
Jiri Musto

# IMPROVING THE QUALITY OF USER-GENERATED CONTENT

ACTA UNIVERSITATIS LAPPEENRANTAENSIS 1001

Jiri Musto

# IMPROVING THE QUALITY OF USER-GENERATED CONTENT

Dissertation for the degree of Doctor of Science (Technology) to be presented with due permission for public examination and criticism in the Auditorium 1316 at Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland on the 2nd of December, 2021, at noon.

Acta Universitatis
Lappeenrantaensis 1001

Supervisor Professor Ajantha Dahanayake
LUT School of Engineering Science
Lappeenranta-Lahti University of Technology LUT
Finland

Reviewers Professor, emeritus Bernhard Thalheim
Institute for Information Technology
Christian-Albrechts-Universität zu Kiel
Germany

Professor Boris Novikov
Department of Informatics
National Research University Higher School of Economics
Russia

Opponents Professor, emeritus Bernhard Thalheim
Institute for Information Technology
Christian-Albrechts-Universität zu Kiel
Germany

Professor Boris Novikov
Department of Informatics
National Research University Higher School of Economics
Russia

ISBN 978-952-335-757-0
ISBN 978-952-335-758-7 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491

Lappeenranta-Lahti University of Technology LUT
LUT University Press 2021

# Abstract

**Jiri Musto**
**Improving the quality of user-generated content**
Lappeenranta 2021
118 pages
Acta Universitatis Lappeenrantaensis 1001
Diss. Lappeenranta-Lahti University of Technology LUT
ISBN 978-952-335-757-0, ISBN 978-952-335-758-7 (PDF), ISSN-L 1456-4491, ISSN 1456-4491

User-generated content is a huge source of information in the modern world. The rise of social media and other user-generated content platforms has enabled the public to create and share an increasing amount of information with others. However, as people share the information with no credible background, the reliability of user-generated content is questionable, and the quality of data and information is uncertain.

This thesis aims to study the underlying issues that reduce user-generated content's data and information quality and presents a solution to improve the overall quality of content. The problems are surveyed through literature and examining existing user-generated content platforms. Most issues relate to using the public as the content provider and not having any proper design decisions to overcome data and information quality flaws. Additionally, the definitions of data and information quality used in existing research are imperfect for the domain of user-generated content, and there is a need for establishing definitions solely for user-generated content.

This research proposes new definitions for data and information quality and presents a platform design that considers the quality of content during design and development to improve the data and information quality of user-generated content. The design enhances the information collection and data curation processes to procure higher quality content from users.

The research has three significant contributions: (1) A comprehensive set of data and information quality characteristics for user-generated content, (2) extension of the development life cycle with data and information quality characteristics for user-generated content platforms, and (3) a framework that integrates quality characteristics into the design to store and assess the reliability and quality of user-generated content.

Keywords: user-generated content, data quality, information quality, quality characteristics

# Acknowledgments

From the beginning of my doctoral studies to the finish line, the journey has been long and full of ups and downs. The emotions I have felt range from desperation to happiness, from disappointment to joy, but while the journey has been rough, I am glad to have traversed it. The last four years have been a valuable experience that I will remember for the rest of my life.

I want to express my gratitude to my supervisor Professor Ajantha Dahanayake for allowing me to undertake the journey and support me in completing my doctoral degree. Your feedback and experience helped me in my research. Without you, I probably would not have even considered postgraduate education.

I thank my reviewers and opponents Professor Boris Novikov and Professor, emeritus Bernhard Thalheim for helping me improve the dissertation. Additionally, I am grateful to Professor, emeritus Thalheim for hosting my two-month visit to Kiel during my studies.

I want to thank my co-workers and fellow doctoral students for their support and answering all the questions I had during my studies. Special thanks to the doctoral school and department secretaries, especially Tarja Nikkinen, for their continuous assistance.

Finally, I wish to thank my family and friends for supporting my academic endeavor and special thanks to my wife, Jaana.

Jiri Musto
December 2021
Lappeenranta, Finland

# Contents

# List of publications

This dissertation is based on the following publications. The rights have been granted by publishers to include the papers in the dissertation.

I. Musto, J. and Dahanayake, A. (2018). Overview of Data Storing Techniques in Citizen Science Applications. In: Benczúr A. et al. (eds) New Trends in Databases and Information Systems. ADBIS 2018. *Communications in Computer and Information Science*, 909.

II. Musto, J. and Dahanayake, A. (2020). Improving Data Quality, Privacy and Provenance in Citizen Science Applications. *Frontiers of Artificial Intelligence and Applications*, 321, pp. 141-160.

III. Musto, J. and Dahanayake, A. (2021). Quality characteristics for user-generated content. *Frontiers of Artificial Intelligence and Applications*. Accepted 2021

IV. Musto, J. and Dahanayake, A. (2021). An approach to improve the quality of user-generated content of citizen science platforms. *ISPRS International Journal of Geo-Information,* 10, pp. 434.

## Author's contribution

Jiri Musto is the main author and investigator in papers I – IV under the supervision of Professor Ajantha Dahanayake. In papers I and II, Musto carried out the corresponding literature and platform reviews after discussions with Prof. Dahanayke. For papers III and IV, Musto designed the citizen science platform and executed the relevant data analysis with feedback from Prof. Dahanayake. Additionally, Musto presented the paper I at the relevant conference.

# Nomenclature

### Abbreviations

| | |
|---|---|
| ALA | Atlas of Living Australia |
| DSR | Design science research |
| ERP | Enterprise resource planning |
| ISO | International Organization for Standardization |
| RSQ | Research sub-question |
| UGC | User-generated content |

# 1 Introduction

Much of the data and information in the modern world is being created by the public through various platforms such as Wikipedia (Wikipedia, 2020), OpenStreetMap (OpenStreetMap, 2021), Facebook (Facebook, 2021), Twitter (Twitter, 2020), Instagram (Instagram, 2021), YouTube (YouTube, 2020), Worldometer (Worldometer, 2020), and iNaturalist (iNaturalist, 2021), to name a few. Platforms where users provide content are called user-generated content (UGC) platforms.

Social media data is being rapidly generated as more than half of the world's population uses social media platforms (Influencer Marketing Hub, 2020; Smart Insights, 2020). For example, Facebook generates four petabytes of data daily, and on average, a third of users' time online is used on social media platforms. It is estimated that users generate 60% of the total amount of data on the internet, and 40% is machine-generated (TechJury, 2020).

The data and information from UGC platforms can be used in healthcare studies (Bordogna *et al.*, 2016), wildlife research (Bayraktarov *et al.*, 2019), customer behavior research (Cai and Zhu, 2015), flood monitoring (Arthur *et al.*, 2018), emergency reporting (Ludwig, Reuter and Pipek, 2015), business influencing (Vincent *et al.*, 2019; Brunt, King and King, 2020), future prediction (Asur and Huberman, 2010), and targeting advertisements and recommendations for potential customers (Ouyang, Li and Li, 2016; Mensah *et al.*, 2020). Utilizing UGC can create value for companies through savings, analytics, and marketing (TINT, 2020).

UGC covers a wide range of subdomains, such as social media, crowdsourcing, and citizen science. Each subdomain includes users in content generation in various forms, and thus, they fall under the general domain of UGC (See *et al.*, 2016). The platforms have a variety of uses for the generated content. For example, Wikipedia, OpenStreetMap, iNaturalist, and Worldometer gather content to share credible and relevant information with other people. On the other hand, Facebook, Twitter, Instagram, and YouTube are meant for sharing subjective thoughts and opinions with other users.

UGC platforms often struggle with data and information quality, and using low-quality data makes the conclusions and results debatable (Leibovici *et al.*, 2017; Xiaojiang, Liwei and Jianbin, 2017; Lansley and Cheshire, 2018). There are instances when content generated by users is unverified or misleading (Syed-Abdul *et al.*, 2013; Goodman and Carmichael, 2020). As a result, Wikipedia is not considered a valid scientific source (Polk, Johnston and Evers, 2015; Wikipedia, 2019), and the public uses OpenStreetMap less than Google Maps (Mooney *et al.*, 2012). Low data quality costs up to $3.1 trillion to the US economy yearly (IBM, 2019).

Data quality has been an ongoing research topic for decades. Some of the most cited works establish the following basic principles of data quality (Redman, 1996; Wang and Strong, 1996; Batini and Scannapieco, 2006, 2016):

- Data quality is multidimensional, consisting of individual characteristics: These characteristics are, for example, accuracy, completeness, credibility, precision, and understandability.
- Data quality characteristics can be grouped into categories*:* Wang and Strong (1996) categorize characteristics into intrinsic, contextual, representational, and accessibility based on what they affect. International Organization for Standardization (ISO) (2008) categorizes characteristics into inherent, inherent and system-dependent, and system-dependent characteristics based on what affects them.
- Data quality is contextual*:* Different domains or contexts require different collections and definitions of data quality characteristics.
- Each characteristic's importance is subjective*:* Each case selects specific characteristics for their definition of data quality. The cases place importance on different characteristics depending on their views, needs, and opinions.
- Data quality is measured through characteristics*:* As data quality comprises individual characteristics, each chosen characteristic must be measured separately to determine a total for data quality.

Most data quality concerns in UGC are related to the fact that people who provide the content are amateurs. There are some ways to alleviate these concerns in UGC, such as:

- Using sensors for collecting data. However, it can be argued that when sensors provide the content, it can no longer be considered UGC as users are no longer part of the content-providing process after they place the sensors (Foody *et al.*, 2015).
- Training the users. Training is possible in subdomains where users are specifically selected, but it can be expensive and not applicable to other projects (Ratnieks *et al.*, 2016).
- Cleaning and filtering data. When social media data is used for analysis, most of the data is cleaned and filtered to increase quality. The original content remains widely unaffected and low quality (Garcia *et al.*, 2017; Leibovici *et al.*, 2017).

Information and data are usually treated as the same concept. However, the two are not interchangeable. Data is quantifiable and measurable without the intent of use, but information requires external perception to be seen as information. Data can be transformed into information through analysis or by giving it a context, and data can be extracted from information (Davenport and Prusak, 2000). In UGC, users provide information, and data is extracted from the information and stored in the platform database. Improving information quality will lead to improved data quality in UGC platforms.

The above-listed points, the wide usage of UGC, and the lack of pertinent data and information quality research have motivated this research to provide more flexible and adaptive approaches for improving data and information quality.

## 1.1 Research questions

This dissertation's primary research question is: *How can the quality of user-generated content be improved by enhancing information collection and data curation processes in user-generated content platforms?*

The following research sub-questions (RSQ) help to answer the main research question:

- RSQ1: *What information collection features in user-generated content platforms influence the quality of content?*
- RSQ2: *How to define quality characteristics and distinguish data and information quality in the domain of user-generated content?*
- RSQ3: *How does the introduction of data and information quality characteristics into information collection and data curation processes influence the quality of user-generated content?*

Table 1 presents how the publications presented in Section 4 relate to the RSQs.

**Table 1.** The relation of research sub-questions and publications

| Research sub-questions | Publication number | Publication channel |
|---|---|---|
| RSQ1 | Publication I | ADBIS 2018, Communications in Computer and Information Science (Musto and Dahanayake, 2018) |
| | Publication II | Frontiers of artificial intelligence (Musto and Dahanayake, 2020) |
| RSQ2 | Publication III | Frontiers of artificial intelligence (Musto and Dahanayake, 2021b) |
| RSQ3 | Publication IV | ISPRS International Journal of Geo-Information (Musto and Dahanayake, 2021a) |

(Musto and Dahanayake, 2018) is a literature review on the current state of the UGC domain with particular attention to the citizen science field, and (Musto and Dahanayake, 2020) extends the literature review by examining citizen science platforms in the field. (Musto and Dahanayake, 2021b) and (Musto and Dahanayake, 2021a) present and evaluate the designed artifacts to resolve the main issue: the quality of user-generated content found in (Musto and Dahanayake, 2018) and (Musto and Dahanayake, 2020).

## 1.2 Methodology

This research follows the design science research guidelines developed by Hevner et al. (2004) by developing an artifact to solve a relevant problem in the research field while using the existing body of knowledge to arrive at an innovative solution. In the end, the

artifact is validated for its relevance in the application domain. It extends the existing knowledge base with the new knowledge formulated for problem-solving in the environment of the research field. This process is illustrated in Figure 1.

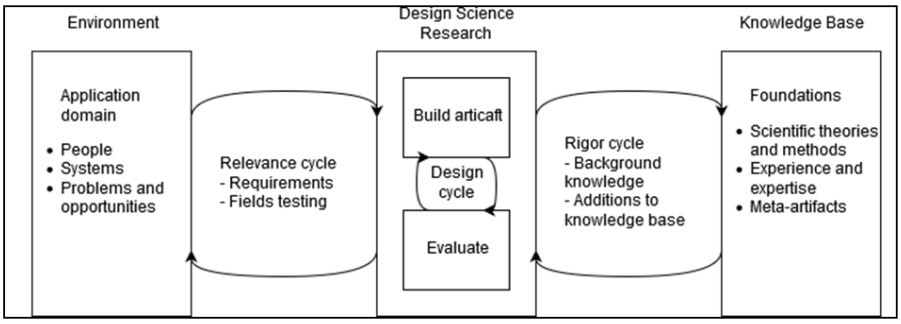


**Figure 1**: Design science research development cycle

(Musto and Dahanayake, 2018) is part of the rigor cycle by building background knowledge on the issues of UGC. (Musto and Dahanayake, 2020) relates to the relevance cycle by combining background knowledge with practical problems from the UGC domain. (Musto and Dahanayake, 2021b) and (Musto and Dahanayake, 2021a) present the design cycle by building and evaluating the artifacts.

## 1.3 Contributions and limitations

Although most of the papers in this dissertation are related to citizen science, the following arguments can be made: UGC is a general term for content created by people. The different sub-domains have many common properties that describe how and what content users create and share. The common properties include: content is provided by regular citizens, content is mostly text with a picture attached, content includes a time and location, and content is reviewed by other users (Krumm, Davies and Narayanaswami, 2008; See *et al.*, 2016). The case of citizen science and its relevant issues are generalized because of the similarities and issues concerning the concept of UGC. There are three significant scientific contributions made in this dissertation:

1. A comprehensive set of data and information quality characteristics defined for UGC.

Data and information quality are highly contextual, and using definitions meant for a different domain can lead to conflicting quality evaluations. As there is a lack of data and information quality definitions in the UGC domain, this research fills the gap by providing proper definitions specifically for the UGC domain. The definitions are developed based on well-established principles from existing research.

2. Extension of the UGC platform's development life cycle with UGC data and information quality characteristics during the platform's requirements acquisition stage to improve the quality of the content collection.

The design and collection processes significantly impact the resulting data and information quality (Wand and Wang, 1996). Integrating quality characteristics into the development life cycle helps designers appropriately consider the chosen quality characteristics and reduce poor design choices to develop superior collection processes.

3. Framework to store and assess the reliability and quality of UGC using quality characteristics.

There is a lack of practical methods for improving the quality of data and information in UGC (Lukyanenko, Parsons and Wiersma, 2016; Ratnieks *et al.*, 2016; Tenkanen *et al.*, 2017; Arolfo and Vaisman, 2018; Ahmouda, Hochmair and Cvetojevic, 2019). Evaluating the quality of acquired content and storing the results enables platform owners and content users to assess the quality of existing content and decide whether the reliability is satisfactory.

When UGC data and information quality are improved, UGC usage and overall utility are increased. This research contributes to society by presenting concepts and practical approaches to improve the data and information quality of UGC. The contributions assist designers and developers of UGC platforms by providing a design process that increases the data and information quality of UGC platforms. Utilizing quality characteristics and assessing them during content acquisition helps platform owners and content users to evaluate the reliability of UGC.

## 1.4 Structure

The dissertation has the following structure: Chapter 2 reviews the background work on data and information quality and UGC. The chapter consists of data quality and information quality, UGC, and receiving reliable content from users.

Chapter 3 presents different research methods and introduces the design science research paradigm and specific research methods used during the research.

Chapter 4 presents the overview of the publications.

Chapter 5 establishes the scientific contributions.

Chapter 6 concludes the dissertation.

# 2 Background

The number of published scientific articles featuring UGC has grown steadily during the last two decades. Figure 2 shows the number of research papers per year from 1991 onwards available in Scopus. Between 1961-1991, only a handful of articles on social media, big data, or crowdsourcing have appeared. Information quality articles amount to less than a hundred, and data quality research papers total almost a thousand.

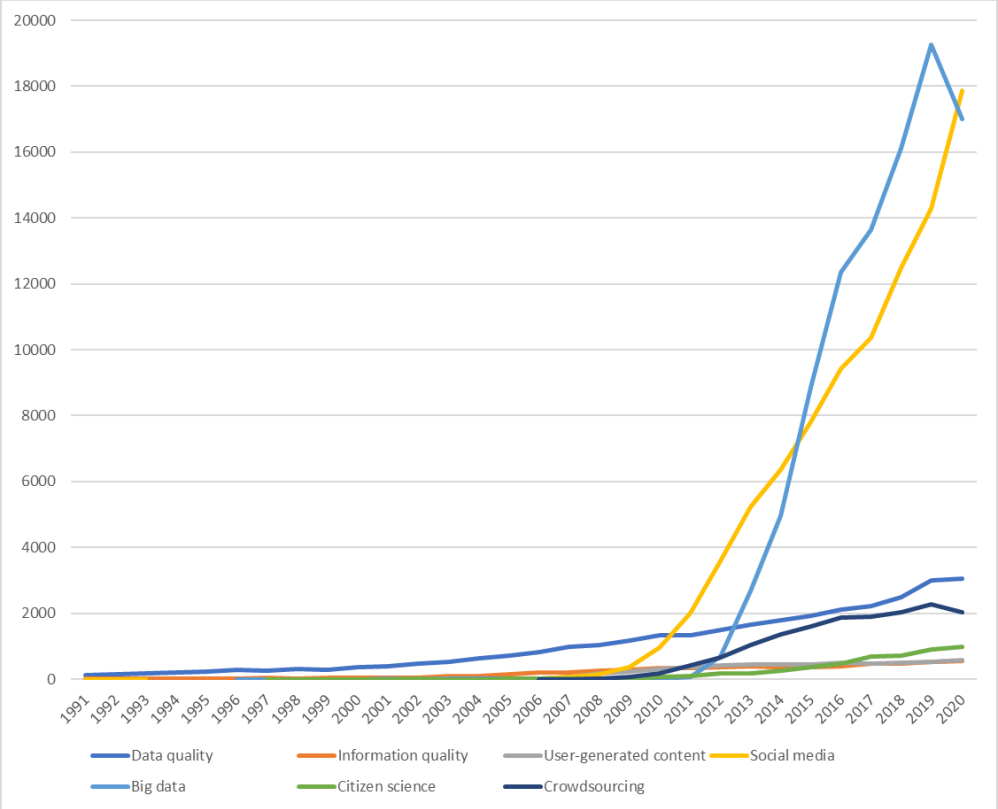


**Figure 2.** The number of articles available in Scopus based on topic-abstract-title search

The number of data and information quality articles has steadily increased during the last three decades. The first appearance of "citizen science" is in 1997, and "UGC" around 2001, while "crowdsourcing" appeared in 2006. Interest in social media increased when Facebook became public access (2006-2007), and a couple of years later in 2011, big data took off. Many publications treat social media content as big data, which explains why they have a similar growth pattern in Figure 2.

## 2.1 Data and information quality

### 2.1.1 Data quality

Wang et al. (1995), Wang and Strong (1996), Wand and Wang (1996), and Strong et al. (1997) have created the foundations for the current data quality research. Wang et al. (1995) develop a framework for analyzing data quality research issues in an organizational context with seven elements: management responsibilities, research and development, production, distribution, operation and assurance costs, personnel management, and legal function. Using the framework, Wang et al. analyze the existing data quality literature and find a need for techniques, metrics, and quality policies to improve data quality. Additionally, they suggest that the link between poor data quality and problem detection procedures needs to be studied.

To develop strategies to enhance data quality, Wang and Strong (1996) survey important data quality characteristics for organizations. Survey responses result in over fifty different quality characteristics. However, the surveyed quality characteristics tend to overlap, and some are deemed less valuable. Based on additional surveys, the fifty characteristics are reduced to fifteen characteristics for data quality: believability, accuracy, objectivity, reputation, value-added, relevancy, timeliness, completeness, appropriate amount of data, interpretability, ease of understanding, representational consistency, concise representation, accessibility, and access security. The characteristics are organized into four categories:

- Intrinsic: Characteristics that affect the quality of data regardless of how data is used.
- Contextual: Characteristics that depend on the purpose of the data for the task at hand.
- Representational: Characteristics related to the format and meaning of data.
- Accessibility: Characteristics that relate to how data can be accessed, used or retrieved.

Wand and Wang (1996) tie completeness, unambiguousness, meaningfulness, and correctness to ontological foundations:. The four generic characteristics are derived from the fifteen data quality characteristics provided by Wang and Strong (1996). Using ontologies, Wand and Wang (1996) provide general guidance on how the characteristics relate to design and production processes, generic reasons for deficiencies, and how to repair them. For example, incomplete data results from loss of information and possible reasons for losing the information are missing states in the information system. The missing states should be repaired by allowing missing cases.

Redman (1996) establishes an alternative way to define data quality characteristics from Wang and Strong (1996). Wang and Strong explore data quality characteristics from a

business employee's perspective, while Redman defines quality from the system's perspective. For example, according to Wang and Strong, accuracy is: "*The extent to which data are correct, reliable, and certified free of error*," while Redman defines accuracy as a measure of proximity of data values v and v'.

Data quality characteristics and their definitions have been examined in the domains of big data (Batini *et al.*, 2015; Firmani *et al.*, 2016) and remote sensing (Batini *et al.*, 2017; Albrecht *et al.*, 2018; Barsi *et al.*, 2019). Each domain brings new challenges to data quality because of the differences in content and usage, and the definitions of the same characteristics vary across domains. For example, accuracy (correctness) of content in big data changes to positional accuracy when tied with locational data.

### 2.1.2 Information quality

Data and information are sometimes conflated. Although there is a relationship between data and information, they have some differences. First of all, information quality is more contextual than data quality (Watts, Shankaranarayanan and Even, 2009). Data is a separate object that can be quantified, but information needs additional knowledge to be presented as information. Much like data quality, information quality is evaluated based on specific characteristics, and the application defines what characteristics are relevant (Bovee, Srivastava and Mak, 2003).

When comparing the data and information quality, there is a significant difference in the relevant quality characteristics. Batini and Scannapieco (2006, 2016) and Wang and Strong (1996) define more than ten characteristics for data quality, while information quality definitions frequently have less than ten. Nicolaou and McKnight (2006) define information quality as currency, accuracy, completeness, relevance, and reliability. Similarly, Nelson et al. (2005) employ accuracy, completeness, currentness, and format as key information quality characteristics. In a model for judging information quality and cognitive authority in web systems, information quality refers to accuracy, goodness, currentness, usefulness, and importance (Rieh, 2002). The most repeating characteristics for information quality are accuracy, currentness, and completeness.

Lee et al. (2002) develop the AIM quality (AIMQ) methodology to assess organizational information quality using data quality characteristics found by Wang and Strong (1996). AIMQ is a questionnaire that collects data on corporate information quality and measures the overall quality of information within the corporate systems. AIMQ is designed to be a practical tool for organizations to investigate, identify problems, and monitor improvements in the organization's information quality.

Information quality is relevant because it affects the intent to use systems. DeLone and McLean (1992) create an information system success model based on system quality and information quality. The updated version (DeLone and McLean, 2003) shown in Figure 3 adds service quality into the model. The system quality in the model refers to the ease

of use, flexibility, reliability, and ease of learning. In the model, information quality is defined as completeness, accuracy, understandability, usability, and timeliness.

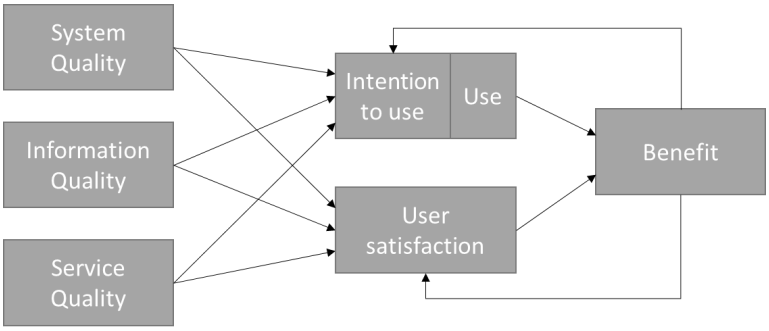


**Figure 3**. Information system success model (DeLone and McLean, 2003)

Using the information system success model, Petters et al. (2013) investigate the specific determinants for information system's success. IT infrastructure, management processes and support, IT planning, trust, competence, and motivation are key determinants that affect information quality. Half of the determinants are specifically related to users and those responsible for the information. This means, that users who create or manage information have the most impact on the success of an information system. Information quality positively impacts user satisfaction, with numerous studies supporting this claim. Information quality can benefit individuals' or organizations' success by increasing productivity and efficiency and improving decision-making (Beebe and Walz, 2005).

### 2.1.3 Traditional content

Corporate data is often described as "traditional" content in comparison to modern web-based content. Traditional content has many specific traits that separate it from web-based content (Trujillo *et al.*, 2015):

- Traditional content is well structured and stored in relational databases.
- The content comes from known sources, such as verified people, that can be considered reliable.
- Content is obtained through platforms or machines made explicitly for acquiring fixed information with accurately defined schemas to minimize non-related content.
- The content is collected with a specific purpose in mind, but the content can be easily used for other purposes.
- The content is reviewed and verified by machines or specifically chosen people.

There are many guides to managing the quality of traditional content compared to managing web-based content. The quality management process is relatively simple

because the content is produced, used, reviewed and maintained within the organization, and many of these steps can be automated (Batini *et al.*, 2006, 2009).

### 2.1.4 Web-based content

Content from the internet can be called web-based content. Web-based content is often unstructured or semi-structured and the content source may only be known at a general level, especially if anonymous internet users provide the content. Web-based content typically does not have a specific purpose other than to be shared with other internet users. There are exceptions, such as citizen science content. If the web-based content is used for other than the original purpose, utilizing the content becomes more complex (Trujillo *et al.*, 2015).

The web-based content is reviewed by other users or by the platform's owner, depending on the platform's purpose. In most cases, the platform acts as a content hub, making others responsible for reviewing the content. Managing the quality of web-based content is challenging for multiple reasons. First, the content is primarily produced outside of the organization by other individuals. Second, mainly outsiders use and review web-based content, and the platform only serves as a hub to store and maintain it (Varlamis, 2010).

### 2.1.5 Quality in practice

In a traditional content scenario, a corporation owns the platform where the content is created, stored, reviewed, and maintained. The content is created and collected within the corporation by machines or employees and the corporation manages the quality as well as the usage of the content. The content can be used within the corporation or given to others. The platform owner, content creator and content user are often the same entity in traditional content.

Several barriers inhibit data quality in organizations, including lack of measurements, training, policies and procedures (Haug and Arlbjørn, 2011). There are multiple data quality assessment and improvement methodologies for use in organizations. The methods fall into two distinct strategies, *data-driven* and *process-driven. Data-driven* approaches improve data quality by gathering new data to replace low-quality data, selecting credible sources, and correcting errors. *Process-driven* techniques improve data collection and analysis processes by controlling or redesigning the collection process and eliminating low-quality data sources (Batini *et al.*, 2009).

On the other hand, in a web-based content scenario, an organization owns the platform where the content is managed, but the content is not created or used by the corporation. Individuals are responsible for creating the content, and third parties use the content. The platform owners are mainly responsible for collecting and storing the content, while outsiders create, review, and maintain it. The underlying issue in managing data quality in web-based content is the separation of responsibilities because the responsibility is

either on the platform owner, the content creator, or the content user (Al Sohibani *et al.*, 2015; Mihindukulasooriya *et al.*, 2015).

Comparing traditional relational database data quality characteristics and geographic information systems' quality characteristics to remote sensing shows that many quality characteristics from traditional content are challenging to transfer directly to remote sensing. For example, completeness is not helpful because remote sensing data is arguably never complete. Remote sensing requires different quality characteristics specific to its usage (Batini *et al.*, 2017).

To improve quality in remote sensing, Albrecht et al. (2018) present the lifecycle of remote sensing data, consisting of four specific phases where each stage includes quality checks. The steps are divided into data acquisition, storage, processing and analysis, and visualization and delivery. The lifecycle is further developed by investigating how different data quality characteristics relate to the specific phases in the lifecycle. Each phase emphasizes different data quality characteristics. For example, during data acquisition, *resolution*, *accessibility* and *spatial accuracy* are important (Barsi *et al.*, 2019).

### 2.1.6 Challenges in quality management in the internet age

1. There are several frameworks but a lack of practical instructions.

Several data and information quality frameworks, processes, and models have been developed (Stang *et al.*, 2008; Mehmood, Cherfi and Comyn-Wattiau, 2009; Tian *et al.*, 2012; Smith *et al.*, 2018; Ayuning Budi *et al.*, 2019). There are frameworks for medical data (Arts, De Keizer and Scheffer, 2002), social media (Tilly *et al.*, 2017), big data (Ge and Dohnal, 2018), remote sensing (Barsi *et al.*, 2019), and healthcare (Bai, Meredith and Burstein, 2018). Many of these list specific steps or items that should be considered when dealing with data quality, but they all share some fundamental issues that make utilization difficult. These issues include having an overly general process or framework that instructs one to "define data quality characteristics" or "perform automatic checks" without practical instructions on how these steps are conducted (Wang, Storey and Firth, 1995; Haug *et al.*, 2013; Hashem *et al.*, 2015). There are examples of how different data quality characteristics are to be defined (Wang and Strong, 1996; ISO, 2008) or measured (Lee *et al.*, 2002; Batini and Scannapieco, 2016), or what techniques can improve data quality (Batini *et al.*, 2009). Even so, these are relatively limited, applied only to specific domains, and require further testing.

2. Many different definitions for the same quality characteristics.

Researchers have developed multiple definitions for data and information quality characteristics (Redman, 1996; Wang and Strong, 1996; Batini and Scannapieco, 2006; ISO, 2008). However, there is no clear consensus on what different characteristics mean

or which ones are essential. In some cases, the definitions of the characteristics are the same, but the characteristic itself is given another term, e.g., believability vs. credibility.

3. Data and information quality characteristics are not universal.

The quality characteristics must be selected and defined for each scenario (Bovee, Srivastava and Mak, 2003; Caballero *et al.*, 2009; Han, Jiang and Ding, 2009). Although there are many definitions for general data and information quality in traditional content, there is a considerable shortcoming in the definitions of data and information quality in the UGC domain.

4. Misunderstanding data and information.

Data and information are often interrelated (Wang and Strong, 1996; Lee *et al.*, 2002; Nelson, Todd and Wixom, 2005), creating confusion amongst readers. Data and information are two different things that have a relationship (Davenport and Prusak, 2000). Data is transformed into information, and information is derived from data, but they require separate definitions and quality characteristics.

5. Differences between traditional content and web-based content.

Traditional content is well structured and produced at a stable rate with known amounts. On the other hand, web-based content is unstructured or semi-structured content generated at irregular rates and amounts, making the storage, review and management of web-based content more unpredictable and complicated (Trujillo *et al.*, 2015).

6. Content acquisition issues.

Traditional content is well-documented and acquired through specified means. The content is produced in a monitored environment by observable sources, which reduces inconsistency, redundancy, incompleteness, and incorrectness compared to web-based content. Web-based content is collected in a constantly changing environment, increasing the amount of redundant and inconsistent content. The number of unknown sources and uncertain origins for content increases the incompleteness and reduces web-based content's correctness and reliability (Varlamis, 2010; Clarke, 2016; Bayona Oré and Palomino Guerrero, 2018).

7. Division of responsibilities between platform owner, content creator, and content user in web-based content.

In traditional content, one entity is responsible for owning the platform as well as creating and using the content. The same entity manages the quality of content and only holds responsibility towards itself, making quality management easy. On the other hand, in web-based content, the platform owner is generally not the content creator nor the content user. Managing content quality is complex with three different entities involved, and the

responsibility may fall on anyone (Varlamis, 2010). The following are examples in web-based content of different entities being responsible for the quality of content:

- In citizen science, the platform owner is responsible
- In Wikipedia, the content creator is responsible
- In social media, the content user is responsible

## 2.2 User-generated content

### 2.2.1 What is user-generated content

UGC has a long history, but the term itself has only been used since the early 2000s (Krumm, Davies and Narayanaswami, 2008; Wyrwoll, 2014). In simple terms, UGC is content created on online platforms by users. Various categories of platforms fall under UGC. These platforms include but are not restricted to:

- Social media
- Citizen science
- Crowdsourcing
- Volunteered geographic information
- Collaborative mapping
- Participatory sensing
- Blogs
- Web pages
- Podcasts
- Reviews

Most UGC research revolves around social media, but some research concerns citizen science, volunteered geographic information, and participatory sensing. In these platforms, amateurs share content for research purposes.

In 2005, Amazon launched the Mechanical Turk crowdsourcing platform, where anyone could recruit labor for data collection. Research results from Amazon Mechanical Turk have revealed that highly reputable users are more likely to provide high-quality data (Peer *et al.*, 2017). Using Mechanical Turk could be considered part of UGC because of its crowdsourcing nature, but determining who is employed can pose a challenge. Experts in the field are likely hired for gathering the content. Additionally, the workers of Mechanical Turk are paid for their contributions, and this monetization scheme is entirely different from other UGC platforms. Mixed research results suggest that the compensation amount may impact the data quality (Buhrmester, Kwang and Gosling, 2011; Litman, Robinson and Rosenzweig, 2015).

Social media has been the subject of research for a long time. Social media platforms are designed for users to connect and share their thoughts with others locally or globally. Social media platforms can be mapped into a matrix based on social presence and self-presentation. Blogs are considered to have high self-presentation but low social presence. In comparison, Facebook has a medium-level social presence. There is some confusion about what social media is, and there is no clear consensus in all cases. For example, sometimes Wikipedia is defined as social media (Kaplan and Haenlein, 2010). There are seven building blocks of social media: sharing, presence, relationships, reputation, groups, conversations, and identity. These blocks are used for defining, classifying, and differentiating social media platforms as well as analyzing and monitoring social media platforms to understand their function and impact (Kietzmann *et al.*, 2011).

In addition to social media, citizen science has become a popular research topic in the 21st century. Citizen science is a field where citizens collect or classify data for research purposes (Elbroch *et al.*, 2011; Lukyanenko, Parsons and Wiersma, 2011; MacKechnie *et al.*, 2011; Hecht and Spicer Rice, 2015; Wiggins and Crowston, 2015). Compared to social media, citizen science platforms are designed for specific content collection purposes. The content is designed to be used for research and establishing facts, thus making data quality more essential for citizen science than social media (Hunter, Alabri and Van Ingen, 2013; Sheppard, Wiggins and Terveen, 2014; Hyder *et al.*, 2015; Lukyanenko, Parsons and Wiersma, 2016; Fritz, Fonte and See, 2017).

In general, UGC follows the principles of web-based content:

- Content is unstructured or semi-structured
- Platform owners do not generate the content
- The responsibility of ensuring quality is ambiguous
- Content is generated at erratic rates

The purpose of the UGC platform affects who is responsible for the quality of content and how it is managed. In crowdsourcing, the responsibility of quality is on the content provider and user, while the platform owner manages the content. On the other hand, in citizen science, the platform owner is responsible for the quality and managing of the content. In social media, the content provider manages the content while the content user is responsible for quality. In social media, the platform owner is typically not responsible for the quality or managing the content outside of enforcing terms of usage (Varlamis, 2010; Clarke, 2016; Bayona Oré and Palomino Guerrero, 2018).

### 2.2.2 Utilizing user-generated content

On UGC platforms that collect videos, such as YouTube, popular videos are more likely to be duplicated and uploaded illegally (Cha *et al.*, 2009). Popular videos integrate the most popular topics and the design of the platforms affects how many views less popular videos gain. For example, the employment of information filtering reduces the number of

views for less popular videos (Cha *et al.*, 2008). Additionally, the content creator's network affects how popular the video will become based on its age. The older the video is, the more impact the social network has (Susarla, Oh and Tan, 2012). This means that popular videos are highly susceptible to losing views because of duplication and with a large network, less popular videos can be successful on these platforms without the loss of views.

The essential qualities of volunteered geographic information and the differences compared to traditional geographic information are reviewed in (Elwood, Goodchild and Sui, 2012). The results show that volunteered geographic information data could complement the professionally gathered data and give new insights and a broader perspective. Similarly, there are advantages to flexible and fast data collection using OpenStreetMap, but issues in heterogeneous data limit the actual usage (Girres and Touya, 2010).

UGC platforms can provide value to search engines, such as Bing and Google. When using search engines, Wikipedia articles appear as one of the top results in most cases, providing massive value to the owning organization. Another UGC platform that proves valuable for the Google search engine is Twitter, particularly when making *trending* or *most popular* queries (Vincent *et al.*, 2019).

### 2.2.3 User-generated content's influence on businesses

An important question regarding UGC is how businesses could utilize it and how UGC affects consumerism (King, Racherla and Bush, 2014). Customer reviews are a form of UGC that can influence businesses. Positive reviews increase hotel room reservations (Ye *et al.*, 2011), and positive or negative UGC in large amounts impact businesses (Tirunillai and Tellis, 2012). On the other hand, there is no conclusive evidence for a similar effect on music sales (Dhar and Chang, 2009).

The possible usage of customer reviews and UGC are investigated in (Ghose, Ipeirotis and Li, 2012) and (Akehurst, 2009). Ghose et al. (2012) experiment with how UGC could be mined and utilized for ranking hotels. Using customer reviews, hotels could be classified by their utility or the best value for money. These classification techniques are used with search engines or platforms that provide hotel booking services. According to Akehurst (2009), mining relevant blogs and linking them to tourism websites increases the number of tourists as they are more likely to trust organizations with proper reviews from actual users. However, problems associated with content mining need to be solved when such a system is developed.

In summary, credible UGC with high information quality positively influences the perceived trust and further usage of a service and content. In turn, it affects word-of-mouth and recommendations (Ayeh, Au and Law, 2013; Filieri, Alguezaui and McLeay, 2015).

UGC has led the production and consumption implosion, leading to capitalism that increases the number of UGC platforms and UGC utilization among various businesses (van Dijck, 2009; Ritzer and Jurgenson, 2010). UGC is considered more effective in influencing consumer behavior than traditional marketing because it is more personalized and directed (Goh, Heng and Lin, 2013).

### 2.2.4 Shortcomings

Much of the existing research in UGC raises data and information quality issues. There are benefits to using UGC (Asur and Huberman, 2010; Tirunillai and Tellis, 2012), but utilizing it without considering the quality of data and information may lead to false results (Becker, King and McMullen, 2015; Jesmeen *et al.*, 2018).

## 2.3 Receiving reliable content from users

### 2.3.1 Issues and challenges

Data quality in UGC is often criticized, and quality is low overall compared to other domains (Brown and Kyttä, 2014; Sadiq and Indulska, 2017; Kaur *et al.*, 2018; Nkonyana and Twala, 2018; Bayraktarov *et al.*, 2019). Users providing the content are considered amateurs who provide untrustworthy or false data (Zhao and Sui, 2017; Haworth *et al.*, 2018; Abdullah-All-Tanvir *et al.*, 2019). Therefore, data and information quality improvement research and methodologies in the UGC domain are exceptionally vital.

Cai and Zhu (2015) survey the data quality challenges in the domain of big data. The biggest challenges with data quality in big data are diversity, volume, rapid change, and no unified data quality standards. Other data quality challenges include context-dependency, subjectivity, quantity, trust, location, aggregation, and distribution (Ludwig, Reuter and Pipek, 2015). Artificial intelligence has been proposed as a tool to evaluate information or implement user filtering to increase the quality (Haralabopoulos, Anagnostopoulos and Zeadally, 2016).

Lukyanenko et al. (2014) denote the issue of using traditional information quality definitions in the UGC domain. Traditional quality focuses on corporate data and information provided with strict rules and restrictions but using similar rules in the UGC context would restrict users too much. Inflexible systems may discourage users from participating and lead to information loss while also preventing the detection of new and undiscovered information. Trying to hold citizens to researchers' standards will only lead to problems with accumulating content (Lukyanenko, Parsons and Wiersma, 2016). Additionally, developers should consider the platform's unlikely uses, such as citizens observing different phenomena than intended (Lukyanenko, Parsons and Wiersma, 2014).

### 2.3.2 Improving data and information quality in user-generated content

Various techniques for improving data and information quality in UGC are available. The techniques are categorized to ex-ante and ex-post methods, before and after content is created (Bordogna *et al.*, 2016),.

One of the most common ex-ante method for improving quality is the reputation model, where users receive some quantifiable attribute for reputation. With a reputation model, every piece of content submitted by users receives an initial score based on how reputable the user is (Guo *et al.*, 2015; Fogliaroni, D'Antonio and Clementini, 2018; Wei *et al.*, 2018; Xiong *et al.*, 2018). Other ex-ante methods for improving content include modifying the data model (Fox *et al.*, 1999; Lukyanenko, Parsons and Wiersma, 2011) or the platform design (Lukyanenko *et al.*, 2019). Traditional citizen science platforms require users to describe and classify an observation accurately, and the users need to have some knowledge level to classify the observation correctly. Using a data model based on attributes would enable a free form input in the user interface rather than strict fill-the-form methods (Lukyanenko, Parsons and Wiersma, 2011).

Design choices and their effects on information quality are investigated in (Lukyanenko *et al.*, 2019). Class-based and instance-based data collection methods are compared using accuracy, precision, and completeness as the quality measures. The data collection methods are evaluated using the NL Nature citizen science platform. Results show that instance-based data collection provides more data and can capture unforeseen pieces of information but the drawback is precision loss. Additionally, completeness and accuracy are not directly affected by the collection method but rather by the user's expertise.

Most ex-post methods for improving data quality involve validation and cleansing (Mezzanzanica *et al.*, 2014; Sun *et al.*, 2018; Bouadjenek, Zobel and Verspoor, 2019). Before analyzing the collected data, traditional techniques are used in data pre-processing (Taleb, Dssouli and Serhani, 2015; Guan *et al.*, 2017). More demanding ex-post methods used extensively in UGC are: peer-review, expert review, and administrator review (Bordogna *et al.*, 2016), where more reputable users or pre-chosen administrators go through dubious data and remove errors or complete the entries to increases data quality.

The most important part for content users in UGC is to receive reliable content. UGC is generally more biased than traditional content because the purpose of UGC is to allow freedom of speech to users. Social media platforms are built to allow users to share subjective content. The reliability of content is based on the quality of the content and the credibility of the content provider. Biases have a significant impact on the credibility of the content provider, and there needs to be a way to determine it. Another source of bias in UGC is the individual who reviews and accepts content as credible (Robertson and Feick, 2016; Burgess *et al.*, 2017; Roman *et al.*, 2017). In Wikipedia, content is moderated by more reputable users, and their biases will impact the content

The quality of content varies drastically depending on the UGC platforms, but the platform should have a way to assess the quality of content. Having an assessment methodology in the platform helps content users believe in and utilize the UGC, although users will decide if they believe in the assessment.

### 2.3.3 Shortcomings

Many research articles identify open issues and challenges in UGC (Chen, Mao and Liu, 2014; King, Racherla and Bush, 2014; Sheppard, Wiggins and Terveen, 2014; Bordogna *et al.*, 2016; Lukyanenko, Parsons and Wiersma, 2016; Mitchell *et al.*, 2017; Xiang *et al.*, 2018). Most of these issues and challenges stem from the initial problems of data and information quality. There are relatively few definitions for quality within the UGC context, and they rely on the existing general data quality research without considering the contextual differences. Only Lukyanenko et al. (2014) mention this mismatch of information quality definitions, but the issue is still open.

Another significant issue in UGC is the reliance on techniques that require human resources to improve data and information quality. Using expert validation or training users to submit higher quality content requires more resources (Bordogna *et al.*, 2016), consuming more than what is available or worth. Improving the collection process to require fewer resources is more appropriate (Lukyanenko *et al.*, 2019). However, the lack of quality definitions in the existing research hinders platforms' design.

## 2.4 Summary

Many research articles related to data and information quality have been published from the 1990s onwards. Data and information quality foundations are based on contributions from existing literature (Redman, 1996; Wang and Strong, 1996; Batini and Scannapieco, 2016). One of the most crucial principles is that data and information quality are multidimensional, requiring specific characteristics to be appropriately defined.

The terms *data* and *information* have been used inseparably and as synonyms. Wang and Strong (1996) present data quality research that is later referred to as information quality research (Lee *et al.*, 2002). Similarly, Nicolaou and McKnight (2006) define data and information as synonyms.

Data and information quality characteristics must be selected based on the domain, and general quality research is not entirely applicable in the UGC domain (Davenport and Prusak, 2000; Bovee, Srivastava and Mak, 2003). Batini and Scannapieco (2016) tackle data and information quality from a general perspective in systems. Redman (1996) investigates data quality from a systems perspective, and Wang and Strong (1996) provide data quality definitions for organizational context. The data and information quality of UGC is still an open issue and requires proper research (Lukyanenko, Parsons and Wiersma, 2014).

The differences between traditional and web-based content restrict what existing research and methodology can be utilized. Quality management in traditional content can focus on selecting reliable sources and gathering new content. In addition, because the content provider, user and platform owner are often the same entity, it is possible to improve the content collection process using policies and rules. In UGC, selecting specific sources is more complicated and sometimes impossible, and the platform owner has minimal influence over the content provider or user, making some quality management techniques impossible to utilize (Bordogna *et al.*, 2016).

In summary, the following are the main shortcomings that need addressing because of the structural and operational differences between traditional content and UGC:

1. The amount of data and information quality research in the UGC domain is low.
2. UGC is more biased compared to traditional content.
3. Lack of distinction between *data* and *information* in research.
4. Lack of unified definitions and standards for data and information quality in UGC.
5. Lack of research to improve the quality of data and information in UGC platforms.
6. Lack of practical solutions for improving data and information quality in UGC.

# 3 Research method

Within the academic community, there exists a wide variety of research methods. This section presents different research methods and explains the most suitable for the research presented in this dissertation.

## 3.1 Research methods

**Action research** is a research method for organizational contexts (Carr and Kemmis, 1986). Canonical action research is a variation on action research for the information systems domain (Davison, Martinsons and Kock, 2004). Action research uses an iterative process from problem diagnosis to planning, intervention, evaluation, and reflection. This process continues until a satisfactory solution has been attained. The process relies on communication between researcher and client during the research.

**Grounded theory** originates from the social sciences and creates new theories from qualitative data (Glaser, Strauss and Strutzel, 1968). The process involves gathering and analyzing data until theoretical saturation. Grounded theory begins with reviewing the literature to select qualitative cases from where the data is collected. Data from cases are constantly compared, and the analysis may lead to new data sources. Although grounded theory is qualitative research, it requires a considerable amount of data for analysis.

**The deductive nomological approach** is a method that heavily relies on existing research. Using the deductive nomological process, the researcher should base their hypothesis on existing theories or laws, making it challenging to conduct research without proper theories within the domain (Hempel, Feigl and Marxwell, 1962). **The hypothetico-deductive (or hypothetico-inductive) approach** is similar to the deductive nomological method but with a slight difference: the hypotheses do not have to be based on existing theories or laws. Instead, they can be based on guesses or personal experiences (Jeffrey and Popper, 1934; Hempel, 1966; Siponen and Klaavuniemi, 2020). This approach makes it easier to enter a research field with no well-established theories.

**Building theories from case studies,** presented by Eisenhardt (1989), has several steps for building theories based on case study:

1. Getting started

Starting case study research requires knowledge of existing literature and, if possible, a sound theory behind the research. Trying to avoid biased opinions is essential at this stage, and the researcher should mainly formulate a research problem and some crucial variables that are essential regarding the issue. However, the relationship between variables and theories should be left out.

2. Selecting cases

Case studies often require multiple cases, but under some specific conditions, single-case studies are valid. Cases should not be chosen randomly but rather replicate previous cases, extend the rising theory or provide examples of opposite situations.

3. Crafting instruments and protocols

Each case study requires data collection, and there must be predefined protocols and possible instruments for data collection. When the protocols are well defined, the case study is more accessible to replication and easier to advance. Instruments for data collection may differ case by case, but they should be as similar as possible to reduce variability. Instruments can be surveys, literature, interviews, or software.

4. Entering the field

When collecting data, factors such as reasons, opportunities, or epiphanies may influence the data collection methods by altering or adding new ways to collect data. Some question the validity of data collection when the techniques have been changed during the process, but modifying the data collection methodology is allowed for theory-building research. The goal is not to generate a summary of data but rather to understand and investigate phenomena. There needs to be some flexibility in the study as the alteration may lead to better theoretical insights.

5. Analyzing data

There is no de facto way to analyzing data, and the most crucial part is that the researcher is highly familiar with each case's data before making any generalizations. During analysis, there are two different analysis opportunities. First, finding some generalizations within the single case data that can be used for cross-case comparison. Another is searching for the patterns between cases. Finding patterns between cases can be done by grouping similar cases and finding differences or grouping by the data source.

6. Shaping hypotheses

To shape hypotheses, theories, or constructs, it is necessary to systematically compare evidence emerging from each case to the created framework. Another important aspect is how the created constructs apply to each case.

7. Enfolding literature

After creating hypotheses, theories, or concepts, they should be compared to existing literature. Examining the similarities and differences between existing literature and developed ideas increase validity and strengthen confidence and generalization.

8. Reaching closure

Reaching closure requires the researcher to know when to stop the case study and iteration between data and literature. When cases provide minimal addition to information and reach theoretical saturation, the case study should be stopped. Saturation is a reason to stop the iteration process as well.

**Design science research (DSR) paradigm** by Hevner et al. (2004) is an iterative process for developing artifacts. It was initially established for information systems but has been adapted to other disciplines (Engström *et al.*, 2020). The goal is to solve an existing unsolved problem by creating an artifact and improving the body of knowledge with insights and explanations of the artifact's results. The artifact can be a system, application, framework, model, or any concrete concept. DSR is an excellent way to research a domain that has fewer theories and existing literature.

**Table 2**. Research method comparison

| Research method | Strengths | Weaknesses |
|---|---|---|
| Action research / canonical action research | - An iterative process that starts with a relevant problem<br>- Can develop an artifact | - Designed for usage in an organizational context<br>- Communication with a client |
| Grounded theory | - Qualitative research<br>- Well established | - Requires a considerable amount of data<br>- Only for building theories through data analysis |
| Deductive nomological approach | - Builds new theories from old theories | - Domain requires theories to be utilized<br>- Only for making theories through data analysis |
| The hypothetico-deductive (or hypothetico-inductive) approach | - Can initiate with guesses or user experience<br>- Iterative process for establishing hypotheses | - Only for building theories through data analysis |
| Building theories from case studies | - Possible to build theories from cases<br>- Good when domain lacks theories | - Only for case studies |
| DSR | - The main principle is to develop an artifact<br>- Iterative process<br>- Good when domain lacks theories<br>- An artifact can be extended to a theory | - General research philosophy |

Table 2. presents the comparison between investigated research methods, processes, and philosophies. Based on the comparison and applicability, DSR by Hevner et al. (2004) is the chosen research philosophy for this research. Other research methods require existing theories from the domain, and their primary output is new theories. To build proper theories, they need to be tested repeatedly, and only after numerous tests can theories be considered valid. The main research output of DSR is an artifact that is not a theory but can be extended into one after repeated testing and evaluation.

## 3.2 Research process

DSR is a research philosophy that has several guidelines:

1. Design as an artifact: The result of the research should produce an artifact. The artifact can be a model, method, or system.
2. Problem relevance: The artifact should be developed regarding a relevant problem based on current issues in the target domain.
3. Design evaluation: The artifact needs to be thoroughly evaluated.
4. Research contributions: The research must provide theoretical or practical research contributions in the target domain.
5. Research rigor: Research needs to be rigorous during the development and evaluation of the artifact.
6. Design as a search process: The search for a practical artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
7. Communication of research: The research should be communicated and shared with an audience.

These different guidelines relate to the three cycles of research in the DSR philosophy presented in Figure 1.

- **Relevance cycle:** The relevance cycle is the starting point of DSR. The selected domain and context provide a relevant problem, requirements, and assessment criteria for the artifact.
- **Design cycle**: The design cycle is the main component of DSR philosophy. It involves building and evaluating the designed artifact in a constant loop. This loop continues until the artifact is validated and the new insights gained can be added to the existing knowledge base.
- **Rigor cycle**: The rigor cycle is necessary to determine the artifact's novelty by examining the existing knowledge. The rigor cycle is also the endpoint of DSR as the validated artifact is added to the knowledge base.

The research process presented in this dissertation is divided into several phases. Each phase has its own research method and relates to the DSR cycles and guidelines.

*Phase 1 – Define a relevant problem:* The first task of this research is to explore a relevant issue to solve. The primary source of data during this phase comes from literature and existing platforms in the selected domain.

*Phase 2 – Artifact design:* After finding a relevant problem, an artifact is designed at a basic level. There is no need to have a complete artifact at this point, but there needs to be a concrete idea of the artifact. Most data comes from literature and existing systems.

*Phase 3 – Artifact development:* During development, relevant literature is investigated, and the design is revised using new knowledge.

