A systematic literature review of open data quality in practice

A Systematic Literature Review of Open Data Quality in Practice

Mohammad Rashid, Marco Torchiano

Politecnico di Torino
Department of Control and Computer Engineering, Turin, Italy
Email: name.surname@polito.it

2016 Open Data Research Symposium

5 October 2016, Madrid, Spain

Abstract

Context: The main objective of open data initiatives is to make information freely available through easily accessible mechanisms and facilitate exploitation. In practice openness should be accompanied with a certain level of trustworthiness or guarantees about the quality of data. Traditional data quality is a thoroughly researched field with several benchmarks and frameworks to grasp its dimensions. However, quality assessment in open data is a complicated process as it consists of stakeholders, evaluation of datasets as well as the publishing platform.
Objective: In this work, we aim to identify and synthesize various features of open data quality approaches in practice. We applied thematic synthesis to identify the most relevant research problems and quality assessment methodologies.
Method: We undertook a systematic literature review to summarize the state of the art on open data quality. The review process starts by developing the review protocol in which all steps, research questions, inclusion and exclusion criteria and analysis procedures are included. The search strategy retrieved 9323 publications from four scientific digital libraries. The selected papers were published between 2005 and 2015. Finally, through a discussion between the authors, 63 paper were included in the final set of selected papers.
Results: Open data quality, in general, is a broad concept, and it could apply to multiple areas. There are many quality issues concerning open data hindering their actual usage for real-world applications. The main ones are unstructured metadata, heterogeneity of data formats, lack of accuracy, incompleteness and lack of validation techniques. Furthermore, we collected the existing quality methodologies from selected papers and synthesized under a unifying classification schema. Also, a list of quality dimensions and metrics from selected paper is reported.
Conclusion: In this research, we provided an overview of the methods related to open data quality, using the instrument of systematic literature reviews. Open data quality methodologies vary depending on the application domain. Moreover, the majority of studies focus on satisfying specific quality criteria. With metrics based on generalized data attributes a platform can be created to evaluate all possible open dataset. Also, the lack of methodology validation remains a major problem. Studies should focus on validation techniques.

Keywords: Systematic Review, Open Data, Data Quality

1. Introduction

In modern digital society, publishing information is a necessity and it is an undeniable fact that services need data. For this, governments, non-governmental organizations, as well as communities take initiatives and increasingly publish their data on the web [R42]. However, this initiatives encounter various barriers; for example, legal restriction for accessing or publishing data. To achieve the full potential in disclosure, data should be freely accessible without any legal issues or any control restrictions. Data published in such a way are referred to as open data.

The use of open data adds values to the services of governmental and non-governmental organizations. According

to the Open Knowledge Foundation[1], and Open Government Data Working Group[2], there are several reasons to have data opened [R32]: (i) transparency (ii) releasing social and commercial value and (ii) citizen participation and engagement.

Those three motivations, while not being the only ones, are common for most open data initiatives. By opening data, stakeholders get the opportunity to scrutinize and reuse the available information in many ways, including identifying patterns in the data and creating new services [R11][R3][R5]. This results into an increased accountability that influences development.

However, due to data quality issues, it is still a major challenge achieving the full capacity of open data initiatives in government or enterprises[Sayogo and Pardo (2012)]. They represent a major obstacle to open data application developers. Due to the diversity of open data, the issues of data quality has become more complex [R62]. An important factor consists in the heterogeneity of data formats; for example, public administrations use various data ranging from images to PDF, CSV files, Excel sheets as well as structured data format – e.g. XML files –, and database records. Also, due to incompleteness and lack of accuracy, the usage of open data inside applications is generally difficult. Moreover the limited number of data quality methodologies and relative validation techniques, hinder the provision of data services. Open data portals suffer from a large number of diverse data structures as well as with no data quality assessment.The majority of open data initiative use linked data principles to share and consume data. Using linked data, creates opportunity by linking with useful information with good provenance [R10]. But the challenge in using linked data is to manage the heterogeneous data together with minimal semantics and structure.

To explore the theme of open data quality we conducted a the systematic literature review (SLR)[Kitchenham and Charters (2007)]. We deduced and present research problems and a list of data quality attributes, we identified several data quality methods and explored the relative validation techniques.

The remainder of this paper is structured as follows: section 2 provides background information about linked data, open government data, data quality. Also provide short details on related works. Section 3 summarizes the methodologies and defines objectives and research questions. Section 4 outlines the results of the review organized according each research question defined in section 3. Section 5 contains the conclusions.

