LUT University

School of Engineering Science

Erasmus Mundus Master's Programme in Pervasive Computing & Communications for sustainable Development PERCCOM

Krishna Teja Vaddepalli

IMPROVING DATA QUALITY IN CITIZEN SCIENCE

Supervisors : MSc. Victoria Palacin (LUT University)

Professor Jari Porras (LUT University)

Examiners : Professor Eric Rondeau (University of Lorraine)

Professor Jari Porras (LUT University)

Associate Professor Karl Andersson (Luleå University of Technology)

This thesis is prepared as part of an European Erasmus Mundus programme PERCCOM - Pervasive Computing & COMmunications for sustainable development.

This thesis has been accepted by partner institutions of the consortium (cf. UDL-DAJ, n°1524, 2012 PERCCOM agreement).

Successful defense of this thesis is obligatory for graduation with the following national diplomas:

• Master in Complex Systems Engineering (University of Lorraine)

• Master of Science in Technology (LUT University)

• Degree of Master of Science (120 credits) –Major: Computer Science and Engineering, Specialisation: Pervasive Computing and Communications for Sustainable Development (Luleå University of Technology)

ABSTRACT

LUT University

School of Engineering Science

Master's Programme in PERCCOM

Krishna Teja Vaddepalli

Title of the work – Improving Data Quality in Citizen Science

Master's Thesis

87 pages, 11 figures, 5 tables, 2 appendix

Examiners : Professor Eric Rondeau (University of Lorraine)

Professor Jari Porras (LUT University)

Associate Professor Karl Andersson (Luleaå University of Technology)

Keywords: Data Quality, Citizen Science, Illu framework

Context: Citizen Science is a growing field in today's technology driven world, where participants collect information or observations of particular phenomena in multitudes of domains. As citizen science deals with a lot of people submitting data, it is prone to data quality issues due to multiple factors such as inaccurate, incomplete or invalid data which might lead to unintended results. **Goal**: To identify the attributes that define data quality and also provide a set of mechanisms that needs to be followed in order to achieve better data quality standard. **Methods**: This thesis studies different kinds of issues that occur in citizen science projects, the attributes of data quality that are affected by the issues, mechanisms which can help in solving them. **Result**: At the end, framework **Illu** was proposed which suggests a set of mechanisms which if followed can improve the data quality in citizen science projects. **Conclusion**: In conclusion, it can be said that citizen science projects are prone to unreliable data if the researchers and scientist conducting the studies do not take into account the data quality aspects and incorporate the solutions to tackle the issue.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This thesis is part of the Erasmus Mundus Master programme in Pervasive Computing and Communication for Sustainable Development (PERCCOM) of the European Union (Kor, A.L. *et al.*, 2019).

I would like to take this opportunity to thank the PERCCOM Selection committee, the host universities, such as University of Lorraine, Lappeenranta University of Technology, ITMO University, Luleå University of Technology and Leeds Buckett University, and especially Professor Eric Rondeau for the efforts that have been invested in PERCCOM.

I would like to express my deep gratitude to my supervisors MSc. Victoria Palacin and Professor Jari Porras for all kinds of support, meaningful feedbacks and encouragement throughout this master thesis.

A special thanks to everyone who have helped me and encouraged me in finishing me this thesis.

LIST OF SYMBOLS AND ABBREVIATIONS

API	Application Programming Interface
AQ	Air Quality
DQ	Data Quality
EJB	Enterprise Java Beans
EU	European Union
FAQ	Frequently Asked Questions
GDPR	General Data Protection Regulation
GPS	Global Positioning System
GQM	Goal Question Metric
HTTPS	HyperText Transfer Protocol Secure
LUT	Lappeenranta University of Technology
NASA	National Aeronautics and Space Administration
NPDES	National Pollutant Discharge Elimination System
PBMS	Predatory Bird Monitoring Scheme
PGAS	Probability Greedy Anonymization Scheme
PM	Particulate Matter
PS	Participatory Sensing
SDK	Software Development Kit
SWAMP	Surface Water Ambient Monitoring Program
TBN	The Birdhouse Network
UK	United Kingdom
USA	United States of America

Improving Data Quality in Citizen Science

1 Introduction

Any project that involves people to observe, monitor, report any phenomenon using a standard method can be called a citizen science project. The number of citizen science projects has increased largely from around 370 in 2015 to around 700 in 2018 (Nature, 2019). One main reason that can be credited for this increase is the rise of mobile devices across the world. By the end of 2018, the total number of mobile subscriptions stood around 7.9 billion (Ericsson.com, 2019). It has been seen and predicted that the number of smart devices per user has been increasing many fold over the years (Statista, 2019) making it easier for people to participate in scientific research where such devices can be used as tools to capture and record scientific observations (Maisonneuve, Stevens and Ochab, 2010).

Even though technological advancement such as internet, mobile phones, and applications have made it easier for people to get involved in scientific research, people involved in the project might not have any form of formal scientific training. Their view of problems differs compared to the view of scientists making their observational data quality non standard. There are many challenges that need to be addressed in order to use this data as evidence for scientific research that is valid, fruitful and acceptable (Bonter and Cooper, 2012). Observations made by people for a research project can have many anomalies. Absence of standardised models for citizen science projects, lack of hypothesis (Silvertown, 2009), lack of motivation, insufficient training (Hunter, Alabri and van Ingen, 2012), overwhelmed, inattentive, digital immigrants (Budde *et al.*, 2017) are some of the anomalies which can lead to incomplete or inaccurate data collection.

Data quality is one of the most serious issues which needs to be addressed for a meaningful research (Hochachka *et al*. 2012), (Kosmala *et al.* 2016), (Williams et al 2018). Though there is a lack of systematic methods in civic sensing to address those issues (Lukyanenko, Parsons and Wiersma, 2016), the challenges citizen science projects face can be addressed by adopting

standardized processes (Bonney *et al.* 2009) aimed to improve data quality. Validating data collected, addressing the issues of data quality would help in making citizen science a widely accepted scientific practice (Dickinson, Zuckerberg and Bonter, 2010), (Crowston and Prestopnik, 2013).

The objective of this work is to 1) identify and analyze different issues related to data quality that are prevalent in citizen science projects and 2) develop mechanisms and metrics which could help in improving the data quality in citizen science projects, so as to make citizen science research more authentic, valid and useful for scientific research on different domains..

To achieve these objectives, more than dozens of the citizen science projects were reviewed from their literature, archives of databases, and interviews of experts in the field. Interviews of researchers and developers involved in participatory sensing projects were conducted to identify the issues and challenges that they have faced. This helped in building a table of different issues found and used or probable solutions to solve the issue. In order to validate the solutions, we have developed a citizen sensing platform named **SENSEI .** Sensei is a participatory sensing movement involving different sectors of the society ranging from researchers to individuals who worked together to co-create civic technologies so as to monitor environmental issues of common interest. Later the same platform was provided to school students (**DO-IT**) for testing a few other mechanisms. This helped us in creating the framework **'ILLU'** (Fig 1), which translates to **home** in my native language, Telugu.

Sensei is a platform available for both web and android platforms. After identifying the issues and probable mechanisms, Sensei platform was developed to test the working and to justify our concerns in choosing those mechanisms. Thus, it can be stated that this platform is developed embedding a multitude of mechanisms which help in increasing the trust in data quality. It contains both the software level mechanisms and also includes the mechanisms related to participant selection, metric selection for evaluation, etc.. With some findings regarding the citizen science projects using hardware devices for collecting observations, this platform was

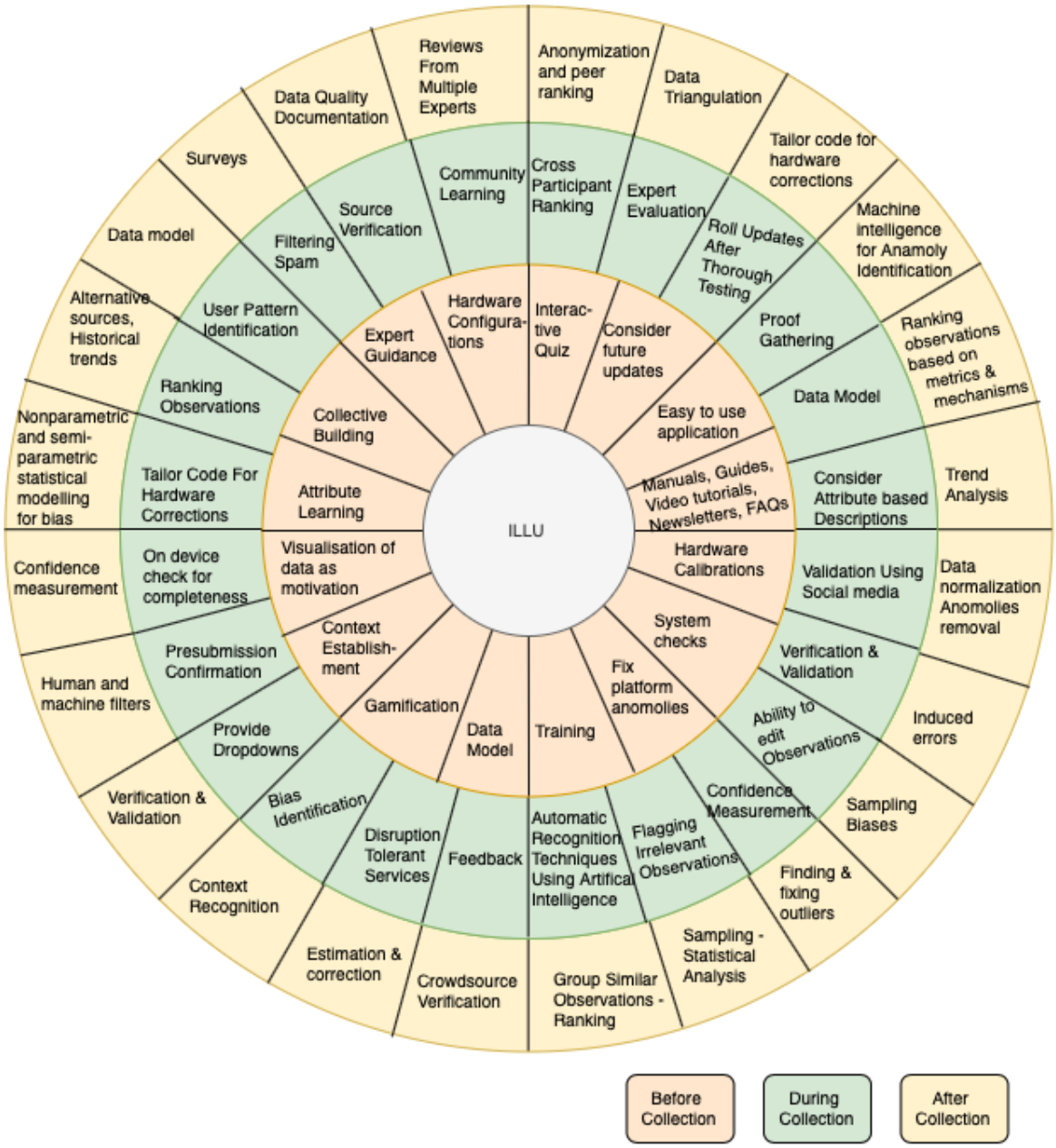
designed to check the anomalies that can occur due to issues in hardware by use of a flic button as a hardware add-on equipment to send signal to the mobile application in order to create an observation. Each step from designing the user interfaces to selection of participants was done considering a lot of parameters to make a fruitful study. A few new issues were identified during the course of study and identified mechanisms were embedded and released over planned over-the-air updates.

In order to make sure that the public can collect and submit accurate data requires researchers to incorporate three critical aspects : 1) clear protocols for data collection, 2) simple data forms, and 3) support documents to help participants understand the protocols and submit their information (Bonney *et al.*, 2009). A framework Illu for improving data quality in citizen science projects was designed, developed and iterated during this process. Illu comprises 61 number of metrics (see Figure 1) that were identified and tested during the environmental sensing initiative and validated through literature and expert interviews. This framework is important because it provides a set of steps that may be followed by researchers to overcome the challenges of data quality and establishes a set of guidelines or processes that makes the observations valid and trustworthy for researchers that depend on people for conducting studies and recording observations.

The results of this work show that the main data quality issues citizen science projects face are: Accessibility, Accuracy, Consistency, Completeness, Reliability, Relevancy, timeliness etc. There is no single solution for these issues because they are complex and require action on different levels and time spans. However, this thesis present a useful tool for practitioners and researchers who may want to run a citizen science project. We concluded that citizen science projects, which are an important way of conducting scientific studies of different nature and magnitude, is prone to unreliable data, if the researchers and scientist conducting the studies do not take into account the data quality aspects and incorporate the solutions to tackle the issue. Issues of data quality in citizen science is listed along with the possible solutions in form of different mechanisms that can be applied to mitigate these issues. Additionally, a set of metrics

that define data quality in citizen science is presented and the mechanisms developed that can be applied to attain these metrics of data quality. A framework that proposes a set of steps that should be followed by citizen scientist and researchers to make sure that the observations or phenomena being observed by citizens are accurately captured and analysed for the study or research results to be valid.

Figure 1 : Framework Illu



We recommend future work to focus on finding other relevant attributes of data quality that can be critical in a participatory project. Additionally, future work in this field could look into the possibility of creating automated tools that can evaluate a citizen science project based on the data quality metrics presented in this report. Also, in the future, new data quality attributes can be added to the described framework.

1.1 Data Quality and Society

One main difference between citizen science data and regular scientific studies is that the data collected in citizen science are observations submitted by individuals, which can be termed as the first stage of analysis performed by the people, and thus can have significant variations in the phenomenon being observed (Shirk *et al.*, 2019). Data quality is defined as a multidimensional measure of accuracy, completeness, consistency, and timeliness (Wand and Wang, 1996).

1.2 Problem Statement

People generally lack formal scientific training. Their view problems differs the view of scientists resulting in the reduction of quality of data collected. Data quality issues range from validating, detecting and eliminating a compromised piece of data. There is a lack of systematic methods in civic sensing to address those issues (Lukyanenko, Parsons and Wiersma, 2016).

1.3 Research Questions

The research objectives of this study is to 1) identify and analyze different issues related to data quality that are prevalent in citizen science projects and 2) develop mechanisms and metrics which could help in improving the data quality aspects in citizen science projects, so as to make citizen science research more authentic, valid and useful for scientific research.

In order to achieve these objectives, we designed the following set of questions:
RQ 1 : What are the data quality issues that citizen science projects face?
RQ 2 : How do we measure the data quality in citizen science?

RQ 3 : What are the different mechanisms or metrics available currently for improving the data quality?

2 Literature Review

Definition of citizen science was given by Irwin: *"developing concepts of scientific citizenship which foregrounds the necessity of opening up science and science policy processes to the public"* (Irwin, 1995).

Since then, many other key terms have been used to refer to the involvement of people in research through monitoring; Participatory Sensing (PS), also known as Urban, Citizen, or People-Centric Sensing, can be defined as a form of citizen engagement for capturing the issues in surrounding environment for contributing to finding the solution of specific issues which help in public health and well-being (Maisonneuve, Stevens and Ochab, 2010). People start on their own initiative, or initiated and encouraged by city authorities to collect media, and other data using different tools to monitor the environment and share the collected data to a common storage. The collected data is analysed by people or city authorities which helps in conclusions and action plans are drawn, and actions are taken (Holler *et al.*, 2014). Crowd sourced science, community science, crowd science, civic science, volunteer monitoring are all used as synonyms for citizen science (Doyle *et al.*, 2019)

2.1 Citizen Science

Citizen Science campaigns involve people in the monitoring of a phenomenon of common interest. A recruitment service which helps in selecting the participants considers campaign specifications and recommends participants for involvement in data collection. There might be many specifications involving multiple factors including device capabilities of participant, demographic diversity, etc. However, this work concentrates on a specific set of requirements for recruitment: participants' reputations as data collectors and availability in terms of geographic and temporal coverage (Estrin, 2010).

In recent years, there has been a tremendous increase in citizen science projects - like Galaxy Zoo, eBird, Air Quality monitoring - especially in areas where it requires the distribution of resources across the region (Hunter, Alabri and van Ingen, 2012). Moreover, there are certain domains of scientific studies that are now giving significant importance to citizen science for the purpose of research. Citizen scientists are involved in projects of varied nature covering a wide spectrum of research topics such as climate change, monitoring invasive species, biological conservation, ecological restoration, water quality monitoring, etc. (Silvertown, 2009).

Citizen science has a long history full of scientific and civic achievements contributing to many fields like astronomy, biology and city management. Some examples of citizen science projects include:

- An American ornithologist named Wells Cook, around 1880s, worked on gathering details regarding the arrival and departure of the birds in the spring and the fall (Askham *et al.*, 2013). The program continued till the 1970s. Over 6 million records were gathered during the entire period (Droege, 2007).
- The Birdhouse Network was used to study the knowledge of bird biology of the participants. To study the attitudes of participants towards science and environment, models like Elaboration likelihood Model were used (Brossard *et al.*, 2005).
- Foldit[1], is a computer game to help participants understand of protein folding (Mason and Garbarino, 2016).
- Galaxy Zoo[2] was aimed to study astronomical data by helping to discover new classes of galaxies. Over 250,000 volunteers took part in the experiment (Messenger *et al.*, 2012).

2.2 Data Quality

The concept of data quality differs based on the context it is being used. Because a data resource which may have an acceptable quality level for certain contexts may not be enough in another context (Even and Shankaranarayanan, 2007). Thus, data quality can be defined by the context of

[1] https://fold.it/portal/
[2] http://zoo1.galaxyzoo.org/Default.aspx

use and can be explained in terms of context as fitness for that particular use (Kahn, Strong and Wang, 2002)

However, to understand the concept of data quality and improve it, it is needed to understand what data means and what are the attributes that define the data quality. Though there are hundreds of attributes which directly or indirectly affect the quality, studies have proposed a few important attributes which can be major players when determining the quality (Pipino, Lee and Wang, 2002).

Data quality is a very critical requirement for any project. To make sure that the public can collect and submit accurate data requires researchers to provide three things: 1) clear protocols for data collection, 2) simple data forms, and 3) support documents to help participants understand the protocols and submit their information (Bonney *et al.*, 2009). However, with these safeguards are in place, it was observed that there are concepts which require special attention : issues of bias—a tendency to over report certain observation and to under report others — and a general reluctance of observers to enter data when they see only common phenomenon.