*Phase 4 – Artifact evaluation:* The artifact is evaluated using criteria from the application domain. The artifact is assessed against existing artifacts. Because there are no existing artifacts to evaluate against, the developed artifact is evaluated in practice with specific requirements. During Phase 4, data collection involves case studies and collecting data from other platforms.

*Phase 5 – Research contribution and communication*: After evaluating the artifact and analyzing the results, the results are communicated to academia and added to the body of knowledge.

Table 3. presents the relationship between the three DSR cycles and the five research phases. The table also includes the specific research method and outcome of each phase.

**Table 3**. Research phases, methods, and outcomes mapped to DSR philosophy.

| **Research phase** | **Research method (input)** | **Research outcome (output)** | **DSR philosophy** |
|---|---|---|---|
| Phase 1 | Literature and platform review | Relevant problem(s) | Relevance and Rigor cycles |
| Phase 2 | Literature and platform review | Artifact design(s) | Relevance cycle |
| Phase 3 | Artifact development | Artifact(s) | Design cycle |
| Phase 4 | Case study | Evaluation | Design and Relevance cycles |
| Phase 5 | Dissertation | Contribution | Rigor cycle |

## 3.3 Summary

DSR by Hevner et al. (2004) is the chosen research philosophy for this dissertation. DSR has three distinct cycles and several guidelines that need to be followed during the research process.

# 4 Overview of publications

## 4.1 Publication I: Overview of data storing techniques in citizen science applications

This section presents the overview of the following publication:

Musto, J. and Dahanayake, A. (2018). Overview of Data Storing Techniques in Citizen Science Applications. In: Benczúr A. et al. (eds) New Trends in Databases and Information Systems. ADBIS 2018. Communications in Computer and Information Science, 909.

### 4.1.1 Research background

The research presented in this section is considered a literature review to explore available scientific literature on UGC platforms, citizen science research, and data storage approaches.

UGC garnered a more significant following and interest during the 2010s. More researchers and organizations have started utilizing content provided by users from different platforms. These platforms include social media, crowdsourcing, volunteered geographic information, and citizen science (Ludwig, Reuter and Pipek, 2015; Yan *et al.*, 2017; Arthur *et al.*, 2018). Today's most notable UGC platforms are Twitter, Facebook, YouTube, Instagram, Wikipedia, OpenStreetMap, eBird, and iNaturalist. However, more platforms are constantly being developed.

Social media platforms are the most common UGC platforms. Content provided in social media is primarily subjective without sources or references. Many believe it to be objective truth when a high-profile content provider claims something on a social media platform (Popat *et al.*, 2017; D. Clark, 2020; Pennycook *et al.*, 2020). The primary purpose of social media platforms is to share subjective content, which exhibits how the platform operates. There are few restrictions on what content can be shared, and content is stored in the database without extra processing.

Wikipedia (Wikipedia, 2020) is known as the primary information source for the general public. Wikipedia has specific guidelines on what kind of content can and should be shared. These guidelines include the usage of references and formatting. OpenStreetMap (OpenStreetMap, 2021) is a crowdsourced geographic information platform that is an open-source version of the Google Maps service, where volunteers provide content. There are specific data formats the content must follow in OpenStreetMap.

iNaturalist (iNaturalist, 2021) is a joint initiative by the California Academy of Sciences and the National Geographic Society that started as a Master's thesis project in 2008. iNaturalist.org's citizen science platform has expanded worldwide and is supported by

local organizations that have integrated their content collection with the iNaturalist platform. The platform can be used to collect observations from nature. eBird (Cornell Lab of Ornithology, 2021) is one of the oldest citizen science projects still ongoing. eBird began in 2002 as a regional birding project but has since extended to worldwide coverage. The eBird platform uses various data processing techniques in content moderation including some automated checks as well as having experts review content (Wiersma, 2010; Yu *et al.*, 2012; Kelling *et al.*, 2013). Galaxy Zoo is a citizen science project that began in 2007 where content providers are asked to classify images (Lintott *et al.*, 2010; Tiley *et al.*, 2019) instead of collecting observations.

All UGC platforms operate using the same basic principles. This process is demonstrated in Figure 4.

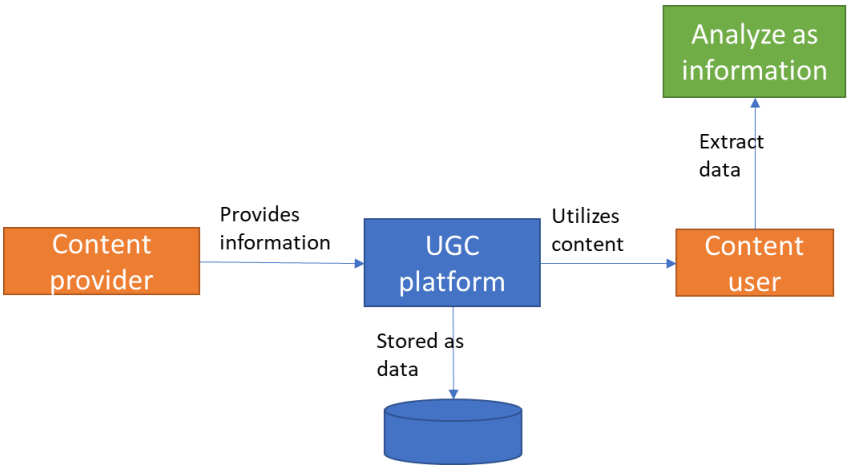


**Figure 4**. UGC platform content process

As Figure 4 shows, every UGC platform has content providers who submit information depending on what content is requested. The content can be extracted as data from the UGC platform either directly from the database or by content mining. Finally, content users can further analyze the data to obtain new information. This process applies to most citizen science platforms, social media platforms, and Wikipedia. Table 4 presents a comparison of different UGC platforms based on their content and users.

**Table 4.** Comparison of different UGC platforms

| | **Citizen science** | **OpenStreetMap** | **Social media** | **Wikipedia** |
|---|---|---|---|---|
| Content provider | Amateur | Knowledgeable | Anyone | Knowledgeable |
| Primary content users | Researchers | General public | Anyone | General public |
| Secondary content users | Anyone | Researchers | Researchers | Anyone |
| Content-type | Subjective observation | Subjective using devices | Subjective opinions | Subjective using references |
| Content format | Text, media | Text, data | Text, media | Text, media |
| Content restrictions | Relevant content, minimal freeform text | Specific data in a particular format | Platform terms of service | Platform-specific guidelines and format |

### 4.1.2 Objective

To identify issues in UGC, a systematic literature review is conducted on citizen science. Because the domain of UGC is vast, but the underlying principles behind the platforms are similar, the scope is limited to citizen science platforms. This publication aims to figure out whether the technology used and its related processes cause any challenges. The following research question is answered in the research: "*What type of data storing techniques and technologies are used in citizen science applications*?"

Over 700 articles are gathered for the systematic literature review using the search terms presented in Table 5. The articles are collected from scientific databases ACM, IEEE, Scopus, Springer, and Web of Science. There are multiple search terms because citizen science has been used with various other terms, such as public participation and volunteered geographic information (See *et al.*, 2016).

**Table 5**. Search terms using the Springer format

| **Query** |
|---|
| "citizen science" AND ("data model" OR "data struct*") |
| "community-based monitoring" AND ("data model" OR "data struct*") |
| "public participation" AND ("data model" OR "data struct*") |
| "volunteer monitoring" AND ("data model" OR "data struct*") |
| "volunteered geographic information" AND ("data model" OR "data struct*") |
| "participatory sensing" AND ("data model" OR "data struct*") |

The articles are filtered based on duplicates, abstracts, and full text using inclusion and exclusion criteria. Finally, fourteen articles are examined in more detail. Figure 5. presents the literature review process and the number of articles left after each step.

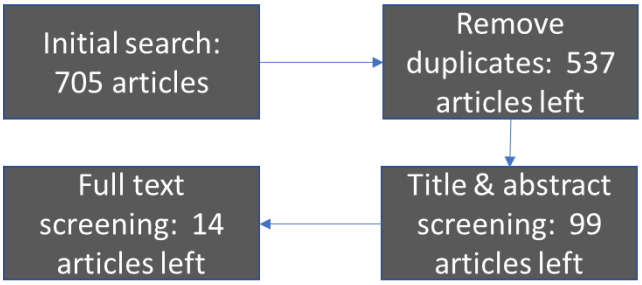


**Figure 5**. Literature review process

When a paper matches at least two inclusion criteria during the screening, it is accepted unless it matches one exclusion criterion. The exclusion criteria are designed to be strict and to filter irrelevant results from the total. 3D modeling-related papers are excluded because using 3D data requires different techniques. The inclusion and exclusion criteria are presented in Table 6.

**Table 6.** Systematic literature review criteria

| Criterion | Inclusion / Exclusion |
|---|---|
| *Title and abstract screening* | |
| Information related to software | Inclusion |
| Information related to database | Inclusion |
| Information related to the database management system | Inclusion |
| Information related to the data model | Inclusion |
| Information related to the data structure | Inclusion |
| No full text available | Exclusion |
| 3D modeling | Exclusion |
| *Full-text screening* | |
| Detailed data model | Inclusion |
| Detailed data structure | Inclusion |
| Application implementation | Inclusion |
| Data management information | Inclusion |
| Design too unrelated to citizen science | Exclusion |

### 4.1.3 Relation to Dissertation's Research Question

The presented publication serves as the foundation for the overall thesis and partially answers the first sub-question of the thesis, "*What information collection features in user-generated content platforms influence the quality of content*?" by examining what data collection techniques and technologies are used in citizen science. The publication's primary purpose is to find frequent data-related practices in the citizen science domain and their effects. The results present frequently used database designs and technologies, relevant topics, and challenges in citizen science. Some suggestions for resolving the presented issues and possible practical models and frameworks are discussed for future development.

### 4.1.4 Research Output and Contribution

During the study, several issues related to citizen science are identified and discussed. Many of the issues are related to the employed technology and how the platforms operate. While the challenges are separate, they have some relation to each other. The challenges are listed in Table 7.

**Table 7**. Challenges of citizen science platforms.

| Challenge | Reason |
|---|---|
| Data quality | Data quality is often low in citizen science because amateurs contribute data (Lukyanenko, Parsons and Wiersma, 2014). Amateur contributions diminish the research's trustworthiness and make data usage more difficult (Elbroch *et al.*, 2011; Kosmala *et al.*, 2016; Leibovici *et al.*, 2017). |
| Provenance | Provenance in data management is highly recommended. Based on National Biodiversity Network guidelines (James, 2006), provenance should be handled in citizen science projects, and some projects take extra care to consider it (Sheppard, Wiggins and Terveen, 2014). Still, provenance management is difficult to perceive when no information regarding provenance is given to users because it is applied in the database. |

| | |
|---|---|
| Standards | Standards have been developed for geographic information (Fonte *et al.*, 2017), open data (OASIS, 2017), sensor data (Huang and Liang, 2014), and biodiversity data (Veen *et al.*, 2012). However, there are no specifically created data standards for citizen science, leading to projects developing different data models and frameworks. Connecting data from one citizen science platform to another is exceedingly difficult if the data follow different standards. |
| Format | Citizen science platforms deal with data in multiple formats. Data is primarily numerical and text data, but multimedia usage has increased. Some projects accept audio and video files (Gouveia *et al.*, 2004; Wiersma, 2010), making storage and quality management arduous. |
| Development methodology | Citizen science projects have a relatively short lifespan compared to other platforms. Citizen science projects are designed to get a working product quickly, use it and discard it. The development methodology often disregards many essential ideas and design concepts, such as handling flawed or erroneous information and focusing only on ease of use (Newman *et al.*, 2010; Freiwald *et al.*, 2018). Getting high-quality data with poorly designed projects is difficult. If the project's lifespan is extended, many of these fundamental problems surface and require an overhaul of the platform. |
| Content user | Citizen science projects focus on gathering data for the specific purpose defined by the project owner. The data is often collected for a singular objective, and possible usage outside of that is not considered. If the project owner decides to release the data to the public or consider using it somewhere else afterward, it may be less valuable or even unusable. |
| Technology | Relational databases are the primary database technology used in citizen science platforms. However, relational databases are often chosen without considering any alternatives or justification for the selection. Some researchers have tested different relational databases without considering NoSQL alternatives (Kotsev *et al.*, 2016), and other researchers have found NoSQL databases better than relational databases (Bonacic, Neyem and Vasquez, 2015). |
| Terminology | The definition of citizen science can be unclear. Various terms have been used to describe citizen science, confusing other researchers (See *et al.*, 2016). |
| Information overflow | Most citizen science projects allow participants to view content submitted by others. With large citizen science projects, the participant may be overwhelmed by the amount of information and have difficulties understanding the content. The projects can employ different filtering mechanisms and search functions to help reduce the amount of information presented (Rees *et al.*, 2011; Havlik *et al.*, 2013). |

Challenges in Table 7 are significant problems in the citizen science domain (Lukyanenko, Parsons and Wiersma, 2014, 2016; Blatt, 2015; Kosmala *et al.*, 2016; Williamson *et al.*, 2016; Steger, Butt and Hooten, 2017; Palacin-Silva and Porras, 2018). Many of the problems relate to *design and development*.

While the presented challenges arise from citizen science, they can be extended outside of the domain and under the umbrella of UGC. *Data quality, format, content user,* and *provenance* are challenges that are easily found in multiple UGC platforms (Demetriou, 2016; Salk *et al.*, 2016; Buntain and Golbeck, 2017; Arolfo and Vaisman, 2018; Ahmouda, Hochmair and Cvetojevic, 2019; Barbosa *et al.*, 2019). Even when there are differences in context and primary usage, these challenges are common to many UGC platforms:

- Content providers are mostly amateurs.
- Content comes in various formats.
- There are no applicable standards for the content in UGC.
- Needs of potential content users are not taken into consideration.
- Provenance is highly recommended but not explicitly implemented.
- An information overflow may happen.

These six challenges influence *data and information quality*. Quality of data and information are crucial in all UGC platforms. Information quality plays a vital role in social media because the content can be used to influence other users (Culotta, 2010; King, Racherla and Bush, 2014; Zhao and Sui, 2017; Lakshen, Janev and Vraneš, 2018; Abdullah-All-Tanvir *et al.*, 2019; Peters, 2020; Twitter Safety, 2020). Information overflow is a more significant issue in social media than citizen science because there is an enormous amount of content. Additionally, there is more contradictory and conflicting content in social media, and separating the truth from false content can be challenging (Buntain and Golbeck, 2017).

The research contributes to understanding the domain of citizen science and UGC. Different data collection and processing techniques have been explored, and several challenges are identified based on the literature review. Many of the challenges from the citizen science domain are common to the overall UGC domain. The identified issues have a significant impact on the quality of collected information and data. These challenges can be resolved or at least mitigated with proper design. Platforms heavily rely on moderators to review content (Rice, 2015; Nevolin, 2017; Heinrich *et al.*, 2019; Truong, de Runz and Touya, 2019). In citizen science, these moderators are reliable people chosen by the project owner, but the same does not apply to other UGC platforms. In Wikipedia, moderators are selected through democracy. In social media platforms, high-level moderators are platform employees who only review content reported by the community. There is no limit to what the community can report, which is why false reports are common in social media (Buntain and Golbeck, 2017; Viviani and Pasi, 2017; Zhao and Sui, 2017; Abdullah-All-Tanvir *et al.*, 2019; Goodman and Carmichael, 2020; Quinn, 2020).

The research presented under this article relates to the first sub-question of the dissertation, "*What information collection features in user-generated content platforms influence the quality of content*?" and partially provides an answer. The results show that the quality of content is a concern in citizen science, and one of the most important aspects is the presentation of the user interface on these platforms. Some information collection features to improve the quality of content have been discussed, and the main objective of this research is to identify common ground among the used features.

## 4.2 Publication II: Improving data quality, privacy, and provenance in citizen science applications

This section presents the overview of the following publication:

Musto, J. and Dahanayake, A. (2020). Improving Data Quality, Privacy and Provenance in Citizen Science Applications. Frontiers of Artificial Intelligence and Applications, 321, pp. 141-160.

### 4.2.1 Research background

Research presented in this section is considered a case study to explore available UGC platforms and related data quality issues.

The results of earlier research (Musto and Dahanayake, 2018) show that data quality, privacy, and provenance are significant issues in citizen science platforms. The results encourage exploring such platforms to explain the issues in more detail and evaluate how they could be resolved.

Privacy is a concern when personal information is combined with location information. Amongst UGC platforms, volunteered geographic information and most citizen science platforms need location information, and there should be some privacy protection with the collected data. There are two popular methods: anonymization (Oliver, Miche and Ren, 2018) and generalizing location (Naghizade *et al.*, 2015). Anonymisation simply removes any personal identifiers from data, and with generalizing location, a broader area is given rather than the exact coordinates. Both methods have been applied in some UGC platforms.

Provenance in UGC platforms is necessary when users can edit already submitted content, and it applies to social media, Wikipedia, and citizen science. Provenance refers to the process of storing and maintaining information where content has originated, who has changed it, and how. When there is no provenance and users can edit content, it is difficult to determine what has been changed and by whom. It can be harmful when a credible user creates content, and someone less credible modifies it to their benefit.

Data quality is considered multidimensional, composed of multiple characteristics (Wang and Strong, 1996; Batini and Scannapieco, 2016). The chosen characteristics vary case-by-case. Each researcher and product owner must decide which characteristics are essential for their definition. Some researchers utilize only a handful of characteristics (DAMA UK, 2013; Immonen, Pääkkönen and Ovaska, 2015; Schmidt *et al.*, 2015; Kosmala *et al.*, 2016; Wiggins and He, 2016; EDM Council, 2017; Arolfo and Vaisman, 2018) while other researchers use dozens of different characteristics (Wang and Strong, 1996; Alabri and Hunter, 2010; Sheppard, Wiggins and Terveen, 2014; Batini and Scannapieco, 2016; Bordogna *et al.*, 2016; Antoniou, 2017; Albrecht *et al.*, 2018). ISO

has established some standards regarding data quality: geographic information data quality 19157 and general data quality model 25012.

ISO 25012 divides data quality into fifteen distinct characteristics. The characteristics are classified into inherent, system-dependent, or both categories. Inherent data quality characteristics are affected by the intrinsic qualities of the data within the application domain. System-dependent quality characteristics are affected by the technology where the data is being used. When a characteristic is classified into both categories, the characteristic can be affected by intrinsic qualities or technology. *Traceability* in the ISO standard is similar to provenance, and *confidentiality* is similar to privacy. The fifteen characteristics and their categories are listed in Table 8.

**Table 8.** ISO data quality characteristics (ISO, 2008)

| Characteristic | Inherent data quality | System-dependent data quality |
|---|---|---|
| Accuracy (syntactic and semantic) | x | |
| Completeness | x | |
| Consistency | x | |
| Credibility | x | |
| Currentness | x | |
| Accessibility | x | x |
| Compliance | x | x |
| Confidentiality | x | x |
| Efficiency | x | x |
| Precision | x | x |
| Traceability | x | x |
| Understandability | x | x |
| Availability | | x |
| Portability | | x |
| Recoverability | | x |

### 4.2.2 Objective

The presented study's primary research objective is to examine how data quality is handled in existing citizen science platforms and discover possible flaws. The platforms are selected through a project hub that gathers information about projects around the globe. Data quality is inspected using the ISO 25012 data quality definition (ISO, 2008). This work continues the research in (Musto and Dahanayake, 2018) to extend the research from literature to practice. The following research questions are explored in the article:

- How well are data quality, privacy, and provenance handled in ongoing citizen science projects?
- How can data quality, privacy, and provenance be improved in citizen science projects?

The research aims to explain which quality characteristics in citizen science platforms are considered less often and reduce overall data quality. Because data comes from amateurs, data credibility is questionable, and data quality plays a vital role in reusing data from citizen science platforms. When the data quality is improved, the projects and researchers gain more accurate results. Before data quality can be improved, the current state needs to be carefully examined.

The citizen science platforms for review are selected through a citizen science project hub called SciStarter (https://scistarter.org/, retrieved 29th June 2021). SciStarter has listed over 1500 projects. Many of the projects have a platform, and SciStarter provides direct links to these platforms. Figure 6 presents the process of selecting a citizen science platform for review.

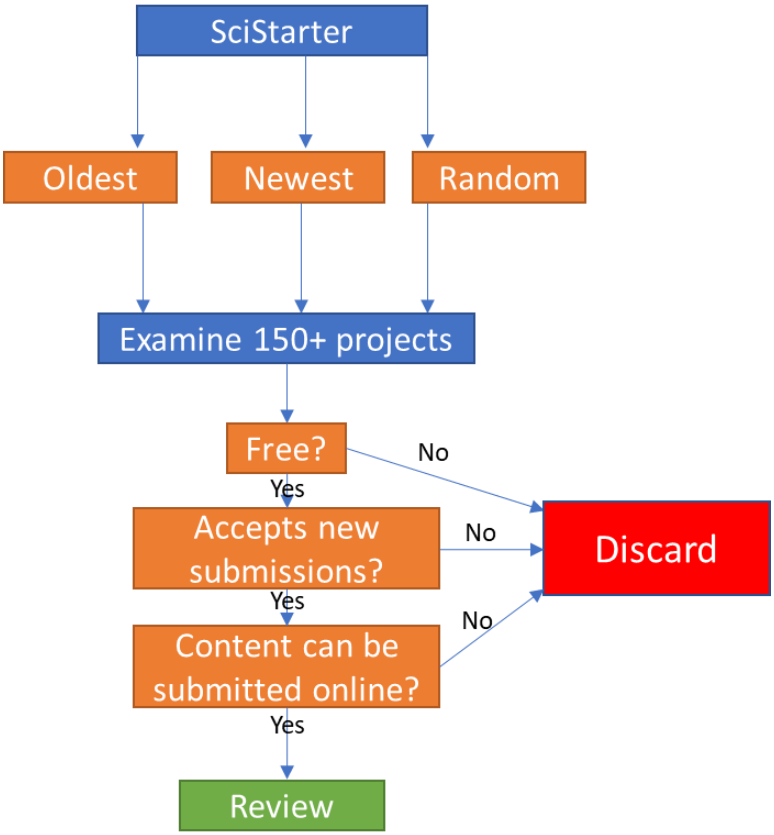


**Figure 6**. Citizen science platform selection process

Numerous examined projects have some limitations that hinder the review of their platform. First, some projects require an initial payment to purchase project-specific training or equipment before participation is allowed. Some projects require the content to be sent via mail or require participation in specific gatherings. Projects that do not

accept new submissions and have finished collecting data cannot be reviewed. In the end, thirty different citizen science platforms and eight project organizing platforms are reviewed. The eight project organizing platforms allow anyone to launch and run their project within the platform.

Each selected platform is reviewed through the user interface. The user interface is one of the essential parts of the data collection process as it is the first barrier that can accept or reject content based on specific criteria (Musto and Dahanayake, 2018). Using the ISO 25012 data quality standard, the user interface is examined with each distinctive characteristic and how well they are handled or considered in the platform.

### 4.2.3 Relation to Dissertation's Research Question

The publication shows how citizen science projects handle data quality and if specific data quality characteristics are adequately considered. Platforms use different methods for ensuring and improving quality based on design choices. Some platforms handle data quality with great care, using specific features in their data collection process, such as basic automatic verification, and other platforms provide low-quality data. This research gathers results from the field and actual projects. Additionally, the focus explicitly targets data quality. Many of the issues and challenges presented in the previous research (Musto and Dahanayake, 2018) relate to the design and development of platforms. These issues impact data quality, one of the significant concerns in citizen science and the pivotal point of this publication.

The results suggest several ideas for improvement based on existing platforms, literature, and other data-intensive fields. This research aims to provide more answers to the first sub-question, "*What information collection features in user-generated content platforms influence the quality of content?*". Additionally, the research results contribute to the second sub-question, "*How to define quality characteristics and distinguish data and information quality in the domain of user-generated content?"* using the ISO 25012 standard for data quality as the definition. As no proper data quality definitions exist for UGC, a well-established standard is used as a starting point to figure out what characteristics are necessary for UGC.

### 4.2.4 Research Output and Contribution

This research's primary output and contribution are the details of why data quality in citizen science and UGC is lacking. The results show that data quality is hugely lacking in specific areas based on individual data quality characteristics, such as completeness. In contrast, there is only a need for minor improvements in other areas.

The research results are presented in Table 9. The table consists of ISO 25012 data quality characteristics, and results are mapped to five categories:

- Best case: The characteristic is well handled, and it improves the data quality, or there is little to nothing that decreases it.
- Good case: The characteristic is adequately handled, but slight improvement could be made.
- Neutral case: The characteristic is partially handled and could be improved.
- Bad case: The characteristic is handled poorly, or little consideration is given.
- Worst case: The characteristic is handled negatively.

The numbers in brackets reflect how many of the reviewed projects considered or implemented the particular characteristic.

**Table 9**. How well data quality characteristics are implemented in citizen science projects

| Data quality characteristic (38) | Best case | Good case | Neutral case | Bad case | Worst case |
|---|---|---|---|---|---|
| **Accuracy (30)** | | Some syntactic and semantic accuracy checks (8) | Syntactic accuracy checks (12) | | No accuracy checks (10) |
| **Completeness (32)** | All fields required (13) | Half or more required (5) | Any number of fields required (2) | Less than half required (11) | Nothing required (1) |
| **Consistency (34)** | 5 (12) | 4 (6) | 3 (7) | 2 (5) | 1 (4) |
| **Credibility (22)** | Sensor (1) | Validity tests (2) | Preliminary test (2) | Moderated (9) | Community (8) |
| **Currentness (31)** | | Observation time is different from upload (30) | Up to the project creator to handle (1) | | |
| **Accessibility (37)** | Maps (and other things) (23) | Analyses / Statistics (3) | Observations only (4) | Reports (3) | Not available (4) |
| **Confidentiality (34)** | | No name (nor location) (9) | Username / anonymous (12) | Location and name are shown with exceptions (6) | Location and name are shown (7) |
| **Efficiency (28)** | 5 (5) | 4 (13) | 3 (6) | 2 (2) | 1 (2) |
| **Precision (30)** | Predefined estimates (1) | Can give estimates (23) | Requires precision (4) | | Up to project creator (2) |
| **Traceability (7)** | | | Can view edits (3) | No provenance (2) | Anyone can edit data (2) |
| **Understandability (29)** | 5 (13) | 4 (9) | 3 (6) | 2 (1) | 1 (0) |
| **Availability (38)** | Can download csv / excel / json file (8) | By contacting (2) | Can download data from previous years (1) | Can download reports (5) | Not available (22) |

Based on the results shown in Table 9, a few characteristics should be highlighted. The worst case in *accuracy* is that no accuracy checks are implemented in the platforms, which applies to a third of the reviewed platforms. Another characteristic is *availability*, where worst case means no data is available for reuse. Even contacting the project owners is not given as an option. *Privacy* (confidentiality) should always be handled with extra care when considering regular citizens' data and information. The worst case in *privacy* is that the content provider's name and location are shown with little regard to possible privacy implications. Finally, mentions of *provenance* (traceability) are found in only seven of the reviewed platforms, and none are in the good or best case. The worst case is that anyone can make edits to content without regard, which creates a massive quality issue and removes any possibility of trusting the data.

Many vital data quality characteristics are overlooked in these platforms, and there are some significant issues. None of the reviewed platforms are found to be perfect. One platform only needed minor improvements, but a considerable portion would require a more extensive overhaul to increase the data quality.

The thirty-eight reviewed platforms include eight citizen science project organizing platforms. Multiple projects are hosted on the same platform, and data quality issues reflect on all hosted projects. Some of these platforms handle quality more poorly than individual projects, and they often rely on the moderation of project creators. Project creators are offered basic tools, but better usage requires knowledge and expertise. A project hosting platform can easily have hundreds or even thousands of projects.

Other researchers' findings support the results of this publication (Alabri and Hunter, 2010; Lukyanenko, Parsons and Wiersma, 2014, 2016; Sheppard, Wiggins and Terveen, 2014; Bordogna *et al.*, 2016; Palacin-Silva and Porras, 2018), but this publication provides a more detailed explanation of what is lacking in data quality. There are several methods for improving the quality of content in citizen science platforms (Bordogna *et al.*, 2016). While some argue that quantity is better than quality (Bayraktarov *et al.*, 2019), this requires a considerable amount of data and still would not be preferable. Some citizen science projects will have fewer participants, making the quality even more important than quantity. This publication provides examples of how specific data quality characteristics could be handled within a platform. These examples are based on existing methods from other projects and available literature. The results contribute to the growing data quality research in citizen science and give more detailed information to project owners about what aspects of quality should be carefully considered.

While the publication investigates citizen science platforms explicitly, these results can be extended to UGC in general because there are many similarities between citizen science and other UGC platforms as mentioned in the previous research (Musto and Dahanayake, 2018).

The most significant differences between citizen science and other UGC platforms are the primary purpose of content and the usage of location information. Most social media

platforms have abandoned location information entirely unless the content provider explicitly gives it. Many UGC platforms outside of citizen science moderate and remove erroneous information (Truong, de Runz and Touya, 2019; Peters, 2020; Quinn, 2020). These differences influence platform implementation details, but general improvements and models can be applied to all UGC platforms.

Both this and the previous publication (Musto and Dahanayake, 2018) contribute to answering the first sub-question, "*What information collection features in user-generated content platforms influence the quality of content?* "

The earlier research (Musto and Dahanayake, 2018) shows a lack of data and development standards in UGC that result in inferior designs and rushed development methodology. Because there are no standards to rely on, each developer designs the platform from scratch, often in a short time. Rapid development leads to creating collection features that do not consider the quality of the content. Many of the design flaws reflect on the user interface of the platform. This research examines different platforms and possible deficiencies in data quality through the user interface. The results show that many platforms require better collection procedures and a user interface that better directs the content provider to deliver higher quality information (Lukyanenko, Parsons and Wiersma, 2011, 2014; Bordogna *et al.*, 2016; Lukyanenko *et al.*, 2019).

The presented research leads to the second sub-question, "*How to define quality characteristics and distinguish data and information quality in the domain of user-generated content?*" by utilizing the ISO 25012 standard for data quality and experimenting to discover whether the standard can be fully applied to UGC. Preliminary results show that while the ISO standard is an excellent general tool, it is not fully applicable to UGC and requires modification.

## 4.3 Publication III: Quality characteristics for user-generated content

This section presents the overview of the following publication:

Musto, J. and Dahanayake, A. (2021) Quality characteristics for user-generated content. Frontiers of Artificial Intelligence and Applications. Accepted 2021.

### 4.3.1 Research background

A systematic analysis of keyword-based article searches in scientific databases gives thousands of results when simply searching "data quality." When another term is included, the results drop to less than 10 % of all articles. The search results are presented in Table 10.

**Table 10**. Results of keyword-based article search

| Search term | Scopus | IEEE | Springer | ACM |
|---|---|---|---|---|
| data quality” | 95069 | 20933 | 50586 | 4892 |
| AND “citizen science” | 1143 | 38 | 393 | 99 |
| AND “big data” | 5547 | 1466 | 3726 | 721 |
| AND “remote sens*” | 8 715 | 2497 | 3672 | 2 |
| AND “crowdsource*” | 2796 | 311 | 1001 | 0 |
| AND “user generated” | 705 | 30 | 574 | 186 |
| AND “social media” | 22327 | 150 | 2262 | 520 |
| “data quality defin*” | 20 | 42 | 59 | 0 |
| “data quality model” | 407 | 123 | 193 | 39 |
| “data quality dimension” | 1154 | 62 | 455 | 49 |
| “data quality characteristic” | 40 | 13 | 86 | 2 |

Data and information quality are defined as a collection of characteristics. In various scenarios, these characteristics have different meanings (Bovee, Srivastava and Mak, 2003; Haug, Arlbjrn and Pedersen, 2009; Haug and Arlbjørn, 2011; Batini *et al.*, 2017). For example, precision in geographic information may refer to locational precision, but precision in healthcare may refer to numerical precision of values. Each quality characteristic used should be explicitly defined based on the usage. However, many researchers and organizations rely on mutual understanding of these terms or use ambiguous definitions (ISO, 2008; Haug and Arlbjørn, 2011; DAMA UK, 2013; Hashem *et al.*, 2015).

There are definitions for general data quality that are primarily accurate in well-structured data from well-established sources, such as corporate data (Wang and Strong, 1996; Batini and Scannapieco, 2006; Lukyanenko, Parsons and Wiersma, 2016). The exact definitions need to be tweaked to be applicable in other domains, such as in geographic information (Sheppard, Wiggins and Terveen, 2014; Batini *et al.*, 2017; Fonte *et al.*, 2017) or UGC (Bordogna *et al.*, 2016; Lukyanenko, Parsons and Wiersma, 2016; Cox, McKinney and Goodale, 2017; Ahmouda, Hochmair and Cvetojevic, 2019). Researchers and organizations rely on existing definitions from general data quality within the UGC

domain, and there are no proper definitions specifically for the UGC domain (Immonen, Pääkkönen and Ovaska, 2015; Spielhofer *et al.*, 2017; Arolfo and Vaisman, 2018; Arthur *et al.*, 2018). There is a need to provide definitions for data quality characteristics in the UGC domain.

Data and information are often conflated (Wang and Strong, 1996; Lee *et al.*, 2002; Pipino, Lee and Wang, 2002; Nelson, Todd and Wixom, 2005; Batini and Scannapieco, 2006, 2016) and they should be separated as two different concepts (Davenport and Prusak, 2000; Watts, Shankaranarayanan and Even, 2009). Data quality characteristics can be without context, while information quality characteristics require knowledge, such as how the information is utilized. Table 11 lists data and information quality characteristics from several researchers that are most commonly used to present an overview of quality characteristics.

**Table 11**. Collection of data and information quality characteristics

| **Data quality** (Redman, 1996; Wang and Strong, 1996; Batini and Scannapieco, 2006; ISO, 2008) | **Information quality** (DeLone and McLean, 1992; Eppler, 2001; Rieh, 2002; Nelson, Todd and Wixom, 2005; Nicolaou and McKnight, 2006; Laumer, Maier and Weitzel, 2017; Fadahunsi *et al.*, 2019) |
|---|---|
| Accuracy | Accuracy |
| Believability, credibility, reputation | Reliability |
| Currentness, timeliness | Currentness, timeliness |
| Objectivity | |
| Value-added | Usefulness, usability |
| | Importance |
| Relevancy | Relevancy |
| Completeness | Completeness |
| Appropriate amount of data, volume | |
| | Goodness |
| Understandability, ease of understanding, interpretability | Understandability, interpretability, comprehensive, presentation |
| Consistency | Consistent |
| Concise representation | Format, concise |
| Accessibility | |
| Access security, confidentiality | Privacy, confidentiality, secure access |
| Provenance | Provenance |
| Granularity | |
| Efficiency | |
| Precision | |
| Compliance | |
| Traceability | |
| Availability | |
| Recoverability | |
| Portability | |

As Table 11 shows, several quality characteristics are used for data and information quality. *Currentness* in data may refer to when the data is stored, but it also refers to when the information has been observed. Similarly, *completeness* of data may mean that database has no missing values. The missing values can be replaced with null values or some other values that provide no information, making information incomplete.

The characteristics collected by Wang and Strong (1996) are used multiple times to define data quality (Strong, Lee and Wang, 1997; Fehrenbacher and Helfert, 2008; Ghasemaghaei and Calic, 2019) and information quality (Lee *et al.*, 2002; Wigand, Wood and Yiliyasi, 2009; Foley, Helfert and Elwood, 2010) without making any distinction between the two.

### 4.3.2 **Objective**

This research aims to find quality characteristics applicable in the UGC domain to answer the research question, "*What are the data quality characteristics of user-generated content?"*

The research aims to select and define a comprehensive set of quality characteristics for UGC. A collection of characteristics is chosen from existing literature within the UGC domain. Each characteristic is cross-examined against various UGC platforms to determine whether the characteristic is essential within the UGC domain.

Table 12 presents the list of initial characteristics selected from Table 11 that are examined within the research. The platforms used to explore the UGC domain are:

- Twitter (Twitter, 2020): Twitter is a popular social media platform where users share short texts called tweets. Tweets can contain multimedia, and other users can comment and like them. Each tweet can have hashtags that work like keywords when searching specific topics.
- Wikipedia (Wikipedia, 2020): Wikipedia is an online encyclopedia maintained and moderated by a large community. More reputable members are chosen as moderators. Wikipedia has a set of guidelines each user should follow, and moderators check and enforce these rules.
- Atlas of Living Australia (ALA) (Atlas of Living Australia, 2021): ALA is an Australian plant and wildlife monitoring platform. ALA uses single observations and datasets from citizens and organizations alike. Singular observations need to be sent through iNaturalist Australia, a citizen science platform integrated with ALA.
- Worldometer (Worldometer, 2020): Worldometer is a crowdsourcing platform that collects information from multiple sources. These sources include user-sent information, news, and organizations. Worldometer is widely referenced as a real-time information provider, especially during the COVID-19 epidemic.

- YouTube (YouTube, 2020): YouTube is a video-sharing platform where anyone can view public videos. Registered users can upload new videos of any type as long as the content does not go against platform's the terms of service. YouTube has similar traits as Twitter, but the significant difference is the content type within the platform.

Each platform is investigated to determine whether a characteristic in the following Table 12 is appropriate for the platform or not. The characteristics that are used in most of the platforms are then defined for the UGC domain.

**Table 12**. Examined quality characteristics

| **Quality characteristics** | | |
|---|---|---|
| Completeness | Compliance | Confidentiality |
| Consistency | Credibility | Currentess |
| Granularity | Objectivity | Precision |
| Privacy | Provenance | Relevance |
| Semantic accuracy | Syntactic accuracy | Timeliness |
| Traceability | Understandability | Usability |
| Value | Volume | |

### 4.3.3 Relation to Dissertation's Research Question

This publication answers the second sub-question, "*How to define quality characteristics and distinguish data and information quality in the domain of user-generated content?*" by first reviewing and selecting quality characteristics for UGC. The results provide an extensive but not exhaustive list of quality characteristics for the UGC domain and evaluate the reviewed platforms' content quality using the defined quality characteristics.

### 4.3.4 Research Output and Contribution

The primary focus of the research is the selection and definition of quality characteristics appropriate for UGC. Based on the cross-examination of characteristics and UGC platforms, some of the characteristics listed in Table 12 are not essential for UGC. Table 13 shows the results of the cross-examination with the following values:

- 1: Platform considers the characteristic when users submit information.
- 0: Platform does not consider the characteristic when users submit information.
- ?: Unclear if the platform considers the characteristic when a user submits information.
- +/-: Situation-dependent if platform considers the characteristic.

**Table 13.** Characteristics mapped to UGC platforms

| Quality characteristics | ALA | Twitter | World-ometer | Wiki-pedia | You Tube | Explanations of the characteristics |
|---|---|---|---|---|---|---|
| Syntactic accuracy | 1 | 0 | 1 | 0 | 0 | User submits information in the syntax expected by the system |
| Semantic accuracy | 1 | 0 | 1 | 1 | 0 | User submits information that follows semantic rules set by the system |
| Completeness | 1 | 0 | 1 | 1 | 0 | The system expects the user to submit a minimum amount of information |
| Consistency | 0 | 0 | 0 | 0 | 0 | Information is consistent in comparison to multiple users input |
| Credibility | 1 | 1 | 1 | 1 | 1 | User's credibility |
| Objectivity | 1 | 0 | 1 | 1 | 0 | User submits objective information |
| Precision | 1 | 0 | 1 | +/- | 0 | Information is detailed |
| Volume | 1 | 1 | 1 | 1 | 1 | Similar information from different sources |
| Compliance | ? | ? | ? | ? | ? | Information is compliant with a standard |
| Currentness | 1 | 1 | 1 | 1 | 1 | Information is current |
| Timeliness | +/- | 0 | 0 | 0 | 0 | Information is from the correct time |
| Privacy | 1 | 1 | 0 | 0 | 1 | Personal information is not displayed |
| Relevance | 1 | 1 | 1 | 1 | 1 | User submits relevant information to the topic |
| Usability | 1 | +/- | 1 | 1 | +/- | Information is usable by others |
| Value | 1 | +/- | 1 | 1 | +/- | Information has value for others |
| Confidentiality | 0 | 0 | 0 | 0 | 0 | Sensitive information is inaccessible |
| Granularity | +/- | 0 | 0 | 0 | 0 | Information is split into specific parts |
| Traceability | 1 | 1 | 1 | 1 | 1 | Information origins are known |
| Provenance | 0 | 0 | 0 | 0 | 0 | Changes to information are known |
| Understandability (or readability) | 1 | 1 | 1 | 1 | 1 | Information is understandable (or readable) |

Table 13 shows that social media platforms such as YouTube and Twitter care less about information's correctness than other UGC platforms. Wikipedia, Worldometer, and citizen science platforms have minimum requirements for the content to be accepted on the platform. Social media platforms are designed to share subjective content, and the lack of consideration for correctness demonstrates this purpose. However, there are many restrictions placed on social media content that are not visible to users, such as the policies of the platform owner. Due to these restrictions, content may be hidden or removed from the platform without warning. ALA, Wikipedia, and Worldometer share more objective

and factual information, so the platforms put more effort into ensuring the correctness of information than social media platforms do.

Based on the results shown in Table 13, some characteristics can be considered non-essential for UGC because none of the reviewed platforms value them. These characteristics are *consistency, compliance, timeliness, confidentiality, granularity,* and *provenance.* The rest of the characteristics are defined in Table 14.

**Table 14.** Quality characteristics for UGC

| Characteristic | Definition |
|---|---|
| Traceability | How well the content is attributed to a specific source and time |
| Credibility | How credible the content is based on who is providing the content |
| Currentness | How promptly content is updated with respect to changes occurring in the real world |
| Relevance | How relevant the given content is to the platform context |
| Accuracy | Accuracy is the closeness of the given content to the expected content. Based on syntactic and semantic accuracy |
| Syntactic accuracy | Closeness of the content syntax that the user provides, depending on the platform context |
| Semantic accuracy | How correctly the information within the content matches the real-world facts |
| Completeness | How complete the content is and whether or not it is missing important information, depending on the platform context |
| Usability | How usable the content is based on the platform context. It is affected by accuracy, completeness, and credibility |
| Value | How useful the content is and whether it provides advantages from its use |
| Understandability (and readability) | How easily the information from the content can be comprehended without ambiguity by a human consumer within the platform context (and how easy written text is to read and comprehend) |
| Objectivity | How unbiased and impartial the content and its information are |
| Privacy | How much of the user's personal information is concealed |
| Volume | The amount of similar information given by multiple users |
| Precision | How detailed the provided content is in the platform context |

The characteristics in Table 14 can be defined in terms of data and information quality characteristics for UGC. Information is defined as the content received from users, and data is defined as content stored in the database. Only precision is situation-dependent. In addition to defining the appropriate quality characteristics for UGC, the characteristics are used to evaluate each described platform's quality compared to the WalkingPaths platform developed in (Musto and Dahanayake, 2021a) using RapidMiner queries. WalkingPaths is a UGC platform prototype designed to integrate the quality characteristics in Table 14 to assess how they could improve a platform's content.

The following datasets are used to compare against WalkingPaths:

- Twitter: A dataset containing tweets related to COVID-19 is used, containing 6012 entities.
- Wikipedia: A dataset of different articles is used, containing 19,797 entities.
- Woldometer: A dataset consisting of COVID-19 data from two weeks is used, containing 2,996 entities.
- YouTube: A dataset of videos mentioning COVID-19 is collected, and it contains 750 entities.

Table 15 presents the quality of WalkingPaths and other UGC platforms using the RapidMiner queries shown in Table 18.

**Table 15**. WalkingPaths and UGC platforms

| Characteristic | WalkingPaths 108 entities | Twitter 6012 entities | YouTube 750 entities | Wikipedia 19797 entities | Worldometer 2996 entities |
|---|---|---|---|---|---|
| Syntactic accuracy | 1.00 | 1.00 | 1.00 | 0.96 | 1.00 |
| Semantic accuracy | 0.96 | 0.93 | NA | 1.00 | 1.00 |
| Completeness | 1.00 | 0.89 | 0.99 | 0.95 | 1.00 |
| Credibility | 0.74 | 0.32 | 0.82 | 0.32 | NA |
| Objectivity | 0.54 | 0.19 | 0.11 | 0.50 | NA |
| Volume | 0.36 | 0.61 | 0.69 | NA | NA |
| Currentness | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Privacy | 1.00 | 0.67 | 1.00 | 1.00 | 1.00 |
| Relevancy | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Usability | 1.00 | NA | NA | 0.85 | 1.00 |
| Value | 0.95 | 0.68 | 0.47 | 0.81 | 0.83 |
| Traceability | 1.00 | 0.66 | 0.67 | 0.67 | 1.00 |
| Understandability | 1.00 | 0.82 | NA | 0.72 | 1.00 |

Overall, WalkingPaths scores similarly to Worldometer. Worldometer has perfect scores in most categories except *value*, which is affected by *credibility*. Compared to other UGC platforms, WalkingPaths is better in *objectivity* and *credibility* than most of the platforms. *Credibility* in social media platforms is based on the number of followers, and *objectivity* is based on positive/negative interactions with the content. However, these are not perfect indicators of *objectivity* and *credibility* and could easily be manifestations of the echo chamber effect (Cinelli *et al.*, 2021). That is why social media platforms' credibility and objectivity scores should be based on something more concrete to give a more accurate and impartial result.

The list provides answers to the research question "*What are the quality characteristics of user-generated content?*" presented in the publication and gives them a definition appropriate for the UGC domain. The characteristics and definitions in Table 14 can be used for evaluating and improving the data and information quality of new and existing UGC platforms. Quality characteristics depend on the domain, and each field needs to define the essential quality characteristics. The quality of UGC is a significant concern when utilizing the content, but the quality can be evaluated and improved with proper definitions. Improving the quality of UGC brings benefits for those who aim to use it.

The presented research has some limitations.

1. Only a limited number of platforms have been examined.
2. The list of quality characteristics is extensive but not exhaustive.

This publication provides an answer to the second sub-question, "*How to define quality characteristics and distinguish data and information quality in the domain of user-generated content?*" of this dissertation.

## 4.4 Publication IV: An approach to improve the quality of user-generated content of citizen science platforms

This section presents the overview of the following publication:

Musto, J. and Dahanayake, A. (2021). An approach to improve the quality of user-generated content of citizen science platforms. ISPRS International Journal of Geo-Information, 10, pp. 434.

### 4.4.1 Research background

UGC is content generated by the general public. Some argue that a person must generate the content, and content created mainly with technology, such as sensors, is not valid. For example, when sensors collect the information and a user is only responsible for placing it, the user technically does not generate the content (Krumm, Davies and Narayanaswami, 2008; Kietzmann *et al.*, 2011; Wyrwoll, 2014).

Citizen science and crowdsourcing platforms impose restrictions on what information users can submit, while social media platforms permit much more. However, even in the case of social media, not everything is allowed. Each platform has its terms of service and other guidelines users need to follow, and other people may report the content to be removed based on copyright claims or the harmful nature of the content (Peters, 2020; Quinn, 2020).

Many businesses and researchers use UGC for various purposes such as monitoring wildlife (Steger, Butt and Hooten, 2017), astronomy (Lintott *et al.*, 2010; Simpson, Page and De Roure, 2014), catastrophe tracking (Middleton, Middleton and Modafferi, 2014; Ahmouda, Hochmair and Cvetojevic, 2019), influenza tracking (Culotta, 2010; Signorini, Segre and Polgreen, 2011), and opinion mining (Mariani, Di Fatta and Di Felice, 2019; Ranjan, Sood and Verma, 2019).

As the number of users using these UGC platforms grow, the weight and usefulness of data and information on these platforms increases. The quality of such content is a significant concern among people who utilize it (Elbroch *et al.*, 2011; Lukyanenko and Parsons, 2011; MacKechnie *et al.*, 2011; Hunter, Alabri and Van Ingen, 2013; Brown and Kyttä, 2014; Sheppard, Wiggins and Terveen, 2014; Cai and Zhu, 2015; Hecht and Spicer Rice, 2015; Kaur *et al.*, 2018). There is no simple way to determine if the data and information are of high quality, if the user who provides the content is credible, or if the content is even authentic. The quality is easily overlooked in UGC (King, Racherla and Bush, 2014; Bordogna *et al.*, 2016; Haworth *et al.*, 2018; Bayraktarov *et al.*, 2019). Improving the quality of data and information would benefit platform owners, users, and third parties aiming to utilize the content.

Batini et al. (2009) organize data and information quality improvement methodologies into *data-driven* and *process-driven*, where *data-driven methods* include gathering new

data or selecting credible sources. *Process-driven methods* redesign or control the collection process. Bordogna et al. (2016) divide the techniques into before (ex-ante) and after (ex-post) content is collected. Combining these two classifications gives an overview of possible ways to improve data and information, as presented in Table 16.

**Table 16**. Quality improvement methods

| | **Data-driven** | **Process-driven** |
|---|---|---|
| **Ex-ante** | Gather new data<br>Selecting credible sources<br>Training users<br>Using standards<br>Moderating | Template<br>Automatic error checking<br>Using reputation models<br>Gamification<br>Expectations on users |
| **Ex-post** | Replace low-quality data<br>Correcting errors in existing data<br>Outlier identification<br>Data comparison | Enrichment<br>Cross-referencing<br>User ranking |

Various platforms use many of the methods listed in Table 16. Most UGC platforms have some sort of *moderation* in place conducted by the community or by selected individuals. *Correcting errors in existing data* or *replacing low-quality data* are often tied to *moderation*.

While the techniques listed in Table 16 are proven to work in many cases, several problems need to be carefully evaluated before using them.

- Training users requires time and money and is often not a feasible option.
- Enrichment and cross-referencing need additional information from other sources to compare against, which may not be available.
- Outlier detection and data comparison can be made against the collected data, but it requires a decent amount of data to be used.
- Selecting credible sources is not a sustainable option in UGC.
- Replacing low-quality data means that there will be less data available.
- Moderating requires an additional workforce.
- Using standards requires there to be standards to be utilized and selecting the most appropriate. Not all domains have standards available.
- Gamification is more related to motivating rather than being able to collect valid information.

*Templates*, *automatic error checking*, and *reputation models* are the best options for UGC platforms. Still, there are no given guidelines on utilizing *templates* or *automatic error checking. Error checking* could be tied to *templates* to make it easier for developers to design the platforms. Most error-checking in existing platforms is about accepting the correct type of information or seeing whether a given time has passed. The automatic checks could be further extended and be based on actual quality characteristics. The

information quality could be improved by improving the data and information collection process using quality characteristics for the template and error checking.

### 4.4.2 Objective

The research presented aims to provide a new method to improve the data and information quality of UGC platforms. By integrating data and information quality characteristics into the platform's design, the information collection process can be modified to improve the quality of information.

Several methods for improving data and information quality in UGC platforms involve users and rely on knowledgeable content providers. There are multiple downsides to such practices, for example, the cost and time, the credibility of other users, and the need for a larger community (Bordogna *et al.*, 2016; Connelly *et al.*, 2016; Fidler and Lavbic, 2017). Another popular method for improving the quality of information is completely removing misinformation used by social media platforms. The process relies on artificial intelligence, and the downside is the possibility of eliminating valid information. For example, YouTube has removed millions of videos from its platform using artificial intelligence, but some videos did not violate any terms of service (Lyons, 2020).

The following quality characteristics obtained from the results in (Musto and Dahanayake, 2021b) are integrated into the design of a platform that collects walking path condition information from the public using these paths:

- Syntactic accuracy
- Semantic accuracy
- Objectivity
- Credibility
- Relevancy
- Value
- Usability
- Currentness
- Completeness
- Volume
- Understandability
- Privacy
- Traceability

The platform is developed using ReactJS (https://reactjs.org/, retrieved 29th June 2021) as the client layer and MongoDB (https://www.mongodb.com/, retrieved 29th June 2021) as the database.

A platform has three parts where quality characteristics are integrated.

1. The client layer consists of the user interface.
2. The server layer consists of the system that operates between the database and the client layers.
3. The database layer consists of the database management system, database, and data model.

Improving the data model can improve the quality of the content because the data model is the basis of how the content is stored (Fox *et al.*, 1999). While some characteristics, such as *accuracy* and *completeness*, can be integrated into the data model, other characteristics, such as *understandability* and *relevancy*, require more information than what the data model can provide. Quality characteristics can be integrated outside the data model and into the system architecture through various checks. These checks can be used to control the content within the user interface and increase the quality (Musto and Dahanayake, 2019).

The classification proposed by Bordogna et al. (2016) is based on the content collection process that includes *before* and *after collection* phases. The process can be extended to cover two additional steps: *during collection* and *presenting the information.* The final version of the content collection process is shown in Figure 7.

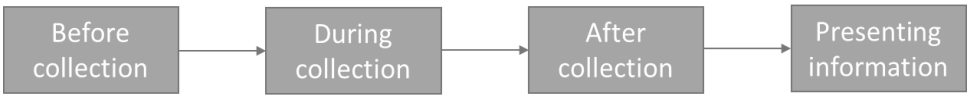


**Figure 7**. Enhanced quality improvement strategy categorization

The quality characteristics can be organized into groups based on Figure 7. Each group is then integrated into the design of the UGC platform based on the following:

- Before collection: Quality characteristics in this group are integrated into the data model and server. The data model is responsible for providing restrictions based on the quality characteristics, and the server can be used to enforce them.
- During collection: Quality characteristics in this group are integrated into the user interface. The user interface is responsible for collecting information related to the characteristics and automatically ensuring that the given information satisfies the requirements for each characteristic.
- After collection: Quality characteristics in this group are integrated into the server. The server is responsible for checking the content before storing it in the database and calculating or adding any relevant metadata.
- Presenting information: Quality characteristics in this group are integrated into the user interface. The characteristics that affect how information is shown on the user interface are essential.