2. Background

2.1. Linked Data (LD)

The linked data approach consists in exposing and connecting data from different sources on the web by means of semantic web technologies. In [R28] Tim Berners-Lee et al. state that linked data ”[..] refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets”.

Tim Berners-Lee in July 2006 [3] outlined the set of Linked Data principles: (1) use URIs as names for things (2) use HTTP URIs, so that people can look up those names, (3) use standard mechanisms to provide useful information when someone looks up a URI, for example RDF (Resource Description Framework) to represent data as graphs and SPARQL (SPARQL Protocol and RDF Query Language) to query Linked Data, and (4) include links to other URIs, to enable the discovery of more things.

Linked Open Data (LOD) approach applies the LD principles to open content. In [R58] the authors mention linked data as, a distributed model for the semantic Web that allows any data provider to publish its publicly available data and meaningfully link them with other information sources over the Web.

2.2. Data Quality (DQ)

Data quality is a cross-disciplinary and multi-dimensional concept [R47][R30]. Data quality, in general, relates to the perception of the ”fitness for use” in a given context. According to Pipino et al. (2002) based on context, quality can be both subjective perceptions and objective measurements. Subjective data quality assessments reflect the needs and experiences of stakeholders. Objective assessments can be task-dependent or task-independent. Task-independent

[1] `https://okfn.org/opendata/`

[2] `http://opengovernmentdata.org/`

[3] `http://www.w3.org/DesignIssues/LinkedData.html`

quality assessment metrics reflects the properties of the data without contextual knowledge of how it will be consumed. Task-dependent metrics, on the other hand, reflect the requirements of the application at hand. ISO/IEC 25012 (2008) defines a general data quality model for data retained in a structured format within a computer system. This model defined the quality of a data product, as the degree to which data satisfy the requirements set by the product owner organization. In general, it is important to identify data quality assurance criteria. These criteria vary with different stakeholders or domain.

2.3. Open Data Quality (ODQ)

Open data provides various means to let anyone to freely access, modify, and share data. Usability of open data is only possible when it can be understood by a human as well as it can be machine processable. For example, publishing open data require the permission of the publisher, granted via an open licence. Making the data accessible, however, does not imply that the users will find the resources usable. Content publisher have to ensure that the resources are credible and discoverable. The credibility is bound to the quality of the actual data, the data sets. The discoverability is bound to the quality of the descriptive data, as well as the metadata.

From a data-driven perspective, open data quality should address the control capability over the ease of access, modifiability, usability and discoverability of data.

2.4. Related Work

Traditional data quality is a thoroughly researched field with several benchmarks and frameworks to grasp its dimensions. In [R30], the authors present a comparative description of various data quality assessment and improvement techniques. However, this study only considered 13 traditional quality assessment methodologies. With increasing popularity in linked data more focused given towards linked data quality assessment methodologies. In **?**, the authors present a comprehensive systematic review of data quality assessment methodologies applied to LOD. They have extracted 26 quality dimensions and a total of 110 objective and subjective quality indicators. However, some of those objectives indicators are dependent on the use case thus there is no clear separation on what can be automatically measured.

But in the case of open data, it should ensure the ease of access, understandability, and discoverability. In short it open data quality ensure "quality control." Due to open data quality broad areas of studies, it includes various quality assessment methodologies. In this work, we explore data quality methodologies applied to open data processes.

3. Research Methodology

This work studies the state of the art on open data quality. It follows the guidelines set out by Kitchenham and Charters (2007); Kitchenham and Brereton (2013) for systematic literature reviews in software engineering. The protocol in a review states the methods that will be used to approach a specific systematic review. The definition of a protocol specifies the research questions addressed in a literature review and also presents the methods that will be used to perform it. Figure 1 shows the protocol and introduces the main steps we followed in this systematic literature review.

In our systematic literature review, the protocol was developed by the first author while the second author validated it. Each step of our protocol is described in more detail below:

1. Construction of research questions (sec.3.1) is the first step to guide the review.
2. Definition of keywords and search string (sec.3.2) used to perform an initial search for papers.
3. Selection of sources (sec.3.3) used to choose the digital libraries that store bibliographic information about papers.
4. Search and selection (sec.3.4) the papers that match the keywords are retrieved, and the relevant ones are selected.
5. Quality assessment (sec.3.5) assigns a level of quality to the selected papers according to a set of quality criteria.