Table 1: Data quality dimensions (Pipino, Lee and Wang, 2002), (Wand and Wang, 1996), (Sabrina, Murshed and Iqbal, 2016)

Attribute	Definition
Accessibility	This attribute explains the level of data which is available and retrievable. The better the retrieval, the better is the accessibility to data. Accessibility becomes the key issue in citizen science as the data shared by the citizens should be accessible by the citizen scientists and also other participants should be able to access the data.
Appropriate amount of data	This attribute explains the amount of data required for analysing the situation or issue. Appropriate amount of data doesn't mean just quantity of data but it deals with the quality of the available amount of data. It helps in analysing the situation in correct methods with less errors.

Believability	The attribute which explains the credibility of data is believability. The importance of this believability comes into act during the process of analysis. The more the data is believable, the better are the results. All the results of the experiment depend on this attribute and hence is considered as very important attribute in the field of data quality.
Completeness	Data is set to be complete if all the required values are filled. Completeness of data helps the system to process information and represent in a meaningful way. It is not tied to any data-related concepts. Less null values means more completeness.
Concise Representation	This attribute explains the extent to which data is compactly represented.
Consistency	This attribute has many references in data - to the values of data, representation of data, physical representation of data. Data is expected to be the same for the same situation. Different values are observed only if there is more than one state in the system matching a real state of the real system.
Reliability	This attribute explains the probability of preventing the errors. The more the data is reliable, the more accurate are the results. Reliability explains the amount of compatibility between expectations and capability. It also explains the ability of the machine to provide the right information.
Interpretability	This attribute explains the extent of clarity in terms of language, symbols, units and definitions. This is one of the key attributes which explains how the stakeholders of the system need to use it. For the analysis to be conducted properly, the users should be able to interpret and enter the data correctly.
Objectivity	This attribute explains the level of unbiasedness in the data. This attributes plays a key role at the citizen level. The more unbiased the citizens are, the more accurate the data is. In a way, it defines the overall quality of the data based on the user perspective.
Relevancy	This attribute of data quality explains how much of the data can be used or error free. More relevant the data, more is the accuracy of analysis. Though it seems to be a single attribute, it is dependent on many attributes like the more unbiased the citizen, the more relevant data is produced.
Reputation	This attribute explains the extent of similarity in the collected data to the original source.

Security	This attribute explains the level to which the data is available for different stakeholders. In a way, this also explains the privacy factor of the user. Providing security to data and privacy to the user helps more users participate in the process. Security also deals with making the data available to different levels of users. All the users would not need all the data to be shown to them. End users would just need data they need while the officials or the scientists need data of many people. Security should take care of all these parameters while not compromising the user privacy.
Timeliness	This attribute explains if the data is out of date and the availability of output on time. There are 3 factors influencing the timeliness : the rate at which the information system is updated in comparison to the real world change, the rate of change in the real world system and the time when the data is being used.
Understandability	This attribute explains the level to which the data can be understood by different stakeholders of the system. It is this aspect that defines the ease of use of the system. It also defines the level to which the data can be comprehended by the analyst.
Value-added	This attribute explains the level to which data can add the value and in what are the advantages of the data collected. It is mainly affected by the relevancy. It affects the reliability of the system.
Traceability	This attribute explains the level to which the data is interpreted, documented, verified and accessible to the stakeholders.

The effects of poor data quality are not just limited to analysis part but also cost a lot in terms of economy. A simple wrong analysis made because of a wrong data would generate huge losses to the enterprise (Strong, Lee and Wang, 1997). Recent research conducted by Gartner found that poor data quality can cost organizations an average of $15 million per year (Moore, 2018).

Pattern of data quality attributes

In order to improve data quality in any system taken into consideration, we first need to determine the methods, techniques and metrics with which we can understand quality of data. This can be done only by employing some kind of measurement on data. In other words, we need

to measure the quality of data being used by a system in order to determine how valuable the information is and what needs to be done to improve the quality (Heinrich and Klier, 2015).

According to Strong, Lee and Wang, there are some patterns found in issues with data quality and they are grouped together with similar elements which provided four types of patterns of data quality issues (Strong, Lee and Wang, 1997). These issues are :

- Intrinsic Data Quality (Accuracy, Objectivity, Believability, Reputation)
- Accessibility Data Quality (Accessibility, Data Security)
- Contextual Data Quality (Value-added, Relevancy, Timeliness, Completeness, Appropriate amount of data)
- Representational Data Quality (Interpretability, Ease of understanding, Representational Consistency, Concise Representation)

Intrinsic Data Quality Patterns :

Intrinsic data quality patterns are mainly caused because of mismatches among sources of same data. Initially, data consumers believe that there might be some kind of conflicts with data from multiple sources which leads to believability issues which leads to the issues of accuracy leading way to poor reputation which reduces the added value of the data. Thus, these four data quality issues are grouped as a pattern (Strong, Lee and Wang, 1997).

Accessibility Data Quality Patterns

Accessibility data quality problems are based on underlying concerns about technical accessibility and data-representation issues which are interpreted by data consumers as accessibility problems, and data-volume issues which are interpreted as accessibility problems (Strong, Lee and Wang, 1997).

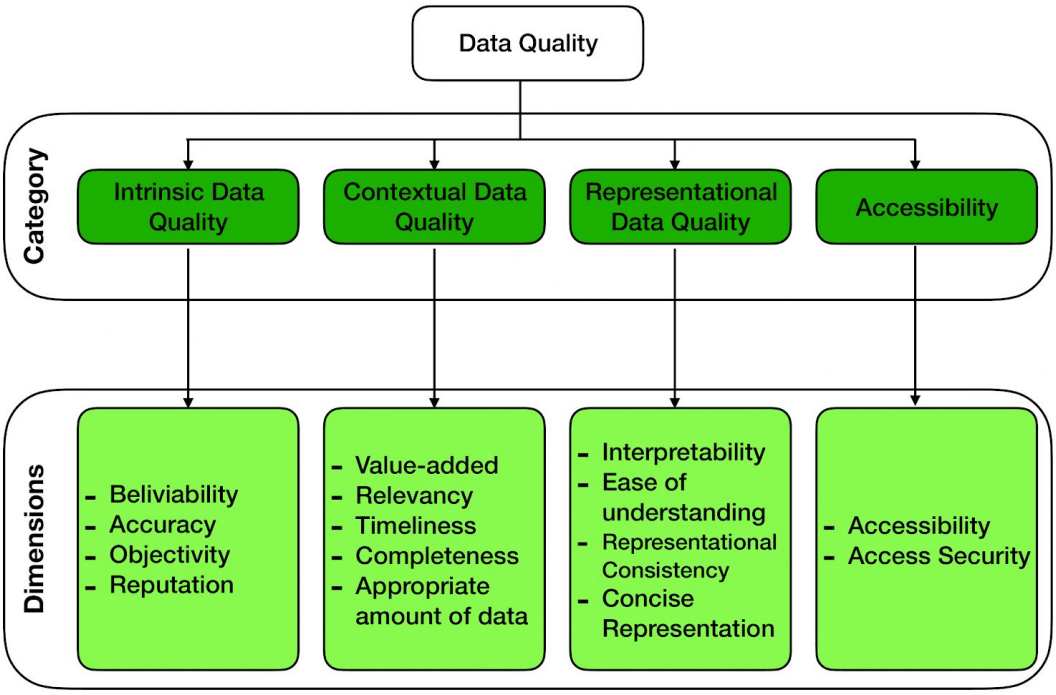


Figure 2 : Patterns of data quality adapted from (Strong, Lee and Wang, 1997)

Contextual Data Quality Pattern

Missing information, inadequately defined or measured data or data that is not properly aggregated would cause issues of data quality of contextual data quality pattern (Strong, Lee and Wang, 1997).

Representational Data Quality Pattern

This pattern helps the human to interpret, and understand the data. Consistency of representation and conciseness of data are aspects of this pattern. Research by Strong, Lee and Wang proposed that this pattern may affect the accessibility data quality pattern too (Strong, Lee and Wang, 1997).

2.3 Assessing Data Quality

Quality is measured generally in range of a number between 0 (poor) and 1 (perfect) (Pipino *et al.* 2002). Data quality problem can be defined as any difficulty that disturbs one or more quality dimensions that makes data completely or largely unfit for use. Data quality project is defined as the actions taken by the organisation to address a Data quality problem given some recognition of poor Data quality by the organization (Heinrich and Klier, 2015)

Some of the key areas where data quality problems arise are: user device handling, activity measurement, environment (Budde *et al.*, 2017).

Data quality can be analyzed using three themes and they are : Data quality metrics, Data quality and testing, Data quality in the software development process (Bobrowski, Marré and Yankelevich, 1998).

A Data quality project can be organised in 3 stages – problem identification, problem analysis and problem resolution. In the phase of problem identification, the organisation would focus on identifying the kinds of issues with the data they have. In the analysis phase, the organisation would plan on how the issue can be resolved, what are the tools and methods available to solve the issue and finally in the resolution phase, actual steps towards solving the issue would be implemented (Strong, Lee and Wang, 1997).

2.4 Measuring data quality and methodologies to measure

Data quality has gained a lot of interest due to the growth of warehouse systems, management support systems, customer relationship management and many other fields (Cappiello *et al.* 2003; Heinrich and Helfert 2003; Kaiser *et al.*, 2007). More recently, it keeps on gaining attention because of the big data era.

Measuring the quality of data will help us to understand its value. We will get to know the value our information and what needs to be done to improve data quality. Also, measuring the quality would help define the goals of a quality improvement strategy (Bobrowski, Marré and Yankelevich, 1998).

An approach to do this is to define the requirements that need to be assessed right from the start like functional and non-functional requirements. This way, as they would be part of the specification, we get to deal with them from the beginning. We get to know what kind of issues we face in the system and be prepared to solve them. A set of metrics may be considered to establish the requirements and check them at different stages of the development process.

Once we measure the quality of our data in alignment with the chosen dimensions, we can decide if our current data satisfies our expectations. Also, we get to know in which dimension it results in failing of which specific aspect, and we also get a clear measure of the badness (Bobrowski, Marré and Yankelevich, 1998).

Methodologies for Measuring Data Quality

As defined by (Wand and Wang, 1996), data quality is multidimensional with accuracy, completeness, consistency and timeliness as its dimensions. The above said dimensions would help in determining the quality of data. These dimensions if properly crafted can help in developing data quality audit guidelines and procedures which help improve the quality of data, help in the data collection process, and in comparing the outcomes conducted by different studies.

Simple techniques like syntax validation, format, values, and checking the validity against schemas can help in improving data quality (Wiggins and He, 2016). Use of history data to compare with current trends can also be used (Welvaert and Caley, 2016). But as the complexity increases with limited historical data, the assessment requires more complex mechanisms. This can be solved to some extent by exploiting social network analysis tools to provide the trust of

data. This was mainly used to solve the issues in Web 2.0 but can be applied to citizen science too (Lukyanenko, Parsons and Wiersma, 2016).

With data being received from multiple sources, accessing is not an issue but maintaining consistency and accuracy is important to make it usable. Unique representation of similar data across the platform helps increase the trust in data (Strong, Lee and Wang, 1997).

As we know the errors can be at any point of the life cycle of the data, errors at the time of creation by volunteers make data useless from perspective of scientists even if the errors are limited (Hunter, Alabri and van Ingen, 2012).

According to (Hunter, Alabri and van Ingen, 2012), the causes of the majority of the errors were due to:

- Lack of validation and consistency checking.
- Lack of automated metadata/data extraction.
- Lack of user authentication and automatic attribution of data to individuals.
- Absence of a data model.
- Lack of data quality assessment measures.
- Lack of feedback to volunteers on their data.
- Lack of graphing, trend analysis and visualization tools.

2.4.1 GQM Methodology

According to (Bobrowski, Marré and Yankelevich, 1998), GQM is a framework for the definition of metrics. GQM is based on the assumption that in order to measure in a useful way, an organization must: Specify goals, Characterize them by means of questions pointing their relevant attributes, Give measurements that may answer these questions (Lavazza, 2000).

This framework uses a top down approach which provides instructions to define metrics, without requiring any knowledge of the specific measures. First a set of dimensions are identified which

are important to define the data quality. Later, questions are formulated characterizing individual dimensions, without any precise definition, in simple language whenever possible. Sometimes, it is impossible to characterize dimensions and thus relevant characteristics are focussed. Finally, metrics to answer these questions, giving us a more precise valuation of the quality of our data are chosen (Caldiera, Rombach, 1994).

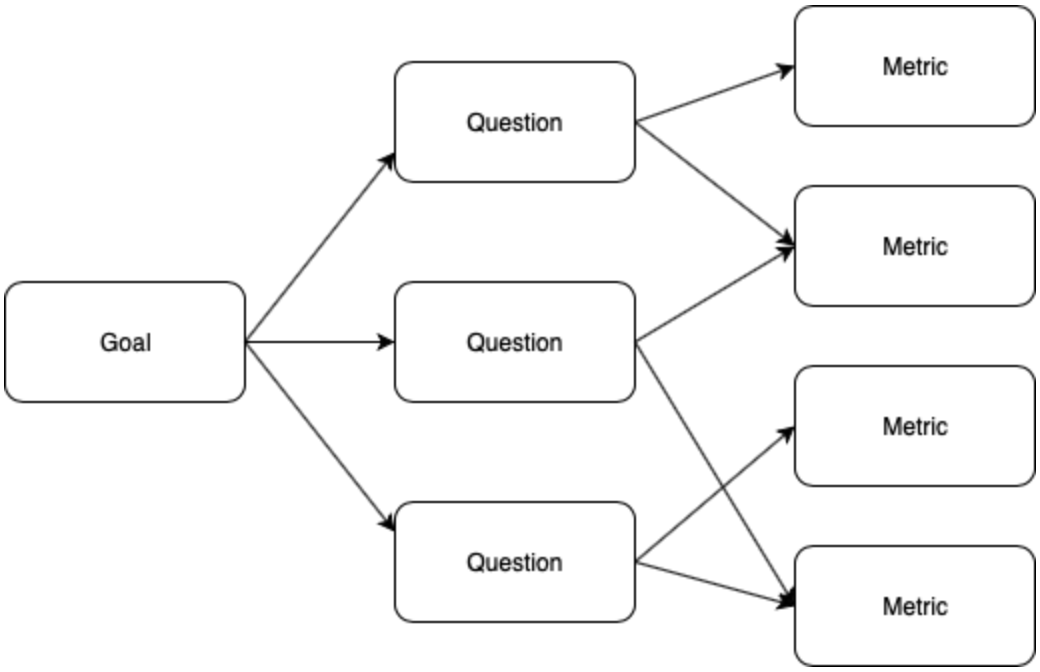


Figure 3 : GQM Methodology

For this thesis, we have used GQM methodology for defining the data quality of the experiment conducted. We have set our Goals at the beginning of the project, we have chosen our Quality metrics, created a few questions, and added mechanisms which would help solving the questions and thus used GQM to define the quality of the data.

2.4.2 Trust Based Methodology

Though some methods like edit checks, database integrity constraints, programmed control of database updates improve the quality of data, it is only limited to certain extent. More control on data quality has to be employed for attaining better results (Strong, Lee and Wang, 1997).

A system called Inferencing Reputation was proposed by (Golbeck and Hendler, 2004) to calculate the reputation based on the user profile similarity using Recommender system. The recommender system is a tree like system where each user and his group act as a branch. This was mainly used for calculating the trust of the user for his movie ratings. In case the user hasn't rated that movie, the system goes to previous step out in the trust network to find the rating given by his connections. This complete process is repeated until a predictive trust is calculated between two users.

Similar to the above method, (Alabri and Hunter, 2010) proposed a different approach which can be used to rank observations and users in citizen science projects. This method suggests that for every observation created, a initial rank is given. When similar observations are submitted, the rank of those observations increases. This way, we get to know how much we can trust that data. This mechanism was used in multiple projects. A few projects have modified this approach and have started to rank participants also to know which participants submit data with more quality (Ren *et al.*, 2015).

To identify the trust factor of the observations being created, we have implemented trust factor metric to rank users and observations in our platform.

2.5 Projects on citizen science

Currently there are many organizations, research groups, scientists and hobbyists who are increasingly employing citizens to observe and record scientific phenomena. These efforts are in a multitude of domains ranging from ecology, space, healthcare and environment to many others. Some of these projects have been running for decades and many others are recent projects that have been made possible by the advancement of information technology based devices such as internet enabled devices.

One of the earliest participatory projects involving the people dates back to the 1880's which has been in the domain of ecology where people reported bird sightings, wildlife and other environmental aspects. It continued until the 1970's and has gathered good amount of data.

