Information Modelling and Knowledge Bases XXXIII
M. Tropmann-Frick et al. (Eds.)

doi:10.3233/FAIA210490

Quality Characteristics for User-Generated Content

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

Abstract. Today, vast amounts of data are collected from the internet, and the general public generates most data using 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. This research examines different user-generated content platforms 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

[1] Corresponding Author, Jiri Musto, School of Engineering Science, Lappeenranta-Lahti University of Technology LUT, Yliopistonkatu 34, 53850 Lappeenranta, Finland; E-mail: jiri.musto@lut.fi.

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 contain 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 differs from 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 review 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 in Table 1 accounts for data quality research's present (2020) status.

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].

$$syntactic\ acc = \frac{\sum_{i}^{K} closeness\ (w_i,V)}{K} \tag{1}$$

Semantic accuracy: How correctly the meaning of values represents real-world facts. An object identification problem where α and β 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].

$$\langle\alpha,\beta\rangle \in \begin{cases} M\ if\ p(M|\underline{x})\ \geq p(U|\underline{x}) \\ \quad U\ otherwise \end{cases} \tag{2}$$

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].

$$completeness = 1 - \frac{T_v}{N_v} \tag{3}$$

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].

$$r_n(g) = \begin{cases} 0, if\ g\ fulfills\ rule\ r_n \\ \quad 1\ else \end{cases}; cons(g) = 1 - \frac{\sum_{n=1}^{N} r_n(g)}{N} \tag{4}$$

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].

$$credibility = \begin{cases} 1 - \frac{dist}{d_{max}} & if\ d(s,e) < d_{max} \\ \quad 0 & otherwise \end{cases} \tag{5}$$

Objectivity: Data is unbiased and impartial, where E is evidence, H is a hypothesis (assumed value), and $p()$ denotes the probability (6)[61].

$$w(E,H,H') = log\ \frac{p(E|H)}{p(E|H')} \tag{6}$$

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].

$$lev_{a,b}(\mathrm{i},\mathrm{j}) = \begin{cases} max(i,j) & if\ min(i,j) = 0, otherwise \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} \end{cases} \tag{7}$$

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].

$$sample\ size = N * \frac{\frac{z^2 * p * (1-p)}{e^2}}{\left[N-1+\frac{z^2 * p * (1-p)}{e^2}\right]} \tag{8}$$

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].

$$compliance\ (degree\ of\ variance\ ) = \frac{\sum(a-\bar{a})(b-\bar{b})}{\sqrt{\sum(a-\bar{a})^2 \sum(b-\bar{b})^2}} \tag{9}$$

Currentness: Currency concerns how promptly data are updated with respect to changes occurring in the real world (10)[23].

$$currentness\ = Age + (DeliveryTime - InputTime) \tag{10}$$

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].

$$timeliness = \max\{0, 1 - \frac{currentness}{volatility}\} \tag{11}$$

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].

$$PrivacyRisk_{(i,j)} = \frac{S_i^a \times V_{(i,j)}^b}{R_{(i,j)}^c} \tag{12}$$

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].

$$SentenceSimilarity\ (Q, Q') = \frac{1}{n}\sum_{1 \le j \le n}(max_{1 \le i \le m} WordSimilarity(w_j, w_i') \tag{13}$$

Usability: A collection of other characteristics characterized by usability aspects, verifiability, imperfection, and integration (14)[67].

$$usability = avg(accuracy + credibility + completeness + currentness + relevance + granularity + accessibility) \tag{14}$$

Value: The extent to which data are beneficial and provide advantages from their use (15).[68]

$$DataValue(t) \ge (GatherCost\ +\ MaintainCost\ +\ AccessCost)/GB/yr\ *\ RetentionPeriod \tag{15}$$

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].

$$confidentiality = \frac{\sum_{i=1}^{n} w_{c_i} * x_{s_i}}{\sum_{i=1}^{n} w_{c_i}} \tag{16}$$

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].

$$granularity = \log|U| - \sum_{i=1}^{n} \frac{|X_i|}{|U|} \log(|X_i|)\ , U\ is\ a\ universal\ \ set\ and\ set\ X \subseteq U \tag{17}$$

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 *λ* is the number of reports (18)[71].

$$Network\ traceability\ entropy\ (NTE), E^{\lambda} = \sum_{\Omega:|\Omega|=\lambda} E(\Omega)/\binom{|R|}{\lambda} \tag{18}$$

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].

$$whyProvenance(Q, I, t) = \{J \in I \mid t \in Q(J)\}$$
$$whereProvenance\ (\{u\}, I, t) = \begin{cases} (A : \emptyset)_{A \in U}, if\ t = u \\ \bot, otherwise \end{cases} \quad (19)$$
$$howProvenance\ (Q, I, t) = \ Q^{K_{How}}\ (I_{How})t$$

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 context-specific content submitted by the public. 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, users' details 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, and 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, and 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. The public uses all introduced platforms, so 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, and 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, and 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, and 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 the 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 1 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 they 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.