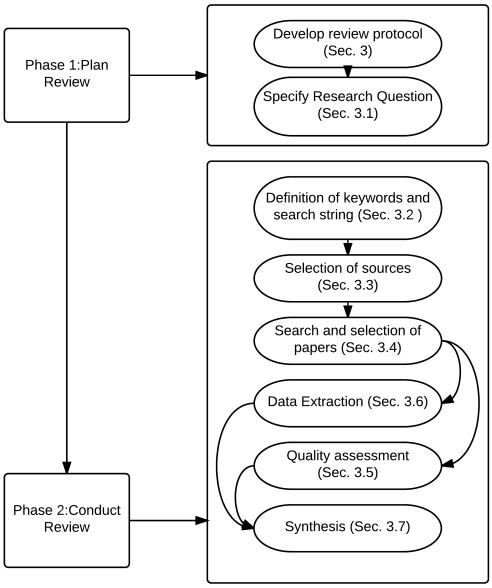

Figure 1: Systematic literature review at a glance based on Kitchenham and Charters (2007); Kitchenham and Brereton (2013)

6. Data extraction (sec.3.6) allows to obtain the relevant information, such as title, authors, problems, etc. from the selected papers.

7. Synthesis (sec.3.7) examines and organizes the data contained in each selected paper.

3.1. Research Questions

The most important part of a systematic review is to specify the research question. Research questions drive the entire systematic review methodology. The goal of our systematic literature review is to understand different features of open data quality. Accordingly, we have defined following four research questions with the relative aims.

RQ1 What challenges and problems have been faced by researchers? To determine which research problems perspectives mainly focused by researchers. Furthermore, report what issues are addressed in practice together with solutions.

RQ2 What frameworks and methods are applied to open data/service quality assessment? To determine which research problems perspectives mainly focused by researchers. Furthermore, report what issues are addressed in practice together with solutions.

RQ3 Which quality attributes are addressed by researchers? To identify the metrics and quality dimensions which are commonly used to evaluate open data quality

RQ4 What kind of validation of frameworks and methods have been performed? To identify which validation methodologies are used in practice for open data quality.

Focus of our study in the domain of open data. For this we extended our RQ3 with the following sub-question:

- What measurement techniques are specific for open data?

3.2. Keywords and search string

The next step of a review is to define a proper search strategy. We need to identify keywords and a search string. As suggested in Kitchenham and Charters (2007), a general approach is to break down the question into individual facets i.e. population, intervention, comparison, outcomes, context, study designs. Then draw up a list of synonyms, abbreviations, and alternative spellings. By taking into account, the main topics found in the research questions a preliminary set of keywords was defined: {*Open Data, Data Quality*}. This set was then extended by searching for synonyms to obtain the final set of keywords reported in Table 1 used to define a search string.

Finally, the search string derived from the final set that we used to look for studies in digital libraries composed by the terms Data AND Quality.

```
("open data" OR "linked data" OR "open government data" OR "linked open data")
AND ("data quality" OR "quality assessment" OR "framework" OR "metrics"
OR "methodology" OR "data services" OR "publishing")
```

3.3. Selection of sources

We performed the initial search for primary studies by using the search strings with the four scientific digital libraries we selected. The selected digital libraries, in Table 2, represent primary sources for computer science research publications. Other sources like Scopus, CiteSeer and Google Scholar were not considered as they mainly index data from the primary sources. However it is not sufficient, for an exhaustive review, resorting only to digital libraries, other sources of evidence must be searched. Therefore at the end of the initial selection of papers we also manually reviewed the reference lists from the selected papers to extend our search space .

Table 1: Search Keywords

Keyword	Synonyms
Open Data	Linked Open Data, Open Government Data, Linked open government data
Data Quality	Information Quality, Data quality Services, Web information quality

Table 2: Sources selected for the search process

Source	URL
IEEExplore	`http://ieeexplore.ieee.org`
SpringerLink	`http://link.springer.com`
ACM Digital Library	`http://dl.acm.org`
Science Direct	`http://www.sciencedirect.com`

3.4. Search and selection

Besides, a set of selection criteria needs to be applied to the studies to identify papers that provide direct evidence about the research question [Kitchenham and Charters (2007); Kitchenham and Brereton (2013)]. The aim of our systematic review is to identify and classify papers that address open data quality aspects and exploit the implicit knowledge of open data. Following we report the inclusion/exclusion criteria that were adopted in this study.

Inclusion Criteria

- Papers presenting research related to open data application, linked data, publishing data, OGD in details based on the particular direction to open data quality.
- Papers from conferences and journals. Also short and workshop papers that fulfill the above criteria.
- Papers published from 2005 to 2015. Open Data is a relatively new concept, therefore, open data adaptation also recent.
- Only papers written in english language.

