Mapping for Change</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stardust @ Home</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>The Great Sunflower Project</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>The Lost Ladybug Project</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Treezilla</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Urban Forest Map (similar to treezilla)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Wisconsin Bat Program</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Zooniverse</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
</tr>
</tbody>
</table>

Table 10: Attributes checking on CS project
Chapter 5 Data quality maturity model for Citizen Science Applications

5.1 Assessment of DQ maturity model

The development of the maturity model is conducted according to design-oriented research. However, The entire development cycle consists of four phases[61]:

1 define scope
2 design model
3 evaluate design
4 reflect evolution

The profitable application of a maturity model went through four phases[61]:

1 select model
2 prepare deployment
3 apply model
4 take corrective actions

Figure 9: Development and application cycle of maturity assessment[61]

In Opposition to the maturity model development cycle and application cycle, that the application of a maturity model always is combined with comprehensive demand because the application cycle is always introducing with need by reason. If we look at the user’s viewpoint the consecutive decision specifications are determined. In this research, the two basic decisions
are identified. First, it should be implemented for CS applications and second, it should be completed to experiment the CS applications through DQ characteristics.

5.2 Maturity stages:

According to [39][40] we have the five levels of Software process Maturity but Maturity stages have also five sub process 1) assemble draft outputs, 2) justified outputs, 3) explained and elaborate outputs, 4) apply disclosure control, and 5) concluded outputs. The following proportions are analysed for each of the sub-processes: a) instructions, processes and methodologies for introducing draft outputs; b) tools, platforms, and systems for introducing draft outputs; and c) research, experience, and information sharing for introducing draft outputs[58].

5.2.1 Designing a CMM for CS applications

The concept of data quality characteristics process maturity has been announced in the literature [12], [3], [6], [33], [40], [39], [41]. Work in this area has been limited to a simple specification of the levels and aligning of some of the activities from CMM, without much serious theory-based development. In this section, we develop a CMM for CS by preparing our work in design to experiment Citizen science applications and integrating knowledge from literature on quality attributes, quality characteristics and maturity models of Data quality.

5.2.2 Maturity Level

In Chapter 4, we have discussed the data quality and maturity. So far from the discussion we have the level of maturity Ad-hoc (less mature), Supported, Managed, Proficient and Optimizing (more mature). Ad-hoc level is normal to deliver low-quality information because their processes are fuzzy. Supported level have tools, platforms, and systems in place but have allow image. Managed have processes, and methodologies in place; have a better understanding quality of information, trustworthy for decision making. Proficient have absorbed standards, best practices, and trends in their activities; have an entire and more right view of the reality and the information composed of high-quality and enables of influence of data exchange and information sharing. The Optimizing level of maturity is made when it can change fast and easily to accurate representation and the information [58].
5.3 Data Quality Maturity Model for CS application

According to [58] we have categorized the level of maturity 0-4 and following chapter 4 we have DQ attributes and characteristics, so we classified level 0 has 0, level 1 has between 0-1, level 2 has than less characteristics followed, level 3 has more than level 2 and level 4 has the most characteristics in the CS applications.

<table>
<thead>
<tr>
<th>Data quality Maturity Level</th>
<th>Contextual data quality</th>
<th>Accessibility data quality</th>
<th>Representational data quality</th>
<th>Intrinsic data quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-Completeness</td>
<td>-Correctness</td>
<td>-Consistency</td>
<td>-Accuracy</td>
</tr>
<tr>
<td>Level 4-Optimizing</td>
<td>-Efficiency</td>
<td>-Accessibility</td>
<td>-Credibility</td>
<td>-Precision</td>
</tr>
<tr>
<td></td>
<td>-Traceability</td>
<td>-understandability</td>
<td>-Confidentialty</td>
<td>-Availability</td>
</tr>
<tr>
<td>Level 3-Proficient</td>
<td>-Efficiency</td>
<td>-Accessibility</td>
<td>-Credibility</td>
<td>-Precision</td>
</tr>
<tr>
<td></td>
<td>-Traceability</td>
<td>-understandability</td>
<td>-Confidentialty</td>
<td>-Availability</td>
</tr>
<tr>
<td>Level2-Managed</td>
<td>Traceability</td>
<td>understandability</td>
<td>Confidentiality</td>
<td>Availability</td>
</tr>
<tr>
<td>Level 1-Supported</td>
<td>0</td>
<td>1</td>
<td>0~1</td>
<td>0~1</td>
</tr>
<tr>
<td>Level o-Ad-hoc</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 11: DQMM for CS application