Table 2: List of different citizen science projects studied

Project	**Aim**	**Quality Assurance Measures**
eBird: A citizen-based bird observation network in the biological sciences (Sullivan *et al.*, 2009)	To collect information about different species of birds and thus contribute to conservation	Checklist-based data entry, request confirmation and details, checklists to prevent mislabelling and misidentification, Automated data quality filters developed by regional bird experts, Local experts review unusual records, flagging, community learning
Citizen Science Noise Pollution Monitoring (Maisonneuve, Stevens and Ochab, 2010)	Investigate how a people-centric approach to noise monitoring can be used to inform government and public about the issue	Hardware calibrations, use of device sensors, normalization
Surface Water Ambient Monitoring Program (Ftp.sccwrp.org, 2019)	To study water quality, toxicity, physical habitat, and benthic macroinvertebrate	Compare data from multiple sources, 6 programs to understand the issues
Air Quality Citizen Science by NASA (Aqcitizenscience.rti.org, 2019)	To study the air quality with low-cost sensors	Compare with satellite images, other sources, use of device sensors
Envirocar (Bröring *et al.*, 2015)	To track driving parameters and calculate the carbon emission	Fixed parameters, expert evaluations
Common Bird Monitoring in Bulgaria (Svetoslav, Iordan and Nikolov, 2017)	To monitor bird breeding	Preset limits, higher value of the visits after sampling data, knowledge of species' range and of individual observer experience, records are validated by scheme organisers or

		local coordinators
Conker Tree Science (Conkertreescience.org.uk, 2019)	To collect the presence of pests on the leaves of plants and their density.	Validating data as subsets of the data, Modelled the error/mis-classification rates and statistically took this into account in the analyses, photo validation
Galaxy Zoo (Jordan Raddick *et al.*, 2013)	To measure the motivations of volunteers participating in online data analysis.	Share the same data to multiple participants and compare their answers, Expert evaluation
Aurorasaurus (MacDonald *et al.*, 2015)	To collect auroral observations made by public and to improve the modelling.	Gamification, use of device sensors for location
Virus Factory (Zooniverse.org, 2019)	To employ citizens to help annotate virus in an image based analysis platform	Automated and manual filters, comparisions with other users
Open Air Laboratories Network (Opalexplorenature.org, 2019)	To allow people learn about local environments	Online data entry system. Online quizzes/tests, observing participants taking the surveys to quantify error rates and identify common mistakes, comparing citizen science data with professionally collected data.
OPAL Bugs Count Survey (Opalexplorenature.org, 2019)	Investigating how the built environment affects the distribution and abundance of terrestrial invertebrates	On upload validation, Expert verification, Photographs proofs, Experimental observations of identification practices and commonly made errors within different sectors of the public.
OPAL Soil and Earthworm Survey	Monitor soil and earthworms in local area	Cleaning survey data and comparing with existing knowledge.

(Opalexplorenature.org, 2019)		
Predatory Bird Monitoring Scheme (PBMS) (The Predatory Bird Monitoring Scheme, 2019)	monitoring concentrations of contaminants in bird carcasses and eggs.	Examination and analysis of samples carried out by experts. Provenance information provided member of public but cross-checked by team members.
Recording Invasive Species Counts (iRecord, 2019)	Monitor invasive species	Data validated by expert, aided by species photograph when provided.
Weather Observations Website (Met Office WOW, 2019)	Cloud based computing platform for collecting and sharing citizen weather observations as an operational service	Meta-data used to generate star ratings. Quality control rules for identifying gross errors. Registered users can flag data that they suspect as erroneous. Special software is used to scan photos and text for inappropriate content

3 Methodology

This work followed the following stages : identification of problem, literature review, case studies, framework development, urban experiment observation, framework validation.

Identification of the problem : Citizens generally lack formal scientific training. They view problems differently than scientists resulting in the reduction of quality of data collected. Data quality issues range from validating, detecting and eliminating a compromised piece of data. There is a lack of systematic methods in civic sensing to address those issues. (Lukyanenko, Parsons and Wiersma, 2016)

Literature review and case studies : With this problem as a base to find the solution, we started to work on what is to be done to solve this issue. We have started to read the previous work reports by various scientists, had some interviews with people working on similar projects of citizen science or data quality, did some case studies of other similar citizen science projects. Literature on various subjects like citizen science experiments, data in citizen science, data quality, data quality for software engineering, attributes of data quality, data quality for citizen science was studied to understand the base of the issue.

Framework development : With some basic idea gained through the above process, we started to work on designing a framework which might solve the issues observed in previous cases. The basic framework had solutions to a few problems but most of them were still a puzzle. We have considered this to be our framework but we were not satisfied with the results. So, we decided to conduct a citizen science experiment to experience the issues firsthand.

Urban experiments observation : We have conducted some workshops inviting people to participate, found their interests, and started developing a citizen science platform combining their interests and our learnings from issues found through literature review, case studies and interviews. Once the platform was ready, we have released it to citizens to use it. It was live from

July 2018 - November 2018. During this period, participants were asked to monitor various elements like 'Nice places in nature, Invasive species, Lost items'. A second phase of this experiment was held in February 2019 with students monitoring sustainable and unsustainable elements.

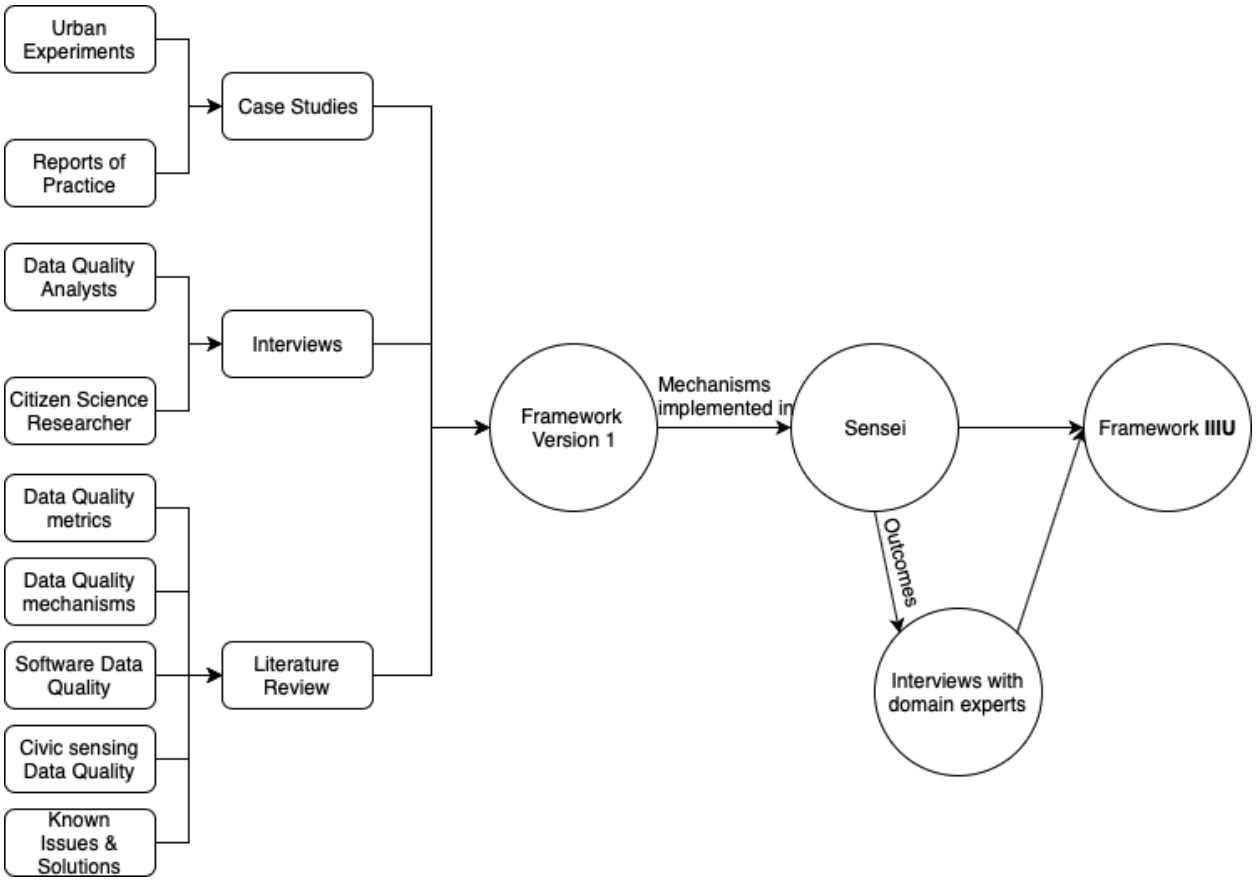
Framework validation : The data collected from the experiments helped us iterate and validate our framework. Also, the interviews held with experts helped us to iterate and validate the framework.

The entire process was inductive research where we expanded our learnings about the issues and solutions. We have released a few updates to further make the platform more efficient. After the completion of the complete testing, we have started analysis of our results and found some interesting patterns which are discussed in the results section.

Table 3 : Methodology for Research Questions

Research Question	Goal	Method	Instrument
RQ 1	What are the data quality issues that citizen science projects face?	Case Study, Literature Review	Observation, Notes from literature
RQ 2	How do we measure the data quality in citizen science?	Literature Review, Case Study, Urban Experiment	Notes from literature, Experiment results
RQ 3	What are the different mechanisms or metrics available currently for improving the data quality?	Interviews, Literature Review	Interview notes, Notes from literature

Figure 4. Methodology



3.1 Case Studies

Citizen science experiments conducted by other people were studied to understand the issues they faced and the mechanisms they employed to improve data quality. With these learnings, we have conducted our own citizen science project called Sensei in city of Lappeenranta. The results section explains the mechanisms used, and findings which led to the development of framework 'Illu'.

3.2 Literature Review

Literature review on data quality, attributes of data quality, data quality in software engineering, citizen science, civic sensing, participatory sensing, data quality in citizen science, methodologies to measure data quality, mechanisms for improving data quality have been analyzed. Most of the data quality papers referred to the late 1990's and early 2000's papers which contained a lot of valuable information. The literature review helped us bridge some gaps and understand many concepts.

3.3 Expert Interviews

Interviews were conducted with people from different fields like the citizen science project teams, big data developers, database admins from different organizations.

1. Interview design : The design of interview was not fixed and varied based on the background of the person being interviewed. The main goal of the interviews was to understand the major data quality issues that they have faced in their projects and how did they solve them. A few interviews were conducted before designing the 'Sensei' application which helped us get some insights of a few mechanisms that can be included. A few interviews took place after the Sensei data was analysed in order to validate our findings by comparing if the mechanisms suggested would be applicable.
2. Interview demographics : A total of 3 big data analysts, 2 citizen science project teams, and 2 database admins were interviewed. Average time spent on the interview was around 30 minutes. Some of the interviews were face to face while most of them happened over skype or calls.

3.4 Framework Design and Iterations

There were 2 main iterations in designing the framework. The first iteration included our learnings from literature review, case studies of other projects and interviews from experts. This knowledge helped us understand a few mechanisms but we had issues computing the capabilities of many mechanisms. So, we have developed a platform including the mechanisms we have

learnt. During this process, we have learnt a few mechanisms that were not presented before which could help in improving the data quality. We have interviewed a few experts to validate the mechanisms we provided to solve those issues. The second iteration included all the mechanisms of first iteration and the new mechanisms we have learnt during the process.

4 Results

The results are computed using the GQM methodology where we first set up a series of goals, prepared questions for each of the goals and tagged metrics which would help us find a solution for each question. Also, we have included our learnings from case studies, literature review and interviews to propose a usable framework.

First, let us go through the platform - case studies, description of the platform, metrics used in building the platform, challenges faced, and then the evaluation criteria we chose.

4.1 Case Studies

In order to understand different mechanisms identified and their limitations, we have conducted two environmental monitoring experiments : Sensei and Do-it involving the public of Lappeenranta city.

4.1.1 SENSEI : Environmental Monitoring Movement in Lappeenranta

SENSEI is a participatory sensing movement involving different sectors of the society like researchers and experts, local organisations, city officials, individuals, and families. They worked together to co-create civic technologies so as to monitor environmental issues of common interest. Three main environmental issues - invasive plant species, abandoned items in nature and nice places were chosen for monitoring. Over 240 local participants have taken part during different stages of this year long process. It included ten community events and workshops. It resulted in generating over hundred solid ideas on issues of common interest, around thirty prototypes were designed and developed, alongside producing hundreds of environmental observations (Palacin *et al.*, 2019).

4.1.2 DOIT

DOIT is a european initiative for developing entrepreneurial skills for young social innovators in an open digital world. It encouraged students from local schools to participate in different

activities. As part of this program, sensei platform was provided to nearly 40 students to monitor aspects like sustainable, and non-sustainable aspects in nature (Doit-europe, 2019).

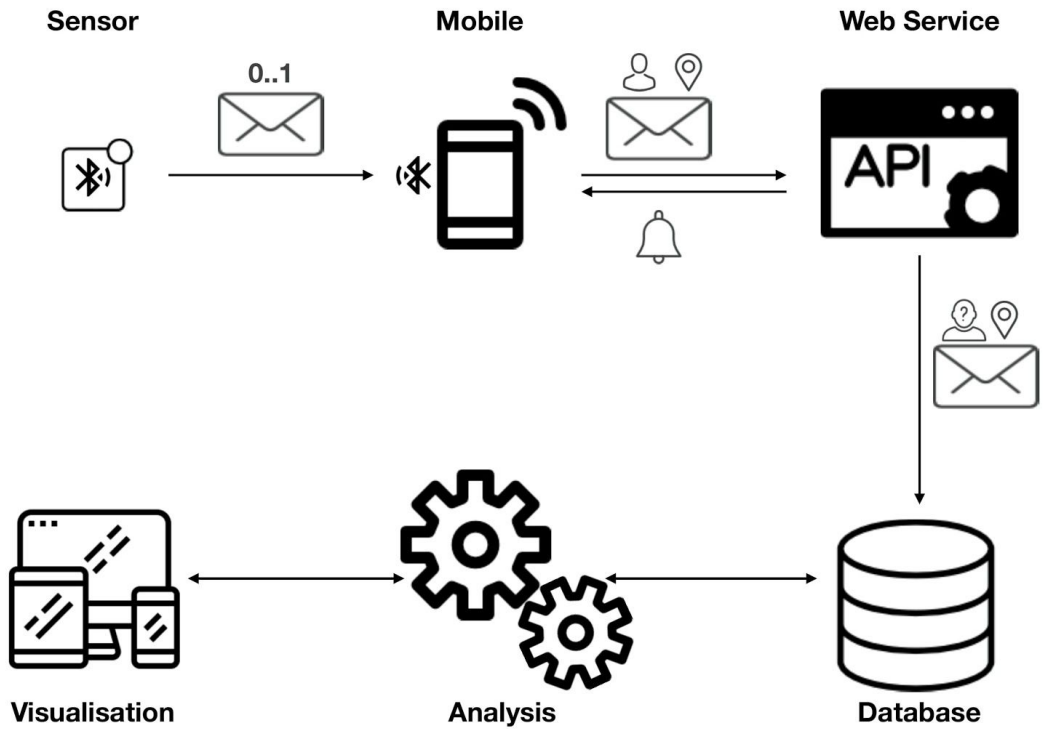
4.1.3 Description of the platform

After many deliberation exercises that were conducted in form of workshops and were attended by Lappeenranta residents, a list of issues were identified that the general public could report using the Sensei platform. Apart from the issues that could be reported via the platform, there are many methods that can be used to interact with the application were also discussed. After the analysis of deliberation outcome a platform called sensei was developed where user can interact with the platform using three different ways - mobile application, website and flic buttons. The main application of flic button is that it connects to the mobile device via bluetooth connection which in turn connects to the platform via mobile application installed on the mobile device. Different metrics were included in the platforms to ensure the usability, security, privacy and hence enhance the quality of data. The main purpose of this platform is to enable users to share their observations for performing analysis.

A flic button, is provided to participants, which interacts with the mobile device via bluetooth which connects to the platform. An android application is installed on the mobile device to capture the clicks on flic buttons and also to take pictures and share it to the platform. The website allows users to create and modify their observations while allowing them to view all observations even without logging in, making the observations accessible to the general public.

A NodeJS based server, which serves the API calls is the main engine of the application and is used by both the mobile application and website.

Figure 5. High level architecture of Sensei project.



4.2 Metrics used in platform

Mobile - The main tools for user interactions are mobile device and the bluetooth enabled clicker. As a result, it was decided to include some metrics on mobile to ensure that it is easy to use and easily adaptable for even a layman user, which helps in maintaining the user engagement with the Sensei applications. The main advantage of using mobile devices is that it helps in enhancing the usability without compromising on the functionality. Some of the major metrics we considered while developing the mobile application were:

Collection of data : Data collection is the crucial part of any crowdsourcing application. The main issue to be considered is that the data is collected from multiple sources and are of different types. Data has to be collected, segregated, sorted and stored for analysis. Collection is one of the

main processes where data quality can be assured and better results can be obtained by properly managing the collection of data

- Users are given liberty to share their observations using mobile application or website.
- In order to create a new observation, it is mandated to authenticate which helped us in removing the fake data to large extent. Studies have shown that authentication helps in reducing the unwanted data. (Alabri and Hunter, 2010)
- Users can also share their observations by clicking the flic buttons. We have utilised the pattern capturing on flic to make it easier for the users to share their observations. For example, single click on the flic records a type of observation while double click triggers a different observation and hold button for third type.
- Also, an option to upload pictures is given to users making the collection more accurate and analysis more effective
- Instead of asking users to give parameters, mainly the location, sensors of mobile have been used to get the location of user. A user confirmation of such data is requested in order to ensure better results.

Quality of data : The purpose of data collection gets defeated if data which is of low quality and cannot be used for analysis is being collected. To ensure the quality of data, we have implemented different metrics at different layers.

- Throughout the platform, anonymous submission of data has been blocked to ensure that data is from a trusted source.
- For clickers the main issue was the loss of connectivity with mobile and the battery life of both mobile and the clickers. To overcome these issues, we have chosen flic buttons for clickers as they use BLE (Bluetooth Low Energy) which consumes far less energy compared to traditional bluetooth and is continuously connected to the device until unpaired or the bluetooth on mobile device is turned off.
- Each type of observation is given a specific predefined click pattern and the clickers helps users create their observations with less hassle.

- In mobile, we have implemented a few metrics which include blocking of the observations if they are submitted continuously with short time sequences from the same place.
- User is expected to give a brief description about the observation before submitting it which helps us understand more clearly about the observation ensuring clarity of data
- For better results, we have asked the users to classify their observations into categories before they submit their observations.
- For improving the accuracy of the location, GPS data from the mobile is used. Also, user is asked to confirm the location before actually saving the data.
- Users are encouraged to provide their observation with an image whenever possible to make the data more believable for analysis. The images tagged with location and the user information help in ensuring accuracy of the data
- To err is human and hence we expected users to record some wrong data. We have an option on the website to edit their observations or to add some images or description to the observation.
- All the collected data is properly tagged, sorted and saved in database and can be retrieved ensuring availability.
- On the server, we have implemented the ranking algorithm which would rank both users and observations based on number of users who have submitted the data which helps in understanding the overall importance of the observation and analysing which user has submitted better results.
- Proper sorting algorithms on the server ensured that data is going to be stored as expected without any problems.
- Encryption of data before sending it over the network was considered but was ignored. This would promise integrity of data.