Table 17 maps the quality characteristics listed previously into the four categories.

**Table 17.** Grouping quality characteristics to the collection process phases

| Characteristic | Before collection (data model, server) | During collection (user interface) | After collection (server) | Presenting information (user interface) |
|---|---|---|---|---|
| Syntactic accuracy | x | x | | |
| Semantic accuracy | x | x | | |
| Objectivity | | | x | |
| Credibility | | x | | |
| Relevancy | | x | | |
| Value | | | x | |
| Usability | | | x | |
| Currentness | | x | | |
| Completeness | | x | x | |
| Volume | | | x | x |
| Understandability | | | | x |
| Privacy | | x | | x |
| Traceability | | x | | |

The categories in Table 17 show the phases of the information collection process in which each characteristic should be carefully considered and integrated. Based on Table 17, the WalkingPaths platform is designed and developed. Figure 8 shows the platform's database schema using a snowflake model (Teorey et al., 2011).

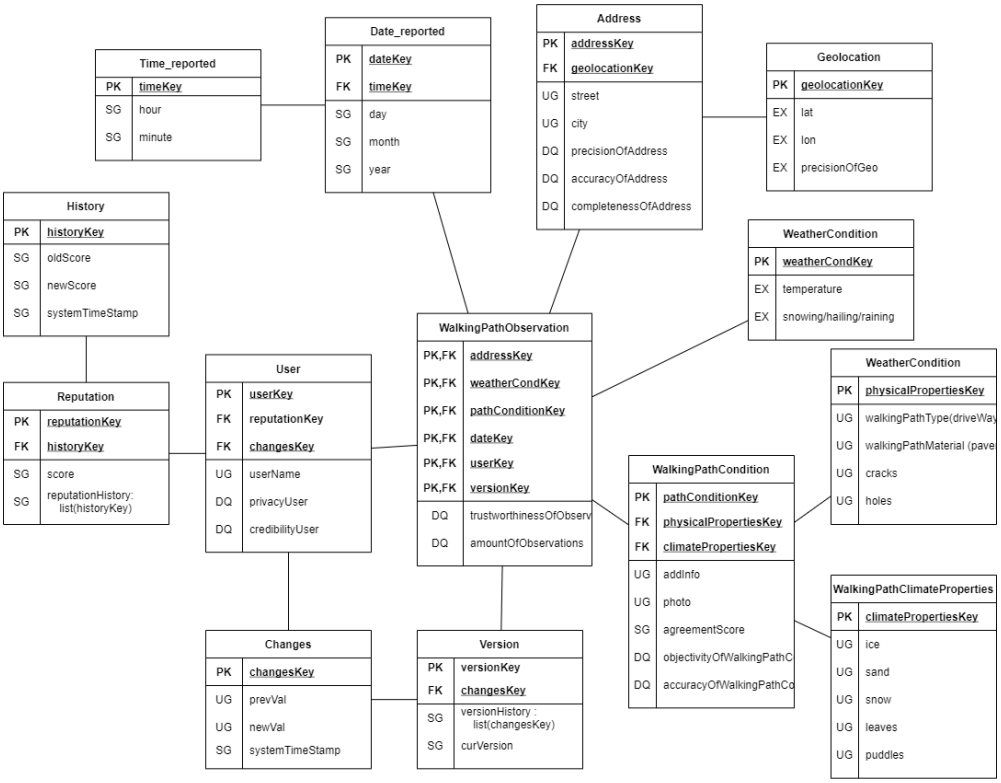


**Figure 8.** Snowflake schema for WalkingPaths

In Figure 8, each attribute has some identifier based on the following:

- PK/FK: Primary key, foreign key
- SG: System generated
- UG: User generated
- DQ: Data quality characteristics

The quality characteristics in the data model are completeness, precision, accuracy, volume, credibility, privacy, objectivity, and traceability. The attributes store the evaluations of each quality characteristic and present them to content users. Figures 9, 10, and 11 present images of the user interface. Figures 9 and 10 show the transition from one view to another using the navigation bar.

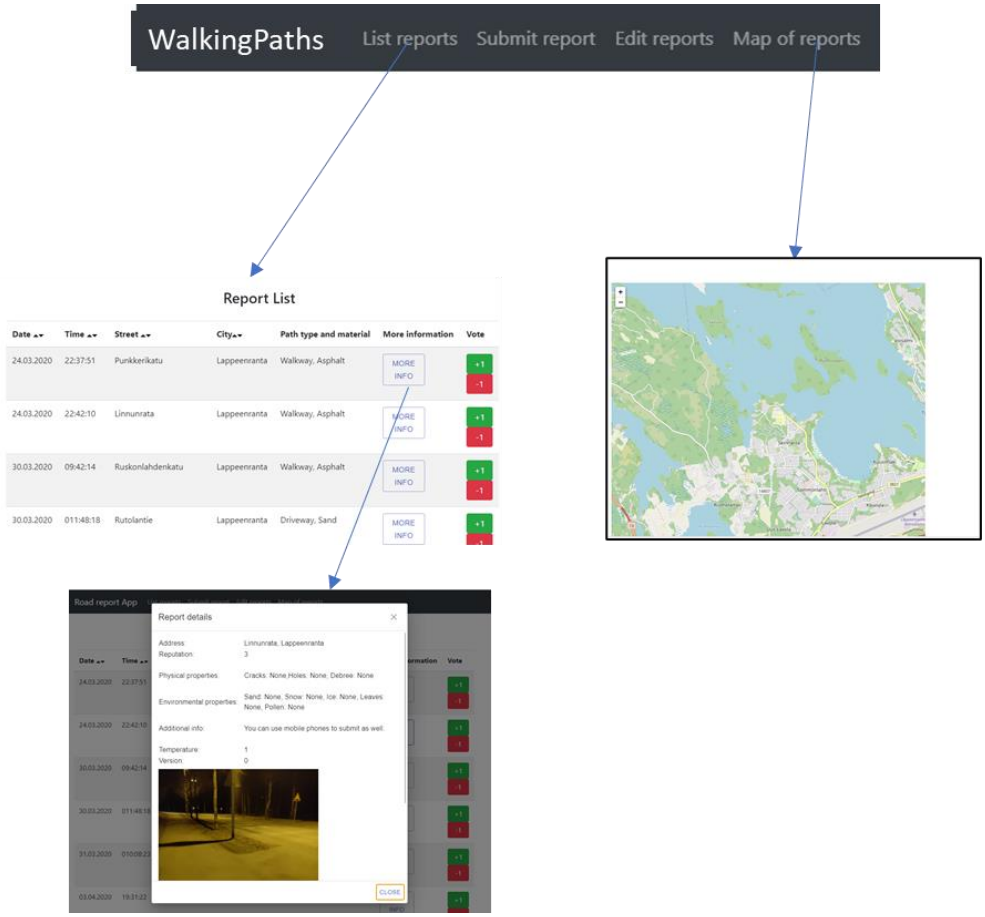


**Figure 9.** List, map, and additional information views

The report list in Figure 9 presents only minimal information, and other relevant information, such as an image, can be viewed by opening the *More info* pop-up. The report list includes up-/downvote buttons that can increase or decrease the validity of a given report.

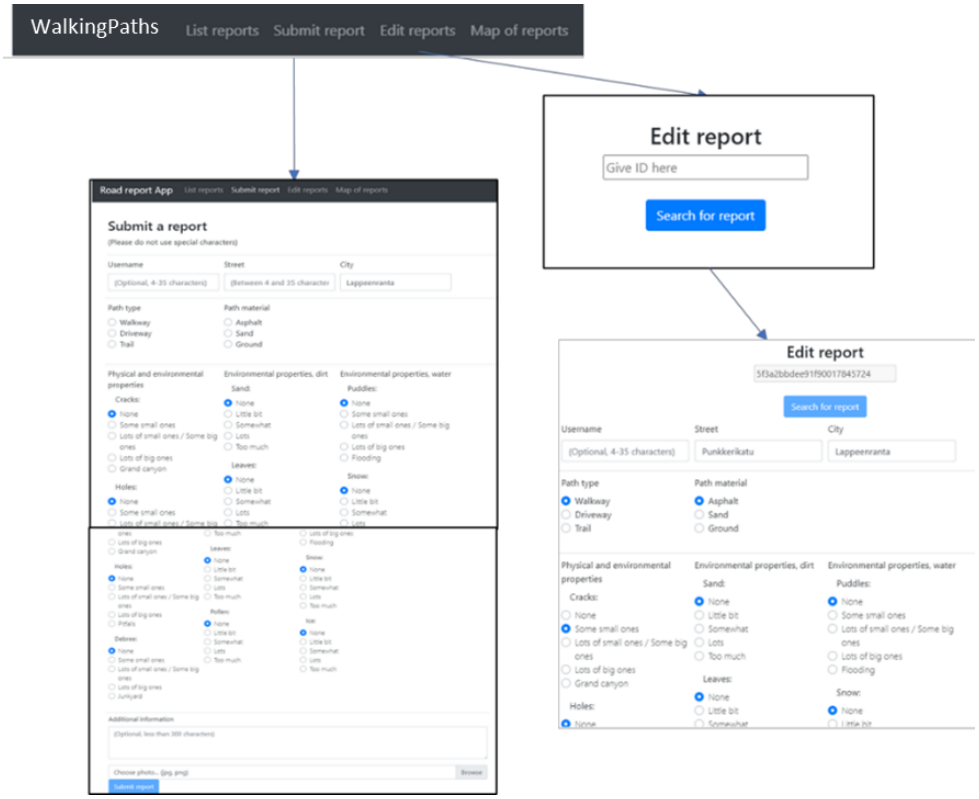


**Figure 10.** Submit and edit views

Figure 10 presents the submit and edit views. As no registration is required on the platform, each content provider is given the ID of a report after submitting it. Previously sent reports can be edited with a correct ID that fetches the current information from the database.

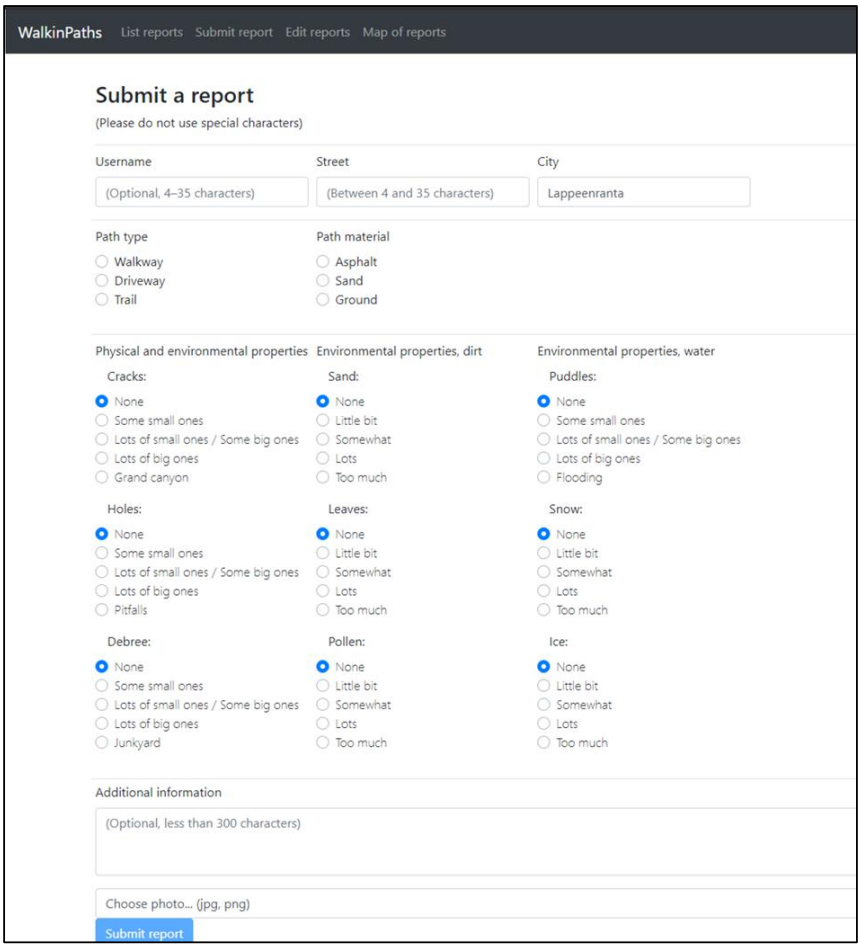


**Figure 11.** Submit report view

Figure 11 shows a more detailed version of the report submission. Most of the information is provided through choice boxes to increase the quality of content and guide the content provider to look for specific details. Content providers can send *Additional information* in the optional text field at the end when necessary.

The quality of data and information for *WalkingPaths* and other UGC platforms are evaluated using the quality characteristics in Table 17. Information quality is assessed through the user interface of each platform. A detailed assessment of information quality through the user interface is complex, so only approximate results are given. Data quality is evaluated using datasets from each platform. Each dataset is subjected to specific queries related to the quality characteristics using a mining and data analysis tool, RapidMiner (https://rapidminer.com/, retrieved 29th June 2021). RapidMiner is a commercial software designed for data mining, analytics, and machine learning. Table 18

presents the general RapidMiner queries for the quality characteristics. Many of the data mining processes are based on the quantitative measures presented in Table 21.

**Table 18.** General RapidMiner queries for DQ characteristics

| Charact eristic | General query | (Data mining) Technique |
|---|---|---|
| Syntactic accuracy | Data entities correspond to the expected syntax and format defined in the dataset. This information is based on the headers and what data is expected, and in what form. | Text/content mining. Compare value syntax to expected (integer, string, date) and filter out incorrect values. Compare the number of correct values to the total number. |
| Semantic accuracy | Data is semantically correct compared to what is expected based on the headers. | Value comparison. Headers define what data should be, for example, "date," "name," "country." Each value is checked to see if they are actual dates, countries, names. |
| Completeness | Each dataset is checked for missing values for completeness. | Filter missing values and compare the amount to total (automated functionality). |
| Credibility | The credibility of the content provider giving the information. | Reputation model and calculation compared to the average score |
| Objectivity | Objectivity is based on how objective given information is. If multiple sources agree on the information, it is more likely to be objective. | Count how many entities from different sources/content providers have the same information and how many are only from singular content providers/sources. |
| Volume | For each dataset, the volume is checked from the number of similar data entities compared to all entities. The similarity is only based on a few attributes. | Count how many entities from different sources/content providers have relatable information based on selected attributes and how many are only from singular content providers/sources. |
| Currentness | Data has given a date/time. Compare that to the time data was extracted from the database. | Content mining and comparison. |
| Privacy | Privacy is measured based on the amount of personal information stored with the data. | Filter out content providers whose possible real names are given and compare them to the total amount (text mining). |
| Relevancy | The relevance of the data to the given context regardless of whether the data is correct or not. | Data comparison to a given relevance factor such as the topic. |
| Usability | Usability is based on the context of usage for each dataset. | Content mining and comparison. |
| Value | Value depends on the user. In this research, value = (Syntactic + Semantic + Credibility + Relevancy + Usability + Understandability) / 6 | Calculation based on other characteristics. |
| Traceability | Each dataset provided attributes for time, location, and content provider, which are checked for traceability. | Count how many entities have a valid time, location, and content provider/source compared to all entities. |
| Understand-ability | Understandability is based on the content of information; in general, readability. Unreadable texts/characters and undefined acronyms reduce the understandability. | Text mining of invalid words. |

### 4.4.3 Relation to Dissertation's Research Question

The presented publication shows how data and information quality can be improved by integrating quality characteristics into the platform's design. This research contributes to the third and final sub-question, "*How does the introduction of data and information*

*quality characteristics into information collection and data curation processes influence the quality of user-generated content?"* The quality of data and information can be increased by improving the information collection and data curation processes based on the chosen quality characteristics. The results show a considerable quality improvement and require less effort from moderators and platform owners afterward. Some changes to the information collection process can reduce and restrict users' freedom, meaning there is a possible trade-off between freedom and quality. Still, the quality can be improved to a certain level by guiding the users to provide higher quality content without restricting their freedom. Introducing data and information quality characteristics into the collection and curation processes positively influences the overall quality without hindering other aspects of the UGC platform.

### 4.4.4 Research Output and Contribution

This research aims to improve data and information quality in UGC platforms by integrating data and information quality characteristics into the design. Incorporating quality characteristics into the design alters the information collection and data curation processes to procure higher quality information from content providers.

The quality of data and information in WalkingPaths is evaluated against other citizen science platforms. The following citizen science platforms are used:

- ALA (Atlas of Living Australia, 2021) is an Australian wildlife and environment data collection platform that collects data from organizations or citizens. Data can be sent as datasets or as singular observations through the iNaturalist Australia integration. A dataset of lace-monitor lizards is downloaded, and it consists of 14,138 entities.
- iNaturalist (iNaturalist, 2021) is a global network of websites operating in multiple countries. iNaturalist provides a platform template that can be utilized and integrated with local environment observation platforms. A dataset of great tit birds is downloaded, and it consists of 39,910 entities.
- Globe at Night (Globe at Night, 2021) is an international citizen science project that collects night sky brightness information and has been going on since 2006. A dataset from 2020 is downloaded, and it consists of 29,507 entities.
- Budburst (Budburst, 2021) is a citizen science project managed by the Chicago botanic garden. The project operates within the United States and observes plants and pollinators. A dataset from Budburst is used for quality evaluation, and it consists of 96,815 entities.

Table 19 presents the data quality of WalkingPaths and other citizen science platforms. These results are achieved using the RapidMiner queries shown in Table 18. The values are between 0 and 1, reflecting the percentage of results.

**Table 19**. WalkingPaths and citizen science platforms

| Characteristic | WalkingPaths 108 observations | ALA 14138 observations | iNaturalist 39910 observations | Globe at Night 29507 observations | Budburst 96815 observations |
|---|---|---|---|---|---|
| Syntactic accuracy | 1.00 | 0.71 | 0.89 | 1.00 | 0.99 |
| Semantic accuracy | 0.96 | 0.80 | 0.90 | 1.00 | 1.00 |
| Completeness | 1.00 | 0.71 | 0.73 | 0.87 | 0.33 |
| Credibility | 0.74 | NA | NA | NA | NA |
| Objectivity | 0.54 | 0.29 | 0.56 | NA | NA |
| Volume | 0.36 | 0.70 | 0.73 | NA | NA |
| Currentness | 1.00 | 0.44 | 0.99 | 1.00 | 0.80 |
| Privacy | 1.00 | 0.80 | 0.98 | 1.00 | 1.00 |
| Relevancy | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Usability | 1.00 | 0.96 | 0.83 | 1.00 | 0.87 |
| Value | 0.95 | 0.74 | 0.72 | 0.78 | 0.79 |
| Traceability | 1.00 | 0.91 | 0.90 | 0.86 | 0.70 |
| Understandability | 1.00 | 0.97 | 0.69 | 0.65 | 0.86 |

Table 19 shows that the overall quality of WalkingPaths is higher than the compared platforms. WalkingPaths achieves better results in most of the characteristics. WalkingPaths is the only platform with a score in *credibility* because it is the only platform that provides a reputation score tied to the users in the dataset. iNaturalist discusses using reputation models, but this is not evident in the dataset. *Volume* is the only characteristic where WalkingPaths performs considerably worse than its counterparts, but this is expected because it is relatively new and short-lived compared to the other platforms. Traceability in other citizen science platforms is reduced because dates and times are missing from the data.

*Semantic accuracy* is reduced in WalkingPaths because of misspelled street names. The mistakes in street names could be resolved by providing a comprehensive list of available cities and suggesting street names during input, similar to how Google Maps operates. However, if the platform is expanded outside of one country, the list of cities and streets will inflate drastically. It can be argued that ALA, iNaturalist, and Budburst perform worse because they collect different kinds of information from WalkingPaths and Globe at Night. Still, the techniques used to develop WalkingPaths can be utilized in any platform that collects any type of observation. As such, the difference in observation types is negligible as long as the underlying principle stays the same.

This publication contributes to the academic community by providing design guidelines for improving data and information quality. The design approach could be improved and extended outside the domain of UGC. Additionally, the presented research also contributes to the field of UGC by presenting a framework for improving data and information quality in new and existing UGC platforms.

The final sub-question, "*How does the introduction of data and information quality characteristics into information collection and data curation processes influence the quality of user-generated content?"* is answered with this publication.

## 4.5 Summary

Section 4 presents four publications that are used to answer the sub-questions defined in Section 1.1:

- SRQ1: *What information collection features in user-generated content platforms influence the quality of content?*

SRQ1 is answered through the results of (Musto and Dahanayake, 2018) and (Musto and Dahanayake, 2020). (Musto and Dahanayake, 2018) presents the common issues from literature, and (Musto and Dahanayake, 2020) examines the field by surveying existing UGC platforms. (Musto and Dahanayake, 2018) is part of the rigor cycle in DSR and surveys the existing knowledge base. (Musto and Dahanayake, 2020) steps into the relevance cycle by combining the literature with practice and presenting relevant data and information quality problems that must be resolved.

- SRQ2: *How to define quality characteristics and distinguish data and information quality in the domain of user-generated content?*

(Musto and Dahanayake, 2021b) examines existing data and information quality research and develops proper definitions specifically for the UGC domain. The publication is part of the design cycle by designing and building an artifact: *the definitions of data and information quality characteristics.* The definitions are evaluated in (Musto and Dahanayake, 2021a).

- SRQ3: *How does the introduction of data and information quality characteristics into information collection and data curation processes influence the quality of user-generated content?*

(Musto and Dahanayake, 2021a) develops and evaluates a new UGC platform design that integrates quality characteristics into the platform. (Musto and Dahanayake, 2021a) is part of the design cycle of DSR developing a new artifact that is evaluated in the publication.

# 5 Scientific contribution

This dissertation provides three scientific contributions that influence the overall data and information quality of UGC:

1. A comprehensive set of data and information quality characteristics defined specifically for UGC.

There are no existing data and information quality characteristic definitions for the UGC domain. Definitions used in the UGC domain are taken from other domains, creating a mismatch between the domain and characteristics.

2. Extension of the UGC platform's development life cycle by using UGC data and information quality characteristics during the platform's requirements acquisition stage to improve the quality of the content collection.

Extending the UGC platform's development lifecycle with data and information quality characteristics makes the necessary quality requirements easier to consider and manage when developing the platform.

3. Framework to store and assess the reliability and quality of UGC using quality characteristics.

Many current methods revolve around improving data and information manually after collecting the content or improving content providers' knowledge by selecting or training them (White *et al.*, 2014; Wu, Li and Wang, 2018; Barachi *et al.*, 2019). These practices require resources and burden content providers. This research provides an alternative solution to improve the quality of data and information. The proposed methodology reduces the need for external resources (i.e., community) and helps increase and maintain a desired level of quality.

## 5.1 Data and information quality

### 5.1.1 Existing data and information quality definitions

Within the domain of UGC, there are no proper data and information quality definitions. Most researchers base their definitions on existing research without adequately considering the context of the applied quality characteristics (Immonen, Pääkkönen and Ovaska, 2015; Spielhofer *et al.*, 2017; Arolfo and Vaisman, 2018; Arthur *et al.*, 2018). UGC is vastly different compared to traditional content and the existing definitions are not adequate. There exist over fifty quality characteristics to choose from, but most characteristics are overlapping and only a portion are chosen and used at a time. Each chosen characteristic should be modified to match the specific domain (Redman, 1996; Wang and Strong, 1996; Batini and Scannapieco, 2006; ISO, 2008).

Table 20 presents four lists of data quality characteristics from well-known literature. Each has been used to define data quality in their respective fields. Some characteristics repeat, and some use a different term to mean the same thing. In the leftmost column, the characteristics are described with the chosen term used in this dissertation.

**Table 20**. Data quality characteristics from literature

| **Chosen term** | **Batini** (Batini and Scannapieco, 2016) | **ISO** (ISO, 2008) | **Redman** (Redman, 1996) | **Wang** (Wang and Strong, 1996). |
|---|---|---|---|---|
| Accuracy | Accuracy | Accuracy | Accuracy | Accuracy |
| Completeness | Completeness | Completeness | Completeness | Completeness |
| Consistency | Consistency | Consistency | Consistency | |
| Credibility | Trust | Credibility | | Believability / Reputation |
| Currentness | Currentness / Timeliness | Currentness | Currency | Timeliness |
| Accessibility | Accessibility | Accessibility | | Accessibility |
| Usability | Usefulness | | | |
| Relevancy | | | Appropriateness | Relevancy |
| Understandability | Readability | Understandability | Interpretability | Ease of understanding / Interpretability |
| Redundancy | Redundancy | | | |
| Efficiency | | Efficiency | Efficient use of memory | |
| Representational consistency | | | Representation consistency | Representational consistency |
| Privacy | | Confidentiality | | Access security |
| Portability | | Portability | Portability | |
| Precision | | Precision | Format precision | |
| Compliance | | Compliance | | |
| Traceability | | Traceability | | |
| Availability | | Availability | | |
| Recoverability | | Recoverability | | |
| Ability to represent null values | | | Ability to represent null values | |
| Format flexibility | | | Format flexibility | |
| Objectivity | | | | Objectivity |
| Value | | | | Value-added |
| Volume | | | | Appropriate amount of data |
| Concise representation | | | | Concise representation |

Each characteristic in Table 20 requires a definition before its fitness for UGC can be evaluated. Table 21 presents existing definitions for the characteristics shown in Table 20.

**Table 21.** Data quality characteristic definitions

| **Data quality characteristic** | **Definition** |
|---|---|
| Accuracy | The closeness between data values $v$ and $v_0$, where $v_0$ is the correct representation of what the data value $v$ aims to represent. Based on syntactic and semantic accuracy (Batini and Scannapieco, 2006). |

| | |
|---|---|
| Syntactic accuracy | The closeness of words in the text to a reference vocabulary. *K* is the number of words, $w_i$ is a word in the text, and *V* is the vocabulary used in the text (Batini and Scannapieco, 2016).<br>$syntactic\ acc = \frac{\sum_i^K closeness\,(w_i,V)}{K}$ |
| Semantic accuracy | How correctly the meaning of values represents real-world facts. An object identification problem where $\alpha$ and $\beta$ are a pair of tuples to be matched, *M* is the set that contains a record of similar existing pair, *U* is the set that represents nonmatch and $\underline{x}$ is a random vector of *n* number of attributes, and *p()* is the probability of matching (Batini and Scannapieco, 2006; Elmagarmid, Ipeirotis and Verykios, 2007).<br>$\langle\alpha,\beta\rangle \in \begin{cases} M\ if\ p(M\|\underline{x}) \geq p(U\|\underline{x}) \\ U\ otherwise \end{cases}$ |
| Completeness | Completeness of a tuple with respect to the values of all its fields where $T_v$ is the number of null values in a tuple and $N_v$ is the total number of values in a tuple (Batini and Scannapieco, 2006; Blake and Mangiameli, 2011).<br>$completeness = 1 - \frac{T_v}{N_v}$ |
| Consistency | Violation of semantic rules defined over (a set of) data items, where items can be tuples of relational tables or records in a file. *g* is the data value, and *N* is the number of rules for *g* (Heinrich *et al.*, 2018).<br>$r_n(g) = \begin{cases} 0, if\ g\ fulfills\ rule\ r_n \\ 1\ else \end{cases}; cons(g) = 1 - \frac{\sum_{n=1}^{N} r_n(g)}{N}$ |
| Credibility | How data are accepted or regarded as true, real, and credible, where *dist* is the distance between the sensor *s* and entity *e*, and $d_{max}$ is the maximum distance acceptable (Firmani *et al.*, 2016).<br>$credibility = \begin{cases} 1 - \frac{dist}{d_{max}} & if\ d(s,e) < d_{max} \\ 0 & otherwise \end{cases}$ |
| Currentness | Currentness concerns how promptly data are updated with respect to changes occurring in the real world (Batini and Scannapieco, 2006).<br>$currentness = Age + (DeliveryTime - InputTime)$ |
| Accessibility | The ability of the user to access the data from his or her own culture, physical status/functions, and available technologies (Batini and Scannapieco, 2006). |
| Usability | A collection of other characteristics characterized by usability aspects, verifiability, imperfection, and integration (Cross and Joana, 2010).<br>$usability = avg(accuracy + credibility + completeness + currentness + relevance + granularity + accessibility)$ |
| Relevancy | The extent to which data are applicable and helpful for the task at hand. *n* is the number of words in a sentence, *m* is the number of characters in a word, and *WordSimilarity* is the similarity between two words between 0 and 1 (Yang, Feng and Fabbrizio, 2006).<br>$SentenceSimilarity\ (Q,Q') = \frac{1}{n}\sum_{1\leq j\leq n}(max_{1\leq i\leq m} WordSimilarity(w_j, w_i'))$ |
| Understandability | The ease with which data can be comprehended without ambiguity and be used by a human information consumer (Dexun *et al.*, 2014).<br>$understand. = -0.33 * Abstraction + 0.33 * Encapsulation + 0.33 * Coupling + 0.33 * Cohesion - 0.33 * Polymorphism 0.33 * Complexity - 0.33 * DesignSize$ |
| Redundancy | Minimality, compactness, and conciseness refer to the capability of representing the aspects of the reality of interest with minimal use of informative resources (Batini and Scannapieco, 2016). |
| Efficiency | The degree to which data has attributes that can be processed and which provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use (ISO, 2008)**.**<br>$Query\ performance = log(row\_count)\ /\ log(index\_block\_length\ /\ 3 * 2\ /\ (index\_length + data\_pointer\_length)) + 1.$ |
| Representational consistency | Coherence of physical instances of data with their formats (Redman, 1996)**.** |

| | |
|---|---|
| Privacy | Data is hidden or concealed from others. *S* is the sensitivity of a data item, and *V* is the visibility in a given context, and *R* is relatedness. *a*, *b* and *c* are real numbers (Senarath, Grobler and Arachchilage, 2019).<br>$PrivacyRisk_{(i,j)} = \frac{S_i^a \times V_{(i,j)}^b}{R_{(i,j)}^c}$ |
| Portability | Degree of effectiveness and efficiency with which a system, product, or component can be transferred from one hardware, software, or other operational or usage environment to another (ISO, 2008). |
| Precision | Precision refers to the amount of detail that can be discerned in space, time, or theme. Using Levenshtein edit distance where *a* and *b* are the given values, *i* and *j* are the indexes(Levenshtein, 1965; Elmagarmid, Ipeirotis and Verykios, 2007).<br>$lev_{a,b}(\mathrm{i,j}) = \begin{cases} max(i,j) & if\ min(i,j) = 0, otherwise \\ min\begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} \end{cases}$ |
| Compliance | Defining and evaluating the compliance between data and schemas is a measure of the relationship (similarity, relatedness, distance, etc.) between two entities. *a* and *b* are values of elements in minimum distance and $\bar{a}$ and $\bar{b}$ are means of all elements (Hulitt and Vaughn, 2010).<br>$compliance\ (degree\ of\ variance\ ) = \frac{\sum(a-\bar{a})(b-\bar{b})}{\sqrt{\sum(a-\bar{a})^2 \sum(b-\bar{b})^2}}$ |
| Traceability | The extent to which data are well documented, verifiable, and easily attributed to a source. *R* is a source, *Ω* is a set of *R*, *E(Ω)* is a measure of uncertainty, and *λ* is the number of reports (Lu *et al.*, 2019).<br>$Network\ traceability\ entropy\ (NTE), E^{\lambda} = \sum_{\Omega:\lvert\Omega\rvert=\lambda} E(\Omega) / \binom{\lvert R\rvert}{\lambda}$ |
| Availability | Property of being accessible and usable upon demand by an authorized entity (ISO, 2008).<br>$Availability = \frac{Runtime}{Total\ operative\ time}$ |
| Recoverability | The degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use. (ISO, 2008). |
| Ability to represent null values | Ability to distinguish neatly (without ambiguities) null and default values from applicable values of the domain (Redman, 1996). |
| Format flexibility | Changes in user needs and recording medium can be easily accommodated (Redman, 1996). |
| Objectivity | Data is unbiased and impartial, where *E* is evidence, *H* is a hypothesis (assumed value), and *p()* denotes the probability (Reiss and Sprenger, 2017).<br>$w(E,H,H') = log\ \frac{p(E\lvert H)}{p(E\lvert H')}$ |
| Value | The extent to which data are beneficial and provide advantages from their use (Wrabetz, 2017).<br>$DataValue(t) \geq (GatherCost + MaintainCost + AccessCost)/GB/yr * RetentionPeriod$ |
| Volume | Appropriate amount of data: the extent to which the quantity or volume of available data is appropriate. Sample size formula where *z* is z-score, *e* is the margin of error, *p* is standard deviation, and *N* is population size (Krejcie and Morgan, 1970).<br>$sample\ size = N * \frac{\frac{z^2 * p * (1-p)}{e^2}}{\left[N-1+\frac{z^2 * p * (1-p)}{e^2}\right]}$ |
| Concise representation | The extent to which data are compactly represented without being overwhelming (Wang and Strong, 1996). |

Data and information require different definitions for characteristics because they are two separate concepts (Davenport and Prusak, 2000; Bovee, Srivastava and Mak, 2003; Batini and Scannapieco, 2016). For example, having a list of temperatures is data, but the data becomes information when given a context, such as predicted temperatures in the upcoming week. Data quality characteristics are essential regardless of the purpose of

data, and information quality characteristics need to consider the purpose. There can be characteristics that are important in both data and information quality. Table 22 presents the definitions for the same characteristics from Table 20 regarding information.

**Table 22**. Information quality characteristic definitions

| **Information quality characteristics** | **Definition** |
|---|---|
| Accuracy | Information is correct and free of errors (Wang and Strong, 1996). |
| Syntactic accuracy | The closeness of words in the text to a reference vocabulary. $K$ is the number of words, $w_i$ is a word in the text, and $V$ is the vocabulary used in the text (Batini and Scannapieco, 2016). |
| Semantic accuracy | How correctly the information represents the real-world facts (Batini and Scannapieco, 2016). |
| Completeness | Information is of sufficient depth and scope for the task at hand (Wang and Strong, 1996). |
| Consistency | Degree of similarity between perceived information (IGI Global, 2021). |
| Credibility | Information is accepted as real and comes from a trusted source (Wang and Strong, 1996). |
| Currentness | How promptly information is updated with respect to changes occurring in the real world (Batini and Scannapieco, 2006). |
| Accessibility | Access to information is restricted (Wang and Strong, 1996). |
| Usability | Based on specified usability aspects. Related to the advantage the user gains from the use of the information (Batini and Scannapieco, 2016). |
| Relevancy | Information is applicable and useful for the task at hand (Wang and Strong, 1996). |
| Understandability | The information is easily comprehended and without ambiguity (Wang and Strong, 1996). |
| Redundancy | The capability to represent the aspects of the reality of interest with minimal use of informative resources (Batini and Scannapieco, 2016). |
| Efficiency | Not applicable to information |
| Representation consistency | Information is identically represented. |
| Privacy | Personal information is hidden. |
| Portability | Not applicable to information |
| Precision | The amount of detail in space, time, or theme within the given information (Elmagarmid, Ipeirotis and Verykios, 2007). |
| Compliance | Information follows given rules and regulations within the context of use. |
| Traceability | How well the information is documented and attributed to a source. |
| Availability | Information is available or easily and quickly retrieved (ISO, 2008). |
| Recoverability | Not applicable to information |
| Ability to represent null values | Not applicable to information |
| Format flexibility | Information can be presented in a variety of formats. |
| Objectivity | Information is unbiased and impartial (Wang and Strong, 1996). |
| Value | The extent to which information is beneficial and provides advantages from its use (Wrabetz, 2017). |
| Volume | The quantity of information is appropriate (Wang and Strong, 1996). |
| Concise representation | The extent to which information is compactly represented without being overwhelming (Wang and Strong, 1996). |

### 5.1.2 Data and information quality characteristics for user-generated content

Each characteristic from Table 21 is inspected for its fitness in measuring data quality in UGC based on importance. The characteristics are evaluated using existing UGC platforms. The evaluation of each characteristic is shown in Table 23 with the following values:

- 1: Platform considers the characteristic important.
- 0: Platform places no importance on the characteristic.
- +/-: The importance of the characteristic is situational.

**Table 23.** Evaluating data quality characteristics in UGC platforms.

| Data quality characteristics | ALA | Twitter | Worldometer | Wikipedia | YouTube |
|---|---|---|---|---|---|
| Syntactic accuracy | 1 | 1 | 1 | +- | 1 |
| Semantic accuracy | 1 | 0 | 1 | 1 | 0 |
| Completeness | 1 | 1 | 1 | 1 | 1 |
| Consistency | 1 | 0 | 1 | 0 | 0 |
| Credibility | 1 | 0 | 1 | 1 | 0 |
| Currentness | 1 | 1 | 1 | 1 | 1 |
| Accessibility | 1 | 1 | 1 | 1 | 1 |
| Usability | 0 | 0 | 0 | 0 | 0 |
| Relevancy | 1 | 0 | 1 | 0 | 0 |
| Understandability | 1 | 0 | 1 | 1 | 0 |
| Redundancy | 0 | 0 | 0 | 0 | 0 |
| Efficiency | 0 | 0 | 1 | 0 | 0 |
| Representation consistency | 1 | 0 | 0 | 0 | 0 |
| Privacy | 0 | 1 | 1 | 1 | 1 |
| Portability | 0 | 0 | 0 | 0 | 0 |
| Precision | 0 | 0 | 1 | 0 | 0 |
| Compliance | 0 | 0 | 1 | 0 | 0 |
| Traceability | 1 | 1 | 1 | 1 | 1 |
| Availability | 1 | +- | 1 | 1 | +- |
| Recoverability | 0 | 0 | 0 | 1 | 0 |
| Ability to represent null values | 1 | 0 | 0 | 0 | 0 |
| Format flexibility | 0 | 1 | 0 | 1 | +- |
| Objectivity | 0 | 0 | 1 | 0 | 0 |
| Value | 0 | 0 | 1 | 0 | 0 |
| Volume | 0 | 1 | 0 | 0 | 0 |
| Concise representation | 0 | 0 | 1 | 0 | 0 |

Table 24 presents the quality characteristics and describes why the characteristic is or is not essential. Three different values are given based on Table 23:

- Yes: Characteristic is essential for data quality in UGC
- No: Characteristic is not critical to data quality in UGC
- Partially: Characteristic is necessary in some specific cases.

**Table 24**. Data quality characteristics and their importance to UGC

| Important for UGC | Data quality characteristics | The reason why characteristic is or is not important |
|---|---|---|
| Yes | Accuracy | Accurate data is essential, especially if the data is stored in a non-relational database without type restrictions. |
| Yes | Syntactic accuracy | Subpart of accuracy. |
| Yes | Semantic accuracy | Subpart of accuracy. |
| Yes | Completeness | Data should be as complete as possible. Even if some specific values are missing, there should be some default data at least |
| Partially | Consistency | If data comes from one platform, consistency is not vital, but inconsistencies exist when extracting data from multiple platforms. |
| Yes | Credibility | The credibility of the content provider is essential, especially in the UGC domain. |

| | | |
|---|---|---|
| Yes | Currentness | The data should be as current as possible in UGC. |
| Yes | Accessibility | Some UGC databases require access by law. |
| No | Usability | The usability of data cannot be determined at the data quality level. |
| No | Relevancy | Relevance of data is unimportant in UGC as the information stored as data should be relevant. |
| Yes | Understandability | Understandability of data is vital as the data needs to be understood to be able to utilize it. |
| No | Redundancy | Many data pieces may be necessary without knowing how it is used, so there is no redundant data. |
| No | Efficiency | Data efficiency is not that important for UGC. It is not meant for repeated access and modification but relatively constant addition. |
| No | Representation consistency | Representational consistency is irrelevant for data quality. |
| Yes | Privacy | As most content providers are citizens and the general public, their private information must be hidden if they so demand |
| No | Portability | Data portability is not that important for UGC as the primary purpose is not to move it around. |
| Partially | Precision | Data precision may sometimes be necessary for UGC, especially with location-related information. |
| No | Compliance | There is a low number of standards within the UGC domain. The variety of content makes data compliance less critical. |
| Yes | Traceability | In UGC, it is necessary to have some information regarding where the content comes from. Additionally, the platform often provides this through metadata. |
| Yes | Availability | Availability of data is vital for UGC if the data is going to be used. From a broader perspective, most UGC data is available for everyone and not just specific entities. |
| No | Recoverability | Recoverability of UGC data is not as important because there should be a regular stream of new data, and recovering old data is not essential. |
| No | Ability to represent null values | The ability to represent null values is not essential for data quality. It is more relevant for the database management system. |
| Partially | Format flexibility | Often, UGC is not predetermined, and the flexibility of storage format may be essential. |
| No | Objectivity | For data quality, it is irrelevant in UGC if it is objective or subjective. Different analyses can be done for subjective data. Thus, objectivity is not a concern. |
| No | Value | The data itself does not provide any inherent value in UGC. The value comes from information. |
| No | Volume | The volume of data is not crucial before analysis. |
| No | Concise representation | Concise representation is not necessary for data quality. The data should be represented as is. |

Similar to data quality characteristics, the information quality characteristics need to be evaluated in UGC platforms. Table 25 presents the evaluation of information quality characteristics in several UGC platforms with similar values as in Table 23.

**Table 25.** Evaluating information quality characteristics in UGC platforms

| Information quality characteristics | ALA | Twitter | Worldometer | Wikipedia | YouTube |
|---|---|---|---|---|---|
| Syntactic accuracy | 1 | 0 | 1 | +- | 0 |
| Semantic accuracy | 1 | 0 | 1 | 1 | 0 |
| Completeness | 1 | 0 | 1 | 1 | 0 |
| Consistency | 0 | 0 | 0 | 0 | 0 |
| Credibility | 1 | 1 | 1 | 1 | 1 |
| Currentness | 1 | 1 | 1 | 1 | 1 |
| Accessibility | 1 | 1 | 1 | 1 | 1 |
| Usability | 1 | +- | 1 | 1 | +- |
| Relevancy | 1 | 1 | 1 | 1 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| Understandability | 1 | 1 | 1 | 1 | 1 |
| Redundancy | +- | 0 | 1 | 0 | 0 |
| Efficiency | 0 | 0 | 0 | 0 | 0 |
| Representation consistency | 1 | 0 | 1 | 0 | 0 |
| Privacy | 1 | 1 | 0 | 0 | 1 |
| Portability | 0 | 0 | 0 | 0 | 0 |
| Precision | 1 | 0 | 1 | +- | 0 |
| Compliance | ? | ? | ? | ? | ? |
| Traceability | 1 | 1 | 1 | 1 | 1 |
| Availability | 0 | 0 | 1 | 1 | 0 |
| Recoverability | 0 | 0 | 0 | 0 | 0 |
| Ability to represent null values | 0 | 0 | 0 | 0 | 0 |
| Format flexibility | 1 | 1 | 0 | 1 | 0 |
| Objectivity | 1 | 0 | 1 | 1 | 0 |
| Value | 1 | +- | 1 | 1 | +- |
| Volume | 1 | 0 | 1 | 1 | 1 |
| Concise representation | 0 | 0 | 1 | 0 | 0 |

Based on the evaluation in Table 25, the essential information quality characteristics can be selected. Table 26 presents the information quality characteristics and reasons they are or are not vital for UGC platforms.

**Table 26**. Information quality characteristics and their importance to UGC

| **Information quality characteristics** | **Important for UGC** | **The reason why characteristic is for information quality** |
|---|---|---|
| Accuracy | Yes | Information accuracy is essential and needs to be evaluated. |
| Syntactic accuracy | Yes | Syntactic correctness of information is vital for information to be understood. |
| Semantic accuracy | Yes | Semantic correctness is a significant component of information correctness. |
| Completeness | Yes | Completeness of information is necessary, especially after analysis. Information from content providers may be incomplete, but it can be completed by appending information from other providers. |
| Consistency | No | Many content providers provide conflicting information as there are no requirements to agree with others. |
| Credibility | Yes | The credibility of information needs to be evaluated regardless of whether the information is from the content provider or analysis. |
| Currentness | Yes | In UGC, information should be as current as possible. |
| Accessibility | Yes | Information on UGC platforms should be accessible by almost anyone. |
| Usability | Yes | Information should be usable by someone, but this can mean the content provider as well. |
| Relevancy | Yes | Information should be relevant within the platform context. |
| Understandability | Yes | Information needs to be understandable. |
| Redundancy | Partially | Redundancy of information may be necessary for some specific UGC platforms, such as citizen science, but mostly it is unimportant. |
| Efficiency | No | Not applicable to information quality. |
| Representation consistency | No | Because UGC platforms accept content in multiple forms and types, the representation does not need to be consistent. |
| Privacy | Yes | Privacy is an integral part of the information, even in UGC. Content providers should not be required to provide private information. Their privacy should be protected when content users use data. |
| Portability | No | Not applicable to information quality. |
| Precision | Partially | With location information, precision can be necessary for UGC. |
| Compliance | No | Regarding information, there are no particular compliance rules. |
| Traceability | Yes | Information sources should be visible. |

| | | |
|---|---|---|
| Availability | No | Availability of information is not as necessary as accessibility. The information may be private, protected, or even removed, making it unavailable on UGC platforms. |
| Recoverability | No | Not applicable to information quality. |
| Ability to represent null values | No | The ability to represent null values is more related to database systems. |
| Format flexibility | Yes | Information in UGC can be provided in many types and formats. |
| Objectivity | Yes | The objectivity of information needs to be somehow discernable, especially in UGC. |
| Value | Yes | Information given by content providers may not provide value as is, but the analysis of such information should. |
| Volume | Yes | The amount of data is vital for proper analysis and relevant information. |
| Concise representation | No | Information representation does not necessarily have to be concise if it is understandable. |

### 5.1.3 Defining the data and information quality characteristics for user-generated content

Table 27 presents the data and information quality characteristics that need to be defined for the UGC domain. These characteristics are found to be crucial for UGC platforms based on available literature and existing UGC platforms.

**Table 27**. The final list of data and information quality characteristics for UGC

| **Data quality characteristics** | **Information quality characteristics** |
|---|---|
| Accuracy | Accuracy |
| Syntactic accuracy | Syntactic accuracy |
| Semantic accuracy | Semantic accuracy |
| Completeness | Completeness |
| Credibility | Credibility |
| Currentness | Currentness |
| Accessibility | Accessibility |
| | Usability |
| | Relevancy |
| Understandability | Understandability |
| | Redundancy |
| Privacy | Privacy |
| Precision | Precision |
| Traceability | Traceability |
| Format flexibility | Format flexibility |
| | Objectivity |
| | Value |
| | Volume |
| Availability | |
| Consistency | |

Tables 28 and 29 present the informal definitions for the UGC domain's data and information quality characteristics.

**Table 28.** Data quality characteristic definitions for UGC

| Data quality characteristics | Informal definition for UGC |
|---|---|
| Accuracy | Accuracy is the closeness of the given content to the expected content. Based on syntactic and semantic accuracy. |
| Syntactic accuracy | The closeness of the content syntax that the user provides, depending on the platform context. |
| Semantic accuracy | How correctly the content matches the real world. |
| Completeness | How complete the content is and whether or not it is missing essential facts depending on the platform context. |
| Consistency | How well the content adheres to the semantic rules. |
| Credibility | How credible the content is based on who is giving the content |
| Currentness | How promptly content is updated with respect to changes occurring in the real world. |
| Accessibility | How well the content is accessible by content users. |
| Understandability | How easily the content can be comprehended without ambiguity by a human consumer within the platform context |
| Privacy | How much of the content provider's personal information is concealed. |
| Precision | How detailed the provided content is in the platform context. |
| Traceability | How well the content is attributed to a specific source and time |
| Availability | How well the content is available for use. |
| Format flexibility | How diverse content can be provided by content providers. |

**Table 29**. Information quality characteristic definitions for UGC

| Information quality characteristics | Informal definitions for UGC |
|---|---|
| Accuracy | Accuracy is the closeness of the given information to the expected information, based on syntactic and semantic accuracy. |
| Syntactic accuracy | The closeness of the information syntax that the user gives, depending on the platform context. |
| Semantic accuracy | How correctly the information matches the real-world facts. |
| Completeness | How complete the information is and whether or not it is missing essential facts. |
| Credibility | How credible the information is based on given sources or content providers. |
| Currentness | How promptly information is updated with respect to changes occurring in the real world. |
| Accessibility | How well the information is accessible by content users. |
| Usability | How usable the information is based on the platform context. It is affected by accuracy, completeness, and credibility. |
| Relevancy | How relevant the information is to the platform context. |
| Understandability | How easily the information can be comprehended without ambiguity by a human consumer. |
| Redundancy | How minimal and conciseness the representation of information is. |
| Privacy | How much of the personal information is concealed. |
| Precision | How detailed the information is in the platform context. |
| Traceability | How well the information is attributed to a specific source and time. |
| Format flexibility | How diverse the information is. |
| Objectivity | How unbiased and impartial the content and its information are. |
| Value | How useful the content is and whether it provides advantages from its use. |
| Volume | The amount of similar information given by multiple users. |

### 5.1.4 **Summary**

Information and data quality are commonly used interchangeably. The primary purpose of this research is to clearly define and separate the two independent concepts that are tied together. Most researchers use definitions for quality characteristics from existing

work that are not designed for UGC. Therefore, each domain needs to have its definitions for quality characteristics. Definitions given in Tables 28 and 29 can be used:

1. *To select characteristics and define data and information quality for a UGC platform.* Each desired quality characteristic acts as a prerequisite for data and information quality that can be used to create quality requirements for the platform during design and development.
2. *To define and evaluate the data and information quality of UGC in an existing platform.* Data and information quality are assessed through individual quality characteristics. The provided definitions will help understand quality and normalize the evaluation process as each characteristic is understood the same way.

Quality is subjective, and not all characteristics have to be used for defining data and information quality, nor are all characteristics equally important. For example, social media platforms may be less interested in the credibility of data and information than citizen science platforms.

The given lists of data and information quality characteristics are not exhaustive even when the original quality characteristics come from well-known and cited literature and are used by many. There are widely accepted terms in specific UGC cases, such as *thematic accuracy* or *positional precision*, which can be used instead of the proposed terminology in particular circumstances.

The listed data and information quality characteristics provide practical benefits for researchers, developers, and content users. Researchers can use the definitions in their research to properly define data and information quality. Developers can create quality requirements for their UGC platforms based on the data and information quality characteristics they want to emphasize. Additionally, developers can utilize the quality characteristics within the design, as discussed in Section 5.2. Finally, content users who want to utilize the content from UGC platforms can evaluate the quality of content based on the data and information quality characteristics.

## 5.2 Content collection in user-generated content

### 5.2.1 Content collection process

Content gathered from content providers has a varying level of quality. Depending on the UGC platform, the content can be highly subjective and have no reliable sources or supporting arguments for the provided information. Social media platforms provide the most subjective content amongst UGC platforms. Social media platforms actively remove content that is against their policies or contains misinformation that may cause extensive harm (Lyons, 2020; Twitter Safety, 2020).

Platforms, such as Wikipedia (Wikipedia, 2020) and Worldometer (Worldometer, 2020), require the content provider to use references that support the information. Moderators or administrators can review the content and its validity by examining the provided references. When false information is found, it can be removed or modified. Some platforms provide reputation scores to content providers based on how reliable they are within the community. Their reputation is affected by their actions and how reliable the content they have provided is. Table 30 presents the usage process in UGC platforms.

**Table 30.** UGC platform content collection process

| Content collection process | What affects quality |
|---|---|
| The content provider submits information in the platform | Quality is affected by what is provided |
| Information becomes content for the platform | Quality is affected by what is accepted by the platform |
| Content is stored as data in a database | Quality is affected by the storage process |
| Content is visible for other platform users | Quality is affected by how the content is presented |
| Data is retrieved from the database by a content user | Quality is affected by how and what data is provided |
| Content user analyzes the data to create information | Quality is affected by the analysis process |

Improving the quality of data and information requires different methods depending on the step in the collection process. When a user provides content, the quality can be improved by limiting what the user can provide, guiding the user with templates or forms, or having specific guidelines about what type of information is requested. After a content provider sends the information, the quality is further influenced by what the platform accepts. Platforms may require a specific type of content and not allow all multimedia types. While multimedia is increasingly common, some UGC platforms do not allow any multimedia and accept only text. When the information is admitted into the platform, the quality can be increased with specific data cleaning methods during the storage process. These methods can also affect how the data can be extracted from the database. Finally, different analysis techniques will affect the quality of the resulting information.

Numerous organizations widely use UGC (Asur and Huberman, 2010; Kaplan and Haenlein, 2010; Goh, Heng and Lin, 2013; Arthur *et al.*, 2018; Kabir and Madria, 2020). Some employ platforms to collect content for a specific purpose, and others collect content from existing platforms, primarily social media platforms. There are various uses for the collected content, but most have a direct or in-direct monetary purpose. For example, gathering content from the public may be less expensive than using the workforce, even with the risk of receiving lower-quality content. Using low-quality data and information may lead to incorrect conclusions and monetary losses. The quality must be at least of an acceptable level, and there should be methods for ensuring that the desired level is reached.