5.3.1 Evaluation of Capability Maturity Model CMM

5.3.1.1 Process Maturity Model

The components of the CMM are procedure areas, precise goals, precise practices, general goals, general practices, traditional work products, sub practices, notes, discipline amplifications, widely wide-spread practice elaborations. CMM can be viewed as a staged or a continuous representation. The staged representation consists of five maturity levels to assist and information system improvement; in addition, its groups together process areas via maturity level, indicating which method areas to implement to achieve every maturity level. Below figure presents the CMM maturity stage architecture:[62]
5.3.1.2 Proposed Maturity Model

Before we discussed about Process Maturity level and then we have found the data quality and maturity capability. So, both levels have similarity in some points. A data Quality Maturity Model (DQMM) could be used to resolve the maturity of the quality of a definite application. Here, we have 39 applications where we check quality characteristics and evaluate them by categorized the quality attributes. More categorically:

- to certify a data quality maturity level for the most optimized characteristics, which matches all the characteristics in the selected attribute and followed Findings Analysis and settlement, assessment and settlement, action performance, Casual analysis and resolution.
- to benchmark existing characteristic based on their quality maturity level and followed Findings Analysis and settlement, Assessment and Settlement, action performance.
- to assess the quality of those characteristics which can managed the level for the CS applications and followed findings Analysis and settlement, Assessment and settlement.
- to assess the maturity of the subjective quality characteristics of the CS applications and followed Assessment and Settlement.
- to compare with other level those, have 0 characteristics for any CS applications and DQ attributes.
Figure 12: Evaluation CMM level
Chapter 6: Discussions

This thesis presented the data quality characteristics and their attributes contribution to citizen science projects and using that knowledge for the improvement of the data quality maturity model (DQMM) for citizen science applications. Quality characteristics mainly comes from one research paper which is the base of this thesis. In citizen science applications most data quality issues relate to the nature of the citizen science applications, i.e. amateurs giving data instead of professional researchers[12].

Data quality characteristics related scientific articles are searched, running keywords on scientific databases. There are 57% journal articles and 39% conference papers and 4% some website related articles screened during this research.

Throughout the research, all the characteristics and categorizations are manually conducted by one researcher and viewing by another researcher. In this case, the papers are reviewed manually based on paper title, abstract and available data on databases by the manual process of screening all the papers. In the thesis, we identified four research questions. In this section, we will discuss those four research questions based on results we found in the study.

RQ1: What are the Data quality problems in Citizen Science application?

The goal of this research question was to identify and describe the problem of citizen science applications based on some level. Maintaining excessive standards in data quality requires appropriate format of data input and management strategies[63]. We have already discussed about the problems and issues in detail in chapter 3. Here, the key points were to identify the researches that report the issues and problems of quality characteristics regarding citizen science applications.

Findings

There are some significant findings respecting research question 1. This is monitored after searching papers that discuss this issue. Some points are highlighted below,

- There are some weakness determined in citizen science applications such as: limited training, knowledge and expertise of contributors and their relative anonymity contributed to negative quality, deceptive or even malicious data being submitted[45].
- Lack of data made some issues for validation and assessment [44]. And those issues are related with the Citizen science application.
- Quality control filters can be implemented in the data entry process to ensure that certain required protocol information is precisely entered. For instance customized data entry types containing only those species commonly located in a specific geographical area at a particular time of year to reduce the probabilities of wrong data entries[63].

Below figure is the identification of problem regarding in CS applications.
RQ2: What are the data quality characteristics specific to citizen science applications?

The goal of this research question is to select quality characteristics and apply those characteristics in the CS application. We have already discussed the quality characteristics in chapter 2 and chapter 3 and those characteristics are applied in the CS application in chapter 4. Here, the key points were to identify the DQ characteristics which are noted in the CS applications.

**Findings**

There are some significant findings regarding research question 2. This is observed after literature review, and the characteristics selected are mostly from reference [12] and the quality attributes are from ([50], [51], [54]), main points are highlighted below,

- The data quality (DQ) characteristics chosen are, Accuracy, completeness, consistency, credibility, correctness, accessibility, confidentiality, efficiency, precision, traceability, understandability, availability[12].
- Then we have chosen data quality attributes such as: 1) intrinsic data quality, the believability or accuracy of the data, 2) contextual data quality, how relevant, timely, and complete the data is, 3) representational data quality, how interpretable and easy to use the data is, and 4) accessibility defines how easy the data is to access from[54], and after studying all the characteristics and attributes, we made some categorization of all the characteristics:
Figure 14: Categorization of DQ characteristics into DQ attributes

- Citizen science projects list is chosen randomly but mostly from [12]. And we have mapped the characteristics in the project list. What kind of project has what kind of characterises and how much it matches and fulfil the DQ attributes?