Exclusion Criteria

- Papers not addressing data quality neither any details related to open data studies.
- Papers are discussing open data/linked open data/OGD but not consider any issue regarding quality assessment.

- Papers that report only abstracts or slides of presentations because the lack of information.
- Grey literature. We do not think that technical reports, unpublished studies, and Ph.D. thesis would add much more information on journal and conference papers.

The selection process consist of three main phases: querying of the selected sources, filtering by title and then filtering by abstract.

1. *Querying of selected sources:* First author queried the primary sources using the search string in order to obtain an initial set of papers.
2. *Filtering by title:* First author filtered the results taking into account only the titles of the papers in the initial set. Then, first author deleted duplicated studies and use defined inclusion set for the first phase according to the following steps:
   (a) Studies included by authors were added to the inclusion set.
   (b) Studies excluded by authors were not added to the inclusion set.
   (c) The decision of inclusion decision for the rest of studies was discussed with second author until agreement was reached.
3. *Filtering by abstract:* First author filtered the studies from the set resulting from step selection by taking into account their abstracts. Then, step selection is repeated in order to obtain a final inclusion set. When confusion arise for first author with some studies, the second author intervened to reach solution.

3.5. Quality assessment

Together with study selection criteria, it is important to provide measures that enable quality assessment of the studies. The quality assessment performed using a set of quality checklist addressing the issues affecting the quality of the studies. We have defined a set of quality criteria that are listed in the checklist provided in Table 3. The assessment for each question is typically scored with values 1, 0.5, and 0, respectively to represent the answers *'yes'*, *'partially'* and *'no'*. The total score was used to get an overall measure of study quality. Quality assessment mainly used in data extraction and synthesis phase to reduce possible bias. The first author assessed the entire set of selected papers based on quality assessment checklist. The second author only intervene when confusion arise. Finally, an agreement on differences was reached by discussion. The quality assessment was done in parallel with data extraction.

Table 3: Quality assessment checklist

Question	Yes	Partially	No
Q1. Did the study clearly describe the challenges and problems that are addressing?	53(84.12%)	10(15.8%)	0(0.0%)
Q2. Did the study consider any related work ?	62(98.4%)	1(1.5%)	0(0.0%)
Q3. Did the study discuss clearly related issues, and provide solution ?	48(76.1%)	12(19.04%)	3(4.76%)
Q4. Did the study explore any features of open data context in details?	56(88.8%)	7(11.1%)	0(0.0%)
Q5. Did the study perform an appropriate experiment to achieve aims of the research?	43(68.2%)	18(28.5%)	2(3.1%)
Q6. Did the study presents a clear statement of findings?	51(80.9%)	10(15.8%)	2(3.1%)

Table 4: Data extraction form

Data Field	Description	Research Question
Title	-	-
Authors	-	-
Year of publication	-	-
Examiner	Name of person who performed data extraction	-
Publication source	-	-
Context	Areas of study: open data process, linked data , open government data and data publishing	-
Research problem	-	RQ1
Criteria and techniques	-	RQ2,RQ3
Validation techniques	-	RQ4
Notes	-	-
Other Information	-	-

3.6. Data Extraction

The goal of data extraction is to record in a structured form the information from studies. For this purpose a data extraction form is used, which also included the quality assessments questionnaire. Data extraction and quality assessment were performed together. Data extraction was conducted on each paper separately [Cruzes et al. (2007)]. Table 4 reports the data extraction form we adopted. The information collected in the data extraction step was stored in a project created with the software NVivo[4].

3.7. Synthesis

The synthesis step is based on the methodology for thematic synthesis described by Cruzes and Dybå (2011). This methodology define codes as descriptive labels applied to segments of text from each study. Based on the research questions, we identified an initial set of codes. Following the first set of codes, we performed a second coding based on the content of selected paper. The second set of coding has been done to get more precise content. Later a further cross-validation on codes was performed by the second author based on a random sample from 30% of the paper. After extracting data by collecting them in the form reported in Table 4 and assigning codes to text segments of each paper, the codes were translated into themes. A model of higher-order themes was created to group papers by themes and to obtain an overall picture; afterward research questions were mapped on the themes. To address the research questions we did not only rely on codes but also on extracted information. Mainly, we summarized the information of the extraction form and counted the papers for each theme.

3.8. Threats to the Validity

For the systematic review and its results, we identified two possible threats to the validity. We could be publication biased as some papers belong to multiple study areas. Publication bias refers to the general problem that positive research outcomes are more likely to be published than negative ones[Kitchenham and Charters (2007); Kitchenham and Brereton (2013)].However, we regard this threat as moderate since we took thematic synthesis approach and considered all possible open data quality study areas.