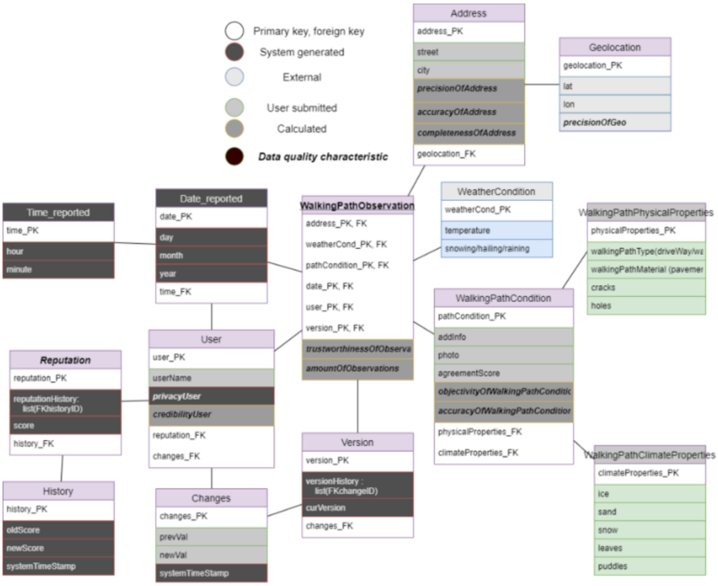
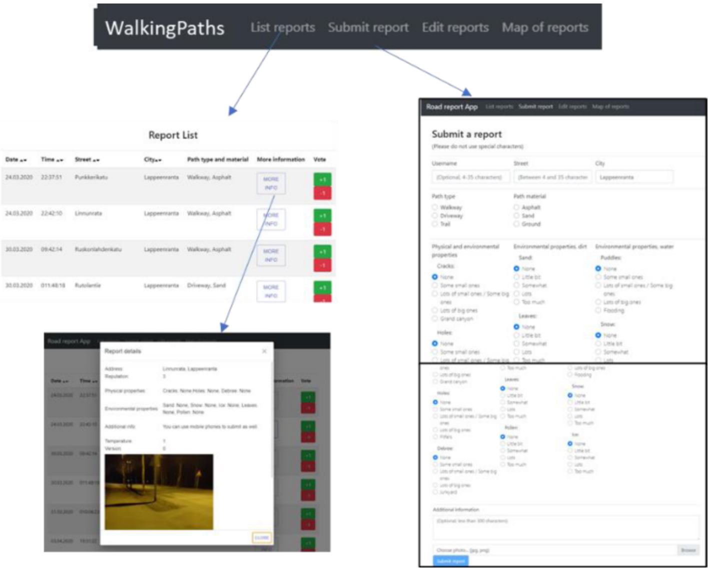


Figure 1. Snowflake schema for WalkingPaths



Report List

Date	Time	Street	City	Path type and material	More information	Vote
24.03.2020	22:37:51	Punkkerikatu	Lappeenranta	Walkway, Asphalt	MORE INFO	+1 -1
24.03.2020	22:42:10	Linnunrata	Lappeenranta	Walkway, Asphalt	MORE INFO	+1 -1
30.03.2020	09:42:14	Ruskonlahdenkatu	Lappeenranta	Walkway, Asphalt	MORE INFO	+1 -1
30.03.2020	011:48:18	Rutolantie	Lappeenranta	Driveway, Sand	MORE INFO	+1 -1

Report details

Address:	Linnunrata, Lappeenranta
Reputation:	3
Physical properties:	Cracks: None Holes: None, Debree: None
Environmental properties:	Sand: None, Snow: None, Ice: None, Leaves: None, Pollen: None
Additional info:	You can use mobile phones to submit as well
Temperature:	7
Version:	0

Submit a report

(Please do not use special characters)

Username: (Optional, 4-35 characters)
Street: (Between 4 and 35 character
City: Lappeenranta

Path type: ○ Walkway ○ Driveway ○ Trail

Path material: ○ Asphalt ○ Sand ○ Ground

Physical and environmental properties
Cracks: ● None ○ Some small ones ○ Lots of small ones / Some big ones ○ Lots of big ones ○ Grand canyon
Holes: ● None ○ Some small ones ○ Lots of small ones / Some big ones ○ Lots of big ones ○ Pitfalls
Debree: ● None

Environmental properties, dirt
Sand: ● None ○ Little bit ○ Somewhat ○ Lots ○ Too much
Leaves: ● None ○ Little bit ○ Somewhat ○ Lots ○ Too much
Pollen: ● None ○ Little bit ○ Somewhat

Environmental properties, water
Puddles: ● None ○ Some small ones ○ Lots of small ones / Some big ones ○ Lots of big ones ○ Flooding
Snow: ● None ○ Little bit ○ Somewhat ○ Lots ○ Too much
Ice: ● None ○ Little bit

Additional information: (Optional, less than 300 characters)

Choose photo... (jpg, png) Browse

Submit report

Figure 2. WalkingPaths list of observations and report submission

Figure 2 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. These decisions should be made during the design phase 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 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, and 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.