RQ3: How to develop a data quality maturity model for citizen science applications?

The objective of this research question is to develop a maturity model related to the data quality characteristics in citizen science applications. Experiment with the data quality maturity model to validate CS applications. We have done this part in Chapter 5.

Findings:

The most significant findings are identified when answering research question 2. After searching papers, when we have selected the DQ characteristics and putting those characteristics into DQ attributes, after that checking the quality Capability[58] on top of those categorized characteristics. This is observed after searching papers, and the maturity model is developed with the help of ([40], [39], [58], [62]), some points are highlighted below,

The DQMM for CS application development process followed following steps during this research:

- The selected DQ characteristics from [12] and putting into the attributes[54] and also taking in account the project list of CS application according to [12] we identified the DQ characteristics of CS applications.
- Using the capability and maturity from [58], we identified the five levels of capabilities: Ad-hoc, Supported, managed, proficient, optimizing.

We evaluate this in such a way so the level 0 has no value and level 4 has the highest quality characteristics categorization matches with CS applications.

RQ4: What are the benefits, outcomes and challenges of data quality Maturity model and validation in citizen science applications?

The objective of this research question is to Identify and describe the main benefits, outcomes and challenges of DQMM (Data quality Maturity Model) for citizen science applications. We will answer this specific question in this chapter as we haven not discussed this question yet.

Findings

Benefits of DQMM in CS applications

Most of the maturity model benefits we have considered in this research come from E-government services. we have selected the most important stages of e-government portals. Moreover, we have ended up that some maturity models are ignoring some essential best
practices and do not provide instructions and suggestions to move to an upper stage of maturity[64]. We have noticed that almost all the maturity models introduced stages such as in[[64], [65], [66]]:

- **Presence**: The idea of this stage is to present citizens information.
- **Interaction**: The idea of this stage is to allow citizens to introduced with the government.
- **Transaction**: The idea of this stage is to allow citizens to complete transactions services over the web.
- **Integration**: The idea of this stage is inclusion of the citizen in the e-government process and sharing information between various companies.

Moreover, we have noticed that the prepared maturity models causes from many limitations [65]:

- The maturity models are ignoring necessary e-government best practices such as: one give-up shop, customer centricity, interoperability, personalization, repayments and e-participation.
- The maturity models are not based totally on a unified and structured model such as: satisfactory practice model.
- The maturity rank of the e-government portals calculated by not providing a way from maturity models.
- For companies to move to an upper stage of maturity, the maturity models are not providing proposals and references [67].

**Outcomes of DQMM in CS applications**

The outcome of the measurements and validation of capability maturity models is achieved through the evaluation of DQMM. In other words, the model’s planning is not responsible for the validation process failure, it should be measured in a standard way and then design a methodology for applying the model that is appropriate to gain the evaluation of validation and measurements.

The model should fulfil the below four high-level requirements [68]:

- Enable each public company to point its current state of maturity and capability in an integral behaviour.
- Enable each public company to compare itself with other companies measured with the same model.
- Suggest suitable necessary roadmaps that public companies can follow to gain their levels of capability and maturity.
- Transfer information about the public companies e-Government readiness to allow the government and policymakers to measured whether they are ready to join the new national e-Government leadership, and to determine improvement programs in case they are not yet prepared.
Challenges of DQMM in CS applications

Citizen science projects contain the public with scientists in collaborative research. The dominant form of citizen science projects, determined in the environmental sciences, focuses on monitoring ecosystems and wildlife populations; volunteers structure a human sensor community for distributed records collection[69]. These particular problems are related with citizen science as follows, regarding to [69]

- The practice of citizen science is frequently virtual due to the fact of it regularly entails an aggregate of spatial, temporal, and physical discontinuities.
- Technology-supported citizen science benefits from economies of scale, however, does no longer usually matter upon large numbers of participants for profitable results.
- While citizen science is no longer probably to replace traditional scientific research, it is already discovering its place as a complement to professionally conducted scientific research in fields like communication biology.
- Citizen science projects are now in particular targeted on scientific research with rising frequency.
Chapter 7 Conclusion

The focus of this research is to understand about data quality characteristics and data quality attributes to device a data quality capability maturity model for citizen science applications. Data quality has been measured through ISO/IEC 25012 standard by dividing data quality into separate characteristics[12]. In this research, we have mostly used ([12], [3], [4], [2], [13], [31], [54], [58]).