We only considered digital libraries as primary data sources. This may create the possible threats to the identification of primary studies. We followed two additional strategies to further decrease the probability of missing relevant papers. First, we validate our search string by using a bibliographic database (SCOPUS) and also individual publishers in the data sources. This approach leads to a high number of duplicates, which we could, however, reliably identify by sorting the documents alphabetically by their title and authors. Second, an explicit manual search of papers done based on references presented in the paper. By this way we ensure we don't miss any relevant studies.

Another threat to the validity regards the selection and data extraction consistency. To increased the validity of the review results, we have conducted selection, and data extraction in parallel and a cross-checking is done on the outcomes between authors after each phase. In the case of data extraction consistency, the 2nd author performed a second extraction on 30% of the selected papers. Moreover disagreement resolved by discussion until a final decision is achieved.

4. Results

The research questions presented in the previous section will guide the presentation of our results. For each RQ we will first present the results of the data synthesis, then we will discuss the results and explore the relative practical implications.

Figure 2 shows the total number of papers retrieved, during the search phase, from each digital library and the number of selected papers. Also, some papers were excluded during data extraction while others were included by examining the references of previously selected papers.

As can be seen in Fig. 2, from the 9,323 retrieved papers, we first discarded duplicates (by ordering them alphabetically by their title and authors) and studies not published in the English language. After applying the inclusion/exclusion criteria, a total of 2,537 papers were found not to be relevant and for 234 publications we were not

[4] `http://www.qsrinternational.com/products.aspx`

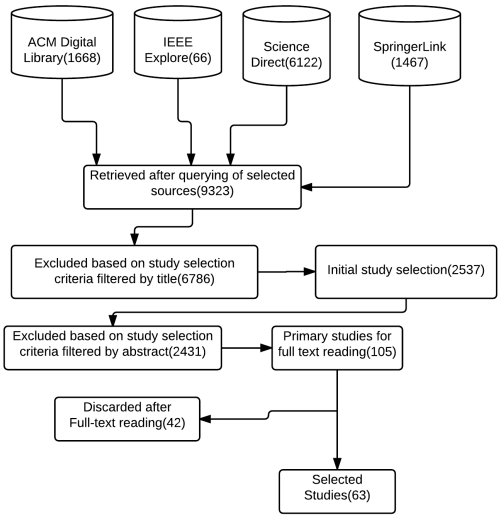

Figure 2: Result of selection phase

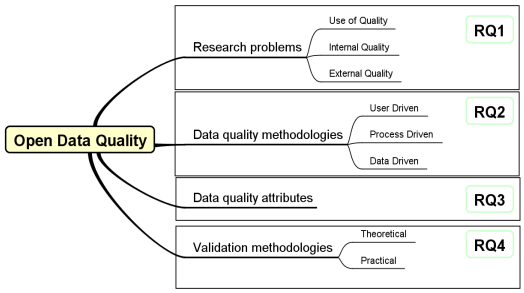

Figure 3: Model of higher order themes of our systematic review

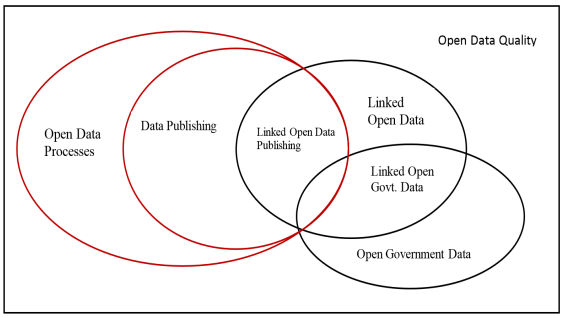

Figure 4: Study Areas

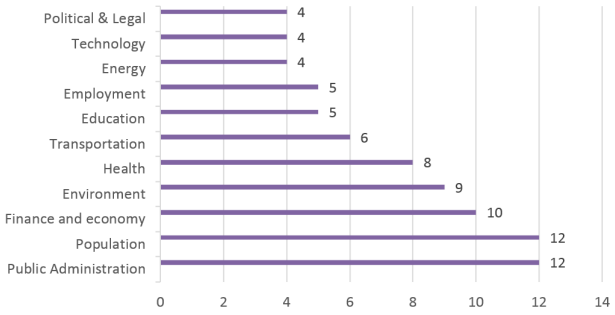

Figure 5: Application Domain

able to obtain a copy of the text. This diminished the pool of papers for full-text reading to 105 papers. In the final pool of primary studies, 63 papers remained after filtering out studies that we found to be irrelevant after assessing the full-text and those that reported on open data quality.