### 5.2.2 Current quality improvement methods

Methods for improving data quality can be divided into two distinct categories (Batini *et al.*, 2009):

- *Data-driven*: Methods for improving data quality, such as correcting errors, replacing low-quality content, gathering new data, and selecting credible sources.
- *Process-driven*: Methods for altering the process to improve data quality, such as gamification, data collection templates, and automatic error checking.

Another classification is to divide the improvement methods into *ex-ante* and *ex-post* categories (Bordogna *et al.*, 2016). These categories include methods for example:

- *Ex-ante:* Methods for improving data quality before the collection, such as training content providers, using standards, moderating, having reasonable expectations of users, and using reputation models.
- *Ex-post:* Methods for improving data quality after the collection, such as outlier identification, content comparison, enrichment, cross-referencing, and user ranking.

Currently, various UGC platforms use some of these methods to improve the quality of their content. Social media platforms focus on freedom of speech and expression, trying not to limit information. However, there are cases where the freedom of speech is limited based on the content. For example, in 2020, Twitter actively removed misinformative tweets relating to COVID-19 (Peters, 2020; Quinn, 2020). Within the rules and policies of Twitter (Twitter, 2021), there is a mention that tweets will be removed if they are encouraging hateful conduct or inciting harm against protected groups. By these rules, demagogy is forbidden by Twitter, and related tweets will be removed. Lesser restrictions by Twitter relate to possibly harmful and pornographic content. Instead of removal, these tweets are hidden, and they can be shown after accepting the warning message.

Citizen science platforms have restrictions on what content providers can input via specific fields and forms that need to be filled. Because the platform targets particular information types, there is no reason to allow complete freedom of speech. There are also some checks on misinformation and incorrect content before sending the information to the platform (Lukyanenko, Parsons and Wiersma, 2011, 2014; Kelling *et al.*, 2013).

Wikipedia has specific format and style for each article that users are required to follow and Wikipedia mandates the usage of reference material. Information without reference material can be accepted to the platform, but it will be flagged as missing a reference (Wikipedia, 2021b).

There are many challenges and issues with the currently used methods.

- Training content providers require resources that not all platforms have.
- Gathering new content from users may not produce better quality.
- Moderating and manually correcting errors requires additional resources.

- Selecting credible sources or using standards is not always possible.
- Cross-referencing and enrichment require other content from outside the platform, and content comparison requires other users to have submitted similar content.

Most of the issues arise from requiring additional resources to correct existing content or influence the content providers. Allocating more resources for improving the quality of content can be a huge hurdle for some platforms and should not be the first-choice method.

The methods from Batini et al. (2009) and Bordogna et al. (2016) that do not have significant issues are: *using collection templates, automatic error checking, reputation models*, and *user ranking*. These methods have been used successfully in UGC platforms, but even in the best-case scenario, the platforms have some problems with quality that can be further improved. In most platforms, the methods are used only to solve some glaring quality-related issues. For example, *automatic error checking* used in most UGC only checks for the most obvious errors, and more subtle problems are manually evaluated. The preliminary design for the platform has a massive influence on how different methods are implemented. When quality is overlooked during the platform's design, the strategies for improving quality will be lackluster and inefficient.

### 5.2.3 **The proposed theoretical framework**

Albrecht et al. (2018) tie data quality characteristics to the life cycle of remote sensing data. The life cycle composes of four phases:

1. Conceptualization
2. Collection
3. Processing
4. Delivering and using

Each specific phase has different data quality checks based on what quality characteristics are relevant. This ideology can be adapted and altered for UGC platforms quite effortlessly. Table 31 presents the combination of the works of Albrecht et al. (2018), Batini et al. (2009), and Bordogna et al. (2016).

**Table 31.** UGC life cycles are tied to quality characteristics and data or process-driven methods

| Phase in the life cycle | Data-driven methods | Process-driven methods | Relevant quality characteristics |
|---|---|---|---|
| Conceptualization | Deciding the objective | Selecting sources | - |
| Before collection | Content requirements | | Syntactic accuracy<br>Semantic accuracy<br>Precision |
| During collection | Standards | Automatic error checking<br>Reputation models<br>User ranking<br>Collection templates | Syntactic accuracy<br>Semantic accuracy<br>Credibility<br>Relevancy<br>Currentness<br>Completeness<br>Privacy<br>Traceability<br>Format flexibility |
| After collection (processing) | Enrichment<br>Manual error correction | Automatic error correction | Objectivity<br>Value<br>Usability<br>Completeness<br>Volume |
| Delivering / using / presentation | Privacy masking<br>Comparison | Cross-referencing | Accessibility<br>Availability<br>Volume<br>Understandability<br>Privacy<br>Format flexibility<br>Redundancy |

Table 31 presents some quality characteristics relevant to each phase, but there may be others that designers may find more appropriate. The quality characteristics depend on the purpose of the platform. Each step in the life cycle provides different methods that influence the design of the data model and the collection process of the platform. Characteristics relevant for specific phases should be integrated into the design by using *data-driven* or *process-driven methods.*

*Conceptualization*: Both the objective of the platform and the desired sources will affect which quality characteristics require more emphasis and careful consideration. If the content providers are platform owner's close acquaintances, they are less likely to give false information and are considered more credible. If the purpose is to collect random information from specific locations, the correctness of the information is easier to examine than if the purpose is to have content providers collect certain information in unexpected places. Selecting the most appropriate source for the task at hand will impact the chosen quality characteristics.

*Before collection:* Throughout the conceptualization of the platform, specific requirements for the content should arise. These requirements considerably affect the design of the data model, what content is expected, and what form the content should take. These requirements relate to *syntactic* and *semantic accuracy* and *precision,* and they can be enforced through the data model or the server layer.

*During collection*: Users provide content through a user interface in the platform. The user interface plays a significant role in what content users can and will send to the platform. The content providers can be guided to provide correct, relevant, current, complete, and accurate information with collection templates and automatic error checking. Reputation models and user rankings can give feedback on how credible a content provider is and how likely they are to provide higher quality content.

*After collection*: When content providers have successfully sent new information to the platform, the content can be checked for possible errors manually or automatically. This error checking should be more focused on the optional content and check for minor mistakes. More significant problems should be rejected during collection. The content can be enriched with information from other sources or content from the platform, such as existing user information or timestamps. User information will help to establish the objectivity and usability of the content.

*Delivering, using, presenting*: The user interface is vital for providing and presenting the existing content. With proper design, it is possible to show information in a more understandable format. The volume of content can be visualized, and content sent by users can be compared against each other.

The categorization given in Table 31 can be used to develop the platform while incorporating the essential quality characteristics into the design. Quality characteristics are integrated into different parts of the platform to ensure that quality stays high enough throughout the whole life cycle of the content. Figure 12 presents the UGC platform life cycle that has been extended from the sensor data life cycle.

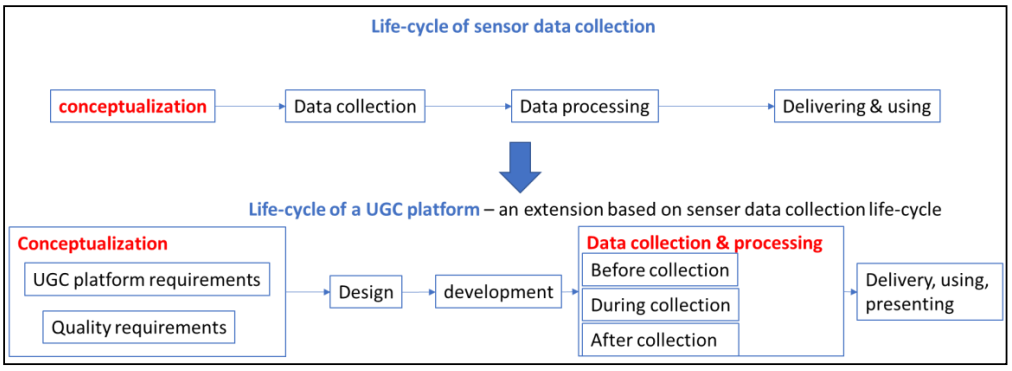


**Figure 12.** Extending sensor data life cycle to UGC platform life cycle

The characteristics listed in Table 31 may not be necessary for each scenario. During the conceptualization, the characteristics must be selected as requirements for the platform based on the purpose. In the design phase, each characteristic can be integrated into the data collection and processing parts:

- Before collection: Data quality characteristics are integrated into the data model that provides content restrictions based on the characteristics.
- During collection: Information quality characteristics are integrated into the user interface. Collection templates and extensive automatic error checking are developed based on the quality characteristics.
- After collection: Data is automatically and manually corrected and moderated, based on the data quality characteristics on the server layer. The quality information can be generated and stored in the database during this phase.

All quality evaluations can be stored in the database to provide a quality evaluation for any content. The quality evaluation can further improve the usage and presentation of content. Storing *quality information* requires the data model to be designed to assess quality characteristics stored in the database. The characteristics can be stored in a separate entity or integrated into the necessary parts in the data model. Figure 13 integrates quality characteristics into a UGC data model as *quality information* entity and as attributes related to quality characteristics. The characteristics are within different entities based on what information is required for the characteristics. The *quality information* and attributes are colored blue to distinguish them from the base data model. Either method can be used for integrating quality characteristics into the data model.

There are benefits and drawbacks to both methods presented in Figure 13. Tying the quality characteristics to the relevant entities makes it easier to find them in the database. Creating a separate entity allows a more straightforward extension regarding the quality characteristics as all characteristics are stored in one entity.

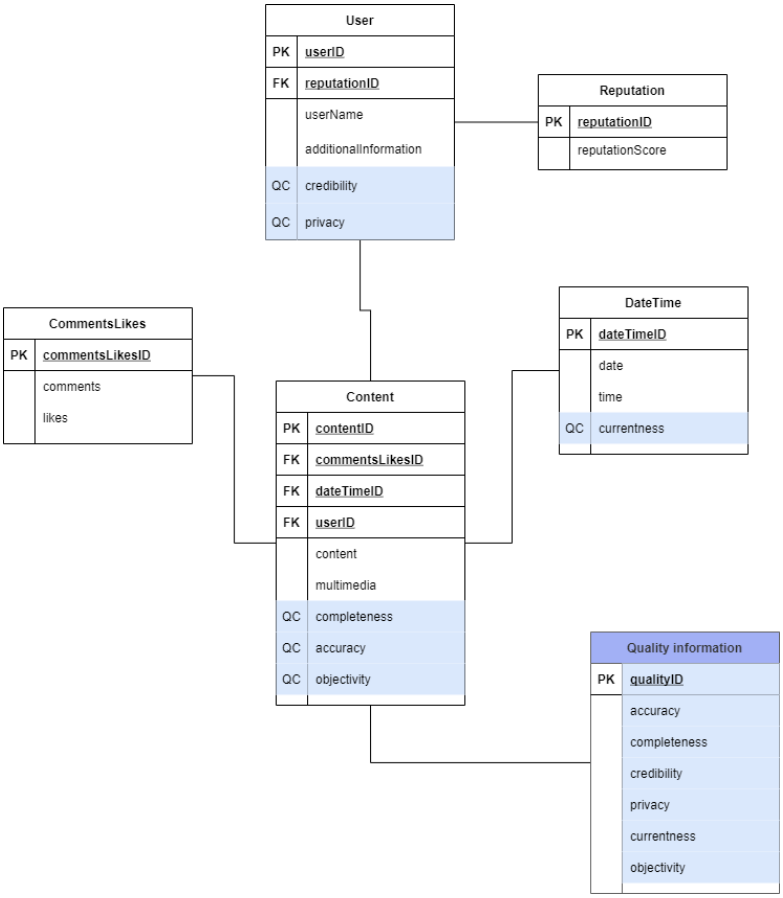


**Figure 13**. Integrating quality characteristics into the data model of the platform

Table 32 presents the benefits and drawbacks of the proposed framework compared to the existing one. It is noteworthy to mention that the design phase is highly vital and determines success or failure. When some quality characteristics are not considered during the design, it may be challenging to add them later. Another issue may arise from the user's perspective and when the user interface feels too restrictive from their point of view. Especially in the case of social media, it is difficult to justify any restrictions on the content. However, the platform can quickly provide templates, and these templates could be developed for different purposes. The social media platform could provide a template to share a link to some source material when sharing factual content.

**Table 32**. Benefits and drawbacks of the proposed framework

| Benefits | Drawbacks |
|---|---|
| - Less external resources required<br>- Works with smaller communities<br>- Less manual work required<br>- Easier to guarantee a specified level of quality<br>- Easier to consider all relevant quality characteristics<br>- Easier to evaluate the quality<br>- Easier to justify the usage of content<br>- Easier to find issues related to quality | - More work during the design<br>- More work during the development of automatic checks<br>- Implementation in an existing platform may be difficult<br>- May require more knowledge on database design |

In most cases, the benefits outweigh the drawbacks, but there are some scenarios where the old design is better. First, if the developer has no interest in the quality of content and the platform will be short-lived, the old model is better as it requires less work initially. Second, if the platform has only a limited number of users, manual labor may be more worthwhile than designing and developing proper automation.

Overall, the new design will promote the data and information quality in UGC platforms and lead to higher quality content that is easier to reuse. The presented theoretical framework is beneficial for developers who create UGC platforms and want to establish proper quality management and ensure a specific level of data and information quality within the platform.

## 5.3 Summary

Data and information have been described as the same concept even though data and information are different (Davenport and Prusak, 2000; Bovee, Srivastava and Mak, 2003). Researchers and organizations have created definitions for data quality for various domains (Redman, 1996; Wang and Strong, 1996; Batini and Scannapieco, 2006; ISO, 2008), but there are no existing data and information quality definitions for the UGC domain.

The quality of data and information is improved using various methods in UGC platforms. Many platforms rely on manual labor via community moderation. Even social media platforms rely on the community to report the content. However, community reports are often misused, and content is removed due to false claims (Notopoulos, 2017; Dodgson, 2020; Haasch, 2020; Worrall, 2020). While using community moderators requires less work from the platform owners, there are many drawbacks to it. Researchers have proposed different methods and techniques to improve quality in several domains, such as associating quality checks to specific phases of lifecycle in remote sensing (Albrecht *et al.*, 2018), modifying the data model in citizen science (Lukyanenko *et al.*, 2019), and training users in volunteered geographic information (Bordogna *et al.*, 2016).

The proposed theoretical framework aims to provide a reusable method to improve the quality of data and information in all UGC platforms, not just some platforms. The

framework is based on existing ideas and techniques and extends them by incorporating quality characteristics to further promote quality and assist with UGC platforms' development.

During platform design, the aims and goals regarding data and information quality need to be defined, and the necessary quality characteristics should be selected and added to the requirements. The characteristics can be selected from Table 27 and defined using the appropriate definitions in Tables 28 and 29. Based on the proposed framework, the necessary quality characteristics can be integrated into the data model or user interface. Integrating quality characteristics will be difficult if data and information quality are not considered during the platform design.

The discussed framework is tested in (Musto and Dahanayake, 2021), and the resulting data and information quality are evaluated against existing UGC platforms. Results show that overall quality is higher than the existing platforms and significantly higher in specific characteristics.

Following groups can benefit and utilize the presented work:

- Researchers: Researchers can use the data and information quality characteristics presented in Tables 27–29. The definitions for individual characteristics must be uniform within a domain, and researchers do not have to rely on definitions from other domains.
- Developers: Developers can select proper data and information quality characteristics as requirements for their UGC platform using Tables 27–29. Additionally, developers can utilize the framework given in Section 5.2.3 to improve the quality of content within the platform.
- Content users: Content users can use the data and information quality characteristics presented in Tables 27–29 to evaluate the quality of content in UGC platforms.

# 6 Conclusion

UGC has increased in popularity over the last decade and researchers and organizations have concerns about the quality of data and information in UGC.

This dissertation aimed to answer the research question "*How can the quality of user-generated content be improved by enhancing information collection and data curation processes in user-generated content platforms?"* The main research question is divided into three sub-questions that have been answered in the reviewed publications in Section 4.

- RSQ1: *What information collection features in user-generated content platforms influence the quality of content?*

The user interface has a significant role in influencing the quality of content. The information collected through the user interface is often of low quality because users are not restricted, or the user interface does not evaluate the given information properly (Newman *et al.*, 2010; Ludwig, Reuter and Pipek, 2015; Arolfo and Vaisman, 2018; Lukyanenko *et al.*, 2019). The user interfaces are not adequately designed to establish boundaries between low and high-quality content, and everything goes through without correcting possible errors. Many UGC platforms rely on moderators or administrators to review content after being collected (Sullivan *et al.*, 2014; Rice, 2015; Lyons, 2020; Cornell Lab of Ornithology, 2021; iNaturalist, 2021; Wikipedia, 2021a). Directing users to provide higher-quality information through the user interface improves the data and information quality (Lukyanenko, Parsons and Wiersma, 2011, 2014; Bordogna *et al.*, 2016; Lukyanenko *et al.*, 2019).

- RSQ2: *How to define quality characteristics and distinguish data and information quality in the domain of user-generated content?*

Several standards and definitions have been established for data and information quality (Redman, 1996; Wang and Strong, 1996; Batini and Scannapieco, 2006, 2016; ISO, 2008, 2013). However, none of the existing literature presents definitions for the UGC domain. Data and information quality need to be defined for each domain separately, and existing definitions from other domains should be avoided (Davenport and Prusak, 2000; Bovee, Srivastava and Mak, 2003; Haug, Arlbjrn and Pedersen, 2009). Data and information quality are distinguished by separating the terms to apply to different parts of the collection process. In UGC, data is a measurable object within a database, and information is the content that users provide. Information requires a context (Bovee, Srivastava and Mak, 2003; Caballero *et al.*, 2009; Han, Jiang and Ding, 2009; Watts, Shankaranarayanan and Even, 2009), and users are asked to provide content for a specific purpose; thus, it is information.

- RSQ3: *How does the introduction of data and information quality characteristics into information collection and data curation processes influence the quality of user-generated content?*

Data and information quality characteristics can be utilized during the design of a UGC platform. The selected characteristics can be used as quality requirements and integrated into the design of the platform. The characteristics can be used to modify the information collection process through the user interface to improve UGC quality. Additionally, characteristics can be applied to the data curation process and used for evaluating the quality of content.

By answering each of the sub-questions, the primary research question is answered. Identifying specific data and information quality characteristics necessary for the platform is the first step to improving the quality of content. Platform owners should decide what characteristics are essential for the platform and discard the rest. Afterward, using the guidelines presented in this research, the characteristics can be integrated into the platform design. Integrating data and information quality characteristics into the platform design through the user interface and data model reduces the amount of low-quality content generated during information collection. Thus, the processed content during data curation is of decent quality and can be further improved by assessing the data quality with individual quality characteristics.

The dissertation process and all four publications can be tied to the three DSR cycles: relevance, rigor, and design cycle. (Musto and Dahanayake, 2018) executes a literature review of the existing knowledge base and is part of the relevance cycle. The results indicate multiple problems in the data and information quality of UGC platforms. (Musto and Dahanayake, 2020) extends these results to practice by surveying existing citizen science platforms to establish relevant problems related to the data and information quality issues revealed in (Musto and Dahanayake, 2018), thus completing the relevance cycle.

(Musto and Dahanayake, 2021b) investigates the existing knowledge base and designs the artifact: *the UGC domain's data and information quality characteristics* that includes definitions for each quality characteristic applied to the target domain. The artifact is tested and evaluated in (Musto and Dahanayake, 2021a) and a second artifact is developed based on the results: *the design framework for enhancing the information collection and data curation processes.* Finally, the second artifact is evaluated in the publication (Musto and Dahanayake, 2021a).

The research presented provides three scientific contributions:

1. A comprehensive set of data and information quality characteristics defined specifically for UGC.
2. Extension of the UGC platform's development life cycle with UGC data and information quality characteristics during the platform's requirements acquisition stage to improve the quality of the content collection.
3. Framework to store and assess the reliability and quality of UGC using quality characteristics.

Carefully considering data and information quality characteristics within the development life cycle leads to better management of quality. Additionally, when quality characteristics are chosen as quality requirements during the design of a platform, they are less likely to be overlooked during development. Setting up functionalities in the user interface based on quality characteristics helps eliminate low-quality information sources and leads to higher-quality content.

There exists an enormous number of UGC platforms. Evaluation of the citizen science platforms shows that some designers and developers have taken great care with their data and information quality. eBird has arguably the best quality control, and iNaturalist has good intermediate-level automatic checks related to quality. Additionally, iNaturalist employs a visible checklist for all observations that indicates what relevant information is missing from the observation. On the other hand, some platforms accept any content through their user interface. Thus, the moderators have extra work when low-quality content is sent to the platform. While the number of malicious users in citizen science platforms may be small, implementing at least rudimentary checks to the user interface drastically reduces the amount of low-quality content.

Social media platforms have shown improvement over the last decade. Existing platforms have discarded information they do not need, such as Twitter no longer collecting location information. They have added new functionalities that improve the quality of information, such as viewing when content was edited. New social media platforms have implemented functionalities for improving information collection while not strictly limiting users. The social media platform Jodel enables users to assign a topic to their content, thus suggesting other users limit information to the specific topics.

There are a few limitations associated with the research conducted in this dissertation.

1. The implemented platform is a citizen science platform. The effects of integrating data and information quality characteristics into other types of UGC platforms are not tested in practice.
2. The integration of quality characteristics is conducted on a new platform rather than an existing platform.
3. The amount of data from WalkingPaths is low compared to other platforms.

4. Not all existing quality characteristics have been examined against all UGC platforms.

In the future, data and information quality characteristics can be integrated into an existing platform, and the effects on the resulting data and information should be evaluated. The effects of singular changes in the design should be evaluated to examine if one design decision provides higher quality content compared to another design decision. Additionally, integrating quality characteristics into a social media platform without reducing user freedom should be assessed. There is also a need to develop proper quantitative measures for each quality characteristic for UGC. Not all measures presented in Table 21 are applicable outside their respective fields and they should be adapted to UGC specifically.

Here are some research questions that can be answered in the future:

- What quality characteristics play a more significant role in the overall quality of data and information in user-generated content platforms?
- How does the developed design affect the usability of user-generated content platforms?
- What is the threshold of user freedom and quality of data and information in user-generated content platforms when limiting content generation based on quality characteristics?

In conclusion, it is necessary to identify and consider data and information quality during the design of UGC platforms. Selecting appropriate quality characteristics as requirements for the platform and integrating them into the design will help ensure the target level of quality. Increasing the overall quality of data and information will encourage others to utilize generated content.

# References

Abdullah-All-Tanvir *et al.* (2019) 'Detecting Fake News using Machine Learning and Deep Learning Algorithms', in *2019 7th International Conference on Smart Computing and Communications, ICSCC 2019*

Ahmouda, A., Hochmair, H. H. and Cvetojevic, S. (2019) 'Using Twitter to Analyze the Effect of Hurricanes on Human Mobility Patterns', *Urban Science*, 3(3), p. 87.

Akehurst, G. (2009) 'User generated content: The use of blogs for tourism organisations and tourism consumers', *Service Business*, 3(1), pp. 51–61.

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

Albrecht, F. *et al.* (2018) 'Providing data quality information for remote sensing applications', in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, pp. 15–22.

Antoniou, V. (2017) 'Assessing VGI Data Quality', in *Mapping and the Citizen Sensor*. Ubiquity Press, pp. 137–163.

Arolfo, F. and Vaisman, A. (2018) 'Data Quality in a Big Data Context', *Advances in Databases and Information Systems*, LNCS(11019), pp. 159–172.

Arthur, R. *et al.* (2018) 'Social sensing of floods in the UK', *PLoS ONE*, 13(1).

Arts, D. G. T., De Keizer, N. F. and Scheffer, G.-J. (2002) 'Defining and improving data quality in medical registries: A literature review, case study, and generic framework', *Journal of the American Medical Informatics Association*, 9(6), pp. 600–611.

Asur, S. and Huberman, B. A. (2010) 'Predicting the future with social media', in *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010*, pp. 492–499.

Atlas of Living Australia (2021) *Open access to Australia's biodiversity data*. Available at: https://www.ala.org.au/ (Accessed: 17 March 2021).

Ayeh, J. K., Au, N. and Law, R. (2013) '"Do We Believe in TripAdvisor?" Examining Credibility Perceptions and Online Travelers' Attitude toward Using User-Generated Content', *Journal of Travel Research*, 52(4), pp. 437–452.

Ayuning Budi, N. F. *et al.* (2019) 'Developing information quality model and measuring

information quality for further improvement: A case of ERP system of a state owned company', in *Proceedings - 2018 4th International Conference on Computing, Engineering, and Design, ICCED 2018*, pp. 46–51.

Bai, L., Meredith, R. and Burstein, F. (2018) 'A data quality framework, method and tools for managing data quality in a health care setting: an action case study', *Journal of Decision Systems*, 27, pp. 144–154.

Barachi, M. E. *et al.* (2019) 'A Novel Quality and Reliability-Based Approach for Participants' Selection in Mobile Crowdsensing', *IEEE Access*, 7.

Barbosa, W. L. *et al.* (2019) 'Data quality problems identified in the bioclimatic data collection process - A survey [Problemas de Qualidade de Dados Identificados no Processo de Coleta de Dados Bioclimáticos - uma revisão da literatura]', in *Iberian Conference on Information Systems and Technologies, CISTI*

Barsi, Á. *et al.* (2019) 'Remote sensing data quality model : from data sources to lifecycle phases', *International Journal of Image and Data Fusion*, 00(00), pp. 1–20.

Batini, C. *et al.* (2006) 'A comprehensive data quality methodology for web and structured data', *2006 1st International Conference on Digital Information Management, ICDIM*, 1(3), pp. 448–456.

Batini, C. *et al.* (2009) 'Methodologies for data quality assessment and improvement', *ACM Computing Surveys*, 41(3), p. 16.

Batini, C. *et al.* (2015) 'From data quality to big data quality', *Journal of Database Management*, 26(1), pp. 60–82.

Batini, C. *et al.* (2017) 'Data quality in remote sensing', in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, pp. 447–453.

Batini, C. and Scannapieco, M. (2006) *Data quality : concepts, methodologies and techniques*. Springer

Batini, C. and Scannapieco, M. (2016) *Data and Information Quality: Dimensions, Principles and Techniques*. Cham: Springer International Publishing (Data-Centric Systems and Applications)

Bayona Oré, S. and Palomino Guerrero, C. (2018) 'Big data: Applications and challenges', in *Proceedings of the 32nd International Business Information Management Association Conference, IBIMA 2018 - Vision 2020: Sustainable Economic Development and Application of Innovation Management from Regional expansion to Global Growth*, pp. 6237–6244.

Bayraktarov, E. *et al.* (2019) 'Do big unstructured biodiversity data mean more knowledge?', *Frontiers in Ecology and Evolution*, 7(JAN), pp. 1–5.

Becker, D., King, T. D. and McMullen, B. (2015) 'Big data, big data quality problem', in *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pp. 2644–2653.

Beebe, N. L. and Walz, D. (2005) 'An empirical investigation of the impact of data quality and its antecedents on data warehousing success', in *Association for Information Systems - 11th Americas Conference on Information Systems, AMCIS 2005: A Conference on a Human Scale*, pp. 15–21.

Blake, R. and Mangiameli, P. (2011) 'The effects and interactions of data quality and problem complexity on classification', *Journal of Data and Information Quality*, 2(2).

Blatt, A. J. (2015) 'The benefits and risks of volunteered geographic information', *Journal of Map and Geography Libraries*, 11(1), pp. 99–104.

Bonacic, C., Neyem, A. and Vasquez, A. (2015) 'Live ANDES: Mobile-cloud shared workspace for citizen science and wildlife conservation', in *Proceedings - 11th IEEE International Conference on eScience, eScience 2015*. Institute of Electrical and Electronics Engineers Inc., pp. 215–223.

Bordogna, G. *et al.* (2016) 'On predicting and improving the quality of Volunteer Geographic Information projects', *International Journal of Digital Earth*. Taylor & Francis, pp. 134–155.

Bouadjenek, M. R., Zobel, J. and Verspoor, K. (2019) 'Automated assessment of biological database assertions using the scientific literature', *BMC Bioinformatics*, 20(1).

Bovee, M., Srivastava, R. P. and Mak, B. (2003) 'A conceptual framework and belief-function approach to assessing overall information quality', *International Journal of Intelligent Systems*, 18(1), pp. 51–74.

Brown, G. and Kyttä, M. (2014) 'Key issues and research priorities for public participation GIS (PPGIS): A synthesis based on empirical research', *Applied Geography*, 46, pp. 122–136.

Brunt, C. S., King, A. S. and King, J. T. (2020) 'The influence of user-generated content on video game demand', *Journal of Cultural Economics*, 44(1), pp. 35–56.

Budburst (2021) *Budburst: An online database of plant observations, a citizen-science project of the Chicago Botanic Garden. Glencoe, Illinois.* Available at: https://budburst.org/ (Accessed: 27 April 2021).

Buhrmester, M., Kwang, T. and Gosling, S. D. (2011) 'Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data?', *Perspectives on Psychological Science*, 6(1), pp. 3–5.

Buntain, C. and Golbeck, J. (2017) 'Automatically Identifying Fake News in Popular Twitter Threads', in *Proceedings - 2nd IEEE International Conference on Smart Cloud, SmartCloud 2017*, pp. 208–215.

Burgess, H. K. *et al.* (2017) 'The science of citizen science: Exploring barriers to use as a primary research tool', *Biological Conservation*, 208, pp. 113–120.

Caballero, I. *et al.* (2009) 'Tailoring data quality models using social network preferences', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5667 LNCS, pp. 152–166.

Cai, L. and Zhu, Y. (2015) 'The challenges of data quality and data quality assessment in the big data era', *Data Science Journal*, 14, pp. 1–10.

Carr, W. and Kemmis, S. (1986) 'Becoming Critical: Education Knowledge and Action Research', in

Cha, M. *et al.* (2008) 'I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System Meeyoung', pp. 14–15.

Cha, M. *et al.* (2009) 'Analyzing the video popularity characteristics of large-scale user generated content systems', *IEEE/ACM Transactions on Networking*, 17(5), pp. 1357–1370.

Chen, M., Mao, S. and Liu, Y. (2014) 'Big data: A survey', in *Mobile Networks and Applications*. Kluwer Academic Publishers, pp. 171–209.

Cinelli, M. *et al.* (2021) 'The echo chamber effect on social media', *Proceedings of the National Academy of Sciences of the United States of America*, 118(9).

Clarke, R. (2016) 'Big data, big risks', *Information Systems Journal*, 26(1), pp. 77–90.

Connelly, R. *et al.* (2016) 'The role of administrative data in the big data revolution in social science research', *Social Science Research*, 59, pp. 1–12.

Cornell Lab of Ornithology (2021) *eBird - Discover a new world of birding*. Available at: https://ebird.org/home (Accessed: 9 March 2021).

Cox, A. M., McKinney, P. and Goodale, P. (2017) 'Food logging: an information literacy

perspective', *Aslib Journal of Information Management*, 69(2), pp. 184–200.

Cross, I. and Joana, P. (2010) *Evaluating the Usability of Aggregated Datasets in the GIS4EU Project*. Available at: https://www.directionsmag.com/article/2130 (Accessed: 15 May 2020).

Culotta, A. (2010) 'Towards detecting influenza epidemics by analyzing Twitter messages', in *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*. New York, New York, USA: ACM Press, pp. 115–122.

D. Clark, M. (2020) 'DRAG THEM: A brief etymology of so-called "cancel culture"', *Communication and the Public*, 5(3–4), pp. 88–92.

DAMA UK (2013) *The Six Primary Dimensions for Data Quality Assessment - Defining Data Quality Dimensions*. Available at: https://www.dqglobal.com/wp-content/uploads/2013/11/DAMA-UK-DQ-Dimensions-White-Paper-R37-1.pdf (Accessed: 9 January 2019).

Davenport, T. and Prusak, L. (2000) 'Working knowledge: how organizations manage what they know', *Ubiquity*, 2000, p. 6.

Davison, R., Martinsons, M. and Kock, N. (2004) 'Principles of canonical action research', *Information Systems Journal*, 14.

DeLone, W. H. and McLean, E. (1992) 'Information Systems Success: The Quest for the Dependent Variable', *Inf. Syst. Res.*, 3, pp. 60–95.

DeLone, W. H. and McLean, E. (2003) 'The DeLone and McLean Model of Information Systems Success: A Ten-Year Update', *J. Manag. Inf. Syst.*, 19, pp. 9–30.

Demetriou, D. (2016) 'Uncertainty of OpenStreetMap data for the road network in Cyprus', in *Proceedings of SPIE - The International Society for Optical Engineering*

Dexun, J. *et al.* (2014) 'Functional Over-Related Classes Bad Smell Detection and Refactoring Suggestions', *International Journal of Software Engineering & Applications*, 5(2), pp. 29–47.

Dhar, V. and Chang, E. A. (2009) 'Does Chatter Matter? The Impact of User-Generated Content on Music Sales', *Journal of Interactive Marketing*, 23(4), pp. 300–307.

van Dijck, J. (2009) 'Users like you? Theorizing agency in user-generated content', *Media, Culture & Society*, 31(1), pp. 41–58.

Dodgson, L. (2020) *YouTubers' Channels Are Being Held Hostage With Fake Copyright Claims*. Available at: https://www.insider.com/youtubers-channels-

are-being-held-hostage-with-fake-copyright-claims-2020-6 (Accessed: 16 May 2021).

EDM Council (2017) *Data Quality Dimensions*. Available at: https://cdn.ymaws.com/edmcouncil.org/resource/resmgr/featured_documents/BP_DQ_Dimensions_Oct17.pdf (Accessed: 9 January 2019).

Eisenhardt, K. (1989) 'Building theories from case study research', *Academy of Management Review*, 14, pp. 532–550.

Elbroch, M. *et al.* (2011) 'The Value, Limitations, and Challenges of Employing Local Experts in Conservation Research', *Conservation Biology*, 25(6), pp. 1195–1202.

Elmagarmid, A. K., Ipeirotis, P. G. and Verykios, V. S. (2007) 'Duplicate record detection: A survey', *IEEE Transactions on Knowledge and Data Engineering*, 19(1), pp. 1–16.

Elwood, S., Goodchild, M. F. and Sui, D. Z. (2012) 'Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice', *Annals of the Association of American Geographers*, 102(3), pp. 571–590.

Engström, E. *et al.* (2020) 'How software engineering research aligns with design science: a review', *Empirical Software Engineering*, 25(4), pp. 2630–2660.

Eppler, M. (2001) 'A Generic Framework for Information Quality in knowledge-intensive Processes', *http://www.alexandria.unisg.ch/Publikationen/54874*

Facebook (2021) *Facebook*. Available at: https://www.facebook.com/ (Accessed: 21 June 2021).

Fadahunsi, K. P. *et al.* (2019) 'Protocol for a systematic review and qualitative synthesis of information quality frameworks in eHealth', *BMJ Open*, 9(3).

Fehrenbacher, D. D. and Helfert, M. (2008) 'An empirical research on the evaluation of data quality dimensions', in *Proceedings of the 2008 International Conference on Information Quality, ICIQ 2008*

Fidler, M. and Lavbic, D. (2017) 'Improving information quality of Wikipedia articles with cooperative principle', *Online Information Review*, 41(6), pp. 797–811.

Filieri, R., Alguezaui, S. and McLeay, F. (2015) 'Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth', *Tourism Management*, 51,

pp. 174–185.

Firmani, D. *et al.* (2016) 'On the Meaningfulness of "Big Data Quality" (Invited Paper)', *Data Science and Engineering*, 1(1), pp. 6–20.

Fogliaroni, P., D'Antonio, F. and Clementini, E. (2018) 'Data trustworthiness and user reputation as indicators of VGI quality', *Geo-Spatial Information Science*, 21(3), pp. 213–233.

Foley, O., Helfert, M. and Elwood, L. (2010) 'The impact of diverse information systems environments on information quality - a design science approach', in *4th European Conference on Information Management and Evaluation, ECIME 2010*, pp. 77–85.

Fonte, C. C. *et al.* (2017) 'Assessing VGI Data Quality', in *Mapping and the Citizen Sensor*. Ubiquity Press, pp. 137–163.

Foody, G. M. *et al.* (2015) 'Accurate Attribute Mapping from Volunteered Geographic Information: Issues of Volunteer Quantity and Quality', *The Cartographic Journal*, 52(4), pp. 336–344.

Fox, T. L. *et al.* (1999) 'Maintaining Quality in Information Systems', *Journal of Computer Information Systems*, 40(1), pp. 76–80.

Freiwald, J. *et al.* (2018) 'Citizen science monitoring of marine protected areas: Case studies and recommendations for integration into monitoring programs', *Marine Ecology*, 39.

Fritz, S., Fonte, C. C. and See, L. (2017) 'The role of Citizen Science in Earth Observation', *Remote Sensing*, 9(4).

Garcia, D. *et al.* (2017) 'Big data analytics and knowledge discovery applied to automatic meter readers', *Advances in Industrial Control*, (9783319507507), pp. 401–423.

Ge, M. and Dohnal, V. (2018) 'Quality management in big data', *Informatics*, 5(2).

Ghasemaghaei, M. and Calic, G. (2019) 'Can big data improve firm decision quality? The role of data quality and data diagnosticity', *Decision Support Systems*, 120, pp. 38–49.

Ghose, A., Ipeirotis, P. G. and Li, B. (2012) 'Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content', *Marketing Science*, 31(3), pp. 493–520.

Girres, J.-F. and Touya, G. (2010) 'Quality Assessment of the French OpenStreetMap Dataset', *Transactions in GIS*, 14(4), pp. 435–459.

Glaser, B. G., Strauss, A. L. and Strutzel, E. (1968) 'The Discovery of Grounded Theory: Strategies for Qualitative Research', *Nursing Research*, 17, p. 364.

Globe at Night (2021) *International citizen-science campaign to raise public awareness of the impact of light pollution*. Available at: https://www.globeatnight.org/ (Accessed: 27 April 2021).

Goh, K. Y., Heng, C. S. and Lin, Z. (2013) 'Social media brand community and consumer behavior: Quantifying the relative impact of user- and marketer-generated content', *Information Systems Research*, 24(1), pp. 88–107.

Goodman, J. and Carmichael, F. (2020) *US election 2020: 'Rigged' votes, body doubles and other false claims*, *BBC News*. Available at: https://www.bbc.com/news/54562611 (Accessed: 25 October 2020).

Gouveia, C. *et al.* (2004) 'Promoting the use of environmental data collected by concerned citizens through information and communication technologies', *Journal of Environmental Management*, 71(2), pp. 135–154.

Guan, Z. *et al.* (2017) 'A survey on big data pre-processing', in *Proceedings - 2017 5th International Conference on Applied Computing and Information Technology, 2017 4th International Conference on Computational Science/Intelligence and Applied Informatics and 2017 1st International Conference on Big Data, Cloud Compu*, pp. 241–247.

Guo, B. *et al.* (2015) 'TaskMe: A cross-community, quality-enhanced incentive mechanism for mobile crowd sensing', in *UbiComp and ISWC 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pp. 49–52.

Haasch, P. (2020) *TikTok Teens, K-Pop Stans Organizing to Report Donald Trump's Twitter*. Available at: https://www.insider.com/tiktok-teens-mass-report-president-trump-twitter-instagram-kpop-stans-2020-6 (Accessed: 16 May 2021).

Han, J., Jiang, D. and Ding, Z. (2009) 'Assessing data quality within available context', in *Data Quality and High-Dimensional Data Analysis - Proceedings of the DASFAA 2008 Workshops*, pp. 42–59.

Haralabopoulos, G., Anagnostopoulos, I. and Zeadally, S. (2016) 'The challenge of improving credibility of user-generated content in online social networks', *Journal of Data and Information Quality*, 7(3).

Hashem, I. A. T. *et al.* (2015) 'The rise of "big data" on cloud computing: Review and open research issues', *Information Systems*, 47, pp. 98–115.

Haug, A. *et al.* (2013) 'Master data quality barriers: An empirical investigation', *Industrial Management and Data Systems*, 113(2), pp. 234–249.

Haug, A. and Arlbjørn, J. S. (2011) 'Barriers to master data quality', *Journal of Enterprise Information Management*, 24(3), pp. 288–303.

Haug, A., Arlbjrn, J. S. and Pedersen, A. (2009) 'A classification model of ERP system data quality', *Industrial Management and Data Systems*, 109(8), pp. 1053–1068.

Havlik, D. *et al.* (2013) 'Robust and Trusted Crowd-Sourcing and Crowd-Tasking in the Future Internet', in *IFIP Advances in Information and Communication Technology*. Springer New York LLC, pp. 164–176.

Haworth, B. T. *et al.* (2018) 'The good, the bad, and the uncertain: Contributions of volunteered geographic information to community disaster resilience', *Frontiers in Earth Science*, 6.

Hecht, J. and Spicer Rice, E. (2015) 'Citizen science: A new direction in canine behavior research', *Behavioural Processes*, 110, pp. 125–132.

Heinrich, B. *et al.* (2018) 'Assessing data quality – A probability-based metric for semantic consistency', *Decision Support Systems*, 110, pp. 95–106.

Heinrich, B. *et al.* (2019) 'Data quality in recommender systems: the impact of completeness of item content data on prediction accuracy of recommender systems', *Electronic Markets*

Hempel, C. (1966) 'Philosophy of Natural Science'

Hempel, G., Feigl, H. and Marxwell, G. (1962) 'Deductive-Nomological vs. Statistical Explanation'

Hevner, A. R. *et al.* (2004) *Design science in information systems research, MIS Quarterly: Management Information Systems*

Huang, C. Y. and Liang, S. (2014) 'A sensor data mediator bridging the OGC Sensor Observation Service (SOS) and the OASIS Open Data Protocol (OData)', *Annals of GIS*, 20(4), pp. 279–293.

Hulitt, E. and Vaughn, R. B. (2010) 'Information system security compliance to FISMA standard: A quantitative measure', *Telecommunication Systems*, 45(2–3), pp. 139–152.

Hunter, J., Alabri, A. and Van Ingen, C. (2013) 'Assessing the quality and trustworthiness of citizen science data', *Concurrency Computation Practice and Experience*,

25(4), pp. 454–466.

Hyder, K. *et al.* (2015) 'Can citizen science contribute to the evidence-base that underpins marine policy?', *Marine Policy*, 59, pp. 112–120.

IBM (2019) *Extracting business value from the 4 V's of big data | IBM Big Data & Analytics Hub*. Available at: https://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data (Accessed: 27 January 2021).

IGI Global (2021) *What is Information Consistency | IGI Global*. Available at: https://www.igi-global.com/dictionary/information-consistency/14370 (Accessed: 25 May 2021).

Immonen, A., Pääkkönen, P. and Ovaska, E. (2015) 'Evaluating the Quality of Social Media Data in Big Data Architecture', *IEEE Access*, 3, pp. 2028–2043.

iNaturalist (2021) *A Community for Naturalists · iNaturalist.org*. Available at: https://www.inaturalist.org/ (Accessed: 29 March 2021).

Influencer Marketing Hub (2020) *42 Essential Social Media Statistics for 2020*, *Influencer Marketing Hub*. Available at: https://influencermarketinghub.com/social-media-statistics-2020/ (Accessed: 28 April 2020).

Instagram (2021) *Instagram*. Available at: https://www.instagram.com/ (Accessed: 21 June 2021).

ISO (2008) 'ISO/IEC 25012:2008 Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model'. ISO. Available at: https://www.iso.org/standard/35736.html (Accessed: 9 January 2019).

ISO (2013) 'ISO 19157:2013 - Geographic information — Data quality'. Available at: https://www.iso.org/standard/32575.html (Accessed: 22 January 2021).

James, T. (2006) 'Improving wildlife data quality: guidance on data verification, validation and their application in biological recording. Guidance manual', *undefined*

Jeffrey, R. and Popper, K. (1934) 'The Logic of Scientific Discovery', *Econometrica*, 28, p. 925.

Jesmeen, M. Z. H. *et al.* (2018) 'A survey on cleaning dirty data using machine learning paradigm for big data analytics', *Indonesian Journal of Electrical Engineering and*

*Computer Science*, 10(3), pp. 1234–1243.

Kabir, M. Y. and Madria, S. (2020) 'CoronaVis: A Real-time COVID-19 Tweets Data Analyzer and Data Repository'. Available at: http://arxiv.org/abs/2004.13932 (Accessed: 4 May 2021).

Kaplan, A. M. and Haenlein, M. (2010) 'Users of the world, unite! The challenges and opportunities of Social Media', *Business Horizons*, 53(1), pp. 59–68.

Kaur, J. *et al.* (2018) 'Systematic literature review of data quality within openstreetmap', in *Proceedings - 2017 International Conference on Next Generation Computing and Information Systems, ICNGCIS 2017*, pp. 159–163.

Kelling, S. *et al.* (2013) 'E Bird: A human/computer learning network to improve biodiversity conservation and research', *AI Magazine*, 34(1), pp. 10–20.

Kietzmann, J. H. *et al.* (2011) 'Social media? Get serious! Understanding the functional building blocks of social media', *Business Horizons*, 54(3), pp. 241–251.

King, R. A., Racherla, P. and Bush, V. D. (2014) 'What we know and don't know about online word-of-mouth: A review and synthesis of the literature', *Journal of Interactive Marketing*, 28(3), pp. 167–183.

Kosmala, M. *et al.* (2016) 'Assessing data quality in citizen science', *Frontiers in Ecology and the Environment*, 14(10), pp. 551–560.

Kotsev, A. *et al.* (2016) 'Next generation air quality platform: Openness and interoperability for the internet of things', *Sensors (Switzerland)*, 16(3).

Krejcie, R. V and Morgan, D. W. (1970) 'Determining Sample Size for Research Activities', *Educational and Psychological Measurement*, 30(3), pp. 607–610.

Krumm, J., Davies, N. and Narayanaswami, C. (2008) 'User-Generated Content', *IEEE Pervasive Computing*, 7(4), pp. 10–11.

Lakshen, G. A., Janev, V. and Vraneš, S. (2018) 'Challenges in quality assessment of Arabic dbpedia', in *ACM International Conference Proceeding Series*

Lansley, G. and Cheshire, J. (2018) 'Challenges to representing the population from new forms of consumer data', *Geography Compass*, 12(7).

Laumer, S., Maier, C. and Weitzel, T. (2017) 'Information quality, user satisfaction, and the manifestation of workarounds: a qualitative and quantitative study of enterprise content management system users', *European Journal of Information Systems*, 26(4), pp. 333–360.

Lee, Y. W. *et al.* (2002) 'AIMQ: A methodology for information quality assessment', *Information and Management*, 40(2), pp. 133–146.

Leibovici, D. G. *et al.* (2017) 'On data quality assurance and conflation entanglement in crowdsourcing for environmental studies', *ISPRS International Journal of Geo-Information*, 6(3).

Levenshtein, V. I. (1965) 'Binary Codes Capable of Correcting Deletions, Insertions and Reversals', *Societ physics, Doklady*, 10, pp. 707–710.

Lintott, C. *et al.* (2010) *Galaxy Zoo 1 : Data Release of Morphological Classifications for nearly 900,000 galaxies ⋆, Mon. Not. R. Astron. Soc*

Litman, L., Robinson, J. and Rosenzweig, C. (2015) 'The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk', *Behavior Research Methods*, 47(2), pp. 519–528.

Lu, X. *et al.* (2019) 'A Universal Measure for Network Traceability', *Omega (United Kingdom)*, 87, pp. 191–204.

Ludwig, T., Reuter, C. and Pipek, V. (2015) 'Social Haystack: Dynamic Quality Assessment of Citizen-Generated Content during Emergencies', *ACM Trans. Comput.-Hum. Interact*, 22.

Lukyanenko, R. *et al.* (2019) 'Expecting the unexpected: Effects of data collection design choices on the quality of crowdsourced user-generated content', *MIS Quarterly: Management Information Systems*, 43(2), pp. 634–647.

Lukyanenko, R. and Parsons, J. (2011) 'Rethinking data quality as an outcome of conceptual modeling choices', in *ICIQ 2011 - Proceedings of the 16th International Conference on Information Quality*, pp. 244–258.

Lukyanenko, R., Parsons, J. and Wiersma, Y. (2011) 'Citizen Science 2.0: Data Management Principles to Harness the Power of the Crowd', in *DESRIST*. Springer, Berlin, Heidelberg, pp. 465–473.

Lukyanenko, R., Parsons, J. and Wiersma, Y. (2014) 'The IQ of the crowd: Understanding and improving information quality in structured user-generated content', *Information Systems Research*, 25(4), pp. 669–689.

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

Lyons, K. (2020) *YouTube took down more videos than ever last quarter - The Verge*. Available at: https://www.theverge.com/2020/8/25/21401435/youtube-videos-moderators-filters-human-appeals (Accessed: 3 May 2021).

MacKechnie, C. *et al.* (2011) 'The role of "Big Society" in monitoring the state of the natural environment', *Journal of Environmental Monitoring*, 13(10), pp. 2687–2691.

Mariani, M., Di Fatta, G. and Di Felice, M. (2019) 'Understanding Customer Satisfaction with Services by Leveraging Big Data: The Role of Services Attributes and Consumers' Cultural Background', *IEEE Access*, 7, pp. 8195–8208.

Mehmood, K., Cherfi, S. S.-S. and Comyn-Wattiau, I. (2009) 'Data quality through conceptual model quality - Reconciling researchers and practitioners through a customizable quality model'

Mensah, S. *et al.* (2020) 'A Probabilistic Model for User Interest Propagation in Recommender Systems', *IEEE Access*, 8, pp. 108300–108309.

Mezzanzanica, M. *et al.* (2014) 'Improving data cleansing accuracy a model-based approach', in *DATA 2014 - Proceedings of 3rd International Conference on Data Management Technologies and Applications*, pp. 189–201.

Middleton, S. E., Middleton, L. and Modafferi, S. (2014) 'Real-time crisis mapping of natural disasters using social media', *IEEE Intelligent Systems*, 29(2), pp. 9–17.

Mihindukulasooriya, N. *et al.* (2015) 'An analysis of the quality issues of the properties available in the Spanish DBpedia', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9422, pp. 198–209.

Mitchell, N. *et al.* (2017) 'Benefits and challenges of incorporating citizen science into university education', *PLoS ONE*, 12(11).

Mooney, P. *et al.* (2012) 'Citizen generated spatial data and information: Risks and opportunities', in *Proceedings of the 2012 International Conference on Industrial Control and Electronics Engineering, ICICEE 2012*, pp. 1990–1993.

Musto, J. and Dahanayake, A. (2018) *Overview of data storing techniques in citizen science applications, Communications in Computer and Information Science*

Musto, J. and Dahanayake, A. (2019) 'Integrating data quality requirements to citizen science application design', in *11th International Conference on Management of Digital EcoSystems, MEDES 2019*

Musto, J. and Dahanayake, A. (2020) 'Improving data quality , privacy and provenance in citizen science applications', *Frontiers in Artificial Intelligence*, 321, pp. 141–160.

Musto, J. and Dahanayake, A. (2021a) 'An Approach to Improve the Quality of User-Generated Content of Citizen Science Platforms', *ISPRS International Journal of Geo-Information*, 10(7), p. 434.

Musto, J. and Dahanayake, A. (2021b) 'Quality characteristics for user-generated content', *Frontiers in Artificial Intelligence*

Naghizade, E. *et al.* (2015) 'How private can i be among public users?', in *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1137–1141.

Nelson, R. R., Todd, P. A. and Wixom, B. H. (2005) 'Antecedents of information and system quality: An empirical examination within the context of data warehousing', *Journal of Management Information Systems*, 21(4), pp. 199–235.

Nevolin, I. (2017) 'Crowdsourcing Opportunities for Research Information Systems', in *Procedia Computer Science*, pp. 19–24.

Newman, G. *et al.* (2010) 'User-friendly web mapping: Lessons from a citizen science website', *International Journal of Geographical Information Science*, 24(12), pp. 1851–1869.

Nicolaou, A. I. and McKnight, D. H. (2006) 'Perceived information quality in data exchanges: Effects on risk, trust, and intention to use', *Information Systems Research*, 17(4), pp. 332–351.

Nkonyana, T. and Twala, B. (2018) 'Impact of poor data quality in remotely sensed data', *Advances in Intelligent Systems and Computing*, 668, pp. 79–86.

Notopoulos, K. (2017) *How Trolls Locked My Twitter Account For 10 Days, And Welp*. Available at: https://www.buzzfeednews.com/article/katienotopoulos/how-trolls-locked-my-twitter-account-for-10-days-and-welp (Accessed: 16 May 2021).

OASIS (2017) *OASIS Open Data Protocol (OData) TC | OASIS*. Available at: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=odata (Accessed: 29 March 2019).

Oliver, I., Miche, Y. and Ren, W. (2018) 'A model for metricising privacy and legal compliance', in *Proceedings - 2018 International Conference on the Quality of Information and Communications Technology, QUATIC 2018*, pp. 229–237.

OpenStreetMap (2021) *OpenStreetMap*. Available at: https://www.openstreetmap.org/#map=5/65.453/26.069 (Accessed: 21 April 2021).

Ouyang, S., Li, C. and Li, X. (2016) 'A peek into the future: Predicting the popularity of online videos', *IEEE Access*, 4, pp. 3026–3033.