Usability : Usability is the main concern with many of the applications of today. Generally, it is observed that retention rates are high (around 62%) after first use of application and most of the applications are used for less than 11 times (TechCrunch, 2019). To overcome such issues, we

have heavily concentrated on developing a platform that can be interesting to interact with. We have contacted a group of 50 people multiple times and have collected their interests and the ways they interact with different devices.

- Based on their feedback, we have made the application available for both mobile and web platforms. Also, as many participants are interested to interact with clickers rather than mobile devices, we have come up with a solution that uses flic buttons.
- To increase the engagement of users, the user experience has been heavily researched upon. Hundreds of designs were considered for both mobile and web platforms. Many prototypes were given to a few volunteers to test. Finally, we come up with a design that is easy to use and interact.
- Load times are crucial when it comes to dealing with large group of users. The higher the load time, the more the users go away (Nah, 2004). We have come up with different techniques (lazy load of images, loading map grids instead of the whole map etc) to keep the load times shorter and make the page load faster even with slow internet connection.
- Showing users what others have shared would enable users to know what is happening around and interact more with the platform. For enabling this feature, we have allowed users to upload pictures of what they see and they are tagged with the location which is shown on the map.
- All the above mentioned things keeps the user start using the application, but what keeps them going is the uniqueness of it. We have also introduced challenges in the platform where a user can challenge some of his peers making them more involved and interested to use the platform.

Security of user data and privacy : With the increasing issues of privacy and security around the world, we have given this as our major point of concern.

- Login is made mandatory for users to interact with the platform. This ensures no anonymous data is logged.
- Token based login is used so as to make the API calls more secure. No data about the user (like username, nickname etc) is used in API calls. Only the token details are

attached to the data making it more secure. So even if data is intercepted by a cracker, he wouldn't be able to get the information of the user who generated the data.

- To ensure further protection, new tokens are requested by mobile devices in regular intervals of time making the API calls more secure.

4.3 Evaluation

A total of 413 observations were submitted during the lifecycle of the project, out of which 96 were submitted using the clickers and 318 were submitted using the application.

Figure 6 and 7 shows the number of observations recorded by each user using different methods, such as clickers and mobile.

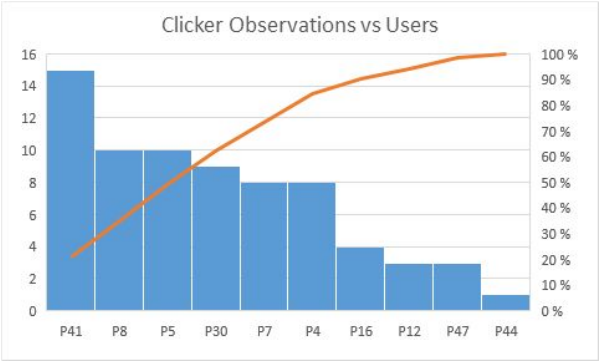


Figure 6. Observation submitted by users using clickers

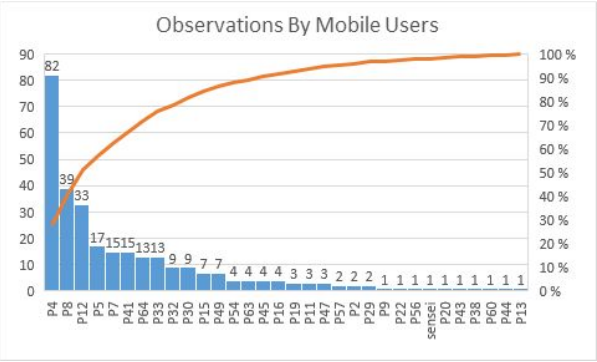


Figure 7. Observations submitted by user using mobile

The observations recorded were of two types, public observations and private observation, where public observations were the ones that were also shown on the platform for other users to view, while the private observations were not shown to other users. The number of public observations were 404 and number of private observations were 9.

Moreover, the users of Sensei platform had different categories under which they could record their observations, such as invasive species, lost items, nice places etc. Figure 8 and Figure 9 below shows the share of different observation types using mobile and clickers.

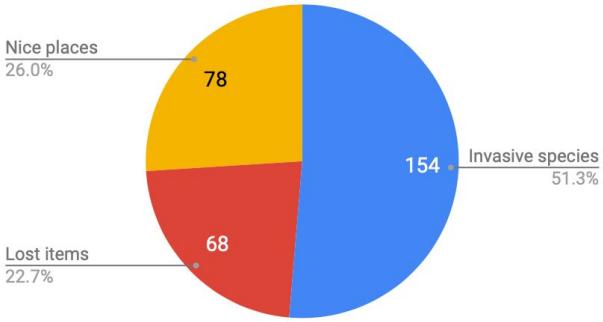


Figure 8. Share of observation type using clickers

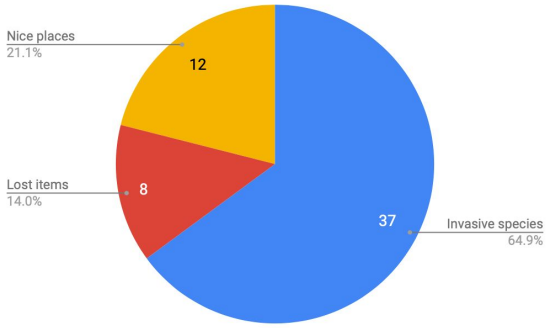


Figure 9. Share of observations type using mobile

Once the experiment was done and we had collected the data, the first thing we had to do was to clean our data. Later, we have used a few set of metrics which were decided before the experiment with the help of GQM methodology to help us evaluate the data. We have chosen a set of dimensions of data quality and had our evaluation criteria drawn up. We then came up with the most suitable metric that could help us identify the validity of the dimension in the data. Once we had our metrics ready, we needed a set of calculations which can help us implement the metrics we have considered. For that, we have used a few calculations and some checks where necessary. We have coded each of the metric with a value and have used them in calculations.

Figure 10. List of different data quality metrics selected for evaluating by Victoria and Krishna

Dimension	Evaluation Criteria	Code	Metric
Accuracy	Is the observation accurate? (reasonable location, input values and rightfully classified)	I1	Observation Image and description match each other
Believability	Is the observation trustworthy and reasonable? (observation info)	I2	Location reasonable, image makes sense, description matches the image or observation
Reputation	Rank of the observation	I3	Rank
Accessibility	Is the observation accessible from the SENSEI front end?	A1	Number of observations accessible from mobile and web; Is it stored consistently on the database
Data Security	Is the observation public or private? What are the types of observations?	A2	Number of public and private observations, Number of private observations, Editing ability (own observations), view ability (other observations)
Value-added		C1	How many observations were classified under subtypes, How many observations had a description,How many observations had an image; frequency of submission
Relevancy	Is the observation usable error free? ; Classified properly; duplicate values	C2	Number of observations well classified; Number of duplicate observations; Proper segregation (types, subtypes, etc)
Completeness	Is the observation complete?	C3	Number of observations with description only, Number of observations with location only, Number of observations with description, image and location
Interpretability	Are the observation details clear? (description, image, classification)	R1	Observation Image and description match each other
Representational Consistency	Is the observation value changing in different situations?	R2	Number of consistent observations

Figure 11. List of different parameters for evaluating data quality attributes Elaborated by Victoria and Krishna

Calculations	Code	Reference
Rank	A	I3
Number of observations accessible from web	B	A1, R2
Number of observations accessible from mobile	C	A1, R2
Database consistency	D	A1
Number of public observations	E	A2
Number of private observations	F	A2
Number of observations with type	G	C1, C2
Number of observations with subtype	H	C1, C2
Frequency of submission per user	I	C1
Number of duplicate observations	J	C2
Number of observations with description only	K	C3
Number of observations with location only	L	C3
Number of observations with description, image and location	M	C3
Number of consistent observations	N	R2, C2
Number of observations that were updated after submission	O	A1, A2
Observation Image and description match each other	P	I1, C3, R2
Reasonable Location	Q	I2, C3
Image makes sense	R	I2
Editing ability (own observations)	S	A2
View ability (other obs)	T	A2
Observations with the right type	U	C2, R2
Observations with the right subtype	V	C2, R2

We have used the above method to rank the individual observations which helped us to understand the quality of the data. This process was inspired by GQM methodology which states us to identify the goal, define the questions and use metrics to answer the questions.

4.4 Analysis

4.4.1 RQ 1: What are the data quality issues that citizen science projects face?

When developing any project, the developing team should be ready to face a few issues. Some of the issues that a developer must consider and solve when working on projects involving citizens

for scientific research are - hardware calibration, participant's digital literacy, connectivity issues, draining of battery, crash of application, laws and regulations in that region etc.

Based on the research we have grouped the issues based on the type of issue. The groups are : Hardware issues, Participant issues, Issues due to biases, Validation issues, Loss of data, Issues that affect quality and miscellaneous issues.

4.4.1.1 Hardware Issues

Some citizen science projects use certain hardware devices to collect data, or to aid people to collect required scientific data. These hardwares may have issues of their own, such as faulty construction or component, overuse, damaged, etc. Some of the most relevant issues that should be considered in citizen science projects related to hardware are as follows:

1. ***Hardware Calibration:*** In projects using hardware sensors for observations, the chances of hardware discrepancies increases many folds. For example, the hardware could lose connectivity or may give incorrect data (Maisonneuve, Stevens and Ochab, 2010).
2. ***Location Issues:*** Most of the mobile GPS sensors are not fully accurate with location. In many cases, there is at least 10-15m radius issues. If not tackled properly, there might be issues like believability and accuracy.
3. ***Hardware Connection Failures :*** In projects using connected hardware for observations, the network hardware should be robust. Failure of which will result in uncaptured observations (Budde *et al.*, 2017).
4. ***Battery Issues :*** When participants use devices, this is one of the most common issues faced. If not dealt properly there might be lesser quantity of data (Guo *et al.*, 2015).
5. ***Data connectivity issues :*** Most of the countryside is not properly connected. When the participant wants to submit an observation, there might be a situation where there is no data connectivity. This leads to loss of valuable observations. Not considering this issue would lead to loss of observations (Guo *et al.*, 2015).

4.4.1.2 Participant Issues

Many citizen science projects require that the participants are well familiar with the field of study often they may still need the participants to be able to use some device or may require them to undergo certain training or workshop (Ren *et al.*, 2015). Since, success of any citizen science project depends on the quality of participants, therefore it is imperative that the issues related to participants are to be addressed. These participants may have issues of their own, such as literacy, data entry, data verification etc. Some of the most relevant issues that should be considered in citizen science projects related to participants are as follows:

1. ***Participant Selection :*** Data on these platforms is mainly dependent on participants. Without proper selection of participants, the results would not be accurate, making the whole process useless (Ren *et al.*, 2015).
2. ***Digital literacy:*** Users should be trained with the platform and functionalities being offered. Lack of this will lead to participants not being able to use the platform to its full potential.
3. ***Errors during entry or submission :*** Platform requiring detailed classification of observation being submitted could have issues of errors during submission. Wrong classification, wrong media selection can be examples. If proper mechanisms are not enforced, issues like inaccuracy might arise.
4. ***Improper classification :*** Improper classifications or irrelevant descriptions lead to this type of errors. Not considering this issue will cause inconsistency in data (Crowston and Prestopnik, 2013).
5. ***Suspicious submissions (based on time) :*** One important aspect to consider is the time spent to submit an observation. Optimal time can be measured by participants and anything far less than that can be considered an issue.
6. ***Copyright images :*** There might be situations where a copyrighted image is used to explain a situation instead of submitting an original image. If not dealt properly, there might be issues with accountability (Maisonneuve, Stevens and Ochab, 2010).

7. ***Privacy & accountability :*** Posting media (images, video, sound recordings etc) of other people without permission explains this issue. There might be issues with law if not dealt (Maisonneuve, Stevens and Ochab, 2010).
8. ***Submitting multiple issues in single observation :*** This is similar to classification but at a different scale. In this type, the participant classifies it properly but the image has multiple issues in it. For example, one image containing two different things A & B might be classified as only A. Not dealing this might create believability and accuracy issues.
9. ***Misinformed submission / backup submission :*** The participant, if not informed, might create multiple submissions without knowledge that he is creating them. This creates a lot of duplicate data to deal with (Kim, Mankoff and Paulos, 2013).

4.4.1.3 Issues due to biases

Humans have their biases and human biases can lead to faulty results if these biases are inadvertently introduced into the system by the researchers or the participants. Thus it is necessary to check that biases are not introduced in citizen science projects.

1. ***Opinion based observations :*** When participants submit data, it is not the original situation that is submitted but the interpretation of the citizen that is submitted. Not correcting this issues lead to inaccuracy. For some people, the situation represents something while for others it might represent a completely different thing (Crowston and Prestopnik, 2013).
2. ***Lack of geographical spread of participants :*** This is similar to biased data but at geographical level. Participants are not spread across equally. so there would be a lot of areas with no data or less data to evaluate. not dealing with this would create biased analysis (Jaimes, Vergara-Laurens and Raij, 2015).

4.4.1.4 Behavioural biases

1. ***Blind spots :*** Participants choosing specific geographical zone to submit observations instead of spreading across the region causes this kind of issues. It also occurs when all

the participants submit observations of one particular type instead of different types of observations.

4.4.1.5 Validation Issues

In citizen science projects, it is of great importance that the data collected by people is accurate and valid. Since, all the participants may not have the same level of expertise or familiarity with the system or concept, they may provide incorrect data points and thus their observations may not be accurate. Thus it is required that the data being provided by the participants are validated against certain standard. There are certain factors that needs to be taken care when validating user generated data, which are defined as follows:

1. ***Ill defined Metrics :*** Metrics definition is one of the key elements which define the quality of data. If they are not defined properly, the quality of data would be very bad.
2. ***Lack of mechanisms to validate metrics :*** Having metrics alone won't solve the issues. There should be proper mechanisms required to validate the data.
3. ***Unclear Context :*** Citizen projects must aim to educate its participants about the context of the study. Lack of this understanding leads to irrelevant observations or corrupted observations.

4.4.1.6 Issues that cause loss of data

In citizen science projects or in any other scientific study, loss of valid data could be a critical factor for determining the success and validity of the scientific study being conducted. There are many stages in a citizen science project where the data being generated is lost due to many possible factors ranging from human error to hardware issues. Some of the issues that must be considered when designing a citizen science projects are as listed below.

1. ***Forgot to submit :*** There might be some situations where the participant is interested to submit an observation but might get distracted due to some other incident. Not solving this issue might lead to loss of important observations.
2. ***Blind Spots :*** This issues might arise when most of the participants are interested in monitoring one particular type ignoring the rest. This would leave other types to have low or no results.

3. ***Battery issues :*** When participants use devices, this is one of the most common issues faced. If not dealt properly there might be lesser quantity of data.
4. ***Data connectivity issues :*** Most of the countryside is not properly connected. When the participant wants to submit an observation, there might be a situation where there is no data connectivity. This leads to loss of valuable observations. Not considering this issue would lead to loss of observations.
5. ***System updates :*** Updates aimed to solve one issue might create other issues. Releasing the updates without proper testing will lead to loss of participant motivation.

4.4.1.7 Issues that affect data acquisition

Data acquisition is the most critical aspect of citizen science project as the concept of using participant based model for collecting observation is the soul of any citizen science project. Since, in a citizen science project, the data being collected by the observers are dependent on the individual perception of the phenomena being recorded, it is required that a standard suit is employed to acquire the data and to do the same it is required that issues listed below are considered.

1. ***Incomplete data :*** It refers to the state where all the required information is not present. This might cause issues like incompleteness (Wiggins and He, 2016).
2. ***Duplicates :*** The participant may submit the observation multiple times in the absence of suitable mechanism to solve this issue. Not providing solution might lead to unnecessary observations (Wiggins *et al.*, 2011).
3. ***Geographical Spread :*** This is similar to biased data but at geographical level. Participants are not spread across equally. so there would be a lot of areas with no data or less data to evaluate. not dealing with this would create biased analysis (Jaimes, Vergara-Laurens and Raij, 2015).
4. ***Inconsistency :*** Improper classifications or irrelevant descriptions lead to this type of errors. Not considering this issue will cause inconsistency in data (Guo *et al.*, 2015).
5. ***Lack of resources to validate :*** At observation level, the participant might submit only a few details ignoring most of the required data. There would be a lot of empty spaces in

these kind of situations. This would lead to incompleteness of data.

6. ***Spamming :*** If an intruder keeps submitting similar observation multiple times, it can be considered as spamming. Another type of spamming is notifying participant when not necessary which might distract.
7. ***Issues during export of data :*** Many case studies state that they had issues while exporting data. Some of the issues include not having all the metadata required, or details of software version while observations captured. If not dealt with this issue, the analysis will never be fruitful (Wiggins and He, 2016).
8. ***Accidental Submissions :*** It refers to the participant submitting the observation inadvertently. Sometimes this might be due to the device being accidentally triggered. Without proper mechanisms to solve this, there might be issues like duplicacy, unwanted data (Kim, Mankoff and Paulos, 2013).

4.4.1.8 Miscellaneous issues

Apart from the above mentioned issues that are categorized under different types, there are some other issues which cannot be classified into any of the above mentioned types but are still very relevant in a citizen science project. These issues are also very significant and should thus be considered by the researchers when developing a citizen science project.

1. ***Application Crashes :*** When a lot of participants subscribe and want to submit their observations at the same time, there is a possibility that the machine might crash if it is not properly designed.
2. ***Language translation issues :*** When dealing at a global or multi-national scale or multi-culture environment, there should be a focus on having multiple languages to help users. If the platform doesn't have the ability to understand colloquial forms of expressions, this issue might arise. This might lead to confusions
3. ***Security of platform :*** Insecure platform can be vulnerable source to hackers to create fake data. Lots of data quality issues arise if proper security is not implemented
4. ***Character sets and encoding (Emojis etc) :*** Every platform may not be robust to understand different character sets and encodings. Either the user should be informed

about this or there should be mechanisms that would not allow the users to use different keyboards.