Overall, 105 papers were assessed as an initial paper list. In addition, an evaluation of quality was conducted on each initially selected paper according to quality criteria presented in section 3.5. We used the *median* average quality of each study. To select most of the relevant papers, we considered quality score 0.5 as a threshold. Finally, thought a discussion between the authors, 63 paper were included in the final set of selected papers. The final list of selected papers is presented in appendix Appendix A.

We divided the selected papers into four areas and counted the related papers. The four areas we identified are: open government data (31 papers), open data processes (25 papers), open data publishing (20 papers), and linked open data (15 papers). It is important to notice that the areas are not mutually exclusive, any paper could present multiple study areas (figure 4). However we considered OGD theme as an usecase of open data quality.

Later, based on data extraction (§ 3.6) and synthesis (§ 3.7) phases we mapped each area to to higher order themes presented in figure 3. Each higher order theme corresponds to a particular research question and it contains multiple themes that help to further specify the research questions.

In addition to the general area of the papers, we also looked into the application domain. Here we considered

domain as different categories of the dataset used in open data applications. Open data heterogeneity makes research to focus on the particular domain. However, open data has a diverse set of application domains. It is useful to identify most common application domain used in the various studies. Based on an initial list of domains from [R15], where the authors provide the most common application domain, we identified a total of 11 domains from selected papers. In figure 5 we present the frequency of the application domains in the selected papers.

4.1. RQ1: Research Problems

To address RQ1, we applied the thematic synthesis (§ 3.7) approach to divide the research problems into three higher order themes, such as issues related to Internal Quality and External Quality. In table 5 we present a summary of the research problems with the possible solution from the selected papers. In the table, research problem column presents the selected problems from the studies on each issue. Details of each theme with possible solutions discussed below.

Table 5: Model view of Research problem with Possible solution

Theme **Issues**	**Research Problems**	**Possible Solution(PS)**	**References**
Internal Quality(IQ)			
Metadata	(IQ1) Lack of standard	Participate in the harmonization of metadata between Open Data catalogs(PS1)	[R30], [R43]
	(IQ2) Incomplete metadata	Gather metadata needs from re-users; implement mechanisms to trace the provenance and use of datasets.(PS2)	[R51], [R46], [R43], [R6]
	(IQ3) Too much vocabularies	Using controlled vocabularies (PS3)	[R47]
	(IQ4) Unstructured meta data	Using open data standards (PS4)	[R51], [R46], [R33], [R47], [R43]
Data structure	(IQ5) Data available in heterogeneous formats	Publish datasets in various formats(PS1)	[R54], [R33], [R19], [R23], [R49]
	(IQ6) Out-dated data	Update data frequently with versioning (PS4)	[R49]
	(IQ7) Machine readability	Use non-proprietary and machine processable data format (PS1)	[R19], [R23], [R31]
	(IQ8) Data as primary and bias	Clarify the context of the data creation process (PS2)(PS3)	[R31], [R23], [R49]
	(IQ8) Lack of data completeness	Possible solution is to use a constant or a decision tree to determine the missing data (PS3)	[R54], [R16], [R47], [R19], [R49]
	(IQ9) Lack of data accuracy	Evaluate data to identify semantic and syntactic error in data (PS2)(PS3)	[R54], [R4], [R33], [R23], [R49]
	(IQ10) Insufficient analysis techniques	Support data quality assessment with standard guidelines (PS3)	[R16], [R30]
	(IQ11) Lack of validation techniques	Support appropriate validation techniques (PS3)	[R38], [R51]
External Quality(EQ)			
Publishing	(EQ1)Risk regarding interoperability ,scalability and usability	Provide support using generic schema (PS6)	[R52], [R47], [R55], [R21], [R14], [R56]
	(EQ2)Frequency of updating data	Data provenance(PS4)(PS6)	([R57], [R21], [R62])
	(EQ3)Irregularity of deployed platform	Use standard guideline(PS6)	([R52], [R29], [R50])
	(EQ4)Lack of categorization facilities	Using the data cataloging methodology(PS5)	([R60], [R62])
Access	(EQ5)Risk regarding automated data reading	Use machine processable format(PS7)	([R37], [R34])
	(EQ6)Suitability of data for release	Use data publishing standards(PS6)	([R2], [R52], [R58], [R36], [R62], [R34])
	(EQ7)Lack of API	Support data publishing through an API capable of reporting access and use(PS7)	[R60], [R62]
	(EQ8)Lack of discoverability	Use metadata evaluation to support searchability (PS2)(PS5)(PS6)	[R57], [R55], [R50], [R56]

4.1.1. Internal Quality

This theme pertains the research problem related to data quality in various studies. We identified issues related to metadata together with the data structure as common research problems for implementing and assessing data quality. The most common problem we found is the lack of completeness and accuracy together with unstructured metadata.