Palacin-Silva, M. and Porras, J. (2018) 'Shut up and take my environmental data! A study on ICT enabled citizen science practices, participation approaches and challenges', in *Proceedings of the 5th International Conference on Information and Communication Technology for Sustainability - ICT4S2018*, pp. 270–288.

Peer, E. *et al.* (2017) 'Beyond the Turk: Alternative platforms for crowdsourcing behavioral research', *Journal of Experimental Social Psychology*, 70, pp. 153–163.

Pennycook, G. *et al.* (2020) 'Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention', *Psychological Science*, 31(7), pp. 770–780.

Peters, J. (2020) *Twitter will remove misleading COVID-19-related tweets that could incite people to engage in 'harmful activity'*, *The Verge*. Available at: https://www.theverge.com/2020/4/22/21231956/twitter-remove-covid-19-tweets-call-to-action-harm-5g (Accessed: 25 October 2020).

Petter, S., Delone, W. and McLean, E. R. (2013) 'Information systems success: The quest for the independent variables', *Journal of Management Information Systems*, 29(4), pp. 7–62.

Pipino, L. L., Lee, Y. W. and Wang, R. Y. (2002) 'Data Quality Assessment', *Communications of the ACM*, 45(4), pp. 211–218.

Polk, T., Johnston, M. P. and Evers, S. (2015) 'Wikipedia Use in Research: Perceptions in Secondary Schools', *TechTrends*, 59(3), pp. 92–102.

Popat, K. *et al.* (2017) 'Where the truth lies: Explaining the credibility of emerging claims on the web and social media', in *26th International World Wide Web Conference 2017, WWW 2017 Companion*. International World Wide Web Conferences Steering Committee, pp. 1003–1012.

Quinn, M. (2020) 'Twitter removes tweet shared by Trump with false coronavirus statistics - CBS News', *CBS News*. Available at: https://www.cbsnews.com/news/twitter-removes-trump-tweet-false-coronavirus-statistics/ (Accessed: 25 October 2020).

Ranjan, S., Sood, S. and Verma, V. (2019) 'Twitter Sentiment Analysis of Real-Time Customer Experience Feedback for Predicting Growth of Indian Telecom Companies', in *Proceedings - 4th International Conference on Computing Sciences, ICCS 2018*. Institute of Electrical and Electronics Engineers Inc., pp. 166–174.

Ratnieks, F. L. W. *et al.* (2016) 'Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers', *Methods in Ecology and Evolution*, 7(10), pp. 1226–1235.

Redman, T. C. (1996) *Data quality for the information age*. Artech House

Rees, E. E. *et al.* (2011) 'Advancements in web-database applications for rabies surveillance', *International Journal of Health Geographics*, 10.

Reiss, J. and Sprenger, J. (2017) 'Scientific Objectivity', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Winter 201. Metaphysics Research Lab, Stanford University

Rice, R. M. (2015) 'Ensuring the quality of volunteered geographic information: A social approach [Qualitätssicherung von freiwillig generierten Geodaten: Ein sozialer Ansatz]', *Kartographische Nachrichten*, 2015(3), pp. 123–130.

Rieh, S. Y. (2002) 'Judgment of information quality and cognitive authority in the Web', *Journal of the American Society for Information Science and Technology*, 53(2), pp. 145–161.

Ritzer, G. and Jurgenson, N. (2010) 'Production, Consumption, Prosumption', *Journal of Consumer Culture*, 10(1), pp. 13–36.

Robertson, C. and Feick, R. (2016) 'Bumps and bruises in the digital skins of cities: unevenly distributed user-generated content across US urban areas', *Cartography and Geographic Information Science*, 43(4), pp. 283–300.

Roman, L. A. *et al.* (2017) 'Data quality in citizen science urban tree inventories', *Urban Forestry and Urban Greening*, 22, pp. 124–135.

Sadiq, S. and Indulska, M. (2017) 'Open data: Quality over quantity', *International Journal of Information Management*, 37(3), pp. 150–154.

Salk, C. F. *et al.* (2016) 'Assessing quality of volunteer crowdsourcing contributions: lessons from the Cropland Capture game', *International Journal of Digital Earth*, 9(4), pp. 410–426.

Schmidt, M. *et al.* (2015) 'The Danish National patient registry: A review of content, data quality, and research potential', *Clinical Epidemiology*, 7, pp. 449–490.

See, L. *et al.* (2016) 'Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information', *ISPRS International Journal of Geo-Information*, 5(5), p. 55.

Senarath, A., Grobler, M. and Arachchilage, N. A. G. (2019) 'A model for system developers to measure the privacy risk of data', in *HICSS*

Sheppard, S. A., Wiggins, A. and Terveen, L. (2014) 'Capturing quality', in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. New York, New York, USA: ACM Press, pp. 1234–1245.

Signorini, A., Segre, A. M. and Polgreen, P. M. (2011) 'The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic', *PLoS ONE*. Edited by A. P. Galvani, 6(5), p. e19467.

Simpson, R., Page, K. R. and De Roure, D. (2014) 'Zooniverse: Observing the world's largest citizen science platform', in *WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web*. New York, New York, USA: Association for Computing Machinery, Inc, pp. 1049–1054.

Siponen, M. and Klaavuniemi, T. (2020) 'Why is the hypothetico-deductive (H-D) method in information systems not an H-D method?', *Inf. Organ.*, 30, p. 100287.

Smart Insights (2020) *Global social media research summary August 2020 | Smart Insights*. Available at: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ (Accessed: 27 January 2021).

Smith, M. *et al.* (2018) 'Assessing the quality of administrative data for research: A framework from the manitoba centre for health policy', *Journal of the American Medical Informatics Association*, 25(3), pp. 224–229.

Al Sohibani, M. *et al.* (2015) 'Factors That Influence the Quality of Crowdsourcing', *Advances in Intelligent Systems and Computing*, 312, pp. 287–300.

Spielhofer, T. *et al.* (2017) 'Data mining twitter during the UK floods Investigating the potential use of social media in emergency management', in *Proceedings of the 2016 3rd International Conference on Information and Communication Technologies for Disaster Management, ICT-DM 2016*

Stang, J. *et al.* (2008) 'A generic data quality framework applied to the product data for naval vessels', in *Proceedings of the 2008 International Conference on Information Quality, ICIQ 2008*

Steger, C., Butt, B. and Hooten, M. B. (2017) 'Safari Science: assessing the reliability of citizen science data for wildlife surveys', *Journal of Applied Ecology*, 54(6), pp. 2053–2062.

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

Sullivan, B. L. *et al.* (2014) 'The eBird enterprise: An integrated approach to development and application of citizen science', *Biological Conservation*, 169, pp. 31–40.

Sun, W. *et al.* (2018) 'Data processing and text mining technologies on electronic medical records: A review', *Journal of Healthcare Engineering*, 2018.

Susarla, A., Oh, J. H. and Tan, Y. (2012) 'Social networks and the diffusion of user-generated content: Evidence from youtube', *Information Systems Research*, 23(1), pp. 23–41.

Syed-Abdul, S. *et al.* (2013) 'Misleading health-related information promoted through video-based social media: Anorexia on youtube', *Journal of Medical Internet Research*, 15(2), p. e30.

Taleb, I., Dssouli, R. and Serhani, M. A. (2015) 'Big Data Pre-processing: A Quality Framework', in *Proceedings - 2015 IEEE International Congress on Big Data, BigData Congress 2015*, pp. 191–198.

TechJury (2020) *How Much Data Is Created Every Day in 2020? [You'll be shocked!]*. Available at: https://techjury.net/blog/how-much-data-is-created-every-day/ (Accessed: 27 January 2021).

Tenkanen, H. *et al.* (2017) 'Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas', *Scientific Reports*, 7(1), p. 17615.

Tian, J.-W. *et al.* (2012) 'A quality evaluation model for enterprise information portals', *International Journal of Advancements in Computing Technology*, 4(23), pp. 153–160.

Tiley, A. L. *et al.* (2019) 'Kross-Sami: A direct IFS comparison of the Tully-Fisher relation across 8 Gyr since z ≈ 1', *Monthly Notices of the Royal Astronomical Society*, 482(2), pp. 2166–2188.

Tilly, R. *et al.* (2017) 'Towards a Conceptualization of Data and Information Quality in Social Information Systems', *Business and Information Systems Engineering*, 59(1), pp. 3–21.

TINT (2020) *46 Mind-Blowing Stats About User-Generated Content (2020 Edition) - The TINT Blog*. Available at: https://www.tintup.com/blog/user-generated-content-stats-study/ (Accessed: 27 January 2021).

Tirunillai, S. and Tellis, G. J. (2012) 'Does chatter really matter? Dynamics of user-

generated content and stock performance', *Marketing Science*, 31(2), pp. 198–215.

Trujillo, G. *et al.* (2015) *Virtualizing hadoop: how to install, deploy, and optimize hadoop in a virtualized architecture*

Truong, Q. T., de Runz, C. and Touya, G. (2019) 'Analysis of collaboration networks in OpenStreetMap through weighted social multigraph mining', *International Journal of Geographical Information Science*, 33(8), pp. 1651–1682.

Twitter (2020) *Twitter*. Available at: https://twitter.com/home (Accessed: 12 April 2020).

Twitter (2021) *The Twitter rules: safety, privacy, authenticity, and more*. Available at: https://help.twitter.com/en/rules-and-policies/twitter-rules (Accessed: 16 May 2021).

Twitter Safety (2020) *Expanding our policies to further protect the civic conversation*. Available at: https://blog.twitter.com/en_us/topics/company/2020/civic-integrity-policy-update.html (Accessed: 25 October 2020).

Varlamis, I. (2010) 'Quality of content in Web 2.0 applications', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6278 LNAI(PART 3), pp. 33–42.

Veen, L. E. *et al.* (2012) 'A semantically integrated, user-friendly data model for species observation data', *Ecological Informatics*, 8, pp. 1–9.

Vincent, N. *et al.* (2019) *Measuring the Importance of User-Generated Content to Search Engines, Proceedings of the International AAAI Conference on Web and Social Media*

Viviani, M. and Pasi, G. (2017) 'Credibility in social media: opinions, news, and health information-a survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(5), p. e1209.

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

Wang, R. Y., Storey, V. C. and Firth, C. P. (1995) 'A Framework for Analysis of Data Quality Research', *IEEE Transactions on Knowledge and Data Engineering*, 7(4), pp. 623–640.

Wang, R. Y. and Strong, D. M. (1996) 'Beyond accuracy: What data quality means to data consumers', *Journal of Management Information Systems*, 12(4), pp. 5–34.

Watts, S., Shankaranarayanan, G. and Even, A. (2009) 'Data quality assessment in context: A cognitive perspective', *Decision Support Systems*, 48(1), pp. 202–211.

Wei, X. *et al.* (2018) 'Data Quality Aware Task Allocation with Budget Constraint in Mobile Crowdsensing', *IEEE Access*, 6, pp. 48010–48020.

White, D. L. *et al.* (2014) 'The vanishing firefly project: Engaging citizen scientists with a mobile technology and real-time reporting framework', in *Proceedings of the 5th ACM SIGSPATIAL International Workshop on GeoStreaming, IWGS 2014*, pp. 85–92.

Wiersma, Y. F. (2010) 'Birding 2.0: Citizen science and effective monitoring in the web 2.0 world [Ornithologie 2.0: La science citoyenne et les programmes de suivi à l'ère d'internet 2.0]', *Avian Conservation and Ecology*, 5(2).

Wigand, R. T., Wood, J. and Yiliyasi, Y. (2009) 'Information quality issues in the mortgage banking industry', in *Proceedings of the 2009 International Conference on Information Quality, ICIQ 2009*

Wiggins, A. and Crowston, K. (2015) 'Surveying the citizen science landscape', *First Monday*, 20(1).

Wiggins, A. and He, Y. (2016) 'Community-based data validation practices in citizen science', in *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. New York, New York, USA: ACM Press, pp. 1548–1559.

Wikipedia (2019) *Wikipedia:Academic use*. Available at: https://en.wikipedia.org/wiki/Wikipedia:Academic_use (Accessed: 7 May 2020).

Wikipedia (2020) *Wikipedia*. Available at: https://www.wikipedia.org/ (Accessed: 12 April 2020).

Wikipedia (2021a) *Wikipedia:About - Wikipedia*. Available at: https://en.wikipedia.org/wiki/Wikipedia:About (Accessed: 26 April 2021).

Wikipedia (2021b) *Wikipedia:Policies and guidelines - Wikipedia*. Available at: https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines (Accessed: 16 May 2021).

Williamson, K. *et al.* (2016) 'Data sharing for the advancement of science: Overcoming barriers for citizen scientists', *Journal of the Association for Information Science and Technology*, 67(10), pp. 2392–2403.

Worldometer (2020) *Worldometer - real time world statistics*. Available at: https://www.worldometers.info/ (Accessed: 12 April 2020).

Worrall, W. (2020) *YouTube Has a Massive False Copyright Claim Problem*. Available at: https://www.ccn.com/youtube-has-massive-false-copyright-claim-problem/ (Accessed: 16 May 2021).

Wrabetz, J. (2017) *Measuring the economic value of data*, *Network World*. Available at: https://www.networkworld.com/article/3221387/measuring-the-economic-value-of-data.html (Accessed: 17 May 2020).

Wu, D., Li, H. and Wang, R. (2018) 'User characteristic aware participant selection for mobile crowdsensing', *Sensors (Switzerland)*, 18(11).

Wyrwoll, C. (2014) *Social Media*, *Social Media*. Wiesbaden: Springer Fachmedien Wiesbaden

Xiang, Z. *et al.* (2018) 'Assessing reliability of social media data: lessons from mining TripAdvisor hotel reviews', *Information Technology and Tourism*, 18(1–4), pp. 43–59.

Xiaojiang, F., Liwei, Z. and Jianbin, L. (2017) 'Measurement for social network data currency and trustworthiness', in *2017 2nd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2017*, pp. 1–5.

Xiong, J. *et al.* (2018) 'MAIM: A Novel Incentive Mechanism Based on Multi-Attribute User Selection in Mobile Crowdsensing', *IEEE Access*, 6, pp. 65384–65396.

Yan, Y. *et al.* (2017) 'Monitoring and assessing post-disaster tourism recovery using geotagged social media data', *ISPRS International Journal of Geo-Information*, 6(5).

Yang, F., Feng, J. and Fabbrizio, G. Di (2006) *A Data Driven Approach to Relevancy Recognition for Contextual Question Answering*. Available at: http://www.ask.com/ (Accessed: 15 May 2020).

Ye, Q. *et al.* (2011) 'The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings', *Computers in Human Behavior*, 27(2), pp. 634–639.

YouTube (2020) *YouTube*. Available at: https://www.youtube.com/ (Accessed: 8 May 2020).

Yu, J. *et al.* (2012) 'Automated data verification in a large-scale citizen science project: A case study', in *2012 IEEE 8th International Conference on E-Science, e-Science 2012*

Zhao, B. and Sui, D. Z. (2017) 'True lies in geospatial big data: detecting location

spoofing in social media', *Annals of GIS*, 23(1), pp. 1–14.

# Publication I

Musto, J., and Dahanayake, A.
**Overview of Data Storing Techniques in Citizen Science Applications**

In: Benczúr A. et al. (eds) New Trends in Databases and Information Systems. ADBIS 2018.

Reprinted with permission from
*Communications in Computer and Information Science*
Vol. 909, pp. 231-241, 2018

# Overview of data storing techniques in citizen science applications

Jiri Musto[1] and Ajantha Dahanayake[1]

[1] Lappeenranta University of Technology, Lappeenranta 53850, Finland
Jiri.Musto@lut.fi, Ajantha.Dahanayake@lut.fi

**Abstract.** Interest in citizen science and the number of citizen science projects have increased considerably during the last decade. Citizen science revolves around gathering data and using it. This means, that data storing is a vital part of any citizen science project and can affect the success or failure. Many researches focus on the citizen side, while the data side is often left out. This study aims to fill the gap by trying to find the current data storing practices in the field of citizen science. A systematic literature review was conducted and multiple similarities in data storing and management techniques were identified between different citizen science projects. Results show that most projects used a traditional relational database to store data, a separate web interface to add, use, modify, and access the data, and data validation was left to users by having them vote on existing data. Data models always considered the data provider (citizen) but left out the end user in their design. The systematic literature review has limitations, as the used databases and opinions of the researcher can affect the screening of papers.

**Keywords:** Citizen Science, Data Storing, Data Management

## Introduction

Citizen science has had different definitions over the years [1]. Oxford dictionary defines citizen science as “The collection and analysis of data relating to the natural world by members of the general public, typically as part of a collaborative project with professional scientists.” [2]

However, in this research citizen science is defined as a field where citizens partici-pate in collecting, analysing or reporting data for a particular purpose. The purpose can be for example scientific research, to provide information for other citizens, or to provide data for government. The citizens can be the end users as well but not neces-sarily. The following terms have been associated with or used as synonyms to citizen science: crowdsourcing, community-based monitoring, public participation, volunteer monitoring, volunteered geographic information (VGI) [1, 3–5].

The concept of citizen science is not particularly new as it has been around for decades. However, it did not come into a wider usage until the first web based citizen science project eBird launched [6]. EBird was launched by the Cornell Lab of Orni-thology in New Zealand in 2002 [7]. Other ongoing projects can be easily found for example from SciStarter [8] and Zooniverse [9]. There are multiple different projects with different topics and different interaction types.

In Table 1, there is a collection on different projects and most are found through the aforementioned websites. The interaction type is classified into one of three pos-sibilities depending on the primary end user of data: citizen-citizen, citizen-government, and citizen-researcher.

**Table 1.** Citizen science projects

| Project | Observation | Interaction type |
|---|---|---|
| Galaxy Zoo | Citizens observe and identify objects in space | Citizen – Researcher |
| Fossil Finder | Citizens observe fossils in Kenya | Citizen – Researcher |
| ISEEChange | Citizens observe weather sightings | Citizen – Researcher |

| | | |
|---|---|---|
| Globe at Night | Citizens observe constellations | Citizen – Citizen |
| iNaturalist | Citizens observe nature | Citizen – Citizen |
| WildPaths | Citizens observe wildlife road crossing | Citizen – Government |
| FixMyStreet [10] | | |
| Citizens observe roads | Citizen – Government | |

These projects collect data and involve citizens in different ways, not just by going out and gathering data with sensors.

Data gathering and storing are a vital part of citizen science. That makes databases, database management systems (DBMS), data models and structures important for citizen science projects.

The purpose of this literature review is to find out what are the current data models, structures, databases, and database management systems used in different fields of citizen science.

The main research question is "What type of data storing techniques and technologies are used in citizen science applications?" This question can be answered by answering the following sub-questions:
What type of databases are common in citizen science applications?
What variety do data models and structures have in citizen science applications?

These questions are answered by doing a systematic literature review on citizen science applications.

## Literature review on citizen science data structures

This research used the systematic literature review method to find relevant articles related to citizen science databases, data structures, or data models. Six different queries were used to narrow down the possible results. The search results were from 2008 onwards as the area of citizen science had not been largely researched before it. A single query for finding out different data models or structures was created. However, as mentioned in the introduction, citizen science has multiple terms associated with it so the same query was used with six different terms. The used queries can be seen in Table 2 in the Springer format.

**Table 2.** Queries to databases – Springer format.

| Query |
|---|
| "citizen science" AND ("data model" OR "data struct*") |
| "community-based monitoring" AND ("data model" OR "data struct*") |
| "public participation" AND ("data model" OR "data struct*") |
| "volunteer monitoring" AND ("data model" OR "data struct*") |
| "volunteered geographic information" AND ("data model" OR "data struct*") |
| "participatory sensing" AND ("data model" OR "data struct*") |

This research used all queries to five different databases and obtained 705 scientific records. The used databases and total results can be seen in Table 3.

**Table 3.** Number of results for each scientific database.

| Scientific database | Number of results |
|---|---|
| ACM | 31 |
| IEEE | 21 |
| Scopus | 534 |

| | |
|---|---|
| Springer | 90 |
| Web of Knowledge | 27 |
| Total | 705 |
| Duplicates | 168 |

After removing the duplicates from the total results, 537 records were left for the systematic literature review. During the first iteration, the records were screened based on title and abstract using inclusion and exclusion criteria, which can be seen in Table 4. The record had to have at least two inclusion criteria and zero exclusion criterion.

**Table 4.** Inclusion and exclusion criteria for first screening.

| Criterion | Inclusion / Exclusion |
|---|---|
| Information related to software | Inclusion |
| Information related to database | Inclusion |
| Information related to database management system | Inclusion |
| Information related to data model | Inclusion |
| Information related to data structure | Inclusion |
| No full text available | Exclusion |
| 3D modeling | Exclusion |

After the first screening, 99 articles were included in the second screening. Again, two inclusion criteria and zero exclusion criterion had to be met to pass the screening. The criteria can be seen in Table 5.

**Table 5.** Inclusion and exclusion criteria for second screening.

| Criterion | Inclusion / Exclusion |
|---|---|
| Detailed data model | Inclusion |
| Detailed data structure | Inclusion |
| Application implementation | Inclusion |
| Data management information | Inclusion |
| Design too unrelated to citizen science | Exclusion |

Finally, 14 articles were chosen for this literature review.

## Results

The final 14 articles provided a variety of information regarding the state of citi-zen science data models, data structures, used databases and how these components are applied in applications.

Veen et al. [12] present an object-oriented relational data model for citizen science observation. Their approach was compared to Darwin Core, Access to Biological Collection Data (ABCD), and R2000. The model was found to have an improved integration and ease-of-use, but it imposed more constraints on data and data pro-vider.

Lukyanenko et al. [13] proposed an attribute-based model to replace an instance-based structure. This model was meant to increase participation in citizen science projects. The benefit of this model is that, the user is not required to have exact knowledge on the things surveyed as they can use attributes to describe what they have seen. With this, more amateurs can join the citizen science project and contrib-ute

information. The limitation in this model is that not all things can be expressed through attributes such as general categories like animals or vehicles. In a later re-search, Lukyanenko et al. [14] noted that an attribute or instance -based data struc-ture can potentially be better than class-based data structures as they can lead to higher quality, accuracy and flexibility. The attribute-based design can have chal-lenges like having too many attributes to cover.

Cuong et al. [15] created a generic data model for crowdsourcing (see Fig. 1) to cope with data uncertainty. This model was applied to an emergency response set-ting, where citizens can inform about possible emergencies. The model uses a voting system to have citizens vote on reports to measure their validity. While the model itself could be applied to a variety of situations, there is no detailed data structure. The application of the data model and a detailed data structure design is up to the user of the model. User can follow the data model to design the general relations and entities but specific attributes are outside of the data model.

Zhao et al. [16] created a model and prototype for a citizen VGI system. The model considers the user's reputation and trustworthiness to assess the reliability and quality of the data that user feeds to the system. Their prototype was implemented on a PostgreSQL database. While their design considers the trustworthiness of a user, it does not necessarily mean that the information given by the most trustworthy per-son is always the best. Another problem is, that the system needs an initial reputation model to be accurate and this initial model is based on registration information of the user. This can lead to privacy issues.

Sheppard et al. [17] tackled the challenges of provenance in citizen science. They created a review and rating for each data report users gave. Other users could con-tribute to the original submission by adding new information such as pictures. Addi-tionally, multiple historical versions on the same event can be stored in the database for other uses. The researchers argued that while their model creates more difficulties to new users and possible performance issues, their model's flexibility and prove-nance capabilities are worth the additional complexity. Most of the performance issues can be managed through various technical means.

Sofos et al. [18] created a new framework for VGI to be used with collaborative network-based concepts. These concepts include a reference network and the usage of networked devices. They were added to reduce errors in measurements to increase data quality and credibility. For example, if multiple participants gathered data about the same location from different viewpoints, the measurement results could be com-pared against each other and corrected appropriately. The test design was composed of a web application that is usable with a desktop or a smartphone and a database using MySQL. Their results show that the networking reduced the amount of errors by a considerable amount and additionally reduces the number of collected data points to achieve same results.

Bröring et al. [19] designed a platform for mapping car-sensor data. Their design uses a NoSQL database and the data model was inspired by the Observations & Measurements standard of the Sensor Web Enablement initiative at the Open Geo-spatial Consortium. The backend was created as a separate entity from the frontend so that the data could be accessed with Representational state transfer (REST) APIs [20]. These work with HTTP protocol using POST or GET requests to return a JSON formatted data from the backend to the frontend. The key contribution was the de-sign of the system that can be reused in collecting sensor data from cars [19].

Kotsev et al. [21] created an architecture for a service-enabled sensing platform. They tested different hardware and software in their research. Most notably, they compared three different relational databases (SQLite, H2 and PostgreSQL) and nothing outside of relational databases.

In 2011, Rees et al [22] created a web application for monitoring rabies. Their de-sign split the system into an SQL Server with a relational database management sys-tem and a web API for a frontend. The design also included the ability to monitor citizen participation regionally. With this, the system owners can design incentives to encourage more people to participate in the areas where citizen participation is de-creasing. The data integration and storage architecture are used for the development of similar databases.

Havlik et al. [23] designed a new framework for citizen science. In their design, a NoSQL database CouchDB with a document-oriented database was employed. They found out that this design increased

flexibility in data model and as the database was easy to replicate on frontend and backend, it increased the system's usability on unstable networks. Additionally, the design handles the location issue when trying to find information by creating an area of interest defined by the user. This area of in-terest allows user to receive messages and information related to that particular area and it eliminates the need of exact location.

Sheppard [24] created a web application framework for VGI using Django and Py-thon for backend and an object-relational model database GeoDjango. The frame-work is meant to be used in different VGI projects and relatively easy to use and configure. It does not follow any standard data format but instead, it allows the sys-tem developers to create their own transformation rules specific to their usage. The application employed an HTML5 frontend, but Sheppard also took into account the possibility of creating a mobile app from HTML5 code. With this, there is no need to design a separate mobile app and the workload is reduced.

Huang and Liang [25] tackle the problem of integrating Open Geospatial Consor-tium Sensor Observation Service (SOS) and OASIS Open Data Protocol (OData) together. SOS defines an open standard protocol and data model for sensor devices and sensor data [26]. On the other hand, OData defines a web protocol for querying and updating data [27]. OData is meant for general public and SOS is meant for used with sensors only. Having different data models and structures make it difficult to use and export data to other sources, so it was necessary to create an adapter be-tween two different models.

Soranno et al. [28] designed a database to handle a massive amount of data cen-tered around lakes in the US. The data was collected from multiple different sources and integrated into one relational database. Lakes are the core identifier in their rela-tional database model and other information such as data originator, climate, topog-raphy and road density around the lake area are connected to the core identifier. The design has to handle data for 50 000 different lakes. While the study is not directly related to citizen science but rather data re-use, it provided good insights. The re-searchers found out that a long or vertical data matrix format is more flexible for storage and manipulation. Their conclusion is that when integrating datasets from different sources, numerous steps are required from experts. The data can be in dif-ferent formats and have different data structures across the datasets so that they first need to be transformed into one generalized structure to effectively upload them into one database. Another step is to automate this process, if the amount of datasets surpass the amount researchers can manually process.

In Table 6, a collection of the reviewed articles, their respective database, frontend design, application of provenance, and the observation usage can be seen. The ob-servations-column shows that the designs and applications have been used in a varie-ty of different fields but all of them revolved around spatial data. Six of the reviewed articles did not apply the designed data model in an actual application, and one of the models was a data integration model for two different data models.

**Table 6.** Database and frontend implementations from the reviewed articles.

| Article | Database and DBMS | Frontend | Provenance | Observations |
|---|---|---|---|---|
| Veen et al. [12] | Relational | - | Yes | Species |
| Huang and Liang [25] | - | - | - | - |
| Zhao et al. [16] | Relational, PostgreSQL | ArcGIS | Yes | Geographical information |
| Rees et al. [22] | Relational, RageDB | Web interface | Yes | Rabies |
| Sofos et al. [18] | Relational, MySQL | HTML / JavaScript / PHP | Yes | Land |
| Kotsev et al. [21] | Relational, PostgreSQL | Web interface | Yes | Soil moisture |
| Soranno et al. [28] | Relational, PostgreSQL | - | Yes | Lakes |
| Sheppard et al. [17] | - | Web interface | Yes | Weather example |

| | | | | |
|---|---|---|---|---|
| Lukyanenko et al. [13] | Relational | - | Yes | Flora and fauna |
| Cuong et al. [15] | - | - | Yes | Emergency example |
| Bröring et al. [19] | NoSQL, MongoDB | Web interface | Yes | Car data |
| Havlik et al. [23] | NoSQL, CouchDB | HTML5 / JavaScript | Yes | Trees and pollen |
| Lukyanenko et al. [14] | Relational | - | - | - |
| Sheppard [24] | Relational, GeoDjango | HTML5 / JavaScript | - | Geographical information |

Most projects employed their frontend with HTML and JavaScript. Some only mentioned employing a web interface without any specific programming language but it can be assumed, that HTML and JavaScript were used. One project used the ArcGIS engine as the user interface for testing but did not create an actual frontend. With web technology it is easier to reach multiple users and upkeep the frontend compared to a traditional desktop application or even a mobile application.

In Table 7 are four additional projects listed from Brovelli's [29] presentation for comparison with the reviewed articles and it shows similar results to the reviewed articles. There are many other DBMSs that were not found in citizen science usage but could be considered as viable solutions such as SQLite, Apache Cassandra, HBase, Scalaris, and OrientDB among others.

**Table 7.** Project examples from Brovelli's [29] presentation

| Article | Database and DBMS | Frontend | Provenance | Observations |
|---|---|---|---|---|
| Land cover validation game | Relational, MySQL | Angular.js / Web interface | - | Land |
| Osaka bike parking report | Relational, PostgreSQL | JavaScript / Web interface | - | Bike parks |
| Via Regina | NoSQL, CouchDB | Cross platform | - | Buildings |
| PoliCrowd 2.0 | Relational, PostgreSQL | Cross platform | - | Participatory platform |

Figures 1 and 2 show examples of data models and structures used for citizen science applications. Both figures show data models from their respective articles, while figure 1 has a more general model and figure 2 has a UML representation of the data model.

There are similarities in both models such as the *Worker* in figure 1 and *Contributor* in figure 2. These have their own *reputation*, which is shown to other users and they are linked to reputation or a report (*GeoEvent, Answer*). The *rewarding model* in figure 1 is similar to *Assess* in figure 2. In those places, the other citizens vote for the validity and accuracy of the report.

Aside from similarities, there are some glaring differences as well. In figure 1, there is an answer template mentioned in the model, which would be useful to get more accurate data. Additionally, in figure 1 the end user is taken into account as the *Requester* of data while in figure 2, there is no mentioning of the probable end user.

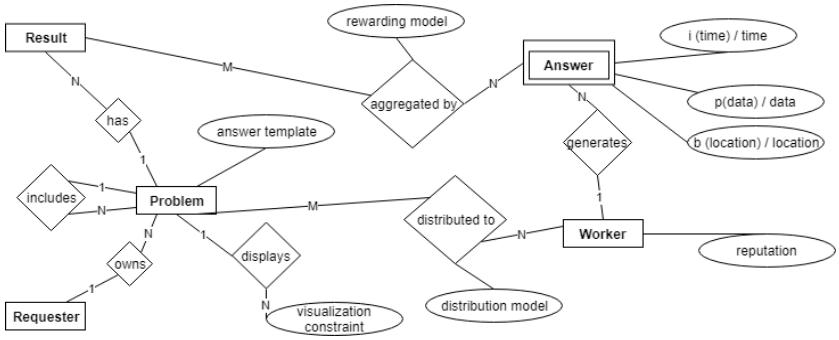


**Fig. 1.** Data model for crowdsourcing [15]

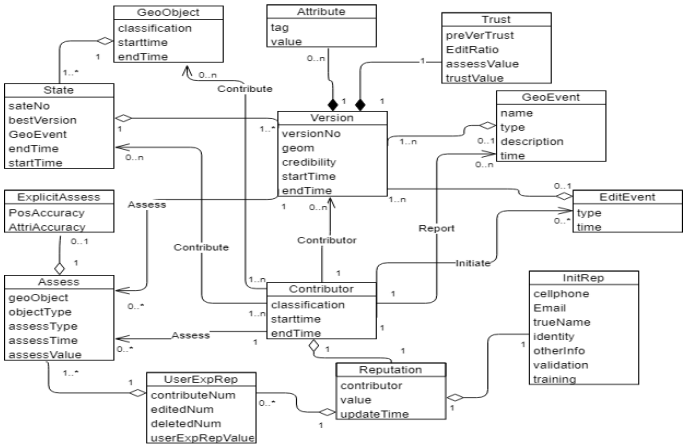


**Fig. 2.** UML representation of a data model for volunteered geographical information with trust-related information [16]

## Implications and discussion

This literature review demonstrated few recurring themes in citizen science applica-tions. First, most applications employ a relational database for their design. This can be explained with the fact, that alternative ways were not common knowledge until recently. Some of the relational database models implement an object-oriented rela-tional model but they are still relational databases [13, 16, 18, 21, 22, 24, 28]. Havlik et al. [23] used a NoSQL database for their design and found it better than a relation-al database. On the other hand, Kotsev et al. [21] tested three different relational databases for their own work and disregarded NoSQL options entirely. This shows that not all database options have been evaluated properly and most database mod-els are designed solely on relational databases. The relational models can be trans-formed into NoSQL models, but they might not be as effective if they are not opti-mized.

In one article outside the literature review, a citizen science project was constructed using SQL Server. The project observed mammals, birds, reptiles and amphibians. The project used mobile app and excel spreadsheet for data uploading. While the original version used SQL server, the researches decided to change in the future to MongoDB to allow easier data processing and integration from multiple sources and faster content delivery to users [30].

Another recurring design in newer publications is that the database and backend are separate from the frontend and the data is accessed through REST APIs or some remote interface. This improves the usability of most designs as the backend is not tied down to a frontend and can then be used in multiple different scenarios. Howev-er, it does impose some challenges if the remote interface is not following any particu-lar standards or there is not enough documentation on how it works [19, 21–24].

Most of the papers designed their system so that other users can rate or improve the pre-existing data. The voting system can be combined to a reputation model that shows which users are trustworthy and provide more data that is accurate. This im-proves the data quality and accuracy as long as most users behave correctly and do not abuse the rating system for their own benefit. When correcting the data, it can be modifiable or require a new submission. A new submission is easier for the database but can be confusing for users, so it can be easier if the existing data can be modified by the owner or by other users. This can lead to issues on access and misuse, as other users need to have access to the data to be able to modify it [12, 15–19, 21, 23].

Finally, many models and applications consider or implement provenance in their design. Provenance is important because knowing the origin of data increases the possibility of validating the data. This increases data quality and accuracy, which will reflect on the usage of the data. The models and applications handle provenance either explicitly or directly by design. Most models record the location and time in-formation for each data. When that data is combined to specific users, the prove-nance of data is fulfilled [12, 13, 15–19, 21–23, 28].

## Conclusions

This article presents a systematic literature review on citizen science databases, data models and data structures. The research question for this literature review was “What type of data storing techniques and technologies are used in citizen science applications?” This question was answered by answering the following sub-questions:
What type of databases are common in citizen science applications?
What variety do data models and structures have in citizen science applications?

Citizen science covers a variety of research fields, which means that the data needed for each project change. The most common database for citizen science applications was a relational database, most notably PostgreSQL. An object-oriented relational model was employed in few cases, but the dominant option was a traditional relational data model. In two cases, a NoSQL solution was used.

All data structures and models had some common ground. Data provenance was taken into account in most data models. Most had location and time data included in their data model and in all cases, the spatial data had some information about the originator. Most papers included some type of voting system for increasing data validity and accuracy. Other users could rate the existing data and give scores depending on how accurate that data was. This does raise some issues with misuse and harassment.

In the newer papers, the backend and frontend were separated from each other to create simplicity and modularity. This also adds flexibility to modifications and reuse of either the front or the backend.

From these points, we can conclude that relational databases are the most common databases in citizen science applications and almost every data model have data provenance, separate backend and frontend, and a rating system to have participants rate each other’s data.

In most cases, the end user of the data is not actually taken into account in the design of the data model or structure and only the participant is considered. This is a limited view of the whole problem if the other end of the spectrum is not taken into account. The designed data structure and model might be easy for participants to add new data but the end users who use that data might have difficulties if they are not taken into account when creating the design. These end users can vary from software designers, to scientists or normal citizens. In the worst case, the end users has to be an expert to be able to use the collected data. There should be a relatively easy way to use the collected data to get more benefits from the citizen science project.

There are few limitations to this research. Although the starting sample is quite big, the reduction to only 14 papers gives limitations to the generalisability of this research. Another limitation is the researcher's opinion on related articles. While the articles were screened based on specific criteria, it is up to the researcher to see if an article fills a criterion or not.

In the future, case projects will be selected and see if they have these similarities implemented in their systems and if other similarities or differences arise. Afterwards, these common features will be closely inspected to see how they are designed and how they could be designed to improve efficiency and overall quality in citizen science databases.

## References


1. See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., et al.: Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. ISPRS International Journal of Geo-Information (5), (2016)
2. Definition of citizen science in English by Oxford Dictionaries, https://en.oxforddictionaries.com/definition/citizen_science, last accessed 2018/01/26
3. United States Environmental Protection Agency Office of Water. Starting Out in Volunteer Monitoring, https://www.epa.gov/sites/production/files/2015-10/documents/2009_06_12_monitoring_volunteer_startmon.pdf (2012)
4. Conrad C., Hilchey K.: A review of citizen science and community-based environmental monitoring: issues and opportunities. Environmental Monitoring and Assessment (176), 273–291 (2011)
5. SciStarter - What is citizen science, https://scistarter.com/, last accessed 2018/04/25
6. New Zealand eBird, http://ebird.org/content/newzealand/. last accessed 2018/01/26
7. Wikipedia eBird, https://en.wikipedia.org/EBird/. last accessed 2018/01/26
8. SciStarter, https://scistarter.com/. last accessed 2018/04/20
9. Zooniverse, https://www.zooniverse.org/. last accessed 2018/04/20
10. FixMyStreet, https://www.fixmystreet.com/, last accessed 2018/04/26
11. OpenStreetMap, https://www.openstreetmap.org/. Accessed 26 Apr 2018
12. Veen L., Van Reenen G., Sluiter F., Van Loon E., Bouten W.: A semantically integrated, user-friendly data model for species observation data. Ecological Informatics (8), 1–9 (2012)
13. Lukyanenko R., Parsons J., Wiersma Y.: Citizen science 2.0: Data management principles to harness the power of the crowd. In: Jain H., Sinha A., Vitharana P, DESRIST 2011, LNCS, vol. 6629, pp. 465–473. Springer, Heidelberg (2011)
14. Lukyanenko R., Parsons J., Wiersma Y.: The IQ of the crowd: Understanding and improving information quality in structured user-generated content. Information Systems Research (25), 669–689 (2014)
15. Cuong T., Mehta P., Voisard A.: DOOR: A data model for crowdsourcing with application to emergency response. In: Giaffreda R., Cagáňová D., Li Y., Riggio R., Voisard A., IoT 360 2014, LNICST, vol. 151, pp. 265–270. Springer Verlag (2015)
16. Zhao Y., Zhou X., Li G., Xing H.: A spatio-temporal VGI model considering trust-related information. ISPRS International Journal of Geo-Information (10), (2016)
17. Sheppard S., Wiggins A., Terveen L.: Capturing Quality: Retaining Provenance for Curated Volunteer Monitoring Data. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, ACM, New York (2014)
18. Sofos I., Vescoukis V., Tsakiri M.: Applications of volunteered geographic information in surveying engineering: A first approach. In: Cartwright, W., Gartner, G., Meng, L., Peterson, M.P., AGILE 2015, Lecture Notes in Geoinformation and Cartography, vol. 217. pp 53–72 (2015)
19. Bröring A., Remke A., Stasch C., Autermann C., Rieke M., Möllers J.: enviroCar: A Citizen Science Platform for Analyzing and Mapping Crowd-Sourced Car Sensor Data. Transactions in GIS (19), 362–376 (2015)
20. Fielding R., Taylor R.: Architectural styles and the design of network-based software architectures. University of California, Irvine Doctoral dissertation (2000)
21. Kotsev A., Pantisano F., Schade S., Jirka S.: Architecture of a service-enabled sensing platform for the environment. Sensors (Switzerland) (15), 4470–4495 (2015)
22. Rees E., Gendron B., Lelièvre F., Coté N., Bélanger D.: Advancements in web-database applications for rabies surveillance. International Journal of Health Geographics (10) (2011)

23. Havlik D., Egly M., Huber H., Kutschera P., Falgenhauer M., Cizek M.: Robust and Trusted Crowd-Sourcing and Crowd-Tasking in the Future Internet. In: Hřebíček J., Schimak G., Kubásek M., Rizzoli A., ISESS 2013, Environmental Software Systems. Fostering Information Sharing, pp 164–176. Springer, Heidelberg, (2013)
24. Sheppard S.: wq: A modular framework for collecting, storing, and utilizing experiential VGI. In: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, ACM, California (2012)
25. Huang C-Y., Liang S.: A sensor data mediator bridging the OGC Sensor Observation Service (SOS) and the OASIS Open Data Protocol (OData). Annals of GIS (20), 279–293 (2014)
26. OGC - Sensor Observation Service, http://www.opengeospatial.org/standards/sos, last accessed 2018/04/19
27. OData - the Best Way to REST, http://www.odata.org/, last accessed 2018/04/20
28. Soranno, P., Bissell, E., Cheruvelil, K., Christel, S., Collins, S., Fergus, C., et al.: Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. GigaScience (4) (2015)
29. Brovelli M. Citizen Generated Content and FOS participative platforms VGI, Lecture Notes, https://earth.esa.int/documents/973910/2642313/MB3.pdf (2016)
30. Bonacic C., Neyem A., Vasquez A.: Live ANDES: Mobile-Cloud Shared Workspace for Citizen Science and Wildlife Conservation. In: IEEE 11th International Conference on e-Science 2015, IEEE, Germany, pp 215–223 (2015)

# Publication II

Musto, J., and Dahanayake, A.
**Improving Data Quality, Privacy and Provenance in Citizen Science Applications**

Reprinted with permission from
*Frontiers of Artificial Intelligence and Applications*
Vol. 321, pp. 141-160, 2019

# Improving data quality, privacy and provenance in citizen science applications

Jiri MUSTO[a,1] and Ajantha DAHANAYAKE[a]
[a]*School of Engineering Science, Lappeenranta-Lahti University of Technology, Lappeenranta, Finland*

**Abstract.** Citizen science project applications can be created in two different ways: to use an existing platform or to build them from scratch. When creating applications for citizen science projects, data quality, privacy and provenance are important characteristics of such applications. Therefore, the objective of this research is to find out how well data quality, privacy and provenance are handled in ongoing citizen science projects. A number of citizen science projects have been compared against ISO/IEC 25012 data quality standard characteristics. Results show that data quality is mostly lacking in the areas of accuracy, privacy, provenance and availability. Projects have not implemented decent accuracy checks when giving data, projects show the real name and location of a participant and such data should not be available to others. Management of provenance is not found in many projects but where it is found, provenance is either really well or extremely poorly handled. Many of these issues could be easily solved with proper data management and testing. At the end of this article, multiple suggestions are given to improve data quality, provenance and privacy in citizen science project applications.

**Keywords.** citizen science, data quality, data management

## 1. Introduction

The number of citizen science projects has been increasing each year with more scientists trying to tap into the power of citizen participation. Citizen science projects have a varied lifetime ranging from less than a year to decades.

Currently there are two common ways to create a citizen science project application. One way is to create and publish your own application, web or phone, and the other is to create your project on a citizen science project platform. Both ways have their own merits and challenges. Creating a project on a platform is easy and fast but not all platforms are suitable for all projects. There may be features the project owner wants and the platform does not provide. On the other hand, creating an application from scratch is time consuming and getting visibility and participants can be more difficult compared to a platform. There are also services, such as SciStarter [1], where people can list their project and put a link to their own website to gain visibility and participants.

Regardless of the chosen method, multiple things should be considered beforehand and during the project. The most important things to consider are what data to collect, how to collect it, for what is it collected and from whom it should be collected. These four things affect the initial data quality and the application can help to increase or decrease the quality of the final data. For example, the data quality will not be high if participants are completely new but with a good application, researchers can help participants to gather higher quality data.

There are many different platforms for citizen science projects, such as Zooniverse [2], EpiCollect5 [3], CitSci.org [4], and CrowdCrafting [5], and thousands of different projects across and outside the various platforms. All these projects have tens of thousands of citizens participating and hundreds of thousands of observations combined. Some platforms have limitations to the project types they support as they follow a specific template when creating new projects. Other platforms are more open with little restrictions to project types, but they have restrictions on the different attributes you can collect. For example, Zooniverse [2] is a platform where you can create data classification projects. These are projects

[1] Corresponding Author: Jiri Musto, School of Engineering Science, LUT University, Yliopistonkatu 34, 53850 Lappeenranta, Finland; E-mail: jiri.musto@lut.fi.

where volunteers look at pictures or texts and classify information based on specific criteria the project owner has given.

In this research, citizen science is defined as "a field where citizens participate in collecting, analyzing, or reporting data to or with scientists." A citizen science application platform is a service where scientists can create projects without creating their own independent application. All the data collection, analysis or reporting happens through the platform.

This research aims to answer the following research questions:

How well is data quality, privacy and provenance handled in ongoing citizen science projects?
How can data quality, privacy and provenance be improved in citizen science projects?

The objective of this research is to answer the research questions and in doing so, improve the data collection in citizen science projects. The contributions of this research are to provide a more detailed explanation on what exactly is lacking in data quality in terms of citizen science applications and to provide improvement suggestions based on currently used methods. With the knowledge from this research, the data collection can be improved and new projects will create higher quality data, and the higher quality data will lead to more accurate results and research.

Because citizen science is an expanding area of science and research with new projects launching frequently, it is important to find out possible problem areas and methods to handle them appropriately. These results might not affect the current ongoing projects, but they can help the future projects to be of higher quality.

The paper is structured as follows. Section 2 introduces related work. Section 3 presents data quality definition and research methods while Section 4 describes the results of this research and proposes improvements. Section 5 discusses the results and implications. Section 6 concludes the paper.

## 2. Related work

Data quality has been under research in citizen science and other related fields such as in databases theory and applications, volunteered geographic information (VGI), participatory sensing, and crowd sensing. VGI related geographical data quality has been compared against the ISO 19157 standard. The standard alone has been found to be lacking and additional quality measures have been given to complement the standard [6].

Most citizen science data quality issues relate to amateurs giving data instead of professional researchers. According to a survey [7], inadequacy of participants is the main concern of many citizen science projects. The inadequacy leads to poor and non-reliable observations, which in turn leads to low quality data. Similarly another survey [8] found out, that data quality is a big concern and common methods to increase quality are expert validation and volunteer training. The survey suggests that automated tests and user-friendly tools can also help improve the quality of data.

In [9], data quality improvement strategies for VGI have been divided into two separate categories: before and after submitting data. Possible strategies for improving data before submission are volunteer training, automated accuracy checking, community validation, and moderating. Similar strategies have been used in many different citizen science projects. Training may not always be possible especially with low budget projects as pointed out by [10]. In the same way, community-based validation may not work if the project is not popular enough. Best option would be to minimize the information that is most likely to be incorrect, thus increasing quality.

Strategies for improving data after submission are ranking the contributors, data mining, external knowledge, enrichment, and fusion [9]. In [11], the ranking of volunteers has been concluded to be useful if used correctly by the scientist as the high quality volunteers' data is more useful than low quality volunteers' data.

Volunteers can be ranked in different ways and another research [12] provides social trust metrics to a citizen science project to improve quality of data. These metrics are used for users and data to determine their reputation. The metrics are derived from the participant's status (researcher, student, volunteer) and from the community's opinion on the participant's data submissions. With the social metrics, the community can weed out potential dummy users and filter the data when going through it.

Another quality validation method has been tested in [13] where other participants' data are compared against each other. A similar method of comparison for increased quality has been employed by Stardust @ Home -project. One research [14] points out, that majority of sites may have quality control working in the background but based on a review of the sites alone, most had no quality control at all. If the quality measures are not transparent, they cannot be evaluated, and all the data associated with unknown quality measures is difficult to accept.

After the implementation of General Data Protection Regulation (GDPR), privacy has been under spotlight especially within EU. Location data is often seen as a problem by itself, but a survey [15] shows that combination of other data aside from location can be a threat to privacy. For example, the location of individuals and many other things can be deciphered from the combination of pictures and environmental data.

In [16] it is stated that privacy concerns are different from individual to another and some might be okay with sharing their location continuously while others may not. When employing privacy techniques this difference in preference should be considered. Many other solutions revolve around anonymizing, offering individual preferences for privacy, or hiding / masking sensitive locations [15]. Privacy has also been looked from animals' perspective in [17], where the researchers explain a case where possible criminals, such as poachers, can use the gathered data for misdeeds such as killing the animals others are observing. This reversed point of privacy is often overlooked but a valid problem people might face. It may not be the most common in citizen science and highly dependent on the field of study, but it is worth considering when creating citizen science projects.

Provenance in citizen science is not widely researched but some have tried to implement provenance within citizen science. Data model for provenance and increased data integration has been proposed in [18]. The data would be integrated behind the scenes for the users and different versions of the data would be saved in the database. Another research [19] shows, that provenance can be implemented in community validation systems as well. When another participant votes on an observation the information is saved and if the vote is later changed, this change will be saved as well.

UK National Biodiversity Network (NBN) has guidelines for data validation and quality. Whenever possible, data should be verified by professionals and the status of verification should be mentioned somewhere with the data entry and during data management, details of provenance should be maintained. The provenance of data should be clearly documented for data users to be able to make their own judgement about validity [20].

Aside from data quality and privacy, data accessibility has been found to be a challenge in citizen science projects. Publishing only raw data does not help others to explore and analyze relevant information if they are not experts in data management [21].

## 3. Research Methods

### *3.1. Research Design*

Citizen science projects have been evaluated against different data quality characteristics. The characteristics come from the ISO/IEC 25012 standards' data quality model [22,23], which divides the attributes into three separate categories:

Inherent data quality: accuracy, completeness, consistency, credibility, currentness
Inherent and system-dependent data quality: accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability
System-dependent data quality: availability, portability, recoverability

Most of the characteristics are considered in this research. Participant privacy is mapped under confidentiality and availability is considered from open data perspective. Compliance, portability and recoverability are the only characteristics that are not considered. Compliance is not considered because it is difficult to find out if any standard has been followed without seeing documentation of the project and possibly their data model. Portability and recoverability are not considered because they relate to the system, have little to no effect on others, and are difficult to measure without access to the underlying system.

ISO standard is chosen as the data quality model because ISO is an international standardization organization that has created many widely used standards. There are other data quality models defined such as the six primary dimensions from DAMA UK [24] and similar dimensions can be found from Experian [25], Blazent [26], Enterprise Solutions [27], EDM Council [28]. Finally, an extensive data quality characteristics can be found from [29]. All models have similar core characteristics. What sets ISO apart from most of the other definitions is the classification of these characteristics between inherent and system-dependent characteristics.

### *3.2. Data quality characteristics*

*Accuracy*: Data accuracy is divided into syntactic and semantic accuracy. Syntactic accuracy means that the given data matches the syntax in the data model (number is number, text is text, date is date etc.). Semantic accuracy means the given data's semantics matches the semantics in the data model such as when asking a location, the location exists / makes sense.

*Completeness*: Data completeness means how complete the submitted data is against all possible attributes of the data entity. The attributes refer to all mandatory and optional fields the participant can fill, and these are put against what the participant submits.

*Consistency*: Data consistency means how consistent data is across the tables and database. If the same data entity is in different tables, the data should be consistent in all of them. Additionally, the data entities within the same table should be consistent and not have varying data in same attributes. In this research, consistency is checked from syntactic accuracy and completeness as there is no access to the database itself.

*Credibility (validity)*: Data credibility or validity means how credible a data entity is. If the data is gotten through sensors, the credibility is measured by the credibility of those sensors. If the data is gotten through human observation, the credibility is measured by the credibility of that human. Researchers are thought to always give credible data and only their methods are questioned but if a citizen gives data, the credibility of the human is questioned.

*Currentness*: Data currentness means the degree of which data attributes are of the right age. For example, the difference of observation to the difference when submitting an observation. If the time is different, both times should be mentioned.

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

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

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

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

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

*Understandability*: Understandability refers to how understandable the data is, how well can people understand and interpret the data when reading, and how well the data is represented with appropriate symbols, language and units.

*Availability*: How available the data is for authorized and unauthorized users / applications. In this research, availability refers to if data can be downloaded from the project.

### *3.3. Collecting data*

The data has been collected by going through different citizen science projects and their applications. Some of the projects have been found from different articles and most have been found through SciStarter – website [1], which describes over 1400 different projects. Projects have been looked through by:

- Arranging projects by their starting date and going through 100+ oldest and newest projects
- Going to random pages between the project range and going through 30+ projects

The most important criteria for choosing the project are: free, still accepting new data submissions and data submission can be done online. If their publications are found in the project website, those are read through to find information related to data quality characteristics.

A total number of 30 applications and 8 application platforms have been chosen for this research. A test user has been registered to each project when it is possible and different features have been tested with the test user such as looking at data, submitting data, editing data, and creating projects (on platforms). Features have also been tested without logging in to see if it is possible to do the same things without registering.