Table 4 : Different issues faced in citizen science projects and their possible solutions

Issue	Solution
Blind spots	Algorithmic predictions based on opendata, satellite data, geographically targeted notifications etc. can be used in some cases to remove blind spots, Nonparametric and semi-parametric statistical modeling for bias
Forgot to submit	Personalised reminders (geographical, time based, activity based, notifications) are useful to encourage users to submit their observations.
Issues during export of data	Keeping track of metadata and version history along with the data while exporting
Lack of resources to validate - observation level	Ask users to submit all mandatory deciding factors by making critical fields mandatory
Lack of resources to validate - data level	This issue cannot be solved completely but predictions based on nearby locations, satellite data
Spamming	Mechanisms such as authenticating users and Flagging and blocking accounts.
Lack of mechanisms to validate metrics	GQM methodology helps in creating mechanisms which can help solve this issue
Hardware calibration	Issues in this domain can be addressed by using tailor made codes for specific hardware or normalize data after collection to remove aberrations. Also, assessment for hardware and collecting diagnostic reports
Unclear Context	FAQs, context awareness, expert guidance, moderator features (asking peers to check if the observation is in context), social connectors (connect with others on social media)
Geographical spread	gamification, opendata, predictions based on neighbouring areas / other sources (satellites), Nonparametric and semi-parametric statistical modeling for bias
Digital literacy	Physical explanations, media instructions, design inclusion features (for physically disabled - visually etc)
Ill defined Metrics	Proper definition of metrics and parameters is a must

Behavioural biases	Bias watchdog mechanism (alerts possible bias in data) could be implemented
Duplicates	Automated flagging or human intervention.
Incomplete data	Making fields mandatory, personalised reminders (to edit observation), automated feature support etc should be provided
Inconsistency	Automated flagging with humans to review the flagged content
Accidental submissions	Check for confirmations - may be immediately after submission or when participant opens the application, Automated flagging (with human in the process)
Copyright images or media	Social Translucence Mechanism - Peer Flagging, moderator, Automated flagging (with human in the process)
Privacy & accountability	Social Translucence Mechanism - Peer flagging, moderator, Automated flagging (with human in the process)
Multiple issues in one image - Lost items + invasive species	Expert review, Observation Ranking, use of AI, Community knowledge, Tagging
misinformed submission / backup submission	Provide Feedbacks on submission or block submit button after being triggered once.
Application Crash	Release after thorough testing. Design to share diagnostics and work on them in case of crashes.
Hardware connection failure	Notify participant, troubleshoot problems, test connection and interactive troubleshooting can be used.
Upgrades (Users unable to submit observations after an update)	Participants should be kept in the loop and application's workability should be checked before releasing new updates.
Character sets and encoding (Emojis etc)	Scripts could be employed to run periodically to clean data based on certain parameters.
Security of platform	A checklist of security protocols should be created and should be employed.
Errors during entry or submission	Features such as embedded validation in design, asking for confirmation, option to edit and use of technology to match error probability
Language translation issues	Providing content and messages in different languages and based on user's preference should be the part of the application feature.
Opinion based observations	Acknowledge that it can't be completely solved and such observations may be flagged for review by experts.

Improper classification	Confirmation check, Expert review, Option to edit, Automated flagging (with human in the end)
Suspicious submissions (based on time)	Confirmation check, Option to edit, Automated flagging
Battery issues	Optimize energy usage, use mobile sensors whenever possible, ask if they want to submit whenever app is opened. Disruption free service - Use local storage when low on battery
Data connectivity issues	Use disruption free service - Capture sensor data and store it in local storage immediately. Share to server once back online.
Location issues	Location confirmation, Tailor code / normalize after collection to manufacture features, Assessment for hardware, collect diagnostic reports, Calibrate hardware
Participant Selection	Training, Expert Guidance, Collective Building

4.4.2 RQ 2: How do we measure the data quality in citizen science?

Each citizen science project is defined for a different purpose. So, it is hard to generalize a set of mechanisms that can be applied to all the projects. Depending on the type of the project and the required outcomes of the project, a set of mechanisms are to be chosen which serve the purpose.

In case of a bird watching experiment, mechanisms like asking the participants to enter what they see instead of the name and species helped the participants participate with more interest and also data acquired was more accurate than the previous data (Sullivan *et al.*, 2009).

When proofs were needed asking citizens to upload pictures of what they see, taking the location information from the device directly instead of asking the citizens to update, would help in increasing the accuracy. Though it would increase some burden on the scientists, accuracy aspect would be solved to a great extent (Mason and Garbarino, 2016).

Applications like Petrowatch which required privacy in terms of location used the concepts of spatial cloaking. They had offered a geographical region instead of a point where multiple users have marked as relevant location (Sabrina, Murshed and Iqbal, 2016).

Over the period of time, multiple mechanisms were proposed by different practitioners of citizen science. Framework part of this paper gives an overview of different mechanisms which if followed can improve the data quality of citizen science projects.

4.4.3 RQ 3: What are the different mechanisms or metrics available currently for improving the data quality?

We have proposed a framework which can help in improving the data quality of citizen science projects. The framework contains 61 mechanisms which are segregated based on the time where they can be applied in the process - before collection, during collection, and after collection (see table 7).

4.4.3.1 Explaining the Framework

Oxford dictionary defines a framework as a basic structure underlying a system, concept, or text. Therefore, it can be said that a framework is a set of principles that can be applied to solve a particular problem. In this project we created a framework for improving the data quality in citizen science projects.

The framework was developed by considering a large amount of literature available, along with interviews, case studies and learnings from implementing sensei platform. The created framework serves as an outcome of this thesis work. It has been observed that there are a lot of data quality issues in the domain of citizen science but these issues can be addressed by employing certain technological solutions and innovative approaches so that they do not occur during the life cycle of the project. In this project, we have listed set of problems and the solutions that can help in solving the problems of data quality in citizen science project. Moreover, we have implemented a set of solutions that can help in improving the quality of data directly or indirectly. These set of solutions constitute the framework. In the framework, the

whole process is applied during three main steps - before collecting the data, during the collection of data, and after the collection of data. Based on the occurrence of the problem, we have created a map which tells us what mechanism can be applied at a point during the process which can help in improving the quality of data.

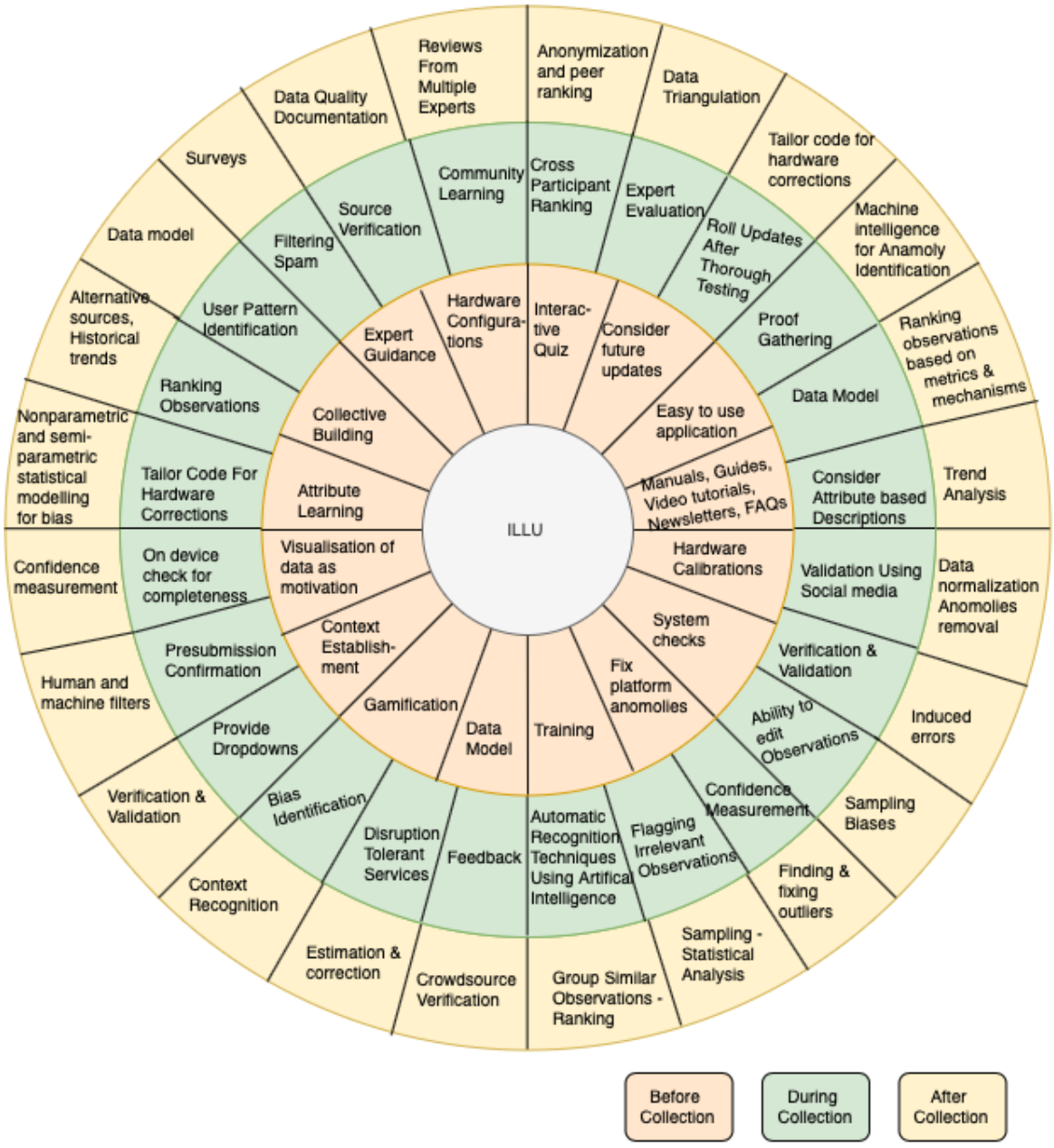


Figure 10 : Framework

4.4.3.2 Before Collection

In the first step of pre data collection, the framework gives a set of processes that should be followed so as to pre-empt the occurrence of certain data quality issues. These preemptive steps deal with users, hardware and other aspects. In the case of users, the preemptive steps to be followed could range from training the citizen scientist, providing expert guidance, manuals, training material, etc. Similarly, in the case of hardware, system checks and hardware calibration should be done beforehand.

Participant Selection & Training : In any citizen science project, this would be the first step. If the proper participants are not selected, the whole process would not be fruitful (Ren *et al.*, 2015). Proper care should be taken when selecting the participants. Their knowledge on the subject should be properly analysed, training should be provided on what is the goal of the project, what are the criteria for monitoring and how to share their observations (Nov, Arazy and Anderson, 2014).

Expert Guidance : As part of training the participants, there are multiple methods suggested. One such method is expert guidance. In this scenario, experts of that field reach out participants and explain to them what is the goal of the project, their tasks, the kind of observations they need to make (Guo *et al.*, 2015). This was followed in a few projects and proved to be a great source of improving data quality.

Collective Building : This is a method where the participants gather in a place and start to collaborate with each other and learn more about the project by their interactions along with experts. This helps participants learn better as knowledge is shared from both the experts and their peer participants (Buytaert *et al.*, 2014).

Attribute Learning : This mechanisms suggests that the participants should be able to understand the attributes of the monitoring subject. In case of bird monitoring, helping the

participants to learn about the structure of bird, their color, average height, pattern of feathers makes them understand better and submit more quality observations (Lukyanenko, Parsons and Wiersma, 2016).

Manuals, Guides, Video tutorials, Podcasts, Newsletter, FAQs : Providing users with manuals, guides, video tutorials or podcasts, newsletters, FAQs regarding the research, subject being monitored and relevant required information helps the users to learn better, refer when needed and thus perform better (Bonney *et al.*, 2009).

Interactive Quiz : This mechanisms helps the organizer to understand if the proper participants are selected and if their training helped them. It also helps participants learn more while interacting. Quiz can be in any form from a one-one talk to a group discussion. This lets the participant understand some crucial data which he might miss otherwise (Bonney *et al.*, 2009).

Context Establishment : Citizen science projects run with a specific goal to achieve. If the participants are not given a proper idea of the context of the project, the data they provide might not serve the purpose. Establishing context to participants is a very crucial process which can decide the outcome of the project.

Gamification : Including the concept of gamification helps to motivate the users. Sometimes, it may be a discouraging factor too. This mechanism involves providing incentives to users in order to motivate them and hence make them submit more observations. If it is not properly implemented, there is a chance to have many irrelevant observations (Budde *et al.*, 2017).

Data Model : Whatever the experiment is, whoever the participants are, without a proper data model, the quality can never be promised. This is the key to determine the quality of data. Without a proper data model, it is almost impossible to get quality data for research. It needs to be considered well before the project starts (Alabri and Hunter, 2010), (Wiggins *et al.*, 2011). This mechanism applies to all stages from starting the project to analysing the data.

Hardware Configuration : Configuring all the hardware required for the platform beforehand helps both the organizers and participants task easier. If not properly planned, there might be issues of platform crashes which might be the reason for losing many valuable observations.

Fix Platform Anomalies : Testing and fixing anomalies plays a key role in not just citizen science but for any project. Releasing the platform with bugs or temporary fixes might create issues to both participants by demotivating them and organizers by loss of observations (Maisonneuve, Stevens and Ochab, 2010).

System Checks : Regular system checks are to be performed to check for any anomalies. Bugs should be collected from the users directly or indirectly and fixes should be released. Regular checks should be performed to make sure of the process moving smoothly. All the algorithms need to be tested with multiple use cases before the actual application is released without which the cost to fix the application would be higher than the cost of developing it and parallely results in loss of data (Slaughter, Harter and Krishnan, 1998).

Hardware Calibrations : There are many cases where citizen science projects depend on sensors or external devices connected. Some examples of sensors include gps sensor for location information, microphone etc. If they are not configured properly, the results would not be accurate (Maisonneuve, Stevens and Ochab, 2010).

Easy to use platform : The design of the application should be based on understanding the user expertise, experience and psychology (Sutcliffe, 2009). The easier the platform to use, the more the people get motivated to use (Sutcliffe and Namoune, 2008). Low battery usage, smooth functioning, not blocking regular functionality to be considered while developing the platform.

Visualisations of data for motivation : This mechanisms states that if participants are shown what their peers are working on or what are the kind of observations they made, it would

motivate them. Also, showing data of peers would give a clear idea of what is to be monitored and establishes a good context in few cases (Sullivan *et al.*, 2009).

Consider future updates : When working on citizen science projects, it is better to have the update plans ready. As it deals with multiple participants from different backgrounds, everyone might not appreciate the changes to the platform. Care should be taken that the updates do not deviate much from the initial versions and if possible, release only the bug fixes.

4.4.3.3 During Collection

In the second step of data collection, the framework recommends a set of processes which, if implemented, can improve data quality in citizen science projects. For example, community learning, cross-participant ranking, proof gathering, attribute based descriptions, ability to edit observations, feedback, disruption tolerant services etc can help participant in reporting the observations accurately. Similarly, source verification, bias identification, tailor made codes can help remove ambiguity in data collection. Validation, spam filtering, flagging irrelevant observations, Automatic recognition techniques, ranking observations done on machine side promise the quality.

Community Learning : This mechanism works great for both organizers and participants. It asks the participants to share their observations, even if they are not sure of the classification and then asks co-participants to have a discussion on what could the observation possibly represent. This way, there is a data quality check by peers before actual data quality check happens. Also, by this method, participants share their observations on all their findings without the issue of their observation being ignored for incomplete data or other reasons (wiggins and He, 2016).

Cross Participant Ranking : The prerequisite for this mechanism is that the data should be available to all other participants. If a participant considers the observation is valid, they can give a positive feedback to that observation or incase he feels there is some issue, they can leave a

negative feedback. Based on this, the data quality can be calculated. This mainly helps in identifying the attribute of timeliness (Wiggins and He, 2016), (Wiggins *et al.*, 2011).

Expert Evaluation : This mechanisms can be applied to almost all citizen science projects. This mechanism works during and after collection of data. Experts of the field should be allowed to evaluate the data and give their ratings. This helps in removing unwanted or unqualified data to a great extent (Roman *et al.*, 2017)

Roll updates after thorough testing : Issue caused by lack of this was personally experienced in 'Sensei'. If proper testing is not done before the release of update, it might result in a wide range of issues. Thorough testing of the platform must be done before the actual release of the platform to the participants.

Proof gathering : Either media like photos, videos, sound recordings or description of the observation explaining the observation help in validating the data. This mechanism could help in solving the issue of timeliness and also helps in gathering the proof of events (Roman *et al.*, 2017), (Wiggins *et al.*,2011).

Consider Attribute Descriptions : This might not be applicable to a few projects but when applicable and used, it increases the quality of observation by a great fold. This mechanism suggests the developers to design the platform such that the participant has to describe the attributes of the subject being observed than just a brief description of the subject. This helps in easy analysis of data (Lukyanenko, Parsons and Wiersma, 2016). If properly used, this data can be used to train machine learning algorithms too.

Validation using social media and crowdsourcing : This mechanism is not valid for all the cases. It suggests the use of hashtags and trends on social media to compare with the received data for quality checks. Also, as social media is not completely reliable, it suggests to crowdsource the validation of social media so that it can be more useful. If done in real time, this

would help ask the participant further questions when necessary to improve the quality (Kim, Mankoff and Paulos, 2013), (Welvaert and Caley, 2016)

Verification : This mechanism suggests the use of algorithms to verify the observations. If performed in real-time, this can help requesting users to submit more data if required. This check helps in reassuring the data quality (Buddee *et al.*, 2017).

Feedback : Providing feedback about the observations to the participants helps them understand and motivates them. This also has another advantage of avoiding spamming. If feedback is not provided to the user about an observation, the user in a confused state continues to submit the same observation over and over again. This could lead to duplicates (Budde *et al.*, 2017),(Alabri and Hunter, 2010).

Ability to edit observations : The participant might not know what he is submitting at the time of submission but laters learns it correctly or the participant mistakenly wrongly classifies the observation. In both of the above cases, the participant wanted to correct the observation. If an option is provided to edit the observation, it would be useful for the participant to edit the wrongly submitted observations (Wiggins and He, 2016).

Automatic recognition techniques : Using technology to identify the observation and tag it with relevant information helps in improving the data quality. This can also be used for flagging of observations (Kelling *et al.*, 2009), (Wiggins *et al.*, 2011).