According to [11], The assessment of DQ characteristics should be, Data analysis, which examines records schemas and performs interview to attain a complete perception of data and related architectural and administration rules; DQ requirements analysis, which surveys the opinion of data users and directors to identify best issues and set new best targets; identification of critical areas, which selects the most applicable databases and data flows to be assessed quantitatively; method modelling, which provides a model of the processes producing or updating data; size of quality, which selects the best dimensions affected via the quality issues recognized in the DQ requirements analysis step and defines corresponding metrics; dimension can be objective when it is based on quantitative metrics, or subjective, when it is based totally on qualitative critiques by data administrators and users. In this research we have categorized all the characteristics maintaining ISO standard and there after carefully studying and mapping as quality attributes to arrive at the DQMM for citizen science applications. Which is based on those attributes and capability maturity levels.

Besides that, this research also have presented different definition of citizen science applications from ([6], [7], [8], [9], [10]), DQ characteristics, DQ attributes, Data quality in databases [15], [16], Data quality dimensions [17], and data quality frameworks [21].

In this research we have presented maturity models available in the literature. The maturity models and the discussion is handled according to 4 issues[70]:

• The 1st issue consists of maturity models stage names. In this part, we can see that although the maturity models stage names are different from one maturity model to another; but as we studied and then find those names which is suitable for the proposed DQ attributes.

• The 2nd issue consists of maturity models level numbers and the categorization of it.

• The 3rd issue consists of the maturity models stage focus. In this part we can see that the most important stages of maturity can be compile into four noticeable stages as the following: representational, contextual, intrinsic, accessibility.

• The 4th issue consists of the maturity models stages appearances. In this part we terminated that some maturity models are ignoring some important appearances. This could be confirmed by the fact that many maturity models have been made without any input from the introductory models.

To conclude the issues, it is clear from the above that there is a need for a maturity model that combined all the best practices of the compared maturity models in the literature. The maturity
model may combine a large spectrum of CS applications for best practices. In this way, the model would be integrated and cover all the conditions of other application as well.

Limitations

Although this research has step-by-step all the categorization of DQ attributes for accessing to maturity stages, there are some imperfections which can be made for future researches. However, it is possible to do more accurate maturity model research in the future. Furthermore, there will be more opportunity to research the distinct stages as well as maintain process maturity model as well. Moreover, there is a chance to cover process maturity model and capability stages together and then mapping both into DQ characteristics to make a good maturity model that is applicable to any citizen science application.

Future work

In a future research, process maturity model and capability stages both could be in one chart and then there should be some comparison with the DQ attributes categorization for providing better results for the citizen science applications.
8 LIST OF TABLES

Table 1: Research question..................................................................................................................8
Table 2: Scientific databases and their representation of knowledge.............................................11
Table 3: Different definitions of citizen science...............................................................................13
Table 4: Data maturity model classifications according to their scope.........................................19
Table 5: Some factor in user acceptance..............................................................................................23
Table 6: Maturity model for determining the level of data quality management..............................26
Table 7: Mapping of maturity level in data quality expectations and dimensions.................................27
Table 8: A framework of general design principles for maturity models...........................................29
Table 9: Selection of DQ characteristics according to DQ attributes..................................................40
Table 10: Attributes checking on CS project.......................................................................................41
Table 11: DQMM for CS application....................................................................................................45
9 LIST OF FIGURES

Figure 1: The process carried out for collecting relevant scientific articles for the literature study.................................................................12
Figure 2: Data quality dimensions and categories.................................................................17
Figure 3: Data quality measurement framework...............................................................18
Figure 4: Data validation process.....................................................................................22
Figure 5: Data quality domains.........................................................................................24
Figure 6: Five levels of software process maturity.............................................................25
Figure 7: CMMI Model component..................................................................................37
Figure 8: Maturity and data quality...................................................................................40
Figure 9: Development and application cycle of maturity assessment...............................43
Figure 10: Maturity Level.................................................................................................44
Figure 11: Process Maturity Level.....................................................................................46
Figure 12: Evaluation CMM level......................................................................................47
Figure 13: Problem identification in CS applications........................................................49
Figure 14: Categorization of DQ characteristics into DQ attributes.................................49
10 Reference


[14] *An Introduction to Duplicate Detection.*


[19] “No Title.”