## 4. Case: Ongoing citizen science projects

### *4.1. Results*

Thirty-eight applications and platforms have been checked against ISO data quality characteristics. Eight are platforms where others can create new projects. The characteristics have been checked as accurately as possible and the discovered different implementations have been generalized and mapped under corresponding characteristics. These generalized implementations can be found in Table 1. All the results are based on what can be found from the project site and they do not consider what might be happening in the background, invisible to users.

*Accuracy*: Accuracy is handled differently in the projects. Syntax accuracy is checked more often than semantic accuracy. The projects check whether the input is a number or in correct date format. Semantically the accuracy checking is worse as in most cases, one can put random string into the place of residence and the system accepts it. Most projects do check if the date is valid and not from the future but that is mostly the extent of semantic accuracy. Third of the projects provide no syntactic or semantic accuracy checks.

*Completeness*: Most of the projects have some required fields when giving observation data to the system. Only one project has no required fields and you can submit empty observations. In half of the projects, mandatory fields are half or less of all the attributes that the participant can input. In most projects, observation needs to have a name, location and time but all other fields can be left empty. The observation can be edited later to increase the completeness if the observation is submitted while logged in but if that is not possible, the data entry is left incomplete.

*Consistency*: Consistency is measured by the researcher based on accuracy and completeness because many attributes can be either empty or not implying that they are not consistent against each other. Additionally, researcher's own recognition of consistency is considered, so the results can be subjective. Consistency has some varying results. When asking text, the participant can input numbers and sometimes vice-versa. While you can argue that numbers are text, it reduces the consistency if the scientist is expecting a descriptive text and instead gets a random number.

**Table 1.** Generalized implementations corresponding to each data quality characteristic as given in the section 3.2. The number of applications out of the total is in the brackets. See Appendix for a detailed list.

| **Data quality charact. (38)** | **Value 1** | **Value 2** | **Value 3** | **Value 4** | **Value 5** |
|---|---|---|---|---|---|
| Acc. (30) | No accuracy checks (10) | Syntactic accuracy checks (12) | Syntactic and semantic accuracy checks (8) | | |
| Compl. (32) | All fields required (13) | Any number of fields required (2) | Half or more required (5) | Less than half required (11) | Nothing required (1) |
| Consist. (34) | 5 (12) | 4 (6) | 3 (7) | 2 (5) | 1 (4) |
| Credib. (22) | Community (8) | Moderated (9) | Preliminary test (2) | Sensor (1) | Validity tests (2) |

| | | | | | |
|---|---|---|---|---|---|
| Current. (31) | Observation time is different from upload (30) | Up to the project creator to handle (1) | | | |
| Access. (37) | Analyses / Statistics (3) | Maps (and other things) (23) | Not available (4) | Observations only (4) | Reports (3) |
| Conf. (34) | Location and name are shown (7) | Location and name are shown with exceptions (6) | No name (nor location) (9) | Username / anonymous (12) | |
| Effic. (28) | 5 (5) | 4 (13) | 3 (6) | 2 (2) | 1 (2) |
| Prec. (30) | Can give estimates (23) | Predefined estimates (1) | Requires precision (4) | Up to project creator (2) | |
| Trace. (7) | Anyone can edit data (2) | No provenance (2) | Can view edits (3) | | |
| Underst. (29) | 5 (13) | 4 (9) | 3 (6) | 2 (1) | 1 (0) |
| Avail. (38) | By contacting (2) | Can download csv / excel / json file (8) | Not available (22) | Can download reports (5) | Can download data from previous years (1) |

*Credibility*: Credibility or validity is checked in different ways from project to project. In 8 out of 22, projects have experts or moderators to go through the data input and validate it. Similar results come from the use of community, where other users vote on the credibility or validate the data. Only two projects implement automatic credibility / validity testing. Data is checked against predefined values and records such as earlier reports or average values on similar observations. These checks are shown to other users and they tell how well each observation fulfills the checks.

*Currentness*: Currentness is usually handled in the same way but with different precision. The participant is asked to provide the date and time of the observation. Time is handled either as a precise time or as a time window. Sometimes you can omit the time altogether. In one project, the participant can submit the time of the observation but when editing the information later the time cannot be changed even if the date is changed. One project uses predefined approximate values like number ranges and vague descriptions (early morning, afternoon, etc.).

*Accessibility*: Most projects use maps accompanied with simple charts to show their data to users. Maps are mostly used to show where observations have been recorded, and charts are used to show simple and easy to understand analyses or statistics, such as how many similar observations have been submitted during last six months. Some projects go beyond mere charts and show more complex analyses and reports to the users. In specific types of projects, the data is never accessible by outsiders. In these projects, the users are asked to analyze information instead of gathering. Sometimes results are available for outsiders, but the specific analysis data are not. Only 4 out of 36 provided only raw data for others.

*Confidentiality* (privacy): User privacy is split into showing participants' name or their username. Data often has location information tied to it meaning others can see the location and the real name of the participant. This is done in 13 out of 34 projects and the rest use a username or no name. In few projects, the participant name is omitted from downloaded data but in another case, exact opposite is done. Some projects offer the option of anonymity to the participant when submitting data.

*Efficiency*: Efficiency is not measured with any tools and is mostly based on the perception of efficiency: how fast the project website loads, how fast you can browse the data and so on. Most projects are pleasantly fast and there are no long load times, but some projects are slow to start, slow to use and difficult to navigate.

*Precision*: Precision is highly dependent on the specific attribute of the data. Location precision depends on the system the participant uses and what map projection the project uses. In most cases, users

are able to give estimates to participant time. Location precision varies from being able to give a rough area or city to giving exact latitude and longitude values. Many other attributes depend on the participants' knowledge and memory, so they are mainly estimates. Sometimes projects give predefined vague options which to choose from such as certain, quite certain, and not quite sure.

*Traceability* (provenance): Information related to data provenance is scarce, so it is difficult to evaluate within the context of citizen science. As the table shows, 3 out of 7 projects have at least some level of provenance and rest do not. In those three cases, one project shows the latest edit on the data while the other two show a separate list of edits in the whole dataset. It can be argued that the full provenance is extremely rare in citizen science applications but it should be taken into account when creating these types of applications. Four out of seven projects have no provenance and in two of them, anyone can edit the data and no other user can see who has done it, which creates a credibility issue.

*Understandability*: Most projects' data is in an understandable and human readable form especially when the data can be seen on a map or if some analysis charts are shown. When downloading the data, it is often in a .csv-file type that is easy to export into a table. In most cases, the exported data has many attributes the reader might not understand but the most important bits are easy to understand and follow. Only in one project the data is somewhat difficult to understand. Understandability is not measured using any tools and is subjective to the researcher's perception of understandability. Biggest issue with the projects is the lack of any explanation of each attribute for a data entity. Most of them are self-explanatory but some are not. Especially, if the participant is for example an institution and an acronym or abbreviation is used within a dataset instead of the name of the institution. If there had been a mention what that particular attribute could contain, the understandability would have been higher.

*Availability*: The recorded and analyzed data is sometimes made available for download for the users. Roughly, a quarter of the projects have the option to download the observation data in some file format. In few cases, the data can only be downloaded if the project organizers are contacted and then asked their permission. Sometimes even the analysis data is available for download and not just the observation data. In most cases, the observation data is not downloadable and only some analyses or yearly reports can be downloaded.

### *4.2. Suggested Improvements*

*Accuracy*: Data accuracy should be handled with more care. Syntactic accuracy checks are already implemented in most projects. Semantic accuracy is more difficult to check when trying to give freedom to the user, but some things should be checked. First, when a user is asked to give a location such as a country and city, there should be some automated check to see if such a country or city exists. Another method is to provide the user with a predefined set of options the user can select from and if the user cannot find what they are looking for, ask them to fill a different field. When this particular field is filled, it is immediately flagged for possible moderators or community to check. This method can work with cities, countries, species, numbers, observation characteristics and many others if implemented correctly. Implementing this type of semantic accuracy check would lead to higher quality data and it can be used to prioritize flagged entries over normal entries for moderators or community to check.

*Completeness*: Completeness is a double-edged sword in citizen science applications. Scientists want as complete data as possible, but most participants are unwilling or unable to provide such complete data. Participants are often amateurs and using their own time, so they may not be able to give all the information a scientist is asking for. This is the reason why most projects only have some mandatory fields and many optional for the participant to fill and then have the option to edit the submission later. If all the fields are required, there are often less than five different fields for participants to fill. Observation information such as name of observed species or phenomena, and short description of the observation, location, time of observation, and participant information are the most common mandatory fields across all projects. These can be considered as minimum required fields for the observation input. The short description can be free word, selected from a list, or measurement information. Participants should also have a way to modify their submissions as many citizen science projects already enable, but this modification option is only available for logged in users. It could be enabled for anonymous users as well if they are asked to provide contact details such as an email address. The project could send an observation ID and an authentication key via email that could be used to edit the submitted data later.

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

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

While this three-staged plan is not foolproof, it would reduce the amount of work experts have to do and at the same time increase the data quality. More emphasis should be put to the first stage because the second and third stage require more human resources and can be difficult to implement in new, small or unpopular citizen science projects. With only a handful of participants, it is difficult to implement the second stage, while the third stage is difficult to implement if there are no experts with free time available within the project. That is why the first stage of automated validity testing should be emphasized and enhanced as much as possible to reduce the use of human resource on menial labor. This will work in all types and sizes of citizen science projects.

*Currentness*: Currentness is one of the most well-handled characteristic out of all. However, it should never be an exact time that is asked from the participant. Asking a participant to provide an approximate time should be done by giving the participant the option to choose a time window and not just one specific approximation. This would help participants and raise the quality of data. When saying an observation is made approximately at 4 p.m., it can be that the participant remembers it poorly and the observation is actually made around 5 p.m. If the participant is asked to provide a time window, they can say it is between 3-5 p.m. and that will be more accurate as the correct time falls within a time window and is not outside the approximation the participant submits.

*Accessibility*: In some projects, participant is asked to provide data without getting anything out of it. When a birding enthusiast reports a bird sighting, he would probably want to know where other birds can be found within his area of operation. If the data is not accessible in any way, the bird enthusiast might lose interest in the whole project. Therefore, there should always be some way to access the data and most preferably, a visual way and not just a table of raw data. A visual output is more meaningful and easier to understand than just the collection of data. This visual output can be a map, analysis charts or something else if applicable. Even in projects, where participants are asked to analyze pictures there can be a visual output or chart of the current state and results. Accessibility can also provide easier error checking. If there are data points that are outside of an expected range, they are easier to notice when the data is put on top of a map or in a chart.

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

and want to protect their privacy but if the change in settings does not change the situation in any way, people may lose their trust on the system itself.

*Efficiency*: It is difficult to provide exact improvement suggestions without knowing the whole structure, but efficiency should improve when handling of data or the server where the project is, is improved. Handling of data affects the efficiency the most and the handling is related to many things such as the underlying database, database management system and so on. Server itself does not improve efficiency per se; it mainly handles the data faster even if efficiency is poor.

*Precision*: Precision is another double-edged sword. Most participants are not able to provide extremely precise data, like Latin name for an animal, and that can be handled during validation process. There are ways to improve the precision of data for example giving predefined options or fields to the participant or filtering wrong information out of the data. The latter can be done by asking different attributes from the participant and then giving possible options to the user to choose from (participant saw something small, black and flying, was it a crow or raven perhaps?). However, this can create time-consuming tasks for the participants and it needs to be carefully designed and implemented so the participants are not demotivated.

*Traceability (provenance)*: Provenance is something that should be implemented with every citizen science project. Having other users edit data and not knowing who did what creates issues, especially if the data can be edited after being validated and this new change does not trigger a new validation. There can also be users who edit correct data into wrong data and this decreases accuracy. Having no way to know who did it and what the previous information is creates problems not only to the community but to the scientists as well. Implementing full provenance can be difficult, as every data-related task needs to be recorded. Additionally, the provenance should be implemented in a transparent way so that the participants can also see provenance data on some level at least. This transparency not only increases the credibility of the data but can also reduce some misbehavior of participants. For example, if one participant edits correct data into wrong data on purpose, this change could be seen by everyone else. If the project employs a reputation model on their community, this misdeed would reduce reputation and so the transparent provenance would act as a prevention method.

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

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

## 5. Implications and discussion

Data quality has different standards and recommendations across the world. ISO/IEC 25012 standard represents software engineering data quality that works in software engineering in general. ISO 19157 standard that is specifically tailored for geographic information could be used to measure citizen science data quality. However, not all citizen science projects deal with geographic information and there are many

other parts outside the geographic information, so the larger scope of ISO/IEC 25012 standard is more appropriate. ISO 19157 standard had been tested with VGI data and even then it is found to be lacking and needed additional quality measures [6].

This research brought similar results as [7–13,21] where data quality has been found lacking in citizen science and participatory projects. The results imply that while data quality methods have taken some steps forward in citizen science, there are still many things to do before reaching the peak of quality. Some of the presented data quality characteristics have been handled with great care in most projects but some have been left out. eBird [30] is one of the longest running citizen science project applications and it has good quality control over the data. eBird provided best results when compared against the ISO/IEC 25012 data quality standard but even in eBird, there is room for improvement. Other projects can learn from eBird but implementing similar quality control can be difficult as there are more people and participants behind eBird compared to many other projects. As mentioned in [10], having a small number of participants can lead to lower quality data, because there are not enough data points to compare against each other.

Citizen science project platforms offer a fast and easy way to start citizen science projects. People can create new projects in under half an hour and start collecting data. While these platforms help people to start new projects and possibly gain participants through the platform, the platforms often rely on community for validation. Without a good number of participants, it is difficult to ensure the quality of data in a platform-based project. With low number of participants, it is vital to have high quality participants as they can gather higher quality data compared to low quality participants. Often in data related applications, quality is better than quantity [11].

Many of the presented data quality characteristics are vital to perceived data quality. Accuracy, credibility and completeness are the most easily perceived characteristics and often the baseline for perceived data quality. These characteristics are most tied to participants especially if the validation is done by community. The lack of quality control over these characteristics is easily perceived by participants and outsiders creating a lack of trust over the collected data. Using different data quality improvement strategies mentioned in [9] helps the project to gather higher quality data especially when using both before and after data submission strategies. Most of these strategies have also been used in the reviewed applications with good results.

Provenance is a difficult characteristic to perceive as a user and is often overlooked in citizen science applications. As mentioned in [14], there may be things running in the background without the users' knowledge and provenance might be one of them. However, this creates a problem in some situations where other people can edit data and there is no visible record on who did what. Even if there is provenance working in the background, to participants it looks like there is no supervision of editing and can lead to people thinking the project does not care about data quality. Like NBN suggested in their guidelines [20], provenance information should be visible to participants as that can also demotivate misdeeds such as false editing. Combined with the community validation provenance proposed in [19], most of false information and misbehavior can be weeded out. The biggest issue with this setup is the large amount of data that is generated through the provenance of observation data and validation data.

Availability has interesting results in the projects. In the modern world, open data is becoming more popular and it is surprising to find out, that most projects do not allow their data to be downloaded. This can be considered a good thing from security and privacy perspective but at the same time, the projects do allow others to view that same data on site. Not giving an option to download the data does not increase privacy if the data is available on site; it just slows down malicious behavior and obstructs beneficial reuse of data. Malicious people can still crawl through the data on the website without the need of downloading while good-natured people cannot reuse the data for good purposes, as it is not in a downloadable form.

Confidentiality has negative results and surprisingly, two out of eight platforms show both real name and location in the observation data. Even worse, one of the two has all the data downloadable and the downloaded dataset includes names and locations. As pointed out in [15], anonymizing data is one of the most common ways to protect privacy but when other data can be combined together, the anonymization might not be enough. Many projects have the option of uploading pictures and despite anonymization, combining the pictures with locations and other information, the privacy of the participant might be threatened. Some of the reviewed projects have the option of allowing or disallowing the usage of name or location. This ties together to the notion that people have different privacy concerns [16].

Different mapping techniques and programming libraries have increased the number of maps used on websites and in citizen science projects. Maps are an easy way to visualize data, if applicable, and they are easy for others to understand. This has increased the accessibility of data, which has been mentioned to be one of the concerns of citizen science stakeholders [21]. Having other ways to show data than in raw form, increases the interest on the project. There are still ways for projects to improve on accessibility as most projects only had the maps as well as raw data or simple charts. Few projects had taken the extra step of including detailed analyses as well, which increase their accessibility.

## 6. Conclusions

This article presents a case study on data quality in ongoing citizen science projects. The research questions for this case study are:

How well is data quality, privacy and provenance handled in ongoing citizen science projects?
How can data quality, privacy and provenance be improved in citizen science projects?

Data quality has been measured using ISO/IEC 25012 standard by dividing data quality into separate characteristics. As the research shows, some parts of data quality are handled well but some parts are lacking. Data provenance is not found in many projects but it is possible that provenance is handled in the background but there is no mention of it. In two projects, everyone had the ability to edit data and there is no information regarding who did what. Some other characteristics are handled similarly in most projects, such as precision and completeness, as participants can give estimations of data rather than exact data. In most projects, after submission participants have the option to edit their own data if they must add or change something.

All reviewed projects can improve at least some of the data quality characteristics. Some projects, such as eBird [30], have little room for improvement but other projects, such as Bleach Patrol [31], should consider to substantially improve the quality of their data. Semantic accuracy, availability, privacy, and provenance have the poorest quality controls on average.

Different improvement suggestions have been proposed based on different working implementations, such as a three-staged credibility checking that is somewhat similar to what eBird has been using.

As limitations for this research, all the projects have been checked by one researcher and errors may have been made while testing. Some of the characteristics have been evaluated through the researcher's perception and they can be biased. There are also characteristics that are difficult to evaluate without access to the underlying system and database. Despite the limitations, this research provides valuable insight into the current state of ongoing citizen science projects, what parts of data quality are well taken care of and what are lacking.

In a future research, a project that considers the improvement suggestions should be created and tested. To bring out the best results, this project should be an ongoing project that is improved based on the suggestions, and results could be tested before and after the improvements.

## References


[1] SciStarter, Welcome to SciStarter, (2019). https://scistarter.com/ (accessed January 9, 2019).
[2] Zooniverse, Zooniverse, (2018). https://www.zooniverse.org/ (accessed January 9, 2019).
[3] Epicollect5, Mobile & Web Application for free and easy data collection., (2018). https://five.epicollect.net/ (accessed January 9, 2019).
[4] CitSci.org, CitSci.org - Comprehensive citizen science support, (2018). https://www.citsci.org/CWIS438/Websites/CitSci/Home.php?WebSiteID=7 (accessed January 9, 2019).
[5] Scifabric, Science affects all of us, Science needs all of us, Crowdcrafting, (2018). https://crowdcrafting.org/ (accessed January 9, 2019).
[6] C.C. Fonte, V. Antoniou, L. Bastin, J. Estima, J.J. Arsanjani, J.-C.L. Bayas, L. See, and R. Vatseva, Assessing VGI Data Quality, in: Mapping and the Citizen Sensor, Ubiquity Press, 2017: pp. 137–163. doi:10.5334/bbf.g.
[7] M. Schröter, R. Kraemer, M. Mantel, N. Kabisch, S. Hecker, A. Richter, V. Neumeier, and A. Bonn, Citizen science for assessing ecosystem services: Status, challenges and opportunities, *Ecosystem Services*. **28** (2017) 80–94. doi:10.1016/j.ecoser.2017.09.017.
[8] A.W. Crall, G.J. Newman, C.S. Jarnevich, T.J. Stohlgren, D.M. Waller, and J. Graham, Improving and integrating data on invasive species collected by citizen scientists, *Biological Invasions*. **12** (2010) 3419–3428. doi:10.1007/s10530-010-9740-9.
[9] G. Bordogna, P. Carrara, L. Criscuolo, M. Pepe, and A. Rampini, On predicting and improving the quality of Volunteer

Geographic Information projects, *International Journal of Digital Earth*. **9** (2016) 134–155. doi:10.1080/17538947.2014.976774.
[10] R. Lukyanenko, J. Parsons, and Y. Wiersma, Citizen Science 2.0: Data Management Principles to Harness the Power of the Crowd, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, Berlin, Heidelberg, 2011: pp. 465–473. doi:10.1007/978-3-642-20633-7_34.
[11] G.M. Foody, L. See, S. Fritz, M. van der Velde, C. Perger, C. Schill, D.S. Boyd, and A. Comber, Accurate Attribute Mapping from Volunteered Geographic Information: Issues of Volunteer Quantity and Quality, *The Cartographic Journal*. **52** (2015) 336–344. doi:10.1080/00087041.2015.1108658.
[12] A. Alabri, and J. Hunter, Enhancing the Quality and Trust of Citizen Science Data, in: 2010 IEEE Sixth International Conference on E-Science, IEEE, 2010: pp. 81–88. doi:10.1109/eScience.2010.33.
[13] R. Esmaili, F. Naseri, and A. Esmaili, Quality Assessment of Volunteered Geographic Information, *American Journal of Geographic Information System*. **2** (2013) 19–26. doi:10.5923/j.ajgis.20130202.01.
[14] L. See, P. Mooney, G. Foody, L. Bastin, A. Comber, J. Estima, S. Fritz, N. Kerle, B. Jiang, M. Laakso, H.-Y. Liu, G. Milčinski, M. Nikšič, M. Painho, A. Pődör, A.-M. Olteanu-Raimond, M. Rutzinger, L. See, P. Mooney, G. Foody, L. Bastin, A. Comber, J. Estima, S. Fritz, N. Kerle, B. Jiang, M. Laakso, H.-Y. Liu, G. Milčinski, M. Nikšič, M. Painho, A. Pődör, A.-M. Olteanu-Raimond, and M. Rutzinger, Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information, *ISPRS International Journal of Geo-Information*. **5** (2016) 55. doi:10.3390/ijgi5050055.
[15] D. Christin, A. Reinhardt, S.S. Kanhere, and M. Hollick, A survey on privacy in mobile participatory sensing applications, *Journal of Systems and Software*. **84** (2011) 1928–1946. doi:10.1016/j.jss.2011.06.073.
[16] R. Ganti, F. Ye, and H. Lei, Mobile crowdsensing: current state and future challenges, *IEEE Communications Magazine*. **49** (2011) 32–39. doi:10.1109/MCOM.2011.6069707.
[17] R.M. Frey, T. Hardjono, C. Smith, K. Erhardt, and A. "Sandy" Pentland, Secure sharing of geospatial wildlife data, in: Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data - GeoRich '17, ACM Press, New York, New York, USA, 2017: pp. 1–6. doi:10.1145/3080546.3080550.
[18] S.A. Sheppard, A. Wiggins, and L. Terveen, Capturing quality, in: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '14, ACM Press, New York, New York, USA, 2014: pp. 1234–1245. doi:10.1145/2531602.2531689.
[19] A. Wiggins, and Y. He, Community-based Data Validation Practices in Citizen Science, in: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16, ACM Press, New York, New York, USA, 2016: pp. 1546–1557. doi:10.1145/2818048.2820063.
[20] T. James, Improving Wildlife Data Quality, 2011. https://nbn.org.uk/wp-content/uploads/2016/02/NBN-Imp-Wildlife-Data-Quality-web.pdf (accessed January 9, 2019).
[21] M. Palacin-Silva, and J. Porras, Shut up and take my environmental data! A study on ICT enabled citizen science practices, participation approaches and challenges, in: Proceedings of the 5th International Conference on Information and Communication Technology for Sustainability - ICT4S2018, 2018: pp. 270–288. doi:10.29007/mk4k.
[22] International Organization for Standardization, ISO/IEC 25012:2008 Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model, (2008). https://www.iso.org/standard/35736.html (accessed January 9, 2019).
[23] iso25000.com, ISO 25012, *ISO*. (2018). http://iso25000.com/index.php/en/iso-25000-standards/iso-25012 (accessed January 9, 2019).
[24] DAMA UK, The Six Primary Dimensions for Data Quality Assessment - Defining Data Quality Dimensions, (2013). https://www.dqglobal.com/wp-content/uploads/2013/11/DAMA-UK-DQ-Dimensions-White-Paper-R37-1.pdf (accessed January 9, 2019).
[25] Experian, What are Data Quality Dimensions?, (2019). https://www.edq.com/uk/glossary/data-quality-dimensions/ (accessed January 9, 2019).
[26] D. Ortega, Seven Characteristics that Define Data Quality, (2017). https://www.blazent.com/seven-characteristics-define-quality-data/ (accessed January 9, 2019).
[27] Enterprise Solutions, Data Quality Guideline - Information Management Framework, 2018. https://www.enterprisesolutions.vic.gov.au/wp-content/uploads/2018/05/IM-GUIDE-09-Data-Quality-Guideline-1.pdf (accessed January 9, 2019).
[28] EDM Council, Data Quality Dimensions, 2017. https://cdn.ymaws.com/edmcouncil.org/resource/resmgr/featured_documents/BP_DQ_Dimensions_Oct17.pdf (accessed January 9, 2019).
[29] F. Sidi, P.H. Shariat Panahy, L.S. Affendey, M.A. Jabar, H. Ibrahim, and A. Mustapha, Data quality: A survey of data quality dimensions, in: 2012 International Conference on Information Retrieval & Knowledge Management, IEEE, 2012: pp. 300–304. doi:10.1109/InfRKM.2012.6204995.
[30] Cornell Lab of Ornithology, eBird - Discover a new world of birding..., (2019). https://ebird.org/home (accessed January 9, 2019).
[31] Bleach Patrol, (2019). https://www.ldeo.columbia.edu/bleachpatrol/ (accessed January 9, 2019).

## Appendix

**Table 1.** All projects that were reviewed in the research and their corresponding values for the data quality characteristics

| Data quality characteristic | Ancient Tree Inventory | Atlas of Living Australia | BioCollect | Bleach Patrol | BudBurst |
|---|---|---|---|---|---|
| Acc. | Syntactic and semantic checks | Syntactic checks | No accuracy checks | No accuracy checks | Syntactic checks |
| Compl. | All fields required, | Most are mandatory | Half or more required | Less than half required | All fields required |
| Consist. | 5 | 5 | 2 | 1 | 5 |
| Credib. | Data is verified by others | Validity tests | Moderated | - | - |
| Current. | Only submission date | Observation time separate | Observation time is different from upload | Observation time is different from upload | Observation time is different from upload |
| Access. | Map | Maps. charts | Maps (and other things) | - | Analyses / Statistics |
| Conf. | Location and name shown | Location and name shown | Location and name are shown | - | No name (nor location) |
| Effic. | 5 | 4 | 3 | - | 4 |
| Prec. | Estimates | Estimates | Can give estimates | Requires precision | Predefined estimate values |
| Trace. | Anyone can edit | - | - | - | - |
| Underst. | 5 | 5 | 4 | - | 4 |
| Avail. | By contacting | Can download csv | Not available | Not available | Can download csv / excel / json |

**Table 2.** All projects that were reviewed in the research and their corresponding values for the data quality characteristics

| Data quality characteristic | CitSci | CoCoRaHS | CosmoQuest | Crowdcrafting | Earthdive |
|---|---|---|---|---|---|
| Acc. | Syntactic checks | Syntactic and semantic checks | - | - | No accuracy checks |
| Compl. | Any number of fields required | - | All fields required | All fields required | - |
| Consist. | 3 | 5 | 5 | 5 | - |
| Credib. | Moderated | Preliminary test | Preliminary test | Community | - |
| Current. | Observations require date and time, but when editing it, cannot change the time | Observation time is different from upload | - | - | Observation time is different from upload |
| Access. | Analyses / Statistics | Maps (and other things) | Reports | Analyses / Statistics | Observations only |
| Conf. | Location and name are shown | - | No name (nor location) | Username / anonymous | No name (nor location) |
| Effic. | 3 | - | - | - | 1 |
| Prec. | Can give estimates | - | - | Up to project creator | Can give estimates |
| Trace. | No provenance | - | - | - | - |
| Underst. | 4 | 3 | - | 3 | 3 |
| Avail. | Can download csv / excel / json | Not available | Not available | Can download reports | Not available |

**Table 3.** All projects that were reviewed in the research and their corresponding values for the data quality characteristics

| Data quality characteristic | eBird | EpiCollect 5 | Eterna | eTick | fold it |
|---|---|---|---|---|---|
| Acc. | Syntactic and semantic checks | Syntactic checks | - | Syntactic checks | - |

| | | | | | |
|---|---|---|---|---|---|
| Compl. | All fields required | Any number of fields required | - | All fields required | - |
| Consist. | 5 | 5 | - | 5 | - |
| Credib. | Moderated | Moderated | - | - | - |
| Current. | Observation time is different from upload | Observation time is different from upload | - | Observation time is different from upload | - |
| Access. | Maps (and other things) | Maps (and other things) | Not available | Maps (and other things) | Not available |
| Conf. | Location and name are shown | Username / anonymous | No name (nor location) | No name (nor location) | No name (nor location) |
| Effic. | 4 | 4 | - | 4 | - |
| Prec. | Can give estimates | Up to project creator | - | - | - |
| Trace. | - | No provenance | - | - | - |
| Underst. | 5 | 4 | - | 5 | - |
| Avail. | Can download reports | Can download csv / excel / json | Not available | Not available | Not available |

**Table 4.** All projects that were reviewed in the research and their corresponding values for the data quality characteristics

| Data quality characteristic | Globe at Night | iNaturalist | iRecord | iSpot | Laji |
|---|---|---|---|---|---|
| Acc. | Syntactic checks | Syntactic and semantic checks | Syntactic and semantic checks | No accuracy checks | Syntactic checks |
| Compl. | All fields required | All fields required | All fields required | Less than half required | Less than half required |
| Consist. | 5 | 5 | 5 | 2 | 4 |
| Credib. | - | Community | Community | Community | Moderated |
| Current. | Observation time is different from upload | Observation time is different from upload | Observation time is different from upload | Observation time is different from upload | Observation time is different from upload |
| Access. | Maps (and other things) | Maps (and other things) | Maps (and other things) | Maps (and other things) | Maps (and other things) |
| Conf. | Username / anonymous | Username / anonymous | Location and name are shown | Username / anonymous | Location and name are shown with exceptions |
| Effic. | 4 | 5 | 4 | 4 | 4 |
| Prec. | Requires precision | Can give estimates | Can give estimates | Can give estimates | Can give estimates |
| Trace. | - | ? | - | - | - |
| Underst. | 5 | 5 | 4 | 4 | 5 |
| Avail. | Can download data from previous years | Not available | Can download reports | Not available | Can download csv / excel / json |

**Table 5.** All projects that were reviewed in the research and their corresponding values for the data quality characteristics

| Data quality characteristic | Landslides @NASA | Linking Landscapes for Massachusetts Wildlife | MLMP Data Portal | MNHNL Data portal | Nature's Calendar |
|---|---|---|---|---|---|
| Acc. | Syntactic checks | No accuracy checks | Syntactic checks | Syntactic checks | Syntactic and semantic checks |
| Compl. | Less than half required | Nothing is required | Less than half required | Half or more required | All fields required |
| Consist. | 2 | 1 | 3 | 3 | 5 |
| Credib. | - | - | Moderated | - | Community |
| Current. | Observation time is different from upload | Observation time is different from upload | Observation time is different from upload | Observation time is different from upload | Observation time is different from upload |

| | | | | | |
|---|---|---|---|---|---|
| Access. | Maps (and other things) | Reports | Maps (and other things) | Maps (and other things) | Maps (and other things) |
| Conf. | Username / anonymous | - | Username / anonymous | Location and name are shown with exceptions | Location and name are shown with exceptions |
| Effic. | 1 | - | 2 | 2 | 5 |
| Prec. | Can give estimates | Can give estimates | Requires precision | Can give estimates | Can give estimates |
| Trace. | Does show who did latest edit | - | - | - | Anyone can edit data |
| Underst. | 2 | - | 3 | 3 | 5 |
| Avail. | Can download csv / excel / json | Can download reports | Can download reports | Can download csv / excel / json | By contacting |

**Table 6.** All projects that were reviewed in the research and their corresponding values for the data quality characteristics

| Data quality characteristic | NatureSpot | Open Phylo | Project Noah | Riverwatch | Shark Trust |
|---|---|---|---|---|---|
| Acc. | Syntactic checks | - | No accuracy checks | - | No accuracy checks |
| Compl. | All fields required | - | Less than half required | - | Half or more required |
| Consist. | 5 | - | 3 | 5 | 3 |
| Credib. | Moderated | - | - | Sensor | - |
| Current. | Observation time is different from upload | - | Observation time is different from upload | - | Observation time is different from upload |
| Access. | Maps (and other things) | Not available | Observations only | Maps (and other things) | Observations only |
| Conf. | Location and name are shown | No name (nor location) | Real name is optional | - | Location and name are shown |
| Effic. | 4 | - | 3 | 4 | 3 |
| Prec. | Can give estimates | - | Can give estimates | Requires precision | Can give estimates |
| Trace. | - | - | - | - | - |
| Underst. | 4 | - | 4 | 4 | 5 |
| Avail. | Not available | Not available | Not available | Can download csv / excel / json | Not available |

**Table 7.** All projects that were reviewed in the research and their corresponding values for the data quality characteristics

| Data quality characteristic | Social Enterprise Mapping for Change | Stardust @ Home | The Great Sunflower Project | The Lost Ladybug Project | Treezilla |
|---|---|---|---|---|---|
| Acc. | No accuracy checks | - | No accuracy checks | No accuracy checks | Syntactic and semantic checks |
| Compl. | Less than half required | All fields required | Half or more required | Less than half required | Less than half required |
| Consist. | 2 | 5 | 5 | 2 | 3 |
| Credib. | Moderated | Analysis comparisons between participants | - | - | Community |
| Current. | Observation time is different from upload | - | Observation time is different from upload | Observation time is different from upload | Observation time is different from upload |
| Access. | Maps (and other things) | Not available | Maps (and other things) | Maps (and other things) | Maps (and other things) |
| Conf. | Username / anonymous | No name (nor location) | No name (nor location) | Location and name are shown | Username / anonymous |
| Effic. | 4 | - | 3 | 4 | 5 |
| Prec. | Can give estimates | - | Can give estimates | Can give estimates | Can give estimates |

| | | | | | |
|---|---|---|---|---|---|
| Trace. | - | - | - | - | Can view edits |
| Underst. | 5 | - | 3 | 5 | 5 |
| Avail. | Not available | Not available | Not available | Not available | Not available |

**Table 8.** All projects that were reviewed in the research and their corresponding values for the data quality characteristics

| Data quality characteristic | Urban Forest Map | Wisconsin Bat Program | Zooniverse |
|---|---|---|---|
| Acc. | Syntactic and semantic checks | Syntactic checks | - |
| Compl. | Less than half required | Less than half required | All fields required |
| Consist. | 3 | 5 | 5 |
| Credib. | Community | - | Moderated |
| Current. | Observation time is different from upload | Observation time is different from upload | Up to the project creator to handle |
| Access. | Maps (and other things) | Reports | Observations only |
| Conf. | Username / anonymous | Location and name are shown with exceptions | Username / anonymous |
| Effic. | 5 | 3 | - |
| Prec. | Can give estimates | Can give estimates | - |
| Trace. | Can view edits | - | - |
| Underst. | 5 | - | - |
| Avail. | Not available | Not available | Not available |

# Publication III

Musto, J., and Dahanayake, A.
**Quality characteristics for user-generated content**

Reprinted with permission from
*Frontiers of Artificial Intelligence and Applications*
Vol., 2021 (in press)

# Quality characteristics for user-generated content

Jiri MUSTO[a,2] and Ajantha DAHANAYAKE[a]
[a] *Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland*

**Abstract.** Today there are vast amounts of data collected from the internet. The general public generates most data with the use of social networks. There is a need to have a comprehensive approach to characterize the quality of such user-generated data collection from the internet. The data quality characteristics accepted among database and computer science communities have definitions that are not domain-specific. Therefore, there is no clear understanding of the data quality characteristics specific to user-generated content. In this research, different user-generated content platforms are examined against the general data quality characteristics to determine which quality characteristics are essential for user-generated content. The research contributes to a list of definitions of those data quality characteristics specific to user-generated content. These definitions help identify quality characteristics useful for user-generated content platforms and their implementations. The quality of the content of Atlas of Living Australia, Twitter, YouTube, Wikipedia, and WalkingPaths is evaluated to assess the essence of the quality characteristics defined in this research.

**Keywords.** data collection, data quality, information quality, quality characteristics, user-generated content

## 1. Introduction

Content generation involving the general public is a lucrative practice today. Such user-generated content (UGC) instigates heated discussions concerning the quality of the collected data. UGC platforms, such as social media platforms, have over three billion users worldwide, and users are averaging over two hours daily on these platforms. According to an article [1], over a billion stories are created daily on Facebook.

UGC is primarily unstructured content gathered and used for a variety of purposes. Social media platforms such as Facebook, Twitter, and Instagram, crowdsourcing platforms such as Wikipedia and OpenStreetMap, and citizen science platforms such as eBird and iNaturalist, are examples of UGC gathering platforms. UGC has been demonstrated to be used for investigating customer feedback [2,3], monitoring catastrophic environmental effects [4], tracking visitors in protected areas [5], flood research [6], emergency reporting [7], future prediction [8], service quality analysis [9], managing online encyclopedia [10], and targeting advertisements and recommendations for potential customers [11,12].

Social media platforms are designed for connecting users and sharing content within the community. Most users use social media platforms to interact with others and seek information about events, businesses, deals, and products [13]. The content shared on these platforms is mainly subjective. However, social networks have been increasingly used as sources for news among the younger generation, which are easily influenced by good or fake news [14]. Without social networks, such users may lack any knowledge of the surrounding world [15]. Furthermore, the younger generation may not read actual news, and as a result, social media has become their predominant world events and news channel.

Content generated by users is said to pertain to cases of unverified, misleading, or erroneous information that diminishes the credibility and lowers the quality of data [16–18]. Because of this, low data quality is one of the significant concerns in UGC [19] that can lead to poor decisions [20,21], or in rare cases, generate errors that eventually crash the underlying platforms [22].

Researchers and organizations have defined data quality as a collection of dimensions or characteristics [23–25]. This definition has been widely adopted and accepted [26,27]. There are over 40 different data quality characteristics, but many overlap with each other [23,25]. Quality characteristics frequently have a different definition depending on the domain; precision in healthcare has a different definition than

[2] Corresponding Author, Jiri Musto, School of Engineering Science, LUT University, Yliopistonkatu 34, 53850 Lappeenranta, Finland; E-mail: jiri.musto@lut.fi.

precision in geographic information. Consequently, there is no clear consensus and agreement on what characteristics fulfill the data quality in each context and use case [24,26,28–30].

Data quality is essential because a massive amount of content can lead to wrong conclusions if the quality is compromised [31,32]. Some platforms suffer from the abuse of "quantity over quality." One extreme example of such abuse is review bombing, where a group of people collectively gang up on one person or product [33,34]. Review bombing is a significant problem in online shops and reviews sites [35,36].

In order to overcome the ambiguity of UGC's data quality, this research examines the following research question:

*What are the quality characteristics of user-generated content?*

Researchers, organizations, and communities have promoted a plethora of formulations of data quality characteristics. This research aims to establish concise formulations of data quality characteristics for UGC by applying formulations found in [23–25,37] as the base. The works are selected based on their citation count and wide usage among researchers. In addition, this research aims to provide a solution for improving the data and information quality in a citizen science platform by integrating quality characteristics into the design of a platform that collects walking path observations

Because of the influence of UGC in modern businesses [38,39], the data quality of UGC is highly contested. Therefore, this research investigates the formulation of quality characteristics of UGC based on available literature. Formal formulations are based on existing formal definitions when applicable to UGC. When hardly any formal definitions exist, the definitions are formulated based on the context and use cases.

The main contributions of this research are:

- Giving exposure to the current status of data quality in UGC platforms
- Formalization of a comprehensive but not exhaustive list of quality characteristics for the domain of UGC
- A comprehensive list of quality characteristics to choose from during the design and implementation of future UGC platforms with substantially improved data quality in the generated content.

## 2. Background

### *2.1. Data quality research*

Data quality is a widely discussed topic in computer science and database technology. The systematic analysis of keyword-based article searches in scientific databases given in Table 1 accounts for the present (2020) status of data quality research.

The number of articles drastically reduces when the term "data quality" is combined with a keyword. It demonstrates that the actual research on data quality is a fraction of the many articles that mention "data quality" as a loud and popular buzzword.

**Table 1.** Results of keyword-based article search in scientific databases

| Search terms | Scopus | IEEE | Springer | ACM |
|---|---|---|---|---|
| "data quality" | 95069 | 20933 | 50586 | 4892 |
| AND "citizen science" | 1143 | 38 | 393 | 99 |
| AND "big data" | 5547 | 1466 | 3726 | 721 |
| AND "remote sens*" | 8 715 | 2497 | 3672 | 2 |
| AND "crowdsource*" | 2796 | 311 | 1001 | 0 |
| AND "user generated" | 705 | 30 | 574 | 186 |
| AND "social media" | 22327 | 150 | 2262 | 520 |
| "data quality defin*" | 20 | 42 | 59 | 0 |
| "data quality model" | 407 | 123 | 193 | 39 |
| "data quality dimension" | 1154 | 62 | 455 | 49 |

| | | | | |
|---|---|---|---|---|
| "data quality characteristic" | 40 | 13 | 86 | 2 |
| "data quality framework" | 319 | 56 | 109 | 12 |

Some widely cited data quality research works belong to the 1990s [20,25,40], and new research works and standards extend them [23,24,41,42]. Researchers and standards define data quality as:

- Multidimensional, divided into characteristics
- Contextual
- Characteristics' importance is subjective
- Quality is measured through the characteristics.

[25] generalizes the data quality characteristics under four categories: intrinsic, contextual, accessibility, and representational characteristics. ISO standard [24] categorizes data quality characteristics into inherent, inherent and system dependent, and system dependent categories.

Different assessment processes and frameworks have proposed specific steps and metrics to evaluate quality and improvement ideas when quality is low [30]. An extensive survey of existing data quality frameworks is provided in [43]. However, there is a lack of actual assessment or evaluation methodology [27,44]. Some frameworks have implemented data quality evaluation for one specific use-case, such as social media, but the final test only consists of one characteristic [30].

### *2.2. User-generated Content*

Data quality in UGC has been explored since social networking, and social media platforms took off during the 21st century. As data quality is contextual, definitions for each characteristic in UGC can be different from other domains. Moreover, even within the UGC domain, there are different definitions for the same characteristics [45–47].

Data quality in UGC is crucial as regular citizens generate the content. The quality of data in UGC is often questioned as users are not experts. As a result, UGC is more vulnerable to low-quality data compared to other domains [48]. For this reason, some projects use specific tools, like sensors, for data collection to make data more reliable compared to just human-computer interaction [49]. Several methods for improving UGC have been proposed, such as participant selection [50], task allocation [51], and reputation models [52].

## 3. Data quality characteristics

ISO quality characteristics [24] are used as the starting point to develop a list of UGC data quality characteristics. These characteristics are presented in Table 2.

**Table 2.** List of initial data quality characteristics

| **ISO Data quality characteristics** [24] | |
|---|---|
| accessibility | availability |
| completeness | compliance |
| consistency | confidentiality |
| credibility | currentness |
| efficiency | portability |
| precision | recoverability |
| semantic accuracy | syntactic accuracy |
| traceability | understandability |

From the ISO characteristics, *accessibility*, *availability*, *efficiency*, *portability*, and *recoverability* are discarded as they are related to the underlying system and not data itself. The list in Table 2 is further extended to accommodate the UGC domain's data quality characteristics with contributions from domains of general data quality, social media, and big data. These additional characteristics are presented in Table 3.

**Table 3.** Data quality characteristic from other domains

| **Extended data quality** [23,40,53] | **Social media** [5,54] | **Big data** [27,55,56] |
|---|---|---|
| objectivity | privacy | relevance |
| provenance | usability | value |
| timeliness | | volume |

To formulate practical definitions for specific characteristics, it is essential to be clear with the general understanding of the term, limiting misinterpretation. Therefore, the formal data quality definitions of the characteristics listed are formulated using existing literature.

*Accuracy*: Closeness between data values *v* and $v_0$, where $v_0$ is the correct representation of what the data value *v* aims to represent. Based on syntactic and semantic accuracy [23].

*Syntactic accuracy*: Closeness of words in the text to a reference vocabulary. *K* is the number of words, $w_i$ is a word in the text, and *V* is the vocabulary used in the text (1)[37].

$$syntactic\ acc = \frac{\sum_i^K closeness\ (w_i, V)}{K} \quad (1)$$

*Semantic accuracy*: How correctly the meaning of values represents real-world facts. An object identification problem where $\alpha$ and $\beta$ are a pair of tuples to be matched, *M* is the set that contains a record of similar existing pair, *U* is the set that represents nonmatch and $\underline{x}$ is a random vector of *n* number of attributes, and *p()* is the probability of matching (2)[23,57].

$$\langle \alpha, \beta \rangle \in \begin{cases} M\ if\ p(M|\underline{x}) \geq p(U|\underline{x}) \\ U\ otherwise \end{cases} \quad (2)$$

*Completeness*: Completeness of a tuple with respect to the values of all its fields where $T_v$ is the number of null values in a tuple and $N_v$ is the total number of values in a tuple (3)[23,58].

$$completeness = 1 - \frac{T_v}{N_v} \quad (3)$$

*Consistency*: Violation of semantic rules defined over (a set of) data items, where items can be tuples of relational tables or records in a file. *g* is the data value, and *N* is the number of rules for *g* (4)[59].

$$r_n(g) = \begin{cases} 0, if\ g\ fulfills\ rule\ r_n \\ 1\ else \end{cases}; cons(g) = 1 - \frac{\sum_{n=1}^{N} r_n(g)}{N} \quad (4)$$

*Credibility*: How data are accepted or regarded as true, real, and credible, where *dist* is the distance between the sensor *s* and entity *e*, and $d_{max}$ is the maximum distance acceptable (5)[60].

$$credibility = \begin{cases} 1 - \frac{dist}{d_{max}} & if\ d(s,e) < d_{max} \\ 0 & otherwise \end{cases} \quad (5)$$

*Objectivity*: Data is unbiased and impartial, where *E* is evidence, *H* is a hypothesis (assumed value), and *p()* denotes the probability (6)[61].

$$w(E, H, H') = log\ \frac{p(E|H)}{p(E|H')} \quad (6)$$

*Precision*: Precision refers to the amount of detail that can be discerned in space, time, or theme. Using Levenshtein edit distance where *a* and *b* are the given values, *i* and *j* are the indexes (7)[57,62].

$$lev_{a,b}(\mathrm{i},\mathrm{j}) = \begin{cases} max(i,j) & if\ min(i,j) = 0, otherwise \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} \end{cases} \quad (7)$$

*Volume*: Appropriate amount of data: the extent to which the quantity or volume of available data is appropriate. Sample size formula where *z* is z-score, *e* is the margin of error, *p* is standard deviation, and *N* is population size (8)[63].

$$sample\ size = N * \frac{\frac{z^2 * p * (1-p)}{e^2}}{\left[N - 1 + \frac{z^2 * p * (1-p)}{e^2}\right]} \quad (8)$$

Compliance: Defining and evaluating the compliance between data and schemas measure of relationship (similarity, relatedness, distance, etc.) between two entities. Where *a* and *b* are values of elements in minimum distance and $\bar{a}$ and $\bar{b}$ are means of all elements (9)[64].

$$compliance\ (degree\ of\ variance\ ) = \frac{\sum(a-\bar{a})(b-\bar{b})}{\sqrt{\sum(a-\bar{a})^2 \sum(b-\bar{b})^2}} \quad (9)$$

*Currentness*: Currency concerns how promptly data are updated with respect to changes occurring in the real world (10)[23].

$$currentness\ = Age + (DeliveryTime - InputTime) \quad (10)$$

*Timeliness*: Data is sufficiently up to date for the task at hand. *Volatility* is the defined length of how long data remains valid (11)[23].

$$timeliness = \max\{0, 1 - \frac{currentness}{volatility}\} \quad (11)$$

*Privacy*: Data is hidden or concealed from others. *S* is the sensitivity of a data item, and *V* is the visibility in a given context, and *R* is relatedness. *a*, *b* and *c* are real numbers (12)[65].

$$PrivacyRisk_{(i,j)} = \frac{S_i^a \times V_{(i,j)}^b}{R_{(i,j)}^c} \quad (12)$$

*Relevance*: The extent to which data are applicable and helpful for the task at hand. *n* is the number of words in a sentence, *m* is the number of characters in a word, and *WordSimilarity* is the similarity between two words between 0 and 1 (13)[66].

$$SentenceSimilarity\ (Q, Q') = \frac{1}{n}\sum_{1 \le j \le n}(max_{1 \le i \le m} WordSimilarity(w_j, w_i') \quad (13)$$

*Usability*: A collection of other characteristics characterized by usability aspects, verifiability, imperfection, and integration (14)[67].

$$usability = avg(accuracy + credibility + completeness + currentness + relevance + granularity + accessibility) \quad (14)$$

*Value*: The extent to which data are beneficial and provide advantages from their use (15).[68]

$$DataValue(t) \geq (GatherCost + MaintainCost + AccessCost)/GB/yr * RetentionPeriod \quad (15)$$

*Confidentiality*: Data is available to authorized persons when and where needed (especially in the medical field). $W_c$ is the weight of confidentiality for a subsystem, $x_s$ is a dependency score for a subsystem, and *n* is the number of subsystems in an information security system (16)[69].

$$confidentiality = \frac{\sum_{i=1}^{n} w_{c_i} * x_{s_i}}{\sum_{i=1}^{n} w_{c_i}} \quad (16)$$

*Granularity*: Granularity concerns the ability to represent and operate on different levels of detail in data, information, and knowledge located at their appropriate level. Shannon entropy in terms of Hartley entropy for partition granularity (17)[70].

$$granularity = \log|U| - \sum_{i=1}^{n} \frac{|X_i|}{|U|} \log(|X_i|)\ , U\ is\ a\ universal\ \ set\ and\ set\ X \subseteq U \quad (17)$$

*Traceability*: The extent to which data are well documented, verifiable, and easily attributed to a source. *R* is a source, *Ω* is a set of *R*, *E(Ω)* is a measure of uncertainty, and $\lambda$ is the number of reports (18)[71].

$$Network\ traceability\ entropy\ (NTE), E^{\lambda} = \sum_{\Omega:|\Omega|=\lambda} E(\Omega)/\binom{|R|}{\lambda} \quad (18)$$

*Provenance*: Provenance of a resource is a record of metadata containing descriptions of the entities and activities involved in producing and delivering or otherwise influencing a given object. *Q* is a query, *I* is an instance, and *t* is a tuple in *U* (19)[72].

$$\begin{aligned} whyProvenance(Q,I,t) &= \{J \in I \mid t \in Q(J)\} \\ whereProvenance\ (\{u\},I,t) &= \begin{cases} (A:\emptyset)_{A\in U}, if\ t = u \\ \perp, otherwise \end{cases} \\ howProvenance\ (Q,I,t) &= \ Q^{K_{How}}\ (I_{How})t \end{aligned} \quad (19)$$

*Understandability*: The ease with which data can be comprehended without ambiguity and be used by a human information consumer (20)[73].

$$understand. = -0.33 * Abstraction + 0.33 * Encapsulation + 0.33 * Coupling + 0.33 *\ Cohesion - 0.33 * Polymorphism 0.33 * Complexity - 0.33 * DesignSize \quad (20)$$

*Readability*: Reading easiness, the ease of understanding written text using Gunning-Fox index (21)[60,74].

$$readability = \ 0.4 * [\left(\frac{words}{sentence}\right) + 100 * (\frac{complexwords}{words}) \ ] \quad (21)$$

**4. Case studies: user-generated content creating platforms**

Citizen science platforms are famous for using public submitted context-specific content. There are over 1000 citizen science platforms (https://scistarter.org), with content related to wildlife, environment, and city management. [75] gives a detailed overview of close to 100 citizen science platform evaluations.

*Atlas of Living Australia* (ALA) (https://ala.org.au/) is an Australian citizen science platform for plant and wildlife monitoring. ALA has integrated another citizen science platform called *iNaturalist* (https://inaturalist.ala.org.au/), allowing data from iNaturalist to be sent to ALA. Features of ALA can be generalized because most citizen science platforms operate using similar functionalities. In citizen science platforms, citizens send reports with a varying number of fields that often include multimedia. Citizens may give a username when submitting reports, and reports can be updated later. Reports can have automated tests for quality and be voted by the community. Some issues specific to citizen science platforms, such as content submitted by regular citizens making credibility questionable, personal details of users are sometimes shown to the public, and some reports stay incomplete.

*Twitter* (https://twitter.com/) is a social media platform where users share short texts and images called tweets. Users can comment, like or reshare other people's tweets. These actions provide context on how well tweets are received. On Twitter, tweets and accounts can be made private. In addition, users have a number of followers and followed, and tweets can have hashtags that work like keywords. Most content comes from individuals without any source material. Thus it is challenging to define credible information. Tweets are occasionally in another language or nonsensical, and some people make fake accounts pretending to be someone else.

*Worldometer* (https://www.worldometers.info/) is a crowdsourcing platform that collects and aggregates information from multiple sources. The sources vary from news articles and healthcare-operated sites to third-party organizations. Worldometer is widely referenced as a reliable real-time information provider during the Covid-19 pandemic. In Worldometer, information is primarily numbers and based on a source. Worldometer is continuously updated and considered to be reliable based on the sources it uses. The information is presented in text, graphs, and tables. However, some information requires users to

contribute, leading to incompleteness. The credibility of information must be checked before sharing it with the public, and inaccurate information from users requires further administrator reviews.