Flagging irrelevant observations : This mechanism suggest peer participants to flag irrelevant observations and help in control of low quality data. This mechanism has a prerequisite of making the data available to peers participating in the program. The flagging of observations would work better if authentication is made mandatory for flagging (Mashhadi and Capra, 2011).

Disruption Tolerant Services : This mechanism suggests the use of local temporary storage for the observations in case of absence of network or other problems. Once the problem is resolved, the application should send the observation to the server. Using this mechanism would help in saving the observations at times (Guo *et al.*, 2015).

Verify observation on device : If we are able to check for the format, values and a few other parameters on the device before sending it to the server, it would help the user in case of some issue in the observation and also blocks erroneous submissions.

User Pattern Identification : This mechanism suggests that if participant selection is made in such a way that there is more geographical area is covered, the geographical spread of the research would be higher and the geographical bias would be lower (Jaimes, Vergara-Laurens and Raij, 2015). An approach called FOAF (Friend of a friend) helps in understanding the relationship between users (Graves, Constabaris and Brickley, 2007) which if modified can help identify the pattern.

Observation Pattern Identification : Identifying the pattern of the user helps us understand his knowledge on the subject. For example, if a user is submitting similar kind of observations in a locality in a regular fashion, we can identify that the issue of that type is high in that area. Also analysing the observation submission patterns would help us understand the intensity of the problem in a few cases.

Source Verification : Allowing participants to submit the data without proper source identification might result in a lot of unnecessary data lowering the quality of data. Also, it would be impossible to contact the participant who submitted it if the researcher needs to clarify his questions on observation. By using verification techniques like registration and authentication (Alabri and Hunter, 2010) or the device info like IMEI or other sources would help improve the data quality. Using techniques like Friend of a friend (FOAF) can help in identity management (Graves, Constabaris and Brickley, 2007)

Filtering Spam : In case of multiple observations being submitted of the same category, at same time from the same participant, it can clearly be identified as a spam. But there are other cases too where spam might occur. Proper algorithms should be implemented to build spam protection and also notify the participant about the issue immediately to avoid further spamming. Also, use of Attack Resistance Trust metric for filtering out participants who submit fake observations would help enhance the quality of data (Alabri and Hunter, 2010).

Ranking Observations : Ranking the observations based on few parameters would help validating the data faster. One good way to rank would be to check how many similar observations were submitted in the same location around the same time by unique users. This would help us know how accurate the collected data is (Alabri and Hunter, 2010). This mechanism can be implemented both during and after collection of data.

Bias Identification : Biases are of many types. A few participants submit their observations only during the weekends which can be counted as weekend bias. A few participants submit observations of only one category which can be considered as bias towards that monitoring type. Others might be biased to submit only in certain region which can be constituted as regional bias. Citizen science projects generally have a range of bias (Roman *et al.*, 2017). Bias cannot be solved completely but identifying the bias can help identifying the biased results for sampling which can help in analysing the data quality (Bennin *et al.*, 2016).

Tailor code for hardware corrections : Each manufacturer designs the device with different set of chipsets and software. When we use sensors from different manufacturers, there would be slight corrections required for each manufacturer. For example, in the case of using location from mobile sensor, all devices don't give location with same accuracy. In case of citizen science projects using participants devices for observation collection, we have devices from multiple manufacturers. Corrects are needed based on hardware for data collected which would help in producing more accurate data (Maisonneuve, Stevens and Ochab, 2010). One other way to deal with specifically location issues, which we implemented in Sensei, is asking the participant to

confirm the location by showing current location on the map before submitting the observation. This location should be editable by the participant and only after confirming the location, the data should be saved.

Provide dropdowns wherever possible : In case of citizen science projects having to classify things, it would be better to provide participants with a list of options instead of asking the user to write down the classification information. This doesn't work all the time, but using this whenever possible would save a lot of time and also helps the participants to know the options they have to classify.

Ask for supporting information before submitting : It is always a good practice to check if all the data is filled properly before sending the data from the device. In case of missing data, prompting the user to fill the data would help achieving completeness and consistency of data.

Confirm with participant before submission : Showing the details of the observation and confirming with the user before actually submitting the observation helps in reducing the errors and also blocks spamming caused by device accidentally being triggered. This method reduces spam and improves accuracy of data.

4.4.4.4 Post Collection

In the final step of post-data collection, the framework suggests a set of mechanisms which can help in improving and enhancing the data quality. For example, expert analysis of data, verification by crowdsourcing, data quality documentation, preparing data models, ranking observations, statistical analysis, etc can help removing anomalies as well as in making sure the quality of observations being recorded are constantly at an acceptable level.

After data is collected, most of the data quality checks that are used in software engineering can be applied on citizen science data. Apart from the software engineering mechanisms, there are more mechanisms that can be applied on citizen science data.

Data Cleansing : Before starting the analysis of data, cleansing the data filtering all the possible simple errors like format errors, syntaxes can help in eliminating data of poor quality. Hybrid methods of cleansing can be applied to improve the quality (Khoshgoftaar *et al.*, 2006).

Expert analysis from multiple experts : After the data is collected, instead of one expert analysing the observations, if multiple experts were to analyse the observation individually, it would help bringing out more relevant data (Wiggins and He, 2016), (Budde *et al.*, 2017),(Loss *et al.*, 2015).

Crowdsource verification : This mechanism suggests to crowdsource the verification of data collected by the project. This includes the publishing of the collected data to participants who would validate the quality of data. As all data cannot be validated, it suggests to publish what data can be validated by the participating people (Wiggins and He, 2016),(Mitra, Hutto and Gilbert, 2015).

Surveys : This mechanism suggests to take surveys from participants regarding the observations they collected, and few other parameters. This helps in analysing the level to which the participant was able to accurately create his observations (Kananuar *et al.*, 2017)

Data Quality Documentation : Documenting the mechanisms and metrics used to analyze the data quality would be useful for future reference and also in case of re-evaluating the data quality, this documentation would be handy (Wiggins *et al.*, 2011).

Human and machine filters : Filtering the data after collection for the irregularities would help in understanding the outliers and other issues in data. Filtering can be done at machine level by writing queries to fetch specific data or filters applied by humans like filtering out some data which is irrelevant (Wiggins and He, 2016).

Analyze alternate sources and historical trends : This mechanism suggests the use of data from alternate sources like satellite images, other researches being carried out and study the historical trends for analyzing the quality of data. This would help us understand the accuracy attribute of data quality (Hunter, Alabri and van Ingen, 2012).

Verification & Validation : This mechanism suggests that the data collected should be verified and validated before analyzing it. This mechanism can be implemented in multiple ways depending on the kind of project and the mechanisms implemented in previous steps (Budde *et al.*, 2017), (Alabri and Hunter, 2010)

Ranking observations based on metrics and mechanisms : For this mechanism, the prerequisite is to have a data quality model available. Based on the data quality model, each observation is ranked individually and the score of the observation would give the quality of data. This was implemented in our project Sensei to study the data quality.

Sampling - Statistical analysis : This mechanisms suggests the use of sampling and statistical analysis to determine the quality of data collected. Here, a random sample of data is considered from the whole data set and statistical analysis is performed on that data. Similar process is followed for other random sets of data to analyze the quality of data (Wiggins and He, 2016), (Welvaert and Caley, 2016).

Nonparametric and semi-parametric statistical modeling for bias : This model was already being used to identify nonlinear relationships between the response and predict variables, and thus reduce the model bias to avoid model misspecification. Using this model can help identifying and reducing geographical biases (Cheng *et al.*, 2018).

Context recognition : This mechanism helps in removing the data with is not context relevant. For example, a few observations may be completely irrelevant to the context of monitoring but

the participants might have submitted them. By applying a few metrics to remove the observations out of context would help improve the data quality (Budde *et al.*, 2017).

Anonymisation and asking peers to rank : This mechanism is similar to the crowdsourcing but involves a few more steps for personal data anonymization. This mechanism suggests to remove the participant information who submitted the observation and then ask peer participants to rank it. This method will remove the bias that might be caused by friends giving better scores to their friends (Wiggins and He, 2016)

Data Triangulation : Comparing data from multiple sources like satellite data, data from other similar projects, historical data would be useful to improve the data quality. In case of no data available from other sources, this mechanism would not be of much use (Welvaert and Caley, 2016), (Alabri and Hunter, 2010), (Loss *et al.*, 2015), (Wiggins *et al.*, 2011).

Data Normalization and Anomalies removal : This is a regular data quality technique for data quality analyzation of software systems. The main difference would be in normalization. What data to be normalized should be properly chosen instead of normalizing complete data (Budde *et al.*, 2017), (Wiggins *et al.*, 2011).

Finding and fixing outliers : Citizen science data is gathered from multiple sources. Data would be very different compared to data from software engineering perspective. Specific algorithms must be developed to identify the outliers in this data. Developing proper algorithm, identifying the outliers and removing them would contribute a lot to data quality (Budde *et al.*, 2017).

Machine intelligence for anomaly removal : Use of machine intelligence to analyze the data and remove the anomalies is one of the known mechanisms in software engineering. Similar approach can be used for citizen science data with little modifications (Kelling *et al.*, 2009).

In the table below, we present different mechanisms that when applied at different stages help achieve and improve data quality in citizen science projects.

Table 5 : Framework **Illu**

Before	**During**	**After**
Participation selection & Training (Nov, Arazy and Anderson, 2014)	Community Learning (Wiggins and He, 2016)	Data Cleansing (Khoshgoftaar *et al.*, 2006)
Expert Guidance (Guo *et al.*, 2015)	Cross Participant Ranking (Wiggins and He, 2016),(Wiggins *et al.*, 2011)	Expert Analysis from multiple experts (Wiggins and He, 2016), (Budde *et al.*, 2017),(Loss *et al.*, 2015)
Attribute Learning (Lukyanenko, Parsons and Wiersma, 2016)	Expert Evaluation (Roman *et al.*, 2017)	Crowdsource verification (Wiggins and He, 2016), (Mitra, Hutto and Gilbert, 2015)
Collective Building (Buytaert *et al.*, 2014)	Roll updates after thorough testing	Surveys (Kananura *et al.*, 2017)
Gamification (Budde *et al.*,2017)	Proof gathering (media, description) (Roman *et al.*, 2017),(Wiggins *et al.*, 2011)	Data quality Documentation (Wiggins *et al.*, 2011)
Interactive Quiz (Bonney *et al.*, 2009)	Consider attribute based descriptions (Lukyanenko, Parsons and Wiersma, 2016)	human and machine filters (Wiggins and He, 2016)
Context Establishment	Data Model (Alabri and Hunter, 2010),(Wiggins *et al.*, 2011)	Data Model (Alabri and Hunter, 2010),(Wiggins *et al.*, 2011)
Manuals, Guides, Video tutorials, Podcasts, Newsletters, FAQs (Bonney *et al.*, 2009)	Validation using Social Media and crowdsourcing (Kim, Mankoff and Paulos, 2013), (Welvaert and Caley, 2016)	Analyze Alternative Sources, Historical Trends (Hunter, Alabri and van Ingen, 2012)
Data Model (Alabri and Hunter, 2010),(Wiggins *et al.*, 2011)	Verification (Budde *et al.*, 2017)	Verification & Validation (Budde *et al.*, 2017), (Alabri and Hunter, 2010)
Hardware Configurations	Feedback (Budde *et al.*, 2017),(Alabri and	Ranking observations based on

(Maisonneuve, Stevens and Ochab, 2010)	Hunter, 2010)	metrics and mechanisms
Fix platform anomalies (Maisonneuve, Stevens and Ochab, 2010)	Ability to edit observations (Wiggins and He, 2016)	Sampling - statistical analysis (Wiggins and He, 2016), (Welvaert and Caley, 2016)
System checks (Maisonneuve, Stevens and Ochab, 2010)	Automatic recognition techniques (Kelling *et al.*, 2009), (Wiggins *et al.*, 2011)	Nonparametric and semi-parametric statistical modeling for bias (Cheng *et al.*, 2018)
Hardware Calibrations (Maisonneuve, Stevens and Ochab, 2010)	Flagging Irrelevant Observations(Mashhadi and Capra, 2011)	Context recognition (Budde *et al.*, 2017)
Easy to use application	Disruption tolerant services - Temporary local storage for connectivity issues (Guo *et al.*, 2015)	Anonymisation and asking other participants to rank (Wiggins and He, 2016)
Visualisation of data as motivation (Sullivan *et al.*, 2009)	Verify the observation for issues on device	Data Triangulation (Welvaert and Caley, 2016),(Alabri and Hunter, 2010),(Loss *et al.*, 2015),(Wiggins *et al.*, 2011)
Consider future updates	User pattern identification (Jaimes, Vergara-Laurens and Raij, 2015)	Data normalization, Anomalies removal (Budde *et al.*, 2017),(Wiggins *et al.*, 2011)
	Source Verification (Alabri and Hunter, 2010)	Group similar observations - Ranking (Alabri and Hunter, 2010)
	Filtering spam (Alabri and Hunter, 2010)	Finding & fixing outliers (Budde *et al.*, 2017)
	Ranking observations (Alabri and Hunter, 2010)	Tailor code for hardware corrections (Maisonneuve, Stevens and Ochab, 2010)
	Bias identification (Mashhadi and Capra, 2011), (Kosmala *et al.*, 2016)	Machine Intelligence for anomaly removal (Kelling *et al.*, 2009)
	Tailor code for hardware corrections (Maisonneuve, Stevens and Ochab, 2010)	

	Provide dropdowns where ever possible	
	Ask for supporting information before submitting	
	Confirm with participant before submission	
	Observation Pattern Identification	

5 Discussion

This section will discuss how the findings from this thesis work are related to previous studies and theories, what the remaining challenges are, and finally what limitations are imposed in the study.

5.1 RQ 1 : What are the data quality issues that citizen science projects face?

Citizen science projects, if not properly designed, can face a lot of data quality issues at different stages of the project. The issues might cause some simple issues to catastrophic issues making it hard to analyze the data collected.

The primary issue every citizen science project has to face is during the selection of participants. Some common issues include participant selection (Ren *et al.*, 2015), geographical spread (Jaimes, Vergara-Laurens and Raij, 2015), unclear context, blind spots, biases (Crowston and Prestopnik, 2013). Many other issues arise if this step is not properly carried out.

We have provided an overview of the common issues faced by citizen science projects under results (See table 6).

5.2 RQ 2 : How do we measure the data quality in citizen science?

Citizen science projects are run for a specific purpose. Each project has a different set of requirements. It is hard to generalize the measurement technique for data quality in citizen science projects. Some of the measurement techniques frequently used were validated by experts (Bristol Natural History Consortium, 2019), (iRecord, 2019), (Sullivan *et al.*, 2009), Checklists, automated filters (Sullivan *et al.*, 2009), User ranking (Ispotnature.org, 2019), classification (Jordan Raddick *et al.*, 2013). There are multiple ways to measure the quality.

- One way is to find the attributes that are mainly required to define the data quality, find the mechanisms which can help analyze the attributes and rate each observation based on the mechanisms selected.

- Other way is to the use of trust based mechanism where similar observations are grouped together and their accuracy is known.
- Comparing data from other sources like historical data, social networks is another method to measure data quality.

5.3 RQ 3 : What are the different mechanisms or metrics available currently for improving the data quality?

Each citizen science project is defined for a different purpose. So, it is hard to generalize a set of mechanisms that can be applied to all the projects. Depending on the type of the project and the required outcomes of the project, a set of mechanisms are to be chosen which serve the purpose. Over the period of time, multiple mechanisms were proposed by different practitioners of citizen science which helped them solve the issues they were facing. Some common mechanisms include training (Nov, Arazy and Anderson, 2014), proof gathering (Roman *et al.*,2019), (Wiggins *et al.*, 2011), expert analysis from multiple experts (Wiggins and He, 2016), (Budde *et al.*, 2017), (Loss *et al.*, 2015), data triangulation (Welvaert and Caley, 2016), (Alabri and Hunter, 2010), (Loss *et al.*, 2015), (Wiggins *et al.*, 2011). Andrew Wiggins, and Budde studied the citizen science projects along with their colleagues and proposed a few mechanisms which could help solve a few issues (Wiggins *et al.*, 2011), (Budde *et al.*, 2017). But there is no single framework which explains all the issues and their possible solutions. This paper focuses on providing the list of issues of data quality generally faced and their probable solutions. As we have studied multiple projects and also have conducted experiments, the applications of framework would be broad in terms of data quality.

A framework named Illu has been proposed with a set of mechanisms and metrics which can help in improving the data quality in citizen science projects. The mechanisms are classified into pre collection, during collection and post collection to facilitate and ease the flow of using them in the citizen science projects.

5.4 Sustainability Analysis

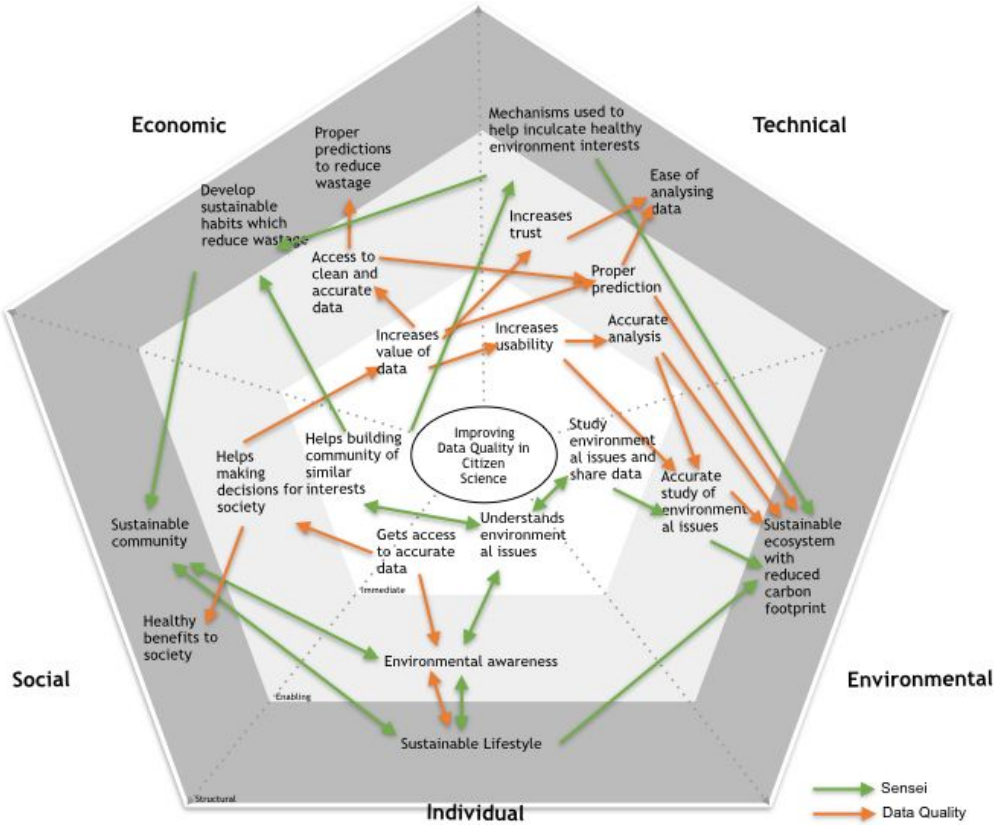
Sustainable development has become a significant term in recent times. The term ‘sustainable development’ is defined as the “development that would satisfy the needs of the present without compromising the ability of future generations to meet their own needs” (Brundtland, Gro, 1985).