Metadata: This issue relates to the lack of metadata information. Metadata is provided to describe the datasets; they are crucial for the retrieval and reuse of datasets.

Data structure: This issue concerns the research problems related to open data structures and methodologies. Here we try to identify research problems based on how different data structures are used in open data applications, issues regarding quality methodologies and validation techniques as well.

Possible Solutions:

(PS1) Use of non-proprietary and machine processable format Many governmental entities still publish data in a large variety of data formats which can also be proprietary [R37][R60]. By using non-proprietary and machine processable formats, government entities can increase accessibility of the data they publish and eventually improve their own accountability [R57].

(PS2) Metadata evaluation It helps to improve interoperability, searchability, and subsequently discoverability in open data applications. A number of efforts in the literature focus on metrics for assessing metadata quality. In [R43] the authors present an implementation of metadata quality metrics on public government data. They evaluate quality by applying five quality metrics, completeness, weighted completeness, accuracy, the richness of information, and accessibility, to three public government data repositories. In [R31] the authors took a user-driven approach where the quality assessment module leverages user-selected meta-data as quality indicators to produce quality assessment scores through user-configured scoring functions.

(PS3) Data quality assessment From traditional data quality methodologies [R30] to semantic web based approach [R25][R19][R33][R34], various techniques can be applied to open data quality assessment. However, these methodologies focus on particular and specific challenges. Quality standards can constitute a guideline to apply data quality assessment in the open data context. For instance, in [R58], [R1], and[R4], the authors used quality standard as a reference for defining a quality assessment methodology. In case of syntactic quality assessment traditional methodologies can be used [R46][R49].

4.1.2. External Quality

This theme explore the issues related to open data publishing. Publishing data regards the method, framework or guidelines used to present information where data can be easily accessible and comprehensible. A publishing platform need to ensure data quality to end users. Publishing open data depends mainly on two issues: how data can be access and the quality of publishing platform. Most common problems we identified is the suitability of data for release together with risk regarding interoperability, scalability and usability.

Publishing: Research problems related to data publishing platform. The quality of a publishing platform depends on how they release the information's to end user. Before releasing data, publishing platforms use various quality assessment methodologies to ensure data integrity and reliability.

Access: This identify issues related to open data access to both humans (end-users) and machines (through re-users). Before publishing data need to process. Here we mainly explore open data accessibility,interoperability and usability issues.

Possible Solutions:

(PS4) Generic schema The heterogeneity of published datasets and their representation is a barrier for open data applications development. Various type of data published in portals such as traffic, touristic, economic, geographical, and environmental data, etc., also they are published in a non-standardized manner, which leads to a large heterogeneity regarding semantics and standards. A number of efforts in the literature approach this challenge by proposing a generic schema. For example, in [R44], the authors propose a minimal schema that is compatible with the predominant data catalog vocabulary and application.

(PS5) Data cataloging Open data catalogs differ widely in scope, terminology, structure, and metadata fields. The authors of [R43] propose a standardized interchange format which could enable machine-readable representations of data catalogs.. With regards to versioning, a solution to the issue is the use of Named Graphs [R10], where the metadata represents the temporal validity of the annotated RDF data. However, this solution is only available for the use of RDF.

(PS6) Publishing standards The main point in external quality is to publish data with proper quality. Also, it has to be discoverable. The discoverability of open data in applications depends much on the quality of the metadata. However, the metadata still requires evaluation for completeness or accuracy. As pointed out in [R21], open data portals success is not only evaluated on the amount of data published, but also on the usability of the data. W3C recommends the use of established open standards and tools, such as XML and RDF as a publishing format [R36]. In the case of EQ1, EQ5, and EQ6 a possible solution would then be publishing data in a machine processable and non-proprietary formats, which follows a publishing standards [R50][R14].

(PS7) Use of API Using REST API, access to updated records can be made easy. All the features offered through the web interface can also be achieved with appropriate API calls [R36]. For example: use API call to get JSON-formatted lists of a sites packages, groups or other objects, get a full JSON representation of a package, resource, etc. In the open data initiatives, utilizing public APIs is understood as one of the principal driving forces for innovation as they enable programmers to explore new uses for the data [R44].