*Wikipedia* (https://www.wikipedia.org/) is an online encyclopedia where registered users create and modify content, and more reputable volunteers act as moderators. Wikipedia requires a source before it accepts content as valid information. In addition, Wikipedia has a specific style that articles must follow. Because community updates and moderates Wikipedia, it is updated fast in the native language compared to translations. Most information is written clearly and understandably.

Nevertheless, there are cases when information is not correct in Wikipedia or correct information is not accepted because of the source. Sometimes, the source material's credibility can be questionable, and volunteer administrators' opinions may be reflected in the accepted content. Few articles are left incomplete because of the lack of contributions.

*YouTube* (https://www.youtube.com/) is a video-sharing platform owned by Google. Anyone can view public videos, but only registered users can upload new videos. Videos are not allowed to infringe any copyright laws, and the content must not be harmful or hateful. YouTube has similar characteristics to Twitter, such as videos have a number of views, and they can be liked/disliked and commented on. As regular citizens make most videos, the information may not be credible, and there is no guarantee of objectivity. It is challenging to validate the official channels from other forms of propaganda, and some users purposefully report videos they do not like.

Each of the introduced platforms has different use-cases and contexts. Content in Wikipedia and Worldometer are meant for public consumption, but their context is different. Content in citizen science platforms is used for research and context changes from one platform to another. Twitter and YouTube are used for connecting with others and sharing subjective content. So Twitter and YouTube have the same context, but the provided content is vastly different. With this in mind, the public uses all introduced platforms. Thus the platforms are expected to have some level of quality in the content.

Table 4 presents the mapping of data quality characteristics listed in Section 3 to the described UGC platforms. Characteristics are examined from the platform's context (credibility relates to the user's credibility). The data quality of UGC is governed by the quality of the content requested from the user. The context defines the limits and requirements for the data quality that the content needs to fulfill. Some characteristics require a specific use-case for the content, such as relevance and value. Each characteristic is given a value as follows:

- 1: The platform takes into consideration by requiring specific content.
- 0: The platform does not take into consideration. The user can submit content without any limitations.
- ?: Unclear if the system considers that characteristic or not.
- +/-: Situation dependent and only applicable to specific use cases.

Table 4 shows that Twitter and YouTube care less about information correctness than the other UGC platforms. Twitter has no regard for completeness, but ALA, Wikipedia, and Worldometer have minimum requirements for submissions. In addition, there are situations when a data quality characteristic needs a degree of variation. In ALA, timeliness is sometimes essential in situations where the information must be from specific periods. When extracting data from the UGC platform, it is beneficial to know the quality of extracted data. When using Twitter and YouTube data, objectivity must be evaluated separately because the platforms place no importance on objectivity.

**Table 4.** Data quality characteristic mapping to platforms that curate UGC

| Data quality characteristics | ALA | Twitter | Worldometer | Wikipedia | YouTube | Explanations of the characteristics In terms of information gathered by the platform |
|---|---|---|---|---|---|---|
| Syntactic accuracy | 1 | 0 | 1 | 0 | 0 | User submits information in the syntax expected by the system |
| Semantic accuracy | 1 | 0 | 1 | 1 | 0 | User submits information that follows semantic rules set by the system |
| Completeness | 1 | 0 | 1 | 1 | 0 | The system expects the user to submit a minimum amount of information |
| Consistency | 0 | 0 | 0 | 0 | 0 | Information is consistent in comparison to multiple users input |
| Credibility | 1 | 1 | 1 | 1 | 1 | User's credibility |

| | | | | | | |
|---|---|---|---|---|---|---|
| Objectivity | 1 | 0 | 1 | 1 | 0 | User submits objective information |
| Precision | 1 | 0 | 1 | +/- | 0 | Information is detailed |
| Volume | 1 | 1 | 1 | 1 | 1 | Similar information from different sources |
| Compliance | ? | ? | ? | ? | ? | Information is compliant with a standard |
| Currentness | 1 | 1 | 1 | 1 | 1 | Information is current |
| Timeliness | +/- | 0 | 0 | 0 | 0 | Information is from the correct time |
| Privacy | 1 | 1 | 0 | 0 | 1 | Personal information is not displayed |
| Relevance | 1 | 1 | 1 | 1 | 1 | User submits relevant information to the topic |
| Usability | 1 | +/- | 1 | 1 | +/- | Information is usable by others |
| Value | 1 | +/- | 1 | 1 | +/- | Information has value for others |
| Confidentiality | 0 | 0 | 0 | 0 | 0 | Sensitive information is inaccessible |
| Granularity | +/- | 0 | 0 | 0 | 0 | Information is split into specific parts |
| Traceability | 1 | 1 | 1 | 1 | 1 | Information origins are known |
| Provenance | 0 | 0 | 0 | 0 | 0 | Changes to information are known |
| Understandability (or readability) | 1 | 1 | 1 | 1 | 1 | Information is understandable (or readable) |

Based on the above analysis and observations, the quality characteristics specific to UGC can be formulated as follows:

*Traceability: How well the content is attributed to a specific source and time.*

Twitter and YouTube record the user and time when content is created. In Worldometer and Wikipedia, the content has a specific source. Wikipedia tracks the user who has added or edited content. Similarly, citizen science platforms track the time created, the place where the content relates, and who submits it.

*Credibility: How credible the content is based on who is giving the content.*

In social media, credibility is subjective even when official channels of credible organizations or people are the creators. Credibility can be based on three factors: number of likes or followers, community opinion based on the comments, and user verification. For Wikipedia and Worldometer, credibility is based on the source material and in citizen science, credibility is based on community opinion and administration.

*Currentness: How promptly content is updated with respect to changes occurring in the real world.*

Twitter is designed for content to be created and shared as soon as possible. On YouTube, most content creators want to create content based on current hot topics. Wikipedia's purpose is to have current facts. Citizen science platforms' purpose is to get current information. Finally, Worldometer is continuously updating its content.

*Relevance: How relevant the given content is to the platform context.*

Worldometer, Wikipedia, and citizen science all have a specific purpose, and all three expect to get relevant content from users. YouTube and Twitter have opinion-based content, and the content always relates to some topics making it arguably relevant.

*Accuracy: Accuracy is the closeness of given content to the expected content. Based on syntactic and semantic accuracy.*

*Syntactic accuracy: Closeness of the content syntax that the user gives depending on the platform context.*

Twitter, Wikipedia, and YouTube all accept various types making information always syntactically accurate. Only Worldometer and citizen science limit what a user can give to ensure syntactic accuracy.

*Semantic accuracy: How correctly the information within the content matches the real-world facts.*

Twitter and YouTube are not interested in semantic accuracy. Worldometer and Wikipedia require sources to check semantic accuracy, and in citizen science, there are limits to what content can be given to have some semantic accuracy.

*Completeness: How complete content is and not missing important information depending on the platform context.*

Social media operates on more opinion-based content, and there is no minimum requirement of what needs to be given. In Wikipedia, short or incomplete information is marked by the platform automatically. Citizen science and Worldometer expect specific information at a minimum before any information can be sent.

*Usability: How usable the content is based on the platform context. It is affected by accuracy, completeness, and credibility.*

On Twitter and YouTube, content created by official channels of organizations is meant to be used by the public. Wikipedia and Worldometer are meant to be used by everyone, and unusable content is quickly removed. On the other hand, citizen science projects are meant for research.

*Value: How useful the content is and provides advantages from its use.*

Citizen science content is meant for research purposes that will lead to some value. Worldometer and Wikipedia are meant to be information sources making their content valuable. Twitter and YouTube provide value when combining a massive amount of content. However, individually, tweets and videos do not provide much value.

*Understandability (and readability): How easily the information from the content can be comprehended without ambiguity by a human consumer within the platform context (and how easy written text is to read and comprehend).*

Wikipedia is meant for the public, and many complex things are explained so that a novice can comprehend. Worldometer provides information in various formats making their content understandable. Citizen science often has maps and graphs to increase understandability. Only social media content can be challenging to understand, but more understandable content will be more popular and promoted.

*Objectivity: How unbiased and impartial the content and its information are.*

Twitter and YouTube are meant for opinion sharing making objectiveness non-essential. Worldometer and Wikipedia require sources to ensure objectiveness. In citizen science, content is subjective but made more objective by using community opinion.

*Privacy: How much of the user's personal information is concealed.*

Worldometer and Wikipedia do not handle private information, and social media platforms allow users to hide their information. In citizen science, content usually includes a location, but users are not required to use their names.

*Volume: The amount of similar information given by multiple users.*

All case platforms want to have a high volume of information. Wikipedia and Worldometer commend having more than one source. When collecting opinions from social media, having multiple people with similar opinions is valuable for researchers. In citizen science, if no one else agrees on a report, it is quickly deemed untrustworthy.

*Precision: How detailed the given content is in the platform context.*

Precision is not considered on Twitter or YouTube. For citizen science, precision is considered whenever there is location-based information given. In Wikipedia, precision is situational, but in most cases, no precision is required. On the other hand, Worldometer does not want any ambiguity in its information; thus, precise information is expected.

The listed characteristics can be used to define the data or information quality characteristics in UGC. Information is the content received from users, while data is the content stored in the database [76]. Only *precision* is not applicable in the context of information.

## 5. Integration Of Quality Characteristics Into The Citizen Science Platform: WalkingPaths

A citizen science web platform called WalkingPaths integrates the essential data quality characteristics listed in Table 2 into its design [76]. The platform is developed using ReactJS for the frontend and NodeJS for the backend with a MongoDB database. Mongoose middleware is used to enforce syntax restrictions on data.

The platform collects walking path information from citizens in Finland. Citizens are asked to fill a simple form consisting of the path's location and condition, and they are given an option to send an image with the observation. The data is collected from March 2020 to August 2020, and the final data set consists of 108 observations.

When integrating quality characteristics into the design, it is necessary to decide where these characteristics should be implemented. Characteristics should be integrated into the data model as well as the user interface. The database may store information related to these characteristics, but the interface is

responsible for checking and enforcing them. Characteristics can be integrated into the user interface by limiting or extracting specific information from the content provider's input. For instance, the address is complete if geolocation exists. Similarly, the characteristics can be added as constraints in the database. A more detailed description of the integration of quality characteristics is found in [76].

Figure 2 shows the database schema using a snowflake model [77] of the platform WalkingPaths. In the center is the fact table *WalkingPathObservation*, and it is connected to several dimension tables. A snowflake schema can be easily transformed into a relational data model. Several data quality characteristics are integrated into the model as separate attributes, and these are bolded and cursive. These include precision, accuracy (syntactic and semantic), completeness, volume, credibility, privacy, objectivity, and traceability. The characteristics can store relevant quality evaluations.

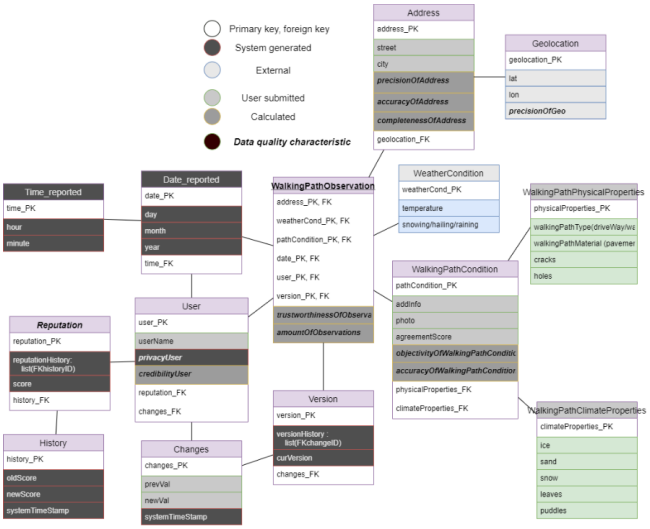


**Figure 2.** Snowflake schema for WalkingPaths

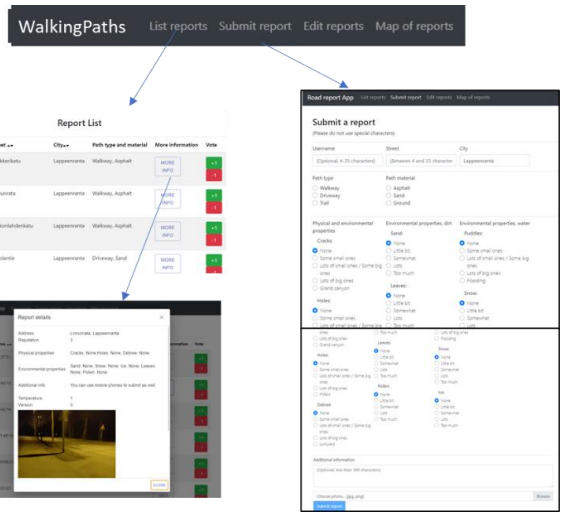


**Figure 3.** WalkingPaths list of observations and report submission

Figure 3 shows the transition using the navigation bar to listing observations and submitting new reports. The observation list only shows minimal details for each report, such as location and time. Users can open a *More information* pop-up -window to reveal other information. Reports can be up-/downvoted, but as the platform does not require registration, some restrictions have been implemented in the voting mechanism to reduce misuse. Most choice boxes in the report window have predetermined values to guarantee each report's completeness. Only two choice boxes do not have a value, but the report cannot be submitted before some value is given to both of them. The usage of choice boxes is an excellent method to increase the report's syntactic and semantic accuracy while enabling the content provider to know what to look for before submitting anything. Finally, additional information can be typed in the text box.

The report editing view is similar to submitting a report with the additional search box for finding existing reports. Map view presents a map where observations are shown as markers. More detailed figures are found in [76].

## 6. Case Study: Data Quality In User-generated Content Platforms

Data quality is evaluated by subjecting a data set from each platform to specific queries related to each quality characteristics presented in Section 4. The queries are performed using the data analytics platform RapidMiner (https://rapidminer.com/), a commercial software designed for data mining, analytics, and machine learning. Table 5 presents the general RapidMiner queries for each of the characteristics. The *value* characteristic for each data set is calculated based on other characteristics to simplify the definition.

RapidMiner query results are given as values between 0 and 1. Values indicate the percentage of correct data entities for each characteristic (conform to the given query). These resulting values are presented in Table 6. Not applicable (NA) results are deemed as zero because if something is not applicable, it does not exist. The number of entities in each data set is given in the headers of Table 6.

**Table 5.** General RapidMiner queries for DQ characteristics

| Characteristic | General query | (Data mining) Technique |
|---|---|---|
| Syntactic accuracy | Data entities correspond to the expected syntax and format defined in the data set. This information is based on the headers and what data is expected, and in what format. | Text/content mining. Compare value syntax to expected (integer, string, date) and filter out incorrect values. Compare the number of correct values to the total number. |
| Semantic accuracy | Data is semantically correct compared to what is expected based on the headers | Value comparison. Headers define what data should be, for example, "date," "name," "country." Each value is checked to see if they are actual dates, countries, names. |
| Completeness | Each data set is checked for missing values for completeness. | Filter missing values and compare the amount to total (automated functionality) |
| Credibility | The credibility of the content provider giving the information. | Reputation model and calculation compared to the average score |
| Objectivity | Objectivity is based on how objective given information is. If multiple sources agree on the information, it is more likely to be objective. | Count how many entities from different sources/content providers have the same information and how many are only from singular content providers/sources. |
| Volume | For each data set, the volume is checked from similar data entities compared to all entities. The similarity is only based on a few attributes. | Count how many entities from different sources/content providers have relatable information based on selected attributes and how many are only from singular content providers/sources. |
| Currentness | Data has given a date/time. Compare that to the time data was extracted from the database | Content mining and comparison |

| | | |
|---|---|---|
| Privacy | Privacy is measured based on the number of personal information stored with the data. | Filter out content providers whose possible real names are given and compare them to the total amount (text mining) |
| Relevancy | The relevance of the data to the given context regardless is the data correct or not. | Data comparison to given relevance factor such as the topic. |
| Usability | Usability is based on the context of usage for each data set | Content mining and comparison |
| Value | Value depends on the user. In this research, value = (Syntactic + Semantic + Credibility + Relevancy + Usability + Understandability) / 6 | Calculation based on other characteristics |
| Traceability | Each data set provided attributes for time, location, and content provider that are checked for traceability. | Count how many entities have a valid time, location, and content provider/source compared to all entities |
| Understandability | Understandability is based on the content of information, in general, readability. Unreadable texts/characters and undefined acronyms reduce the understandability | Text mining of invalid words. |

**Table 6.** Query result of data quality characteristics

| **Characteristic** | **WalkingPaths 108 entities** | **ALA 894 entities** | **Twitter 6012 entities** | **YouTube 750 entities** | **Wikipedia 19 797 entities** | **Worldo-meter 2996 entities** |
|---|---|---|---|---|---|---|
| Syntactic accuracy | 1.00 | 0.95 | 1.00 | 1.00 | 0.96 | 1.00 |
| Semantic accuracy | 0.96 | 0.96 | 0.93 | NA | 1.00 | 1.00 |
| Completeness | 1.00 | 0.72 | 0.89 | 0.99 | 0.95 | 1.00 |
| Credibility | 0.74 | NA | 0.32 | 0.82 | 0.32 | NA |
| Objectivity | 0.54 | 0.23 | 0.19 | 0.11 | 0.50 | NA |
| Volume | 0.36 | 0.42 | 0.61 | 0.69 | NA | NA |
| Currentness | 1.00 | 0.29 | 1.00 | 1.00 | 1.00 | 1.00 |
| Privacy | 1.00 | 0.86 | 0.67 | 1.00 | 1.00 | 1.00 |
| Relevancy | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Usability | 1.00 | 0.79 | NA | NA | 0.85 | 1.00 |
| Value | 0.95 | 0.77 | 0.68 | 0.47 | 0.81 | 0.83 |
| Traceability | 1.00 | 0.76 | 0.66 | 0.67 | 0.67 | 1.00 |
| Understandability | 1.00 | 0.93 | 0.82 | NA | 0.72 | 1.00 |

Results show that chosen platforms do not support all quality characteristics, and Twitter and YouTube performed the worst out of all. These are social media platforms designed for opinion sharing and not for credible data collection and information sharing., It is necessary to consider integrating data quality characteristics into the design during its implementation to accommodate the maximum number of quality characteristics.

Overall, WalkingPaths scored similarly to Worldometer, aside from a few significant aspects. Semantic accuracy is less in WalkingPaths than in Worldometer because there are some misspellings in the addresses given in WalkingPaths. Semantic accuracy could be improved with easy addition to the user interface where a content provider is recommended the address during typing. However, if a similar platform is extended outside of one country, the list of cities and street names would inflate drastically. Other significant differences are credibility, objectivity, and volume that do not apply to Worldometer. Compared to other platforms, WalkingPaths is better in objectivity and credibility and only loses in volume.

WalkingPaths achieved higher scores in everything except volume in comparison to ALA. ALA has been available for many years, so it is understandable for WalkingPaths to have a lower volume score. For completeness, currentness, and traceability, the most significant difference in scores is missing dates and times in ALA data, and a lot of data entities were before the year 2000. In some instances, time formatting changed. ALA data provided some information on the source of observations, but there are no methods to

determine if the source is credible, making credibility unapplicable. While it can be argued that ALA performs worse because it collects different kinds of observations, the same techniques used in the development of WalkingPath can be utilized in any type of observation. The difference in observation types is negligible as both platforms' underlying principle stays the same.

## 7. Discussion

To improve the quality of information, the method of the collection must be improved. The improvement can be made by implementing checks or limits within the user interface to reduce misinformation drastically. In Worldometer, users can only give a limited amount and type of content through the user interface, thus ensuring that the information sent to the system is at least of decent quality. Social media platforms could use specific filters for information searches that are based on different criteria. Twitter already has hashtags implemented, but these are always user-defined. There could be some reserved hashtags that, when used, Twitter could enforce some quality control checks for the content shared while using the specific hashtag.

Another way to improve the collection is to remodel the user interface. Most users give content based on what is asked in UGC platforms. What is asked defines what is received, not just having checks or limits applied to the user interface but designing it to answer specific questions. Even if the input is not limited, most users will unconsciously avoid giving misinformation when answering questions.

Not all users may care about the quality. The content's quality could be evaluated by the application based on the selected quality characteristics. The results of these evaluations could be embedded, for example, as system-generated data to Twitter API. This way, regular users would not see these evaluation results, and they would only be visible in raw Twitter data. Another possibility would be to add an option for regular users to see these evaluation results, similar to the history of edits Facebook has implemented. It is not shown unless selected explicitly by the user.

A platform where quality characteristics of UGC are integrated into the user interface and data model is presented in [76]. The same platform is used in this research to evaluate the design against non-citizen science UGC platforms. The integration of quality characteristics brings advantages and disadvantages to the content provided.

Some of the advantages of implementing quality characteristics are:

- Receiving higher quality content from users
- Determine the quality of content
- Enables content filter for users (if necessary)
- Possible to show others the quality score of a given content (if necessary)
- The quality characteristics implementation can justify reusing data collected from the platform

Some disadvantages are:

- May limit what content users can share
- May limit the way content is shared and used
- May affect how data is stored

Designers and developers of UGC platforms should consider having some data and information quality control implementations. During the design phase, these decisions should be made to fully utilize appropriate methods and ensure the quality of the content shared through the platform. The disadvantages of such an approach, depending on how the characteristics are implemented. For example, when implementing checks for content completeness, it is possible to either require absolute completeness or allow incompleteness. If absolute completeness is required, users cannot submit any incomplete content. Thus, the content they share is limited. If incomplete content is allowed, the user may share this content and later edit it, or the system can mark the content as incomplete for others. It is possible to avoid the disadvantages through design decisions. Currently, Worldometer requires absolute completeness, while Wikipedia allows incomplete content.

The research presented has some limitations, such as:

- Only a limited number of platforms have been examined

- Only the data quality characteristics available in research works have been considered, but the list can be extended by integrating experiences from the practice.
- The definitions presented are only applicable to the UGC domain and are not designed to be used for other domains

## 8. Conclusion

Quality of content is an essential part of any platform that collects content from non-experts with varying levels of expertise and knowledge. Unfortunately, UGC platforms are considered untrustworthy because the quality of content is questionable [16–18].

It is necessary to understand what quality is to improve data quality. Data and information quality must be defined for each domain, and there are no existing definitions for UGC. This research provides an extensive but not exhaustive list of quality characteristics with definitions specifically tailored for UGC. The importance of quality characteristics depends on the platform, and different contexts for the platform will change what characteristics should be emphasized.

Considering and integrating quality characteristics during the design of a platform has been presented in [75,78]. The articles provide general guidelines on how the quality characteristics can be implemented in the design of a platform. A citizen science platform for collecting WalkingPaths information is created to experiment with the proposed methodology, and the quality of collected content is evaluated against existing citizen science platforms [76].

Results show that integrating quality characteristics into the design increases the overall quality of UGC platforms. Most characteristics can be easily integrated into the design without significant changes. This method can be used in any platform and even applied to an existing platform if necessary. The most important part is identifying which characteristics are essential in each platform, and this has to be done by considering the context where the information will be used. The definitions of quality characteristics for UGC are helpful instruments for identifying essential characteristics for a UGC platform's content.

This research contributed to the formulation of specific quality characteristics definitions specifically for the UGC domain that collects content using social networks and web technology. The presented definitions are based on existing definitions of general data quality characteristics but modified for UGC usage. Quality characteristics depend on the context of the platform. Even within the same domain, different contexts for the platform will change what characteristics should be emphasized. This research contributes to building a cumulative tradition of building a sound set of UGC's quality characteristics.

## References


[1] Influencer Marketing Hub. 42 Essential Social Media Statistics for 2020 [Internet]. Influencer Marketing Hub. 2020 [cited 2020 Apr 28]. Available from: https://influencermarketinghub.com/social-media-statistics-2020/

[2] Ranjan S, Sood S, Verma V. Twitter Sentiment Analysis of Real-Time Customer Experience Feedback for Predicting Growth of Indian Telecom Companies. In: Proceedings - 4th International Conference on Computing Sciences, ICCS 2018. Institute of Electrical and Electronics Engineers Inc.; 2019. p. 166–74.

[3] Mariani M, Di Fatta G, Di Felice M. Understanding Customer Satisfaction with Services by Leveraging Big Data: The Role of Services Attributes and Consumers' Cultural Background. IEEE Access. 2019;7:8195–208.

[4] Ahmouda A, Hochmair HH, Cvetojevic S. Using Twitter to Analyze the Effect of Hurricanes on Human Mobility Patterns. Urban Sci. 2019;3(3):87.

[5] Tenkanen H, Di Minin E, Heikinheimo V, Hausmann A, Herbst M, Kajala L, et al. Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. Sci Rep. 2017 Dec 14;7(1):17615.

[6] Arthur R, Boulton CA, Shotton H, Williams HTP. Social sensing of floods in the UK. PLoS One. 2018;13(1).

[7] Ludwig T, Reuter C, Pipek V. Social Haystack: Dynamic Quality Assessment of Citizen-Generated Content during Emergencies. ACM Trans Comput Interact. 2015;22.

[8] Asur S, Huberman BA. Predicting the future with social media. In: Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010. 2010. p. 492–9.

[9] Haryani CA, Hidayanto AN, Budi NFA, Herkules. Sentiment Analysis of Online Auction Service Quality on Twitter Data: A case of E-Bay. In: 2018 6th International Conference on Cyber and IT Service Management, CITSM 2018. IEEE; 2019. p. 1–5.

[10] Bykau S, Korn F, Srivastava D, Velegrakis Y. Fine-grained controversy detection in Wikipedia. In: Proceedings -

International Conference on Data Engineering. 2015. p. 1573–84.
[11] Ouyang S, Li C, Li X. A peek into the future: Predicting the popularity of online videos. IEEE Access. 2016;4:3026–33.
[12] Mensah S, Hu C, Li X, Liu X, Zhang R. A Probabilistic Model for User Interest Propagation in Recommender Systems. IEEE Access. 2020;8:108300–9.
[13] Whiting A, Williams D. Why people use social media: a uses and gratifications approach. Qual Mark Res An Int J. 2013 Aug 30;16(4):362–9.
[14] Viviani M, Pasi G. Credibility in social media: opinions, news, and health information-a survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2017 Sep 1;7(5):e1209.
[15] David CC, San Pascual RS, Torres ES. Reliance on Facebook for news and its influence on political engagement. PLoS One. 2019 Mar 1;14(3).
[16] Polk T, Johnston MP, Evers S. Wikipedia Use in Research: Perceptions in Secondary Schools. TechTrends. 2015 May 1;59(3):92–102.
[17] Syed-Abdul S, Fernandez-Luque L, Jian WS, Li YC, Crain S, Hsu MH, et al. Misleading health-related information promoted through video-based social media: Anorexia on youtube. J Med Internet Res. 2013 Feb 13;15(2):e30.
[18] Goodman J, Carmichael F. US election 2020: "Rigged" votes, body doubles and other false claims [internet]. BBC News. 2020 [cited 2020 Oct 25]. Available from: https://www.bbc.com/news/54562611
[19] Lewandowski E, Specht H. Influence of volunteer and project characteristics on data quality of biological surveys. Conserv Biol. 2015;29(3):713–23.
[20] Redman TC. Data quality for the information age. Artech House; 1996. 303 p.
[21] Warth J, Kaiser G, Kügler M. The impact of data quality and analytical capabilities on planning performance: insights from the automotive industry. In: Wirtschaftsinformatik Proceedings. 2011.
[22] Laranjeiro N, Soydemir SN, Bernardino J. Testing web applications using poor quality data. In: Proceedings - 7th Latin-American Symposium on Dependable Computing, LADC 2016. 2016. p. 139–44.
[23] Batini C, Scannapieco M. Data quality : concepts, methodologies and techniques. Springer; 2006. 262 p.
[24] ISO. ISO/IEC 25012:2008 Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model [internet]. ISO; 2008 [cited 2019 Jan 9]. Available from: https://www.iso.org/standard/35736.html
[25] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. J Manag Inf Syst. 1996;12(4):5–34.
[26] Arolfo F, Vaisman A. Data Quality in a Big Data Context. In: ACM SIGMOD Record. Springer International Publishing; 2018. p. 159–72.
[27] Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Sci J. 2015 May 22;14(0):2.
[28] Batini C, Rula A, Scannapieco M, Viscusi G. From data quality to big data quality. J Database Manag. 2015;26(1):60–82.
[29] DAMA UK. The Six Primary Dimensions for Data Quality Assessment - Defining Data Quality Dimensions [Internet]. 2013 [cited 2019 Jan 9]. Available from: https://www.dqglobal.com/wp-content/uploads/2013/11/DAMA-UK-DQ-Dimensions-White-Paper-R37-1.pdf
[30] Immonen A, Pääkkönen P, Ovaska E. Evaluating the Quality of Social Media Data in Big Data Architecture. IEEE Access. 2015;3:2028–43.
[31] Bayraktarov E, Ehmke G, O'Connor J, Burns EL, Nguyen HA, McRae L, et al. Do big unstructured biodiversity data mean more knowledge? Front Ecol Evol. 2019;7(JAN).
[32] Sadiq S, Indulska M. Open data: Quality over quantity. Int J Inf Manage. 2017;37(3):150–4.
[33] Hall C. Valve fought more than 40 'review bombs' on Steam in 2019 - Polygon [Internet]. Polygon. 2020 [cited 2020 Oct 12]. Available from: https://www.polygon.com/2020/2/6/21126787/steam-review-bombs-policy-effectiveness-valve
[34] Kuchera B. The anatomy of a review bombing campaign - Polygon [Internet]. Polygon. 2017 [cited 2020 Oct 12]. Available from: https://www.polygon.com/2017/10/4/16418832/pubg-firewatch-steam-review-bomb
[35] Hawkins J. Yelp vs Google: How they deal with fake reviews [internet]. 2018 [cited 2020 Oct 12]. Available from: https://searchengineland.com/yelp-vs-google-how-do-they-deal-with-fake-reviews-307332
[36] The Guardian. How TripAdvisor changed travel [internet]. 2018 [cited 2020 Oct 12]. Available from: https://www.theguardian.com/news/2018/aug/17/how-tripadvisor-changed-travel
[37] Batini C, Scannapieco M. Data and Information Quality: Dimensions, Principles and Techniques. Cham: Springer International Publishing; 2016. 500 p. (Data-Centric Systems and Applications).
[38] Vincent N, Johnson I, Sheehan P, Hecht B. Measuring the Importance of User-Generated Content to Search Engines. Vol. 13, Proceedings of the International AAAI Conference on Web and Social Media. 2019 Jul.
[39] Brunt CS, King AS, King JT. The influence of user-generated content on video game demand. J Cult Econ. 2020 Mar 1;44(1):35–56.
[40] Strong DM, Lee YW, Wang RY. Data quality in context. Commun ACM. 1997;40(5):103–10.
[41] Moraga C, Moraga MÁ, Calero C, Caro A. SQuaRE-aligned data quality model for web portals. In: Proceedings - International Conference on Quality Software. 2009. p. 117–22.
[42] Redman TC, Fox C, Levitin A. Data and data quality. Understanding Information Retrieval Systems: Management, Types, and Standards. 2011. 269–284 p.
[43] Cichy C, Rass S. An overview of data quality frameworks. IEEE Access. 2019;7:24634–48.
[44] Lin S, Gao J, Koronios A, Chanana V. Developing a data quality framework for asset management in engineering organisations. Int J Inf Qual. 2007;1(1):100–26.
[45] Bordogna G, Carrara P, Criscuolo L, Pepe M, Rampini A. On predicting and improving the quality of Volunteer Geographic Information projects. Vol. 9, International Journal of Digital Earth. Taylor & Francis; 2016. p. 134–55.
[46] Alabri A, Hunter J. Enhancing the quality and trust of citizen science data. In: Proceedings - 2010 6th IEEE International Conference on e-Science, eScience 2010. IEEE; 2010. p. 81–8.
[47] Lee D. Big Data Quality Assurance Through Data Traceability: A Case Study of the National Standard Reference Data

Program of Korea. IEEE Access. 2019;7:36294–9.
[48] Kaur J, Singh J, Sehra SS, Rai HS. Systematic literature review of data quality within openstreetmap. In: Proceedings - 2017 International Conference on Next Generation Computing and Information Systems, ICNGCIS 2017. 2018. p. 159–63.
[49] Chin MJ, Babashamsi P, Yusoff NIM. A comparative study of monitoring methods in sustainable pavement management system. In: IOP Conference Series: Materials Science and Engineering. 2019.
[50] Xiong J, Chen X, Tian Y, Ma R, Chen L, Yao Z. MAIM: A Novel Incentive Mechanism Based on Multi-Attribute User Selection in Mobile Crowdsensing. IEEE Access. 2018;6:65384–96.
[51] Wei X, Wang Y, Tan J, Gao S. Data Quality Aware Task Allocation with Budget Constraint in Mobile Crowdsensing. IEEE Access. 2018 Aug 30;6:48010–20.
[52] Pang L, Li G, Yao X, Lai Y. An Incentive Mechanism Based on a Bayesian Game for Spatial Crowdsourcing. IEEE Access. 2019;7:14340–52.
[53] Pipino LL, Lee YW, Wang RY. Data Quality Assessment. Commun ACM. 2002;45(4):211–8.
[54] Smith M, Szongott C, Henne B, Von Voigt G. Big data privacy issues in public social media. IEEE Int Conf Digit Ecosyst Technol. 2012;
[55] Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manage. 2015 Apr 1;35(2):137–44.
[56] Chen M, Mao S, Liu Y. Big data: A survey. In: Mobile Networks and Applications. Kluwer Academic Publishers; 2014. p. 171–209.
[57] Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: A survey. IEEE Trans Knowl Data Eng. 2007 Jan;19(1):1–16.
[58] Blake R, Mangiameli P. The effects and interactions of data quality and problem complexity on classification. J Data Inf Qual. 2011;2(2).
[59] Heinrich B, Klier M, Schiller A, Wagner G. Assessing data quality – A probability-based metric for semantic consistency. Decis Support Syst. 2018;110:95–106.
[60] Firmani D, Mecella M, Scannapieco M, Batini C. On the Meaningfulness of "Big Data Quality" (Invited Paper). Data Sci Eng. 2016;1(1):6–20.
[61] Reiss J, Sprenger J. Scientific Objectivity. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. Winter 201. Metaphysics Research Lab, Stanford University; 2017.
[62] Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soc physics, Dokl. 1965;10:707–10.
[63] Krejcie R V, Morgan DW. Determining Sample Size for Research Activities. Educ Psychol Meas. 1970;30(3):607–10.
[64] Hulitt E, Vaughn RB. Information system security compliance to FISMA standard: A quantitative measure. Telecommun Syst. 2010;45(2–3):139–52.
[65] Senarath A, Grobler M, Arachchilage NAG. A model for system developers to measure the privacy risk of data. In: HICSS. 2019.
[66] Yang F, Feng J, Fabbrizio G Di. A Data Driven Approach to Relevancy Recognition for Contextual Question Answering [Internet]. 2006 [cited 2020 May 15]. Available from: http://www.ask.com/
[67] Cross I, Joana P. Evaluating the Usability of Aggregated Datasets in the GIS4EU Project [Internet]. 2010 [cited 2020 May 15]. Available from: https://www.directionsmag.com/article/2130
[68] Wrabetz J. Measuring the economic value of data [internet]. Network World. 2017 [cited 2020 May 17]. Available from: https://www.networkworld.com/article/3221387/measuring-the-economic-value-of-data.html
[69] Gallaher SM. An Approach For Measuring The Confidentiality Of Data Assured By The Confidentiality Of Information Security Systems In Healthcare Organizations [Internet]. University of Central Florida; 2012 [cited 2020 May 17]. Available from: http://library.ucf.edu
[70] Yao MX. Granularity measures and complexity measures of partition-based granular structures. Knowledge-Based Syst. 2019 Jan 1;163:885–97.
[71] Lu X, Horn AL, Su J, Jiang J. A Universal Measure for Network Traceability. Omega (United Kingdom). 2019 Sep 1;87:191–204.
[72] Cheney J, Chiticariu L, Tan W-C. Provenance in Databases: Why, How, and Where. Found Trends R Databases. 2009;1(4).
[73] Dexun J, Peijun M, Xiaohong S, Tiantian W. Functional Over-Related Classes Bad Smell Detection and Refactoring Suggestions. Int J Softw Eng Appl. 2014 Mar 31;5(2):29–47.
[74] Gunning R. The technique of clear writing. Toronto: McGraw-Hill; 1952.
[75] Musto J, Dahanayake A. Improving Data Quality, Privacy and Provenance in Citizen Science Applications. In: Frontiers in Artificial Intelligence and Applications. IOS Press; 2020. p. 141–60.
[76] Musto J, Dahanayake A. An Approach to Improve the Quality of User-Generated Content of Citizen Science Platforms. ISPRS Int J Geo-Information. 2021 Jun 25;10(7):434.
[77] Teorey T, Lightstone S, Nadeau T, Jagadish HV. Business Intelligence. In: Database Modeling and Design. Elsevier; 2011. p. 189–231.
[78] Fox TL, Guynes CS, Prybutok VR, Windsor J. Maintaining Quality in Information Systems. J Comput Inf Syst. 1999;40(1):76–80.

# Publication IV

Musto, J., and Dahanayake, A.
**An approach to improve the quality of user-generated content of citizen science platforms**

Reprinted from
*ISPRS International Journal of Geo-Information*
Vol. 10, pp. 434, 2021

*Article*

# An Approach to Improve the Quality of User-Generated Content of Citizen Science Platforms

**Jiri Musto [1,*] and Ajantha Dahanayake[1]**

— **Citation:** Musto, J.; Dahanayake, A. An Approach to Improve the Quality of User-Generated Content of Citizen Science Platforms. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 434. https://doi.org/10.3390/ijgi10070434

Academic Editors: Gavin McArdle, Biana Schoen-Phelan and Wolfgang Kainz

Received: 28 April 2021
Accepted: 22 June 2021
Published: 25 June 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



[1] Lappeenranta-Lahti University of Technology LUT; firstname.lastname@lut.fi
* Correspondence: jiri.musto@lut.fi

**Abstract:** The quality of the user-generated content of citizen science platforms is discussed widely among researchers. Content is categorized into data and information: data is content stored in a database of a citizen science platform, while information is context-dependent content generated by users. Understanding data and information quality characteristics and utilizing them during design improves citizen science platforms' overall quality. This research investigates the integration of data and information quality characteristics into a citizen science platform for collecting information from the general public with no scientific training in the area where content is collected. The primary goal is to provide a framework for selecting and integrating data and information quality characteristics into the design for improving the content quality on platforms. The design and implementation of a citizen science platform that collects walking path conditions are presented, and the resulting implication is evaluated. The results show that the platform's content quality can be improved by introducing quality characteristics during the design stage of the citizen science platform.

**Keywords:** information quality; data quality; citizen science; user-generated content; characteristics of data and information; empirical study

## 1. Introduction

Citizen science projects have garnered a wide following and usage over the years. In citizen science, non-professional participants contribute to a project by submitting new information or classifying existing data. In the modern digital world, this takes place through an online platform. A notable example of a citizen science project for collecting new information is eBird [1], which has been in operation since 2002, while a notable project for classifying existing data is Galaxy Zoo [2]. To give a more recent example, a project that emerged in 2020 is currently implementing a citizen science classification game within a video game called Borderlands

3; it allows the players of Borderlands 3 to participate in the citizen science project by playing a microbiome mapping game within the actual video game [3].

A wide range of terms have been used to refer to the information provided by citizens, such as volunteered geographic information, participatory sensing, citizen science, and crowdsourcing [4]. This research will focus on using the term "citizen science."

Citizen science projects have a wide variety of use-cases such as astronomy [5], wildlife research [6], flood monitoring [7], smart cities [8], and many more. There are over a thousand different ongoing projects listed in SciStarter [9]. The major challenge and complaint in citizen science and similar domains is the data quality [6,10–20]. As participants who contribute content have no scientific training, there is a risk that they will produce low-quality data [13]. Some researchers have argued, however, that there is no statistically significant difference in quality between content provided by citizens and experts [17], although this is still a minority opinion. Indeed, in [21], over a hundred different citizen science projects have been investigated, and several issues in data quality management have been identified.

Data quality is essential when data is used for research purposes, to form decisions, or to establish facts [22–24]. Data quality is defined as multidimensional (divided into characteristics), and this applies in multiple data quality models with different characteristics and varying definitions [25–27]. Most data quality definitions apply to general data quality. However, data quality depends on the scenario [25,26,28–31], and each domain requires different definitions. Therefore, it can be an extremely limiting factor to demand similar data quality standards from regular citizens as are requested of professionals within organizations [10,32]. The International Organization for Standardization (ISO) has established some standards for data quality for different domains [33–35].

Another issue that needs to be considered when discussing data quality is information quality. Information and data are two separate terms. Information is more contextual than data, but data can be transformed into information when given a context, category, or analysis [36,37]. Data is a separate object that can be quantified, but information requires external knowledge or perception [28]. For example, a list of temperatures is data, but if the temperatures are tied to dates, they are information about the upcoming temperatures. Likewise, a list of dates and locations by itself is data, but if there is the knowledge that the dates and locations indicate where a band has been touring, it becomes information. This distinction between data and information is crucial for improving data quality in citizen science.

In citizen science, citizens provide information to the platform because the citizens are asked to provide specific information. Information received from citizens is mostly unfiltered, and this is affected by what information a citizen can submit. The provided information is affected by the content collection approach, and when the information is stored as data, it undergoes a data curation process that influences its quality. Information quality is the quality of the information received from citizens, while data quality is the quality of content within the database. Data quality is primarily affected by information quality.

Data quality improvement methodologies can be classified into two strategies, *data-driven* and *process-driven.* The *data-driven* approach aims to improve data quality by replacing low-quality data, correcting errors, or selecting credible sources. *Process-driven techniques* mainly *control* or *redesign the collection process* to remove the causes of low-quality data and introduce methods that produce higher-quality data [38]. Thus, *process-driven* methods can be considered to improve the quality of information. In citizen science platforms, *process-driven* methods and correcting errors are feasible strategies to improve quality. Platforms automatically correct minor errors and more significant errors are fixed by the community or moderators [1,17,39,40]. Selecting volunteers before they provide any content or replacing low-quality data are not

recommended methods because holding citizens to professional researchers' standards leads to issues with data acquisition. With greater data requirements, participants with little expertise in the area can feel discouraged, thus reducing the number of contributions. Additionally, citizens observe their surroundings differently and identify phenomena outside the intended purpose, thus providing valuable data [10]. Selecting credible volunteers is possible after content has been collected and the credibility of volunteers has been adequately evaluated [41].

There are general guidelines and methodologies for incorporating data quality into the platform or design process [42–45]. However, the existing guidelines mainly indicate a step in a process where data quality should be considered without explaining how to implement methods for improving data quality concretely. Higgins et al. [45] present a generic infrastructure for citizen science platforms and recommend keeping quality elements as part of the data. As a result, there is no clear understanding of an approach to date that can increase citizen science platforms' content quality.

This research therefore aims to improve the quality of data collected and the quality of the information provided by citizens by integrating quality characteristics into the design of a citizen science platform, as suggested in [21,46]. Incorporating quality characteristics as requirements in creating a platform helps to monitor quality and ensure that the desired quality criteria have been met before any data or information is gathered.

The primary approach to increasing the quality of collected information is using the quality characteristics as prerequisites during the collection process and creating quality controls for the user interface to reduce the amount of low-quality content, as well as integrating quality elements into the data model. The proposed method is evaluated by looking at a platform that collects walking path observations from citizens. The number of participants involved in citizen science projects varies, and the amount of data accumulated may be low, making the quality of data more essential.

This research contributes to the study of improved content quality in citizen science platforms. It aims to improve the content quality by considering data and information quality characteristics during the design and development. This way, the data quality could be enhanced during data collection, and less work is needed for cleaning before analyzing data.

## 2. Quality of citizen science data and information

In 1996, Wang and Strong [25] defined 15 essential quality characteristics for the business domain and divided them into four categories. Redman [27], meanwhile, divides data quality into 12 characteristics and four categories. Information quality can be considered to be similar to data quality, but there are significant differences. Data consists of raw facts, while information is derived from data in a given context. Therefore, information quality can only be determined within a given context, while data quality can be context-independent [28,36,37].

ISO has developed a number of data quality standards for different use cases. ISO 19157 [33] is a data quality standard for geographic information that focuses on positional and thematic accuracy, completeness, and temporal correctness. The applicability of ISO 19157 to volunteered geographic information has been investigated in [47]. While many of the standard's components are suitable, the standard is not perfect because of the difference between authoritative and volunteered geographic information. Volunteered geographic information is heterogenic and lacks data specifications and comparable reference data that are required for the standard. Additional quality indicators, such as reliability, should thus be used to extend the standard. ISO 19157 is not an appropriate standard for citizen science because citizen science platforms collect other information besides just geographical information.

The data quality model ISO 25012 [34] for structured data in computer systems is a more appropriate starting point for citizen science platforms than ISO 19157. The data model divides data quality into 15 characteristics and three categories. Table 1 presents the comparison of three different collections of data quality characteristics and their applicability to information quality based on context-dependency.

The first column in Table 1 determines whether the listed characteristic can be applied to information quality or instead lacks context-dependency. Some characteristics proposed by Wang and Strong can be grouped under one characteristic comparable to ISO or Redman. For example, "interpretability" and "ease of understanding" can be grouped to form "understandability". Similarly, "believability" and "reputation" can be grouped to form "credibility". Some of the characteristics are intended for the underlying system and are therefore not applicable for information quality. Few characteristics can be argued to be or not be suitable for information. These depend on the context of the information, and where and how it is provided.

**Table 1.** Data quality characteristic comparison

| **Applicable for information quality** | **Wang & Strong** [25] | **ISO 25012** [34] | **Redman** [27] |
|---|---|---|---|
| Yes | believability | credibility | |
| Yes | accuracy | accuracy (syntactic and semantic) | accuracy |
| Yes | objectivity | | |
| Yes | reputation | | |
| Yes | relevancy | | appropriateness |
| Yes | value-added | | |
| Yes | timeliness | currentness | currency |
| Yes | completeness | completeness | completeness |
| Yes | an appropriate amount of data | | |
| Yes | interpretability | | interpretability |
| Yes | ease of understanding | understandability | |
| Yes/No | representational consistency | | representation consistency |
| Yes/No | concise representation | | |
| No | accessibility | accessibility | |
| Yes | access security | confidentiality | |
| No | | efficiency | efficient use of memory |
| No | | compliance | |
| Yes/No | | precision | format precision |
| Yes | | traceability | |
| No | | availability | |

| | | |
|---|---|---|
| No | portability | portability |
| No | recoverability | |
| No | consistency | consistency |
| No | | ability to represent null values |
| No | | format flexibility |

*Precision*, *representational consistency*, and *concise representation* rely on the context of information. Concise representation can belong to understandability, but precision and representational consistency are affected by how the information is shown. Representational consistency helps improve understandability, but it may be necessary to have different information representations. Depending on the usage, precision can have a different meaning. (*Format*) *precision* means there are a sufficient number of decimals in a given number. *Locational precision* pertains to how exact the given location information is. For example, geocoordinates are more precise than a street name, and a city is more precise than a country. Precision in location information is not to be confused with *locational accuracy*, which pertains to whether the information is close to the proper location or not. Some citizen science platforms require geocoordinates (high precision), but other platforms settle for knowing the street of the observation or a 5km radius (low precision).

Based on Table 1, the final list of characteristics for information quality is presented in Table 2. *Accuracy* is divided into *syntactic* and *semantic accuracy*. Syntactic accuracy pertains to whether the information is of the correct type and syntax. If the information is textual, syntactic accuracy can be extended to, for example, language. Semantic accuracy means that information is logical and follows semantic consistency. For example, Finland is part of Europe, and it would be semantically wrong to say Finland is part of Asia. In ISO 19157, accuracy is tied to thematic and positional accuracy, and these definitions of accuracy can be used in some cases. However, syntactic and semantic accuracy can be used for locational information. Syntactic accuracy means that the location is provided in the correct format (street name, postal code, latitude, longitude), and semantic accuracy means that the information is semantically solid. Accuracy does not consider how detailed the positional information is; for that, precision can be used.

**Table 2**. Chosen characteristics and their definitions

| Final list | Definition |
|---|---|
| credibility | The source of information is credible. Does not consider the credibility of the information. |
| accuracy (syntactic and semantic) | Information is syntactically and semantically correct, i.e. of correct type and logical. |
| objectivity | Information is objective and not affected by the source's opinions or biases. |
| relevancy | Information is relevant for the topic. |
| value | Information is valuable, provides new insights, benefits. |
| currentness | Information is as recent as possible. |
| completeness | Information is complete and not missing important details. |
| volume | Multiple sources provide similar information. |
| understandability | Information is easy to understand. |

| | |
|---|---|
| privacy (access security) | Information source's privacy is protected i.e. personal information of a citizen is protected. |
| traceability | Information origin can be traced to a user, time and location. |

**3. Case study: Information and data quality in citizen science platforms**

Numerous researchers have identified problematic issues with data and information quality in citizen science platforms. The following features are standard amongst most citizen science platforms:

- Inexperienced users or content providers submit information
- Content can be reviewed by community/moderators
- Content can be text and multimedia
- Content is not only freeform text
- Location and time are part of the information
- Information can be precise or general

Data and information quality should be evaluated based on individual characteristics. The quality of the information received from citizens is vital as it affects the quality of data. Therefore, each platform's user interface needs to be examined from the content provider's perspective to evaluate the information provided by citizens. Unfortunately, it is difficult to accurately assess information quality directly from the user interface.

Four platforms have been selected to evaluate the quality of data and information in citizen science based on the following criteria:

- Platforms collect different observations.
- Are still collecting observations.
- Provide access to data.
- Have a large quantity of data.

iNaturalist [39] is a global network of websites that operates in multiple countries. Each country uses the iNaturalist platform template to collect environmental observations from citizens and connect with existing platforms within the operating country. For example, Laji.fi is integrated with iNaturalist in Finland. A dataset related to the great tit (Parus major) bird is downloaded, and it consists of 39910 entities.

Atlas of Living Australia (ALA) [48] is an Australian environment and wildlife data collection platform. Data is collected in two ways, either from other organizations such as environmental research facilities or from citizens. Citizens can submit data directly using the ALA website or through the citizen science platform iNaturalist Australia that has been integrated with ALA. The lace monitor (Varanus varius) lizard dataset is downloaded, and it consists of 14138 entities.

Globe at Night [49] is an international citizen science project that gathers night sky brightness information from citizens. The project has been ongoing since 2006, and its main base of operations is in the United States. Data from 2020 is downloaded, and the dataset consists of 29507 entities.

Budburst [50] is a citizen science project managed by the Chicago Botanic Garden. The project focuses on observing plants and pollinators within the United States. The whole available dataset is downloaded from the website, and it consists of 96815 entities.

Each of the datasets has location information, but they have different levels of precision. For example, some have exact latitude and longitude coordinates, while others provide a location at only the city or country level.

### *3.1 Information quality evaluation through the user interface*

The quality of information submitted by the citizen is examined through the user interface within the platforms.

Globe at Night offers a generic report interface. Users can set the exact observation time and location or provide the location at the country level. Users are presented with options for the information to be sent, such as different sky conditions and darkness, and exact details are optional. The most challenging aspects of the Globe at Night interface are that there is no option to record who submitted the information and there are only a handful of correctness checks. For example, the location can be set to Africa, and then users can select the United States as the country. Issues like this risk reducing the accuracy and completeness of the information.

ALA uses the iNaturalist user interface for single observations, so ALA and iNaturalist evaluation is combined. The user interface in iNaturalist imposes some restrictions but also offers some options for the users. First of all, the species must be selected from a predetermined list and cannot be arbitrary. Similarly, the time of observation is precisely formatted and checked. Users are given the option to mask or generalize the precise location information. However, the masked location shows correctly in the downloaded data, undermining the utility of the given choice. Location information is provided through a map as coordinates, and the user can use a selection tool to cover a wider area. The location is automatically matched to a specific city and country based on the coordinates, but the user can overwrite this. iNaturalist has a list of quality criteria that are automatically checked for each observation. This list includes items such as the following: has location, picture or sound included, correct place, correct date, the community has identified, and so on. Information pertaining to over half of the items requires the community to form a consensus before it can be verified.

Budburst offers a predetermined list of species that the user selects from. Users can add new species to the list that are somehow moderated. The location information in Budburst can be at the state level as well as in the form of a precise location. Users have limited freeform input when submitting information and each input field has syntax checks, increasing the syntactic accuracy drastically. Each field is also mandatory aside from optional comments, making the completeness high. Many of the information fields can be approximated, so the user does not have to be an expert or be constantly checking what time it is. The biggest downside in Budburst is that users are always anonymized. None of the observations can be tied to a specific user, so it is impossible to determine whether several observations come from one or several different users.

Each of the user interfaces of the platforms has its own positive and negative traits. Overall, iNaturalist has the best user interface regarding information quality. It uses automatic checks related to the information, which increases the accuracy and value of the content. The most significant benefit for iNaturalist is that it provides user information about each observation, which helps to assess the credibility of every user who submits information. Considering the extensiveness of input fields in reports, Budburst has the most user-friendly and accessible user interface when submitting and viewing information.