Supporting sustainability in software engineering would have an impact on our society, economy, and environment as software systems have a heavy impact on our lives. Though there is no standard definition for sustainable software engineering, a tremendous effort is being put on topics related to sustainable development such as network optimization, energy efficient coding (Owusu & Pattinson, 2012), smart grids (Duboc *et al.*, 2019), (Friderikos, Helard, Porras, & Rao, 2014), green IT (Penzenstadler *et al.*,2012).

In order to analyze the sustainability aspects of a project, a sustainability analysis model was proposed by Becker *et al.* (2016). It helps to assess the systemic effects of a software on the five dimensions of sustainability i.e, individual, societal, economical, environmental, and technical . This model focuses on the following three core systemic effects defined by Hilty & Aebischer (2015):

- Immediate effects: Direct effects of the entire life cycle of the software system.
- Enabling effects: Appear from the use of the system over a long period.
- Structural effects: “persistent changes observable at the macro level” (Hilty & Aebischer, 2015).

Figure 11 : Sustainability Analysis



For this thesis, there are two parts of sustainability analysis, one from the view of data quality and the other with respect to Sensei.

Sensei observations : As many mechanisms to improve data quality are included in the platform, they help people to develop healthy environmental interests which can directly trigger the aspects which can reduce the wastage saving them economically and when done in societal level, it not only saves economically but also helps in building sustainable community. Also, as the study is regarding environmental issues and data is circulated among the participants, it helps them understand the environmental issues immediately and they get good awareness in long term helping them inculcate a sustainable lifestyle. As Sensei is a co-created environment, projects like this would help in building communities of similar interests who when start inculcating sustainable habits help form a sustainable community which helps in reducing the carbon footprint of that community and building a sustainable ecosystem.

Data quality observations : Data quality helps citizens to perceive accurate information. In case of environmental studies, accurate information helps citizens to understand the environment which included with developing sustainable habits helps in building a sustainable lifestyle. When this accurate data is provided at a scale of society, it helps the society to make proper decisions providing healthy benefits to the society. Also, as this data is of good quality, it increases the economic value of data reducing the time and money to clean before using it. Also a lot can be saved by proper predictions in terms of reducing the wastage. When the value of data is increased i.e, it is of higher quality, it increases usability, trust, helps make proper predictions and accurate analysis. It also smoothens the process of analysing the data which again helps in proper prediction which helps in reducing the carbon footprint and building a healthy ecosystem.

5.5 Study Limitations

There are a few observed limitations in the work due to the following factors - Sensei platform was only available to use for registered participants in city of Lappeenranta and hence the amount of data captured was limited. The duration of the project was during the summer of 2018 which led to monitoring of specific subjects.

Analysis of data of different projects was done with the assumption that no alterations were performed and original data was made available. Also, for the projects where the goals were not specified, data quality analysis was performed based on the results obtained.

Few of the mechanisms were studied but not practically applied due to few constraints. Also, special focus was applied on attributes of Accuracy, Believability, Reputation, Accessibility, Data Security, Value-added, Relevancy, Completeness, Interpretability, Representational consistency.

In addition to the stated metrics of data quality, there could be other metrics that may be relevant for any other citizen science project, but are not considered in this project as each citizen science

project is different in nature. Thus this framework could be used in case of any generic citizen science project, but may not be completely applicable in a certain specific project.

6 Conclusion

This thesis is set out to answer the research questions proposed at the beginning of the paper. In order to do so, data from dozens of citizen science projects was studied, analyzed and a set of mechanisms were assumed. To validate the mechanisms, a citizen sensing platform named sensei was developed incorporating the mechanisms. The experiment provided with a few more issues which were solved by trial and error method and the solutions were added to the mechanisms list which helped in forming a framework ILLU.

In conclusion, it can be said that citizen science projects are a capable way of conducting scientific studies of different nature and magnitude which was already proved by different studies like Galaxy zoo, eBird engaging thousands of citizens to monitor different aspects of environment. However, citizen science projects are prone to unreliable data if the researchers and scientist conducting the studies do not take into account the data quality aspects and incorporate the solutions to tackle the issue from the beginning of their study.

This study identified 35 major issues that affect data quality in citizen science projects, categorized them into 8 groups based on similarities. Possible set of solutions to each issue was provided.

In later part of the study, we present a framework called "ILLU", which includes mechanisms to tackle data quality issues in citizen science. It was developed using literature from the data quality field, case studies and expert interviews. This framework may be used by practitioners to design their projects to collect data with higher quality. One of the most important conclusions of this work is that though there are many mechanisms applied to improve the quality of data, there might still be issues generated if they are not properly implemented and conveyed to the participants.

References

Alabri, A. and Hunter, J. (2010) 'Enhancing the quality and trust of citizen science data', Proceedings - 2010 6th IEEE International Conference on e-Science, eScience 2010, pp. 81–88. doi: 10.1109/eScience.2010.33.

Aqcitizenscience.rti.org. (2019). Air Quality Citizen Science. [online] Available at: https://aqcitizenscience.rti.org/#/ [Accessed 11 May 2019].

Askham, N. *et al.* (2013) 'The Six Primary Dimensions for Data Quality Assessment', Group, DAMA UK Working, p. 16. Available at: https://www.dqglobal.com/wp-content/uploads/2013/11/DAMA-UK-DQ-Dimensions-White-Paper-R37.pdf

Becker, C., Betz, S., Chitchyan, R., Duboc, L., Easterbrook, S. M., Penzenstadler, B., ...Venters, C. C. (2016). Requirements: The key to sustainability. *IEEE Software*, *33*(1), 56–65. https://doi.org/10.1109/MS.2015.158

Bennin, K., Keung, J., Monden, A., Kamei, Y. and Ubayashi, N. (2016). Investigating the Effects of Balanced Training and Testing Datasets on Effort-Aware Fault Prediction Models. 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC).

Birds.cornell.edu. (2019). [online] Available at: http://www.birds.cornell.edu/citscitoolkit/conference/proceeding-pdfs/Phillips%202007%20CS%20Conference.pdf [Accessed 30 May 2019].

Bobrowski, M., Marré, M. and Yankelevich, D. (1998) 'A software engineering view of data quality', Intl. Software Quality Week Europe …, pp. 1–10. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.41.5713&rep=rep1&type=pdf.

Bonney, R., Cooper, C., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. and Shirk, J. (2009). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. BioScience, 59(11), pp.977-984.

Bonter, D. and Cooper, C. (2012). Data validation in citizen science: a case study from Project FeederWatch. Frontiers in Ecology and the Environment, 10(6), pp.305-307.

Bristol Natural History Consortium. (2019). BioBlitz - Bristol Natural History Consortium. [online] Available at: https://www.bnhc.org.uk/bioblitz/ [Accessed 10 May 2019].

Bröring, A., Remke, A., Stasch, C., Autermann, C., Rieke, M. and Möllers, J. (2015). enviroCar: A Citizen Science Platform for Analyzing and Mapping Crowd-Sourced Car Sensor Data. Transactions in GIS, 19(3), pp.362-376.

Brundtland, Gro, H. (1985). World Commission on environment and development.*Environmental Policy and Law*, *14*(1), 26–30. https://doi.org/10.1016/S0378-777X(85)80040-8

Budde, M., Schankin, A., Hoffmann, J., Danz, M., Riedel, T. and Beigl, M. (2017). Participatory Sensing or Participatory Nonsense?. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(3), pp.1-23.

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T., Bastiaensen, J., De BiÃ¨vre, B., Bhusal, J., Clark, J., Dewulf, A., Foggin, M., Hannah, D., Hergarten, C., Isaeva, A., Karpouzoglou, T., Pandeya, B., Paudel, D., Sharma, K., Steenhuis, T., Tilahun, S., Van Hecken, G. and Zhumanova, M. (2014). Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. Frontiers in Earth Science, 2.

Caldiera, V.R.B.G. and Rombach, H.D. (1994). The goal question metric approach. Encyclopedia of software engineering, pp.528-532.

Cappiello, C., Francalanci, C. and Pernici, B. (2003). Time-related factors of data quality in multichannel information systems. Journal of Management Information Systems, 20(3), pp.71-92.

Cheng, M., Huang, T., Liu, P. and Peng, H. (2018). Bias Reduction for Nonparametric and Semiparametric Regression Models. Statistica Sinica.

Conkertreescience.org.uk. (2019). Home | Conker Tree Science. [online] Available at: http://www.conkertreescience.org.uk [Accessed 10 May 2019].

Crowston, K. and Prestopnik, N. R. (2013) ‘Motivation and data quality in a citizen science game: A design science evaluation’, Proceedings of the Annual Hawaii International Conference on System Sciences, pp. 450–459. doi: 10.1109/

Dickinson, J., Zuckerberg, B. and Bonter, D. (2010). Citizen Science as an Ecological Research Tool: Challenges and Benefits. Annual Review of Ecology, Evolution, and Systematics, 41(1), pp.149-172.

Doit-europe. (2019). Doit Europe. [online] Available at: https://www.doit-europe.net [Accessed 5 Jun. 2019].

Doyle, C., David, R., Li, J., Luczak-Roesch, M., Anderson, D. and Pierson, C. (2019). Using the Web for Science in the Classroom: Online Citizen Science Participation in Teaching and Learning.

Duboc, L, Betz, S, Penzenstadler, B, Akinli Kocak, S, Chitchyan, R, Leifler, O, Porras, J, Seyff, N & Venters, CC 2019, Do we really know what we are building? Raising awareness of potential Sustainability Effects of Software Systems in Requirements Engineering. in *27th IEEE International Requirements Engineering Conference*. 27th IEEE International Requirements Engineering Conference, Jeju Island, Korea, Republic of, 23/09/19.

Ericsson.com. (2019). Q4 2018 update – Ericsson Mobility Report. [online] Available at: https://www.ericsson.com/en/mobility-report/reports/q4-update-2018 [Accessed 16 May 2019].

Estrin, D. (2010) 'Pervasive Computing', 6030(October). doi: 10.1007/978-3-642-12654-3.

Even, A. and Shankaranarayanan, G. (2007) 'Utility-Driven Assessment of Data Quality', The DATA BASE for Advances in Information Systems, 38(2), pp. 75–93. doi: 10.1145/1240616.1240623.

Ftp.sccwrp.org. (2019). [online] Available at: http://ftp.sccwrp.org/pub/download/DOCUMENTS/TechnicalReports/527_SWAMP_SD_StreamAssessmentSynthesis.pdf [Accessed 11 May 2019].

Friderikos, V., Helard, M., Porras, J., & Rao, T. R. (2014). Toward a Smart, Fully Connected Society: An Overview of the 32nd Meeting of the Wireless World Research Forum [From the Guest Editors]. *IEEE Vehicular Technology Magazine*, *9*(3), 24–26. https://doi.org/10.1109/MVT.2014.2335392

Garbarino, J. and Mason, C.E. (2016). The power of engaging citizen scientists for scientific progress. Journal of microbiology & biology education, 17(1), p.7.

Graves, M., Constabaris, A. and Brickley, D. (2007). FOAF: Connecting People on the Semantic Web. Cataloging & Classification Quarterly, 43(3-4), pp.191-202.

Guo, B. *et al.* (2015) 'Mobile Crowd Sensing and Computing: The Review of an Emerging Human-Powered Sensing Paradigm', ACM Computing Surveys, 48(1), pp. 1–31. doi: 10.1145/2794400.

Guo, B., Wang, Z., Yu, Z., Wang, Y., Yen, N., Huang, R. and Zhou, X. (2015). Mobile Crowd Sensing and Computing.

Heinrich, B. and Helfert, M. (2003). Analyzing Data Quality Investments in CRM-A model-based approach.

Heinrich, B. and Klier, M. (2015). Metric-based data quality assessment — Developing and evaluating a probability-based currency metric. Decision Support Systems, 72, pp.82-96.

Hilty, L. M., & Aebischer, B. (2015). ICT for Sustainability: An Emerging Research Field. Retrieved from http://publicationslist.org/data/lorenz.hilty/ref-225/2014_Hilty_Aebischer_ICT_for_Sustainability.pdf

Holler, J. *et al.* (2014) From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence. doi: B978-0-12-407684-6.00001-2.

Hochachka, W.M., Fink, D., Hutchinson, R.A., Sheldon, D., Wong, W.-K., Kelling, S., 2012. Data- intensive science applied to broad-scale citizen science. Trends in Ecology & Evolution 27, 130–137. DOI: 10.1016/j.tree.2011.11.006

Hunter, J., Alabri, A. and van Ingen, C. (2012). Assessing the quality and trustworthiness of citizen science data. Concurrency and Computation: Practice and Experience, 25(4), pp.454-466.

Irwin A (1995) Citizen Science. London: Routledge.

Ispotnature.org. (2019). Home | iSpot Nature. [online] Available at: https://www.ispotnature.org [Accessed 10 May 2019].

Jaimes, L., Vergara-Laurens, I. and Raij, A. (2015). A Survey of Incentive Techniques for Mobile Crowd Sensing. IEEE Internet of Things Journal, 2(5), pp.370-380.

Jordan R, Singer F, Vaughan J, and Berkowitz A. 2009. What should every citizen know about ecology? Front Ecol Environ 7: 495–500.

Jordan RC, Gray SA, Howe DV, *et al.* 2011. Knowledge gain and behavioral change in citizen-science programs. Conserv Biol 25: 1148–54.

Jordan Raddick, M., Bracey, G., Gay, P., Lintott, C., Cardamone, C., Murray, P., Schawinski, K., Szalay, A. and Vandenberg, J. (2013). Galaxy Zoo: Motivations of Citizen Scientists. Astronomy Education Review, 12(1).

Juran, J.M., Gryna, F.M.J., and Bingham, R.S. Quality Control Handbook (3rd ed.). McGraw-Hill Book Co, New York, NY, 1974.4.

Kahn, B. K., Strong, D. M. and Wang, R. Y. (2002) 'Information quality benchmarks: product and service performance', Communications of the ACM, 45(4ve), pp. 184–192. doi: 10.1145/505999.506007.

Kaiser, M., Klier, M. and Heinrich, B. (2007) 'How to Measure Data Quality? - A Metric-Based Approach', ICIS 2007 Proceedings, p. 16. Available at: http://aisel.aisnet.org/icis2007/108.

Kananura, R., Ekirapa-Kiracho, E., Paina, L., Bumba, A., Mulekwa, G., Nakiganda-Busiku, D., Oo, H., Kiwanuka, S., George, A. and Peters, D. (2017). Participatory monitoring and evaluation approaches that influence decision-making: lessons from a maternal and newborn study in Eastern Uganda. Health Research Policy and Systems, 15(S2).

Kelling, S., Hochachka, W., Fink, D., Riedewald, M., Caruana, R., Ballard, G. and Hooker, G. (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies. BioScience, 59(7), pp.613-620.

Khoshgoftaar, T., Folleco, A., Van, J. and Bullard, H.L., 2006. Multiple imputation of missing values in software measurement data.

Kim, S., Mankoff, J. and Paulos, E. (2013). Sensr. Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13.

Kor, A.L., Rondeau, E., Andersson, K., Porras, J. and Georges, J.P. (2019, July) Education in green ICT and control of smart systems: A first hand experience from the International PERCCOM masters programme.

Kosmala, M., Wiggins, A., Swanson, A., Simmons, B., 2016. Assessing data quality in citizen science. Frontiers in Ecology and the Environment 14, 551–560. DOI: 10.1002/fee.1436

Lavazza, L. (2000). Providing automated support for the GQM measurement process. IEEE Software, 17(3), pp.56-62.

Loss, S., Loss, S., Will, T. and Marra, P. (2015). Linking place-based citizen science with large-scale conservation research: A case study of bird-building collisions and the role of professional scientists. Biological Conservation, 184, pp.439-445.

Lukyanenko, R., Parsons, J. and Wiersma, Y. (2016). Emerging problems of data quality in citizen science. Conservation Biology, 30(3), pp.447-449.

MacDonald, E., Case, N., Clayton, J., Hall, M., Heavner, M., Lalone, N., Patel, K. and Tapia, A. (2015). Aurorasaurus: A citizen science platform for viewing and reporting the aurora. Space Weather, 13(9), pp.548-559.

Maisonneuve, N., Stevens, M. and Ochab, B. (2010). Participatory noise pollution monitoring using mobile phones. Information Polity, 15(1,2), pp.51-71.

Mashhadi, A. and Capra, L. (2011). Quality control for real-time ubiquitous crowdsourcing. Proceedings of the 2nd international workshop on Ubiquitous crowdsouring - UbiCrowd '11.

Mason, C. E. and Garbarino, J. (2016) 'The Power of Engaging Citizen Scientists for Scientific Progress', Journal of Microbiology & Biology Education, 17(1), pp. 7–12. doi: 10.1128/jmbe.v17i1.1052.

Messenger, J. C. *et al.* (2012) 'The National Cardiovascular Data Registry (NCDR) data quality brief: The NCDR Data Quality Program in 2012', Journal of the American College of Cardiology, 60(16), pp. 1484–1488. doi: 10.1016/j.jacc.2012.07.020.

Met Office WOW. (2019). Met Office WOW. [online] Available at: https://wow.metoffice.gov.uk [Accessed 10 May 2019].

Mitra, T., Hutto, C. and Gilbert, E. (2015). Comparing Person- and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15.

Moore, S. (2018). How to Create a Business Case for Data Quality Improvement. [online] Gartner.com. Available at: https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/ [Accessed 30 May 2019].

Nadkarni, N. and Stevenson, R. (2009). Symposium 9: Linking Scientists with Nontraditional Public Audiences to Enhance Ecological Thought. Bulletin of the Ecological Society of America, 90(1), pp.134-137.

Nah, F. (2004). A study on tolerable waiting time: how long are Web users willing to wait?. Behaviour & Information Technology, 23(3), pp.153-163.

Nature (2019). No PhDs needed: how citizen science is transforming research. [online] Available at: https://www.nature.com/articles/d41586-018-07106-5 [Accessed 4 Jun. 2019].

Nov, O., Arazy, O. and Anderson, D. (2014). Scientists@Home: What Drives the Quantity and Quality of Online Citizen Science Participation?. PLoS ONE, 9(4), p.e90375.

Old Weather Blog. (2019). Old Weather Blog. [online] Available at: https://blog.oldweather.org [Accessed 10 May 2019].

Opalexplorenature.org. (2019). OPAL Air Survey | OPAL. [online] Available at: https://www.opalexplorenature.org/airsurvey [Accessed 10 May 2019].

Opalexplorenature.org. (2019). OPAL Bugs Count Survey | OPAL. [online] Available at: https://www.opalexplorenature.org/bugscount [Accessed 10 May 2019].

Opalexplorenature.org. (2019). OPAL Soil and Earthworm Survey | OPAL. [online] Available at: https://www.opalexplorenature.org/soilsurvey [Accessed 10 May 2019].

Ottinger, G. (2017) 'Making sense of citizen science: Stories as a hermeneutic resource', Energy Research and Social Science. Elsevier, 31(November 2016), pp. 41–49. doi: 10.1016/j.erss.2017.06.014.

Owusu, F., & Pattinson, C. (2012). The Current State of Understanding of the Energy Efficiency of Cloud Computing. In *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications* (pp. 1948–1953). IEEE. https://doi.org/10.1109/TrustCom.2012.270

Palacin, V., Ginnane, S., Ferrario, M., Happonen, A., Wolff, A., Piutunen, S. and Kupiainen, N. (2019). SENSEI: Harnessing Community Wisdom for Local Environmental Monitoring in Finland. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19.

Penzenstadler, B., Bauer, V., Calero, C., & Franch, X. (2012). Sustainability in software engineering: a systematic literature review. In *16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012)* (pp. 32–41). IET. https://doi.org/10.1049/ic.2012.0004

Pipino, L. L., Lee, Y. W. and Wang, R. Y. (2002) 'Data quality assessment', Communications of the ACM, 45(4), p. 211. doi: 10.1145/505248.506010.

Rantala, M. (2016) 'Data quality analysis in industrial maintenance; theory vs. reality'. Available at: https://www.doria.fi/bitstream/handle/10024/125127/Diplomityö_Rantala_Miika.pdf?sequence=3.

Ren, J., Zhang, Y., Zhang, K. and Shen, X. (2015). SACRM: Social Aware Crowdsourcing with Reputation Management in mobile sensing. Computer Communications, 65, pp.55-65.

Roman, L., Scharenbroch, B., Östberg, J., Mueller, L., Henning, J., Koeser, A., Sanders, J., Betz, D. and Jordan, R. (2017). Data quality in citizen science urban tree inventories.

Sabrina, T., Murshed, M. and Iqbal, A. (2016) 'Anonymization Techniques for Preserving Data Quality in Participatory Sensing', 2016 IEEE 41st Conference on Local Computer Networks (LCN), pp. 607–610. doi: 10.1109/LCN.2016.103.

Shirk, J., Ballard, H., Wilderman, C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B., Krasny, M. and Bonney, R. (2019). Public Participation in Scientific Research: a Framework for Deliberate Design.

Silvertown, J. (2009). A new dawn for citizen science. Trends in Ecology & Evolution, 24(9), pp.467-471.

Slaughter, S., Harter, D. and Krishnan, M. (1998). Evaluating the cost of software quality. Communications of the ACM, 41(8), pp.67-73.

Statista. (2019). Forecast on connected devices per person worldwide 2003-2020 | Statistic. [online] Available at: https://www.statista.com/statistics/678739/forecast-on-connected-devices-per-person/ [Accessed 21 Feb. 2019].

Strong, D., Lee, Y. and Wang, R. (1997). Data quality in context. Communications of the ACM, 40(5), pp.103-110.

Sullivan, B., Wood, C., Iliff, M., Bonney, R., Fink, D. and Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation, 142(10), pp.2282-2292.

Sutcliffe, A. (2009). Designing for User Engagement: Aesthetic and Attractive User Interfaces. Synthesis Lectures on Human-Centered Informatics, 2(1), pp.1-55.

Sutcliffe, A. and Namoune, A. (2008). Getting the message across. Proceedings of the 7th ACM conference on Designing interactive systems - DIS '08.

Svetoslav, S., Iordan, H. and Nikolov, S. (2017). Population trends of common birds in Bulgaria: Is their status improving after the EU accession?. Acta Zoologica Bulgarica, [online] 69(1), pp.95-104. Available at: https://pecbms.info/wp-content/uploads/2018/10/spasov-and-hristov-2017.pdf [Accessed 12 May 2019].

Tcv.org.uk. (2019). [online] Available at: https://www.tcv.org.uk/sites/default/files/172/files/CommunityCitizenScienceguidance.pdf [Accessed 10 May 2019].

TechCrunch. (2019). Nearly 1 in 4 people abandon mobile apps after only one use – TechCrunch. [online] Available at: https://techcrunch.com/2016/05/31/nearly-1-in-4-people-abandon-mobile-apps-after-only-one-use/?guccounter=1&guce_referrer_us=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvLmluLw&guce_referrer_cs=UI8MMEBF-dANOnzaeuotfw [Accessed 8 May 2019].

The Predatory Bird Monitoring Scheme. (2019). The Predatory Bird Monitoring Scheme. [online] Available at: https://pbms.ceh.ac.uk [Accessed 10 May 2019].

The Shark Trust. (2019). Great Eggcase Hunt. [online] Available at: https://www.sharktrust.org/great-eggcase-hunt [Accessed 10 May 2019].

Trumbull DJ, Bonney R, Bascom K, and Cabrel A. 2000. Thinking scientifically during participation in a citizen-science project. Sci Educ-Netherlands 84: 265–75.

Trust, B. (2019). National Bat Monitoring Programme - Our Work - Bat Conservation Trust. [online] Bat Conservation Trust. Available at: https://www.bats.org.uk/our-work/national-bat-monitoring-programme [Accessed 10 May 2019].

Wand, Y. and Wang, R. Y. (1996) 'Anchoring data quality dimensions in ontological foundations', Communications of the ACM, 39(11), pp. 86–95. doi: 10.1145/240455.240479.

Wang, R. and Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), pp.5-33.

Welvaert, M. and Caley, P., 2016. Citizen surveillance for environmental monitoring: combining the efforts of citizen science and crowdsourcing in a quantitative data framework. SpringerPlus, 5(1), p.1890.

Wiggins, A. and He, Y. (2016). Community-based Data Validation Practices in Citizen Science. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16.

Wiggins, A., Newman, G., Stevenson, R. and Crowston, K. (2011). Mechanisms for Data Quality and Validation in Citizen Science. 2011 IEEE Seventh International Conference on e-Science Workshops.

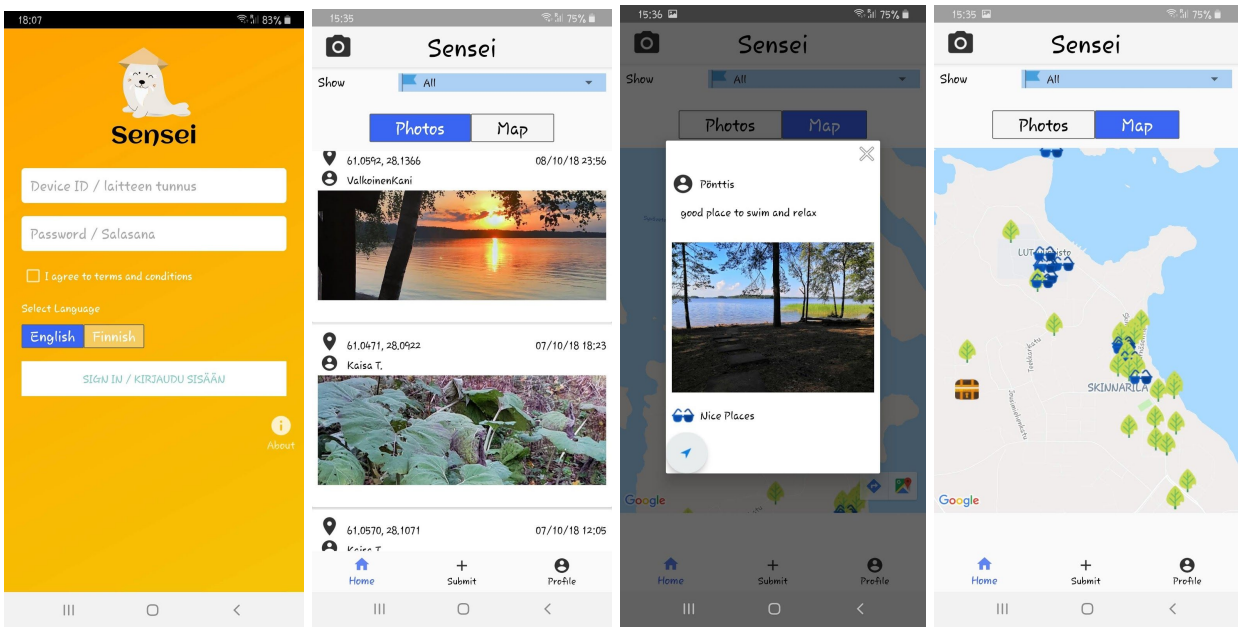
Wixom, B.H. and Watson, H.J. (2001). "An Empirical Investigation of the Factors Affecting Data Warehousing Success," MIS Quarterly, Vol.25, No.1, pp. 17-41.

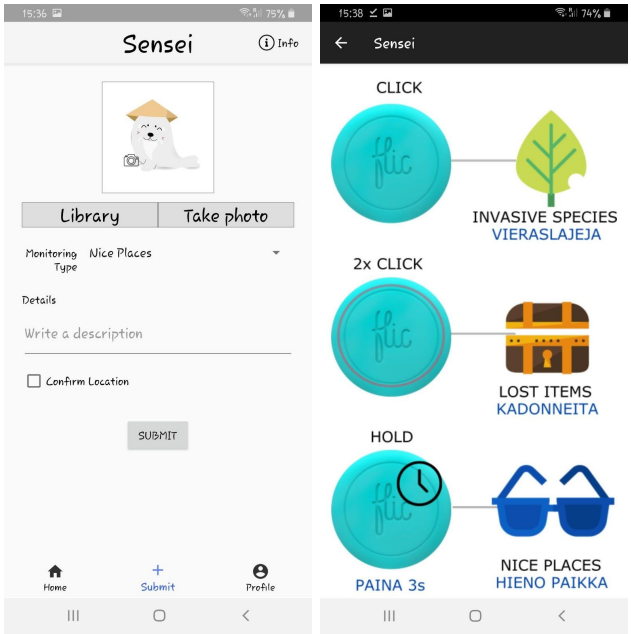
Zooniverse.org. (2019). [online] Available at: https://www.zooniverse.org/projects/markbasham/science-scribbler-virus-factory/about/research [Accessed 11 May 2019].

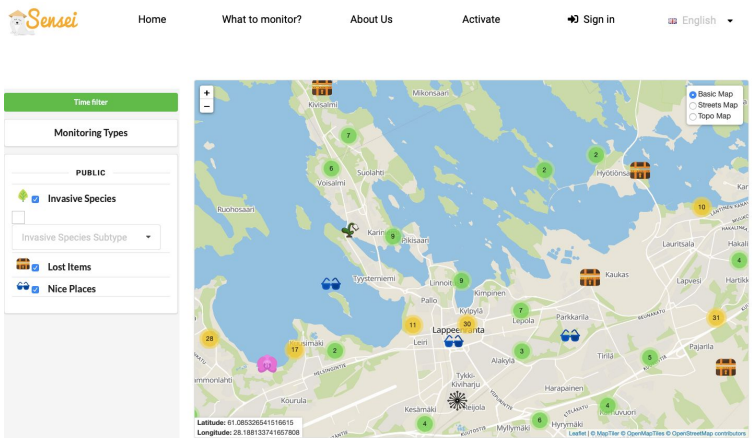
iRecord. (2019). iRecord. [online] Available at: https://www.brc.ac.uk/irecord/ [Accessed 10 May 2019].

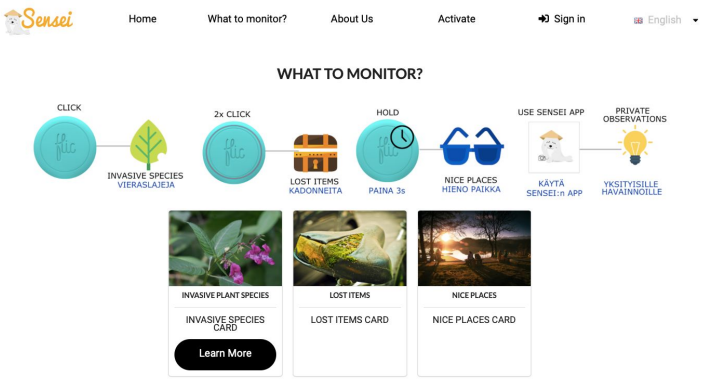
mailto:www-bgs@bgs.ac.uk, B. (2019). mySoil App | British Geological Survey (BGS). [online] Bgs.ac.uk. Available at: https://www.bgs.ac.uk/mysoil/ [Accessed 10 May 2019].

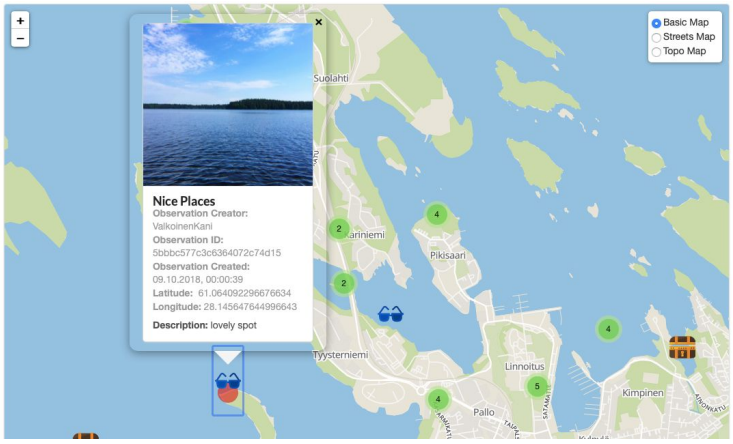
Appendix 1. Screenshot of the from the Sensei platform






Sensei
Home
What to monitor?
About Us
Activate
Sign in
English
Time filter
Monitoring Types
PUBLIC
Invasive Species
Invasive Species Subtype
Lost Items
Nice Places
Basic Map
Streets Map
Topo Map


Sensei
Home
What to monitor?
About Us
Activate
Sign in
English
WHAT TO MONITOR?
CLICK
INVASIVE SPECIES
VIERASLAJEJA
2x CLICK
LOST ITEMS
KADONNEITA
HOLD
PAINA 3s
NICE PLACES
HIENO PAIKKA
USE SENSEI APP
KÄYTÄ SENSEI:n APP
PRIVATE OBSERVATIONS
YKSITYISILLE HAVAINNOILLE
INVASIVE PLANT SPECIES
INVASIVE SPECIES CARD
Learn More
LOST ITEMS
LOST ITEMS CARD
NICE PLACES
NICE PLACES CARD


Basic Map
Streets Map
Topo Map
Nice Places
Observation Creator:
ValkoinenKani
Observation ID:
5bbbc577c3c6364072c74d15
Observation Created:
09.10.2018, 00:00:39
Latitude: 61.064092296676634
Longitude: 28.145647644996643
Description: lovely spot
Suolahti
Pikisaari
Tysterniemi
Linnoitus
Kimpinen
Pallo

Appendix 2. Interview 1 - Citizen science project developer

This team has worked on developing a citizen science project which involved over 80000 participants and ran over a span of 1 year. The main participant group was from North America. It was a web based platform involving a questionnaire.

Interview with developer in this project helped us understand some of the issues faced by his team and the solutions they found to solve those issues specific to citizen science projects. Some of the issues faced by the team were people dropping out, lack of geographical spread of participants, no clear idea of participants interests, problem with tracking the patterns, crash of website, resubmissions of observations.

To solve the above problems, they have released a few versions solving few problems in each version. As this is a website, the releases were easy without the participant requiring to install or update any applications.

To solve the issues of dropping out, they have introduced gamification and incentivization. Though there was dropout after introduction of gamification and incentivization, it was very low compared to before. Also, new participants queued up to participate. To track the patterns of participants, cookies and browser fingerprints were used. Regular updates were released to keep the website from crashing. To block the resubmissions on the website, submit button was blocked after it is clicked once and a notification was sent once the observation is submitted. Issue of geographical spread wasn't solved but data from alternate sources helped to minimize the effects to some extent. Apart from these, they have used the concepts of post-stratification analysis and giving weights with proportions for robustness checks.

Appendix 3. Interview 2 - Database Admins

An admin from IBM and one from cognizant were interviewed to identify different issues they face with the quality of data in their daily life and how do they solve those issues. Their expertise in the field helped us understand a few mechanisms which can be employed in citizen science projects.

There are multiple issues which are found in databases but only few relate to the citizen science projects. The main area where their inputs were used was after collection part of the framework.

At the end of the session, a view of framework was shown to them and they helped us add a few mechanisms which helped in finishing the framework.