### *3.2 Data quality evaluation*

Data quality is evaluated using a dataset from each platform. The evaluation is conducted using the characteristics presented in Table 2. The data quality is evaluated by subjecting each dataset to specific queries related to each quality characteristic using the data analytics platform RapidMiner (https://rapidminer.com/). RapidMiner is a commercial data science platform meant for data mining, analytics, and machine learning.

Table 3 presents the general queries formulation method for RapidMiner usage for each of the characteristics. Value as a quality characteristic is inherently subjective, and so everyone has their own opinion on the value of data. For the sake of simplicity, value in this research is evaluated with the help of other characteristics, but this is by no means the only way to assess the value.

**Table 3**. RapidMiner queries for each characteristic

| # | Characteristic | (Data mining) Technique |
|---|---|---|
| 1 | Syntactic accuracy | Compare values to the expected input and format. Based on the most changing attribute. |
| 2 | Semantic accuracy | Compare values if they are semantically correct based on what is expected. |
| 3 | Completeness | Compare missing values to the total amount of values |
| 4 | Credibility | User reputation if available. |
| 5 | Objectivity | Count how many entities from different sources/content providers have the same information and how many are only from singular content providers/sources. |
| 6 | Volume | Count how many entities from different content providers have relatable information based on selected attributes. Unlike objectivity, the information does not have to be the same, but there must be some similarities, such as location. |
| 7 | Currentness | Given date is later than 31.12.2010. |
| 8 | Privacy | Filter out content providers whose possible real names are given and compare them to the total amount (text mining). |
| 9 | Relevancy | Data comparison to given relevance factor such as the topic. By default, everything is relevant. |
| 10 | Usability | If the content is missing from essential attributes (location or time), deemed unusable. |
| 11 | Value | Calculation based on other characteristics, (Syntactic + Semantic + Credibility + Relevancy + Usability + Understandability) / 6. |
| 12 | Traceability | Count how many entities have a valid time, location, and content provider/source compared to all entities. |
| 13 | Understandability | Text mining of invalid words in specific attributes. |

The following list presents specific adjustments for the general data mining queries:

- Syntactic accuracy:
    - ALA: Based on the *verbatim date* attribute and expected syntax of yyyy/MM/dd hh:mm
    - Globe at Night: No issues in syntactic accuracy.
    - BudBurst: Based on the *country* attribute. The expected syntax is the acronym of a country (*US*), which means, for example, *United States* is incorrect.
    - iNaturalist: Based on the *timezone* attribute. The majority of the values are the country locations, and the minority of values are *UTC* or *Eastern Time*.
- Semantic accuracy:

    - ALA: *Sex* attribute is inspected to determine whether values are *male, female,* or *unknown.*
    - Globe at Night: No semantic issues.
    - BudBurst: No semantic issues.
    - iNaturalist: Timezone is compared to the collection location.
- Credibility:
    - ALA: NA
    - Globe at Night: NA
    - BudBurst: NA
    - iNaturalist: NA
- Objectivity:
    - ALA: The location similarity is tied to the city/county level.
    - Globe at Night: NA
    - BudBurst: NA
    - iNaturalist: Dataset includes an *agreement* attribute that reflects how many other users agree on the observation.
- Volume:
    - Globe at Night: NA
    - BudBurst: NA
- Privacy:
    - Globe at Night: No personal information in data.
    - BudBurst: No personal information in data.
- Traceability
    - Globe at Night: Missing user identification, so traceability is reduced in the calculations.
    - BudBurst: Missing user identification, so traceability is reduced in the calculations.
- Understandability:
    - ALA: Understandability is measured through the *Locality* attribute and mining incomprehensible texts or locations that do not make sense.
    - Globe at Night: *Sky_comment* attribute is mined for non-English texts and incomprehensible values.
    - BudBurst: *location_title* attribute is used to measure understandability by mining incomprehensible texts or locations that do not make sense.
    - iNaturalist: *place_guess* attribute is mined for invalid words to measure understandability.

Table 4 presents the RapidMiner query results as values between 0 and 1, reflecting the percentage of results. Values indicate the percentage of correct data entities for each characteristic (conform to the given query). For example, if a dataset consists of 40 000 entities and 5000 entities are missing a location, the traceability would be 0.96 (3 x 40 000 – 5000 divided by 3 x 40 000). Not applicable (NA) results are deemed as zero when evaluating *value* because if something is not applicable, it does not exist.

**Table 4**. RapidMiner query results

| Characteristic | ALA | iNaturalist | Globe at Night | Budburst |
|---|---|---|---|---|
| Syntactic accuracy | 0.71 | 0.89 | 1.00 | 0.99 |
| Semantic accuracy | 0.80 | 0.90 | 1.00 | 1.00 |
| Completeness | 0.71 | 0.73 | 0.87 | 0.33 |
| Credibility | NA | NA | NA | NA |
| Objectivity | 0.29 | 0.56 | NA | NA |
| Volume | 0.70 | 0.73 | NA | NA |
| Currentness | 0.44 | 0.99 | 1.00 | 0.80 |
| Privacy | 0.80 | 0.98 | 1.00 | 1.00 |
| Relevancy | 1.00 | 1.00 | 1.00 | 1.00 |
| Usability | 0.96 | 0.83 | 1.00 | 0.87 |
| Value | 0.74 | 0.72 | 0.78 | 0.79 |
| Traceability | 0.91 | 0.90 | 0.86 | 0.70 |
| Understandability | 0.97 | 0.69 | 0.65 | 0.86 |

Budburst and Globe at Night do not provide any user information in the dataset, so there is no way to tie any observation to a specific provider. This means that there is no information on how many users provide information and how many come from the same provider. This undermines multiple characteristics, most notably objectivity, credibility, and traceability. On the other hand, the lack of any identifying information increases privacy.

Out of all the datasets, Budburst has the lowest completeness. Many attributes are left empty in the dataset. This reflects the nature of the user interface, which allows users to provide much optional information, such as different pollinators. A better method of storing this information would be to have some default values for each attribute rather than leaving them empty.

While ALA and iNaturalist use the same interface to collect singular observations, ALA also contains datasets to add to their collection via their interface. This difference is reflected in the data extracted from each platform. ALA has more attributes in the dataset than iNaturalist, and each has different names for the same attributes. These differences demonstrate the comparative accuracy and currentness of the data from each dataset. Some data in the ALA dataset is sourced from the 20$^{th}$ century, while all data in iNaturlist has been provided after 2010.

Globe at Night has the fewest attributes in the dataset. This means that there is only a minimal amount of extra information within the data, making its completeness the best out of the tested datasets. On the other hand, the comment attributes include text in multiple languages, which reduces the understandability considerably. Some of the comments are also short acronyms that, without any context, are difficult to understand.

## 4. Integration Of Quality Characteristics Into The Citizen Science Platform: WalkingPaths

The idea of increasing data and information quality in citizen science platforms by integrating quality characteristics into the model is presented in [21] and [46]. [46] provides general guidelines for increasing data quality by not allowing insufficient quality data into the system and [21] provides the initial design for integrating data quality characteristics into the design of a citizen science platform via specific checks or attaching the quality characteristic into the data model. To test and evaluate this idea, a citizen science web platform called WalkingPaths is developed. The platform is developed using ReactJS for the frontend and NodeJS for the

backend with an NoSQL database, MongoDB. Mongoose middleware is used to enforce syntax restrictions on data. The platform integrates the information quality characteristics listed in Table 2 into the platform's design.

The platform collects walking path information from citizens in Finland. Citizens are asked to fill out a simple form consisting of the path's location and condition, and they are given an option to send an image in addition to the observation. The data is collected from March 2020 to September 2020, and the final dataset consists of 108 observations.

*4.1 Platform design*

The quality of information from the content provider depends on the user interface. When integrating quality characteristics into the design, this fact is crucial to consider. Having proper checks and limitations in the user interface will increase the quality of the information received from content providers, increasing the overall data quality within the system. For example, location can be considered complete if a valid address or geolocation is given. The rules for limiting content within the user interface can be received from the database. The data model can require specific data types, and the user interface can limit the possibilities based on these restrictions. If illegal data types are given, the information is not stored, and citizens are asked to modify it.

Data and information quality characteristics can be divided into four categories based on their implementation:

Before collection: Characteristics that should be implemented before collecting information from content providers. These should be integrated into the data model and backend.

- Syntactic accuracy: Within the data model, the syntax of each data is defined. Depending on the chosen database, the syntax is automatically enforced or manually enforced via the backend (NoSQL). In this research, the syntax is evaluated using the middleware Mongoose.
- Semantic accuracy: Semantic accuracy rules come from an expected value. When requiring a date, it is expected to receive a valid date. Semantic rules for content come from the database and can be enforced and checked in the backend or user interface.

During collection: Characteristics that should be implemented during the collection of content. These should be integrated into the user interface.

- Syntactic accuracy: Syntactic accuracy during collection can be enforced by making type checks in the user interface and not allowing incorrect or illegal types to be submitted.
- Semantic accuracy: Semantic accuracy can be increased during the collection by giving the content provider a selection field rather than freeform text fields. Another method is to check if the given information in the field matches specific content, such as asking a country and checking whether a given country exists in the list of countries.
- Privacy: Privacy relates to personal information, and it is up to the developers to decide whether or not to collect personal information. The easiest method of increasing privacy is not collecting personal information, especially in a citizen science platform where location is often necessary. Whenever private information is being collected, clear statements should be made on what is collected and how it is used. In addition, the user needs to be offered the opportunity to consent to their personal information being used and be given the option to delete personal information if it has been collected [51,52].
- Completeness: During collection, completeness can be ensured by not allowing content providers to submit incomplete content. There can be a variable degree of completeness.
- Traceability: Traceability requires information regarding when content is submitted and where it comes from. This information is most easily collected from the user interface when

a content provider is submitting content. For example, the date and time can be stored, and the location and content provider's name can be requested if necessary.

- Relevance: Each platform, especially in citizen science projects, has some specific use case for collecting data. For example, content providers can be restricted to providing only information relevant to the topic during the content collection.
- Credibility: Credibility is related to a content provider's credibility rather than content credibility. Content provider's credibility can be determined in various ways, but the most common method is reputation models. If the content provider has previously submitted high-quality content, then their credibility score will be higher.
- Currentness: When content is submitted and when the observation has been made can be directly taken from the user interface.

After collection: Characteristics that should be implemented after the collection of content should be integrated into the backend.

- Completeness: Completeness of given content can be checked after submission, and the provided data can be marked complete/incomplete. If it is possible to edit the content later, this value can then be updated.
- Objectivity: The objectivity of content can be based on various aspects. One aspect is the content provider and what content is submitted. If the content has an image attached, it is easier to determine objectivity. If a reputable content provider submits the content, it is most likely to be objective. Different objectivity values can be directly attached to the data. Objectivity can also be determined using a voting system in the platform.
- Volume: After content is submitted, similar content can be checked and calculated based on the similarity score. For example, content related to the same location area can be grouped to form general information found on its content.
- Value: The value of content can be determined and calculated on various conditions, and this value score can be attached to the data.
- Usability: The usability of data can be determined and calculated on various conditions, and this usability score can be attached to the data.

Presenting information: Characteristics that should be implemented when presenting the information. These should be integrated into the user interface.

- Privacy: If personal information has been collected, the extent to which this information is shared with others should be evaluated. It is unnecessary to show personal information in most cases, and thus it should be omitted from the user interface. The option to hide personal information could be added for citizens on the platform.
- Volume: Having multiple similar observations or reports in a platform must be indicated in some form. There is a significant difference between one person making a claim and ten people making the same claim. The volume of content can be presented in different ways, depending on how the content is presented in general.
- Understandability: Information should be presented understandably. For example, a list of observations and reports can be a challenging format for understanding the bigger picture, and it is therefore better to use an alternative method for presenting the information. For example, in most citizen science platforms, there is a map that shows different locations. Another approach is to show statistical analysis of specific pieces of data. Regardless of the methodology, each is implemented in the user interface.

Figure 1 shows the database schema developed using the snowflake data model [53] of the platform WalkingPaths. In the center is the fact table *WalkingPathObservation,* and it is connected to several dimension tables. Each dimension table holds a primary key and a foreign key to any

sub-tables. The fact table contains foreign keys from all linked dimension tables and has them as a combined primary key. Thus, a snowflake schema can be easily transformed into a relational data model.

The attributes in the database schema have different initials tied to them:

- PK and FK: Indicates the primary and foreign keys.
- SG: The platform generates the attribute.
- UG: The attribute is user-generated, i.e., given by the user.
- EX: The attribute is obtained from external sources.
- DQ: The attribute stores information related to data quality.

Several data quality characteristics are integrated into the model as separate attributes. These include accuracy (syntactic and semantic), completeness, volume, credibility, privacy, objectivity, and traceability. These characteristics are used for storing relevant quality evaluations and presenting them to others. For this platform, the precision of geolocation is considered necessary. The information regarding how many meters the location may be off of the given coordinates is stored in the database. The quality of data can be described, quantified, and guaranteed more easily using the data quality attributes when sharing the data with others.

*Geolocation* is separated from the address because the coordinates are not mandatory, and the location is extracted from the device directly while the user manually provides the address. *Changes* stores all modifications made by a user to any existing observation within the platform. For example, if a user modifies an observation by adding a photograph or modifying the additional information, the *prevVal* stores the previous content, and *newVal* stores the modified content. Storing the modifications made to any information helps restore "correct" values if observations have been incorrectly changed. Historical data can be used as reference data to improve the quality of new data in some scenarios. For example, if there has never been ice in July, it is unlikely to happen in the present, which means that if someone claims there is ice, the observation can be marked for further inspection and validation.

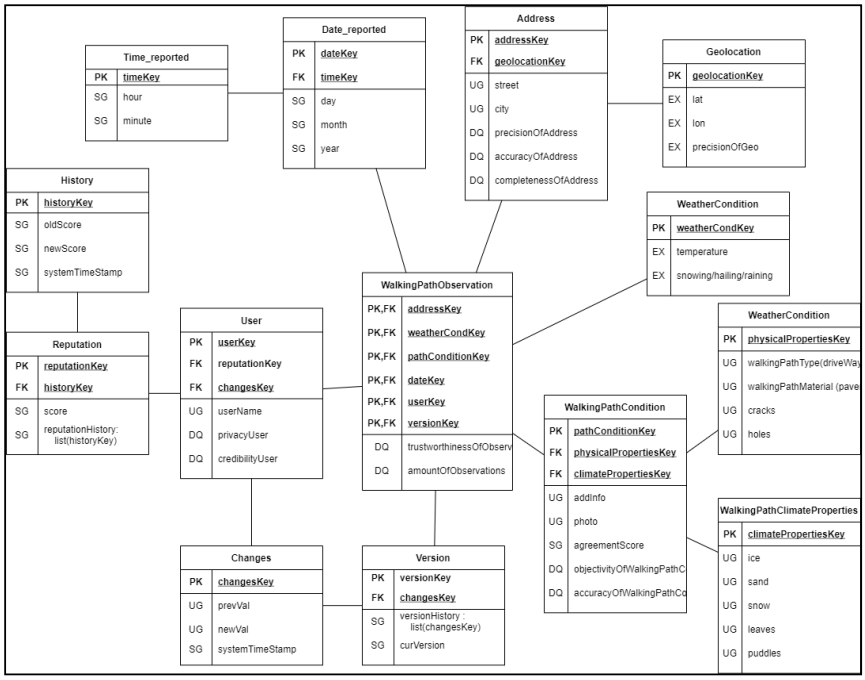


**Figure 1.** A database schema using the snowflake model for WalkingPaths

Figure 2 shows which characteristics relate to each view in the user interface. The characteristics affect how information is collected or shown in the user interface. Different characteristics are required for different views. For example, accuracy is essential in the *New Observation* -view because data is being collected, but accuracy is unnecessary in the *Observations list* -view as the data has already been collected and requires no further refinement of accuracy.

The following characteristics are integrated into the user interface: completeness, privacy, understandability, credibility, objectivity, traceability, accuracy (syntactic and semantic), volume, value, usability, and relevance.

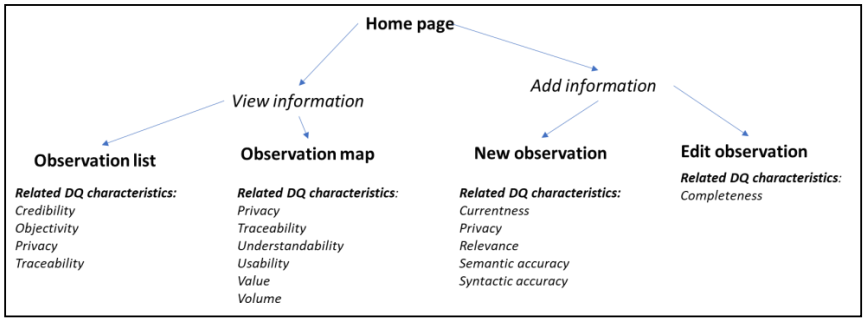


**Figure 2.** Quality characteristics in WalkingPaths views

Figures 3 and 4 show the transition using the navigation bar to each view presented in Figure 2. The observation list only notes minimal detail for each report, such as location and time.

Everything else regarding the report can be viewed by opening the *More information* pop-up window. Each report can be up-/downvoted by anyone. As the platform does not require registration, some other forms of restrictions have been implemented in the voting mechanism to reduce misuse. *Edit observation* view is similar to *New observation* except for the fact that when editing an existing observation the user must provide an ID to retrieve the current information.

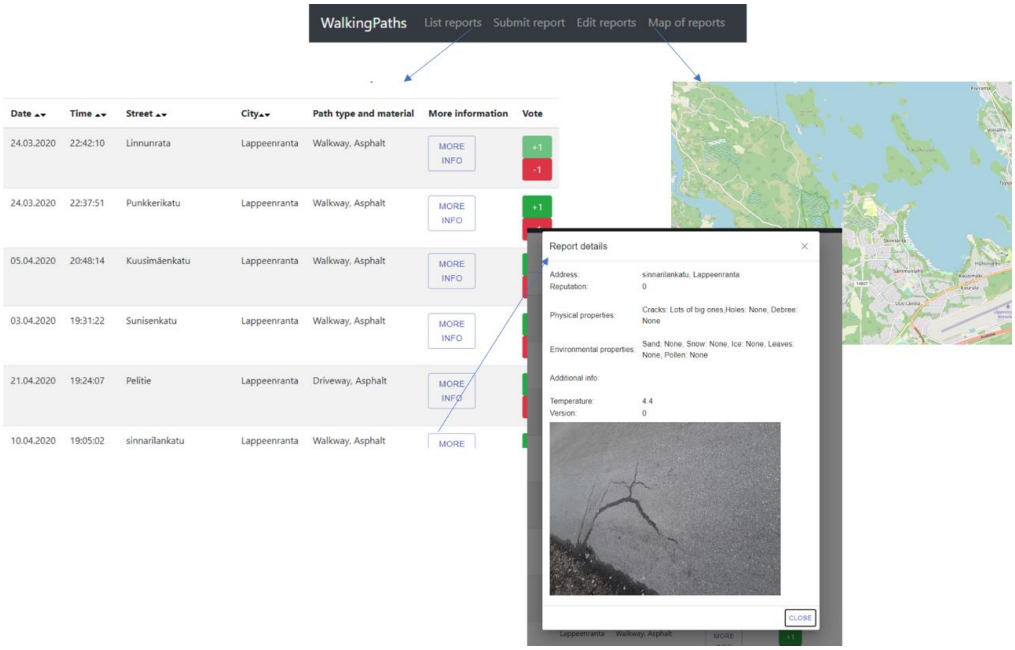


**Figure 3.** WalkingPaths observation list and map

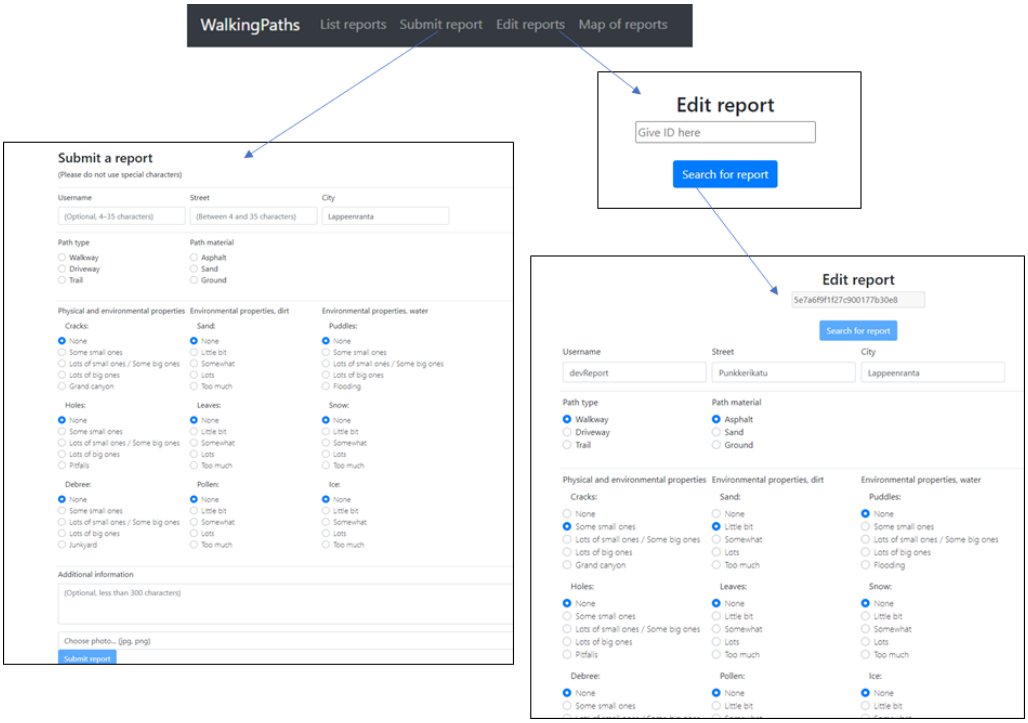


**Figure 4.** The *New observation* and *Edit observation* views of WalkingPaths

Figure 5 presents a higher-resolution view of report submission, as shown in Figure 4. Only four fields are freeform text, and two of them are mandatory. *Username* and *Additional information* are optional fields that can be left empty. Many of the choice boxes in the report window have predetermined values to ensure each report's completeness. Only two choice boxes do not have a value, but the report cannot be submitted before some value is given to both of them.

The usage of choice boxes is an excellent method for increasing the report's syntactic and semantic accuracy. They also enable the content provider to know what to look for before submitting anything. Any additional information can be written in the text box. The user is not required to give an exact location; the city and street names (optional building numbers) are sufficient. There are a few reasons why a precise location is not required:

1. The location precision of a smartphone is inconsistent. The precision varies between smartphone models and, with buildings or trees around the area, the precision decreases. This imprecision may result in placing the actual location on a different street to that which is suggested by the coordinates [54–57].
2. Use of the location requires permission from the user, and not all are willing to give consent.
3. Precise location raises privacy concerns [58].

Marking the disclosure of a precise location as optional information may hinder the exact precision of results, but it increases users' privacy. The downside of this is that if the user does not remember or know the street name, it can affect their willingness to contribute observations.

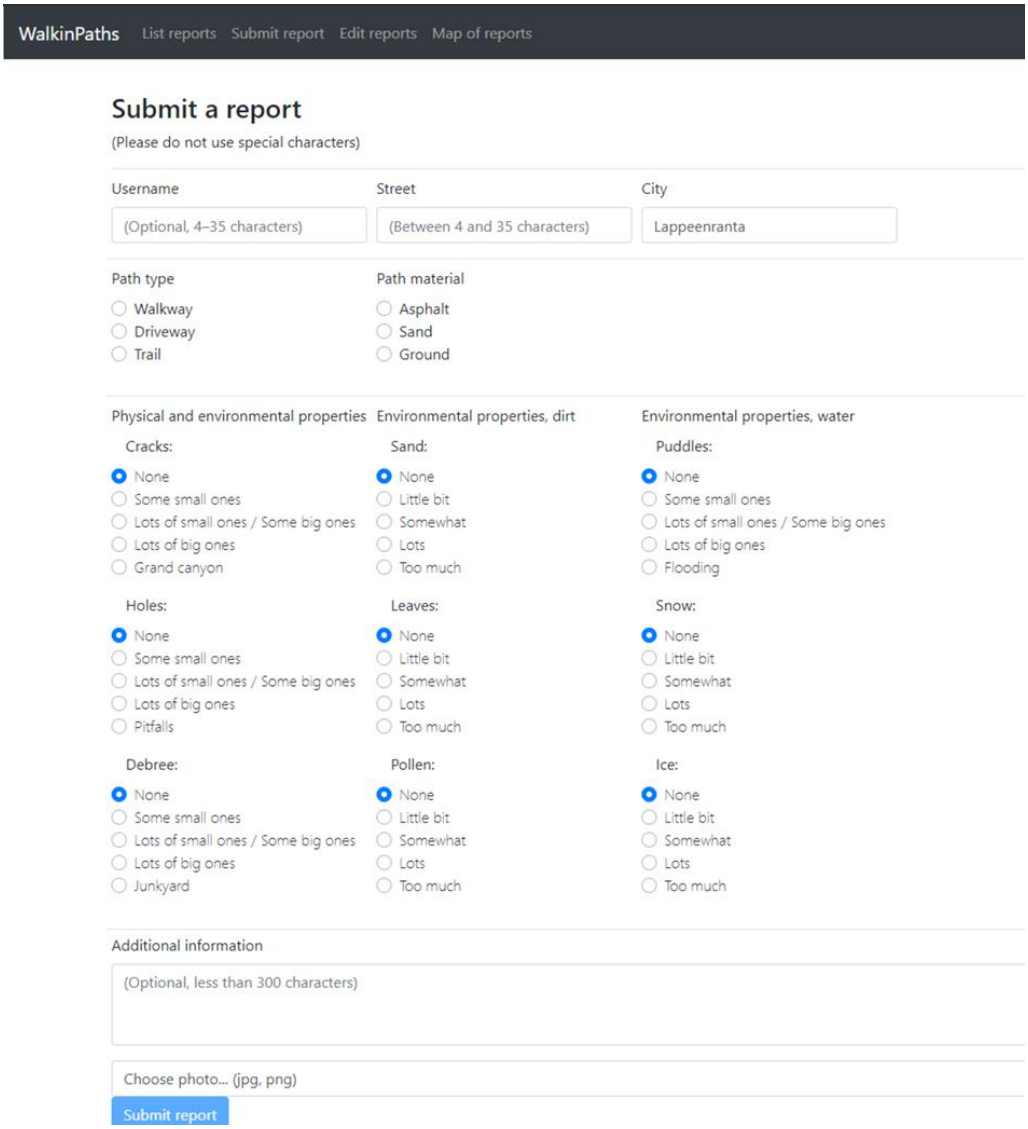


**Figure 5.** A detailed version of the new observation view

*4.2. Evaluation and analysis*

WalkingPaths is subjected to the same RapidMiner queries as other citizen science platforms. Table 5 presents the analysis results for WalkingPaths combined with the previous results from Table 4.

**Table 5**. WalkingPaths compared to other citizen science platforms

| Characteristic | WalkingPaths 108 observations | ALA 14138 observations | iNaturalist 39910 observations | Globe at Night 29507 observations | BudBurst 96 815 observations |
|---|---|---|---|---|---|
| Syntactic accuracy | 1.00 | 0.71 | 0.89 | 1.00 | 0.99 |
| Semantic accuracy | 0.96 | 0.80 | 0.90 | 1.00 | 1.00 |
| Completeness | 1.00 | 0.71 | 0.73 | 0.87 | 0.33 |
| Credibility | 0.74 | NA | NA | NA | NA |
| Objectivity | 0.54 | 0.29 | 0.56 | NA | NA |
| Volume | 0.36 | 0.70 | 0.73 | NA | NA |
| Currentness | 1.00 | 0.44 | 0.99 | 1.00 | 0.80 |
| Privacy | 1.00 | 0.80 | 0.98 | 1.00 | 1.00 |
| Relevancy | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Usability | 1.00 | 0.96 | 0.83 | 1.00 | 0.87 |
| Value | 0.95 | 0.74 | 0.72 | 0.78 | 0.79 |
| Traceability | 1.00 | 0.91 | 0.90 | 0.86 | 0.70 |
| Understandability | 1.00 | 0.97 | 0.69 | 0.65 | 0.86 |

WalkingPaths scored better results than other citizen science platforms in most categories and is the only platform with results pertaining to credibility. Credibility in WalkingPaths is based on a user's reputation, and this reputation is tied to their previous observations and agreement scores. However, the volume in WalkingPaths is the lowest, although this is expected as the project is new. The most significant difference in scores is missing dates and times in other platforms for completeness, currentness, and traceability.

In WalkingPaths, semantic accuracy is affected by misspelled street and city names. This could be quickly resolved by adding a comprehensive list of available cities and suggestions for street names when the content provider starts typing. However, if a similar platform is extended outside of one country, the list of cities and street names would inflate drastically. While it can be argued that ALA, iNaturalist, and Budburst perform worse than WalkingPaths or Globe at Night because they collect different kinds of observations, the same techniques used in the development of WalkingPath can be utilized in any type of observation. The difference in the types of observations is negligible as platforms' underlying principle stays the same.

## 5. **Discussion**

Improving data quality based on the data model is not a new idea [6,32,46,59]. This research aims to improve data and information quality by integrating quality characteristics into the citizen science platform's design, mainly focusing on the data model and user interface. Improving the data model is presented as an excellent option to enhance data quality, while processes that add or modify the data should be examined and improved [46]. Data quality

characteristics can be considered constraints for the data model and the platform's user interface during the design stage.

Integrating quality characteristics into the design of a platform can increase the quality of data and information. Integration into a new platform is easier compared to integrating similar characteristics into an existing platform. If quality characteristics are integrated into the data model of an existing platform, the whole platform needs to be shut down in the worst-case scenario while making these changes. All current data needs to be either discarded or modified to comply with the new data model. Characteristics integrated into the user interface are easier to integrate as they do not require significant changes or modifications to the data model.

The quality of a platform design has a significant impact on the engagement of users. Many different variables affect how engaging a platform is. For example, placing a burden on users by requiring too much detail and information can demotivate citizens from submitting information, and the amount of information users send should be kept to a minimum [32]. Placing too many restrictions on what type of information users can submit may negatively influence their willingness to engage with the platform and continue to contribute. On the other hand, having too much freedom may equally demotivate users as they are unsure what information should be given, and the quality of information is drastically reduced [60,61]. Thus, there are tradeoffs with engagement and platform constraints that should be appropriately balanced. The adverse effects of limitations can be alleviated by masking the rules as guidelines rather than automatically implementing the rules to outright reject information.

Some researchers have investigated how citizen science platforms' data quality can be increased by training citizens [62], using reputation models [63], and using attribute filtering methods for data input [59]. These are excellent choices for increasing the quality of data and information, but they require more from citizens than making changes to the platform would.

There are some limitations to WalkingPath data collection worth mentioning. First of all, the amount of data used for evaluation is small compared to citizen science platforms that have been online for a longer time. Another limitation is that the data is limited to one country. Finally, it would be beneficial to investigate how the integration of quality characteristics into an existing platform affects the quality of data and information and whether the benefits outweigh the costs.

## 6. **Conclusion**

This research presents an approach to improve citizen science platforms' data and information quality by integrating quality characteristics into the platform design. Results show that incorporating quality characteristics into the design increases the overall quality of data compared to existing citizen science platforms. Furthermore, most characteristics can be integrated without significant changes to the design. Some of the characteristics are integrated into the data model, and others are integrated into the user interface. Several are integrated into both by attaching a score to the data entity in the data model.

This research's integration criterion and method are helpful instruments for citizen science platform designers to improve data and information quality. This framework can be used in any platform and even be applied to an existing platform if necessary. The framework presents four categories for classifying the chosen characteristics to aid in deciding whether they should be integrated into the user interface or the data model.

The most important step is identifying which characteristics are essential in each platform, and this has to be done by considering the context in which the information will be used. This research selects frequently used characteristics for data and information quality that can be

utilized in most citizen science platforms. However, the list is not exhaustive, and there may exist some relevant characteristics for specific cases.

Data and information quality are easier to define when quality is split into data and information characteristics. Researchers often base their definition and selection of data quality characteristics on previous research of classical quality models [25–27,34]. These research works assess classical data quality, and are required to be adjusted case-by-case because data quality depends on the scenario [10,29]. Context is vital for information [28,37]. Data quality characteristics that are context-dependent can be transformed into information quality characteristics. This paper investigates data quality characteristics from earlier works and filters out those that do not apply to information to identify data and information quality characteristics related to citizen science platforms.

Some people trust data from citizen science platforms less than other sources because citizens are considered non-professionals who provide inaccurate data [6,22,64,65]. However, this is not necessarily true, and even if it is, there are methods to increase the quality of data in the platform [17,32,59,66].

In the future, a method will be developed so that the characteristics can be implemented into an existing platform to investigate how schema evolution can be accommodated to improve the quality of data in existing citizen science platforms and how quality is improved before and after the integration of quality characteristics.

**Supplementary Materials:** The following are available online at https://www.mdpi.com/article/10.3390/ijgi10070434/s1.

**Author Contributions:** Conceptualization, Jiri Musto and Ajantha Dahanayake; Methodology, Jiri Musto and Ajantha Dahanayake; Supervision, Ajantha Dahanayake; Writing—original draft, Jiri Musto; Writing—review & editing, Jiri Musto and Ajantha Dahanayake. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study is available in (Supplementary Material).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Cornell Lab of Ornithology eBird - Discover a new world of birding Available online: https://ebird.org/home (accessed on Mar 9, 2021).
2. Lintott, C.; Schawinski, K.; Bamford, S.; Slosar, A.; Land, K.; Thomas, D.; Edmondson, E.; Masters, K.; Nichol, R.C.; Raddick, M.J.; et al. *Galaxy Zoo 1 : Data Release of Morphological Classifications for nearly 900,000 galaxies ✦*; 2010; Vol. 000;
3. Waldispühl, J.; Szantner, A.; Knight, R.; Caisse, S.; Pitchford, R. Leveling up citizen science. *Nat. Biotechnol.* **2020**, *38*, 1124–1126, doi:10.1038/s41587-020-0694-x.
4. See, L.; Mooney, P.; Foody, G.; Bastin, L.; Comber, A.; Estima, J.; Fritz, S.; Kerle, N.; Jiang, B.; Laakso, M.; et al. Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS Int. J.*

*Geo-Information* **2016**, *5*, 55, doi:10.3390/ijgi5050055.
5. Simpson, R.; Page, K.R.; De Roure, D. Zooniverse: Observing the world's largest citizen science platform. In Proceedings of the WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web; Association for Computing Machinery, Inc: New York, New York, USA, 2014; pp. 1049–1054.
6. Lukyanenko, R.; Parsons, J.; Wiersma, Y. The IQ of the crowd: Understanding and improving information quality in structured user-generated content. *Inf. Syst. Res.* **2014**, *25*, 669–689, doi:10.1287/isre.2014.0537.
7. Arthur, R.; Boulton, C.A.; Shotton, H.; Williams, H.T.P. Social sensing of floods in the UK. *PLoS One* **2018**, *13*, doi:10.1371/journal.pone.0189327.
8. Liu, X.; Heller, A.; Nielsen, P.S. CITIESData: a smart city data management framework. *Knowl. Inf. Syst.* **2017**, *53*, 699–722, doi:10.1007/s10115-017-1051-3.
9. SciStarter Welcome to SciStarter Available online: https://scistarter.com/ (accessed on Jan 9, 2021).
10. Lukyanenko, R.; Parsons, J.; Wiersma, Y.F. Emerging problems of data quality in citizen science. *Conserv. Biol.* **2016**, *30*, 447–449, doi:10.1111/cobi.12706.
11. Nasiri, A.; Abbaspour, R.A.; Chehreghan, A.; Arsanjani, J.J. Improving the quality of citizen contributed geodata through their historical contributions: The case of the road network in OpenStreetMap. *ISPRS Int. J. Geo-Information* **2018**, *7*, 253, doi:10.3390/ijgi7070253.
12. Leibovici, D.G.; Rosser, J.F.; Hodges, C.; Evans, B.; Jackson, M.J.; Higgins, C.I. On data quality assurance and conflation entanglement in crowdsourcing for environmental studies. *ISPRS Int. J. Geo-Information* **2017**, *6*, doi:10.3390/ijgi6030078.
13. Sheppard, S.A.; Wiggins, A.; Terveen, L. Capturing quality. In Proceedings of the Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14; ACM Press: New York, New York, USA, 2014; pp. 1234–1245.
14. Elbroch, M.; Mwampamba, T.H.; Santos, M.J.; Zylberberg, M.; Liebenberg, L.; Minye, J.; Mosser, C.; Reddy, E. The Value, Limitations, and Challenges of Employing Local Experts in Conservation Research. *Conserv. Biol.* **2011**, *25*, 1195–1202, doi:10.1111/j.1523-1739.2011.01740.x.
15. Haklay, M.; Basiouka, S.; Antoniou, V.; Ather, A. How many volunteers does it take to map an area well? the validity of Linus' law to volunteered geographic information. *Cartogr. J.* **2010**, *47*, 315–322, doi:10.1179/000870410X12911304958827.
16. Mitchell, N.; Triska, M.; Liberatore, A.; Ashcroft, L.; Weatherill, R.; Longnecker, N. Benefits and challenges of incorporating citizen science into university education. *PLoS One* **2017**, *12*, doi:10.1371/journal.pone.0186285.
17. Bordogna, G.; Carrara, P.; Criscuolo, L.; Pepe, M.; Rampini, A. On predicting and improving the quality of Volunteer Geographic Information projects. *Int. J. Digit. Earth* 2016, *9*, 134–155.
18. Medeiros, G.; Holanda, M. Solutions for Data Quality in GIS and VGI: A Systematic Literature Review. *Adv. Intell. Syst. Comput.* **2019**, *930*, 645–654, doi:10.1007/978-3-030-16181-1_61.
19. Torre, M.; Nakayama, S.; Tolbert, T.J.; Porfiri, M. Producing knowledge by admitting ignorance: Enhancing data quality through an "I don't know" option in citizen science. *PLoS One* **2019**, *14*, doi:10.1371/journal.pone.0211907.

20. Dorn, H.; Törnros, T.; Zipf, A. Quality evaluation of VGI using authoritative data-a comparison with land use data in southern Germany. *ISPRS Int. J. Geo-Information* **2015**, *4*, 1657–1671, doi:10.3390/ijgi4031657.
21. Musto, J.; Dahanayake, A. Improving Data Quality, Privacy and Provenance in Citizen Science Applications. In Proceedings of the Frontiers in Artificial Intelligence and Applications; IOS Press, 2019; Vol. 321, pp. 141–160.
22. Bayraktarov, E.; Ehmke, G.; O'Connor, J.; Burns, E.L.; Nguyen, H.A.; McRae, L.; Possingham, H.P.; Lindenmayer, D.B. Do big unstructured biodiversity data mean more knowledge? *Front. Ecol. Evol.* **2019**, *7*, doi:10.3389/fevo.2018.00239.
23. Sadiq, S.; Indulska, M. Open data: Quality over quantity. *Int. J. Inf. Manage.* **2017**, *37*, 150–154, doi:10.1016/j.ijinfomgt.2017.01.003.
24. Lewandowski, E.; Specht, H. Influence of volunteer and project characteristics on data quality of biological surveys. *Conserv. Biol.* **2015**, *29*, 713–723, doi:10.1111/cobi.12481.
25. Wang, R.Y.; Strong, D.M. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–34.
26. Batini, C.; Scannapieco, M. *Data quality : concepts, methodologies and techniques*; Springer, 2006; ISBN 9783540331735.
27. Redman, T.C. *Data quality for the information age*; Artech House, 1996; ISBN 9780890068830.
28. Bovee, M.; Srivastava, R.P.; Mak, B. A conceptual framework and belief-function approach to assessing overall information quality. *Int. J. Intell. Syst.* **2003**, *18*, 51–74, doi:10.1002/int.10074.
29. Haug, A.; Arlbjrn, J.S.; Pedersen, A. A classification model of ERP system data quality. *Ind. Manag. Data Syst.* **2009**, *109*, 1053–1068, doi:10.1108/02635570910991292.
30. Han, J.; Jiang, D.; Ding, Z. Assessing data quality within available context. In Proceedings of the Data Quality and High-Dimensional Data Analysis - Proceedings of the DASFAA 2008 Workshops; 2009; pp. 42–59.
31. Batini, C.; Blaschke, T.; Lang, S.; Albrecht, F.; Abdulmutalib, H.M.; Barsi, Á.; Szabó, G.; Kugler, Z. Data Quality in Remote Sensing. **2017**, *XLII*, 18–22.
32. Lukyanenko, R.; Parsons, J.; Wiersma, Y.F.; Maddah, M. Expecting the unexpected: Effects of data collection design choices on the quality of crowdsourced user-generated content. *MIS Q. Manag. Inf. Syst.* **2019**, *43*, 634–647, doi:10.25300/MISQ/2019/14439.
33. ISO ISO 19157:2013 - Geographic information — Data quality 2013.
34. ISO ISO/IEC 25012:2008 Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model 2008.
35. ISO ISO/TS 8000:2011 - Data quality 2011.
36. Watts, S.; Shankaranarayanan, G.; Even, A. Data quality assessment in context: A cognitive perspective. *Decis. Support Syst.* **2009**, *48*, 202–211, doi:10.1016/j.dss.2009.07.012.
37. Davenport, T.; Prusak, L. Working knowledge: how organizations manage what they know. *Ubiquity* **2000**, *2000*, 6.
38. Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* **2009**, *41*, 16, doi:10.1145/1541880.1541883.
39. iNaturalist A Community for Naturalists · iNaturalist.org Available online: https://www.inaturalist.org/ (accessed on Mar 29, 2021).
40. Kelling, S.; Lagoze, C.; Wong, W.-K.; Yu, J.; Damoulas, T.; Gerbracht, J.; Fink, D.; Gomes, C. E Bird: A human/computer learning network to improve biodiversity conservation and

research. *AI Mag.* **2013**, *34*, 10–20.
41. Rajaram, G.; Manjula, K. Exploiting the Potential of VGI Metadata to Develop A Data-Driven Framework for Predicting User's Proficiency in OpenStreetMap Context. *ISPRS Int. J. Geo-Information* **2019**, *8*, 492, doi:10.3390/ijgi8110492.
42. Shanks, G.; Darke, P. Understanding Data Quality in a Data Warehouse. *J. Res. Pract. Inf. Technol.* **1998**, *30*, 122–128.
43. Cai, L.; Zhu, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci. J.* **2015**, *14*, 2, doi:10.5334/dsj-2015-002.
44. Immonen, A.; Pääkkönen, P.; Ovaska, E. Evaluating the Quality of Social Media Data in Big Data Architecture. *IEEE Access* **2015**, *3*, 2028–2043, doi:10.1109/ACCESS.2015.2490723.
45. Higgins, C.I.; Williams, J.; Leibovici, D.G.; Simonis, I.; Davis, M.J.; Muldoon, C.; Van Genuchten, P.; O'hare, G.; Wiemann, S. Citizen OBservatory WEB (COBWEB): A Generic Infrastructure Platform to Facilitate the Collection of Citizen Science data for Environmental Monitoring ©. *Int. J. Spat. Data Infrastructures Res.* **2016**, *11*, 20–48, doi:10.2902/1725-0463.2016.11.art3.
46. Fox, T.L.; Guynes, C.S.; Prybutok, V.R.; Windsor, J. Maintaining Quality in Information Systems. *J. Comput. Inf. Syst.* **1999**, *40*, 76–80, doi:10.1080/08874417.1999.11647427.
47. Fonte, C.C.; Antoniou, V.; Bastin, L.; Estima, J.; Arsanjani, J.J.; Bayas, J.-C.L.; See, L.; Vatseva, R. Assessing VGI Data Quality. In *Mapping and the Citizen Sensor*; Ubiquity Press, 2017; pp. 137–163 ISBN 978-1-911529-16-3.
48. Atlas of Living Australia Open access to Australia's biodiversity data Available online: https://www.ala.org.au/ (accessed on Mar 17, 2021).
49. Globe at Night International citizen-science campaign to raise public awareness of the impact of light pollution Available online: https://www.globeatnight.org/ (accessed on Apr 27, 2021).
50. Budburst Budburst: An online database of plant observations, a citizen-science project of the Chicago Botanic Garden. Glencoe, Illinois. Available online: https://budburst.org/ (accessed on Apr 27, 2021).
51. General Data Protection Regulation (GDPR) – Official Legal Text Available online: https://gdpr-info.eu/ (accessed on Jun 7, 2021).
52. Bill Text - SB-1121 California Consumer Privacy Act of 2018. Available online: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1121 (accessed on Jun 7, 2021).
53. Teorey, T.; Lightstone, S.; Nadeau, T.; Jagadish, H.V. Business Intelligence. In *Database Modeling and Design*; Elsevier, 2011; pp. 189–231.
54. GPS.gov: GPS Accuracy Available online: https://www.gps.gov/systems/gps/performance/accuracy/ (accessed on Apr 19, 2021).
55. Merry, K.; Bettinger, P. Smartphone GPS accuracy study in an urban environment. *PLoS One* **2019**, *14*, e0219890, doi:10.1371/journal.pone.0219890.
56. Schaefer, M.; Woodyer, T. Assessing absolute and relative accuracy of recreation-grade and mobile phone GNSS devices: A method for informing device choice. *Area* **2015**, *47*, 185–196, doi:10.1111/area.12172.
57. Tomaštík, J.; Saloň, Š.; Piroh, R. Horizontal accuracy and applicability of smartphone GNSS positioning in forests. *Forestry* **2017**, *90*, 187–198, doi:10.1093/forestry/cpw031.
58. De Montjoye, Y.A.; Hidalgo, C.A.; Verleysen, M.; Blondel, V.D. Unique in the Crowd: The privacy bounds of human mobility. *Sci. Rep.* **2013**, *3*, 1–5, doi:10.1038/srep01376.

59. Lukyanenko, R.; Parsons, J.; Wiersma, Y. Citizen Science 2.0: Data Management Principles to Harness the Power of the Crowd. In *DESRIST*; Springer, Berlin, Heidelberg, 2011; Vol. 6629 LNCS, pp. 465–473 ISBN 9783642206320.
60. Wehn, U.; Almomani, A. Incentives and barriers for participation in community-based environmental monitoring and information systems: A critical analysis and integration of the literature. *Environ. Sci. Policy* **2019**, *101*, 341–357, doi:10.1016/j.envsci.2019.09.002.
61. Hobbs, S.J.; White, P.C.L. Motivations and barriers in relation to community participation in biodiversity recording. *J. Nat. Conserv.* **2012**, *20*, 364–373, doi:10.1016/j.jnc.2012.08.002.
62. Fonte, C.C.; Bastin, L.; Foody, G.; Kellenberger, T.; Kerle, N.; Mooney, P.; Olteanu-Raimond, A.-M.; See, L. Vgi Quality Control. In Proceedings of the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences; 2015; Vol. 2, pp. 317–324.
63. Pang, L.; Li, G.; Yao, X.; Lai, Y. An Incentive Mechanism Based on a Bayesian Game for Spatial Crowdsourcing. *IEEE Access* **2019**, *7*, 14340–14352, doi:10.1109/ACCESS.2019.2894578.
64. Blatt, A.J. The benefits and risks of volunteered geographic information. *J. Map Geogr. Libr.* **2015**, *11*, 99–104, doi:10.1080/15420353.2015.1009609.
65. See, L.; Comber, A.; Salk, C.; Fritz, S.; van der Velde, M.; Perger, C.; Schill, C.; McCallum, I.; Kraxner, F.; Obersteiner, M. Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts. *PLoS One* **2013**, *8*, doi:10.1371/journal.pone.0069958.
66. Guo, J.; Liu, F. Automatic data quality control of observations in wireless sensor network. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 716–720, doi:10.1109/LGRS.2014.2359685.

**ACTA UNIVERSITATIS LAPPEENRANTAENSIS**

**963.** OYEWO, AYOBAMI SOLOMON. Transition towards decarbonised power systems for sub-Saharan Africa by 2050. 2021. Diss.

**964.** LAHIKAINEN, KATJA. The emergence of a university-based entrepreneurship ecosystem. 2021. Diss.

**965.** ZHANG, TAO. Intelligent algorithms of a redundant robot system in a future fusion reactor. 2021. Diss.

**966.** YANCHUKOVICH, ALEXEI. Screening the critical locations of a fatigue-loaded welded structure using the energy-based approach. 2021. Diss.

**967.** PETROW, HENRI. Simulation and characterization of a front-end ASIC for gaseous muon detectors. 2021. Diss.

**968.** DONOGHUE, ILKKA. The role of Smart Connected Product-Service Systems in creating sustainable business ecosystems. 2021. Diss.

**969.** PIKKARAINEN, ARI. Development of learning methodology of additive manufacturing for mechanical engineering students in higher education. 2021. Diss.

**970.** HOFFER GARCÉS, ALVARO ERNESTO. Submersible permanent-magnet synchronous machine with a stainless core and unequal teeth widths. 2021. Diss.

**971.** PENTTILÄ, SAKARI. Utilizing an artificial neural network to feedback-control gas metal arc welding process parameters. 2021. Diss.

**972.** KESSE, MARTIN APPIAH. Artificial intelligence : a modern approach to increasing productivity and improving weld quality in TIG welding. 2021. Diss.

**973.** MUSONA, JACKSON. Sustainable entrepreneurial processes in bottom-of-the-pyramid settings. 2021. Diss.

**974.** NYAMEKYE, PATRICIA. Life cycle cost-driven design for additive manufacturing: the frontier to sustainable manufacturing in laser-based powder bed fusion. 2021. Diss.

**975.** SALWIN, MARIUSZ. Design of Product-Service Systems in printing industry. 2021. Diss.

**976.** YU, XINXIN. Contact modelling in multibody applications. 2021. Diss.

**977.** EL WALI, MOHAMMAD. Sustainability of phosphorus supply chain – circular economy approach. 2021. Diss.

**978.** PEÑALBA-AGUIRREZABALAGA, CARMELA. Marketing-specific intellectual capital: Conceptualisation, measurement and performance. 2021. Diss.

**979.** TOTH, ILONA. Thriving in modern knowledge work: Personal resources and challenging job demands as drivers for engagement at work. 2021. Diss.

**980.** UZHEGOVA, MARIA. Responsible business practices in internationalized SMEs. 2021. Diss.

**981.** JAISWAL, SURAJ. Coupling multibody dynamics and hydraulic actuators for indirect Kalman filtering and real-time simulation. 2021. Diss.

**982.** CLAUDELIN, ANNA. Climate change mitigation potential of Finnish households through consumption changes. 2021. Diss.

**983.** BOZORGMEHRI, BABAK. Finite element formulations for nonlinear beam problems based on the absolute nodal coordinate formulation. 2021. Diss.

**984.** BOGDANOV, DMITRII. Transition towards optimal renewable energy systems for sustainable development. 2021. Diss.

**985.** SALTAN, ANDREY. Revealing the state of software-as-a-service pricing. 2021. Diss.

**986.** FÖHR, JARNO. Raw material supply and its influence on profitability and life-cycle assessment of torrefied pellet production in Finland – Experiences from pilot-scale production. 2021. Diss.

**987.** MORTAZAVI, SINA. Mechanisms for fostering inclusive innovation at the base of the pyramid for community empowerment - Empirical evidence from the public and private sector. 2021. Diss.

**988.** CAMPOSANO, JOSÉ CARLOS. Integrating information systems across organizations in the construction industry. 2021. Diss.

**989.** LAUKALA, TEIJA. Controlling particle morphology in the in-situ formation of precipitated calcium carbonate-fiber composites. 2021. Diss.

**990.** SILLMAN, JANI. Decoupling protein production from agricultural land use. 2021. Diss.

**991.** KHADIM, QASIM. Multibody system dynamics driven product processes. 2021. Diss.

**992.** ABDULKAREEM, MARIAM. Environmental sustainability of geopolymer composites. 2021. Diss.

**993.** FAROQUE, ANISUR. Prior experience, entrepreneurial outcomes and decision making in internationalization. 2021. Diss.

**994.** URBANI, MICHELE. Maintenance policies optimization in the Industry 4.0 paradigm. 2021. Diss.

**995.** LAITINEN, VILLE. Laser powder bed fusion for the manufacture of Ni-Mn-Ga magnetic shape memory alloy actuators. 2021. Diss.

**996.** PITKÄOJA, ANTTI. Analysis of sorption-enhanced gasification for production of synthetic biofuels from solid biomass. 2021. Diss.

**997.** MASHLAKOV, ALEKSEI. Flexibility aggregation of local energy systems—interconnecting, forecasting, and scheduling. 2021. Diss.

**998.** NIKITIN, ALEKSEI. Microwave processes in thin-film multiferroic heterostructures and magnonic crystals. 2021. Diss.

**999.** VIITALA, MIRKA. The heterogeneous nature of microplastics and the subsequent impacts on reported microplastic concentrations. 2021. Diss.

**1000.** ASEMOKHA, AGNES. Understanding business model change in international entrepreneurial firms. 2021. Diss.



ISBN 978-952-335-757-0
ISBN 978-952-335-758-7 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491
Lappeenranta 2021