4.2. RQ2: Methods

4.2.1. Results

In RQ2, we focused on identifying the methodologies aimed at assessing data quality. To identify the methods used in the selected papers, we considered the papers describing a data quality assessment and looked at the processes adopted. We identified 16 methods from selected studies and divided them into three categories.

User feedback driven Represents the methods based on user feedback to measure data quality attributes. Usually the measures are obtained by means questionnaires or interviews. The users can be common data consumer or experts.

Using questionnaire: This methods evaluate data quality based on user survey on specific set of questionnaires [R61][R54]. This technique is a practical approach to collect a large amount of data.

Crowd sourcing data quality assessment: Crowd sourcing as a means to handle data quality problems [R51]. This method uses feedback from both expert crowd and common users. Then it process feedback data to apply on different quality assessment criteria. They use expert crowd to discover and classify quality. However, this methodology need human interpretation for particular quality issues.

Data quality rating: This methods based on quality rating such as scorecard [R30] or a five point Likert scale [R18] to provide assessment on data quality. Generally, the source of data is from user feedback. These methodologies provide acceptable parameters and tolerances, as well as a high-level view of the risks associated with data quality issues.

Sieve linked data Quality: In [R31] author present a framework, Sieve which is included as a module in Linked Data Integration Framework (LDIF). The quality assessment module leverages user selected meta-data as quality indicators to produce quality assessment scores through user-configured scoring functions. This methodology use user-configured quality assessment metrics and fusion functions which could provide flexibility. It is also agnostic to provenance and quality vocabularies.

Data-driven Data-driven strategies assess the quality of data by directly evaluating the value of data.

Statistical Distributions: This methods exploit statistical distributions on data such as use of linked data properties and types for evaluating the data quality. In [R46] present two algorithms that evaluate incomplete and noisy Linked Data sets: SDType adds missing type statements, and SDValidate identifies faulty statements. It use statistical distributions of properties and types for enhancing the quality of incomplete and noisy datasets.

Metrics-Driven Framework: In [R54] the authors presents a metric-driven framework for evaluating the inherent quality dimensions of data-sets before they are used. They use quality dimensions specific metrics to apply data quality assessment. This inherent data quality dimensions follows the ISO/IEC 25012 standards. This technique automatically process data using metrics and evaluated based on predictive statistical models for measuring its quality before release [R54]. It also adopted theoretical validation based on Property-based measurement framework.

Using Semantic Web Resources: This methods handled data quality problems using the Semantic Web technology framework, namely using SPARQL on RDF representations [R33][R34]. It mainly handles data quality issues related to the growing amount of data available on the semantic web, namely using SPARQL on RDF representations. Also, use of semantic web reference data to spot incorrect literal values and functional dependency violations.

Italian National Bureau of Census Methodology (ISTAT): [5] It has been designed within the Italian National Bureau of Census to collect and maintain high quality statistical data on Italian citizens and businesses [R30]. This method is strongly focused on formal norms, since it is aimed at regulating data management activities in such a way that their integration can satisfy basic quality requirements [Batini and Scannapieco (2006)]. It provides a variety of simple but effective statistical techniques for quality measurement.

Complete Data Quality Methodology (CDQM): This method is applied by considering existing techniques and tools and integrating them in a framework that can work in both intra- and inter-organizational contexts [R49][R30]. It can also applied to all types of data such as structured, semi-structured and unstructured [Batini and Scannapieco (2006)]. It provide completeness through considering existing techniques and tools and integrating them in a framework. It also flexible since it supports the user in the selection of the most suitable techniques and tools within each phase and in any context [R30].

A Methodology for Information Quality Assessment(AIMQ): The AIMQ method is objective and domain independent technique for quality evaluation [R30]. This is due to the fact its the only information quality methodology which uses benchmarking [Lee et al. (2002)]. It is also objective and domain dependent technique for quality assessment by when using gab analysis techniques as benchmarking.

Meta data quality: In [R43] present a method for meta data quality assessment. They use set of metrics on specific data attributes for evaluating metadata quality. Automatic approach and can apply to a diverse set of data-set. Metrics needs validation to be considered sufficiently viable and reliable.

Process-driven Process-driven strategies assess the quality by evaluating the process or by redesigning the processes that creates the data.

Total Data Quality Management (TDQM): This method devise a practical methods for business and industry to improve data quality [6][R30]. It focus on the customer to understand and evaluation data quality [R49]. In this method considering how it process information [R49] two key steps considered (1) to clearly define what an organization means by data quality and (2) to develop metrics that measure data quality dimensions and that are linked to the organization's goals and objectives [Kovac et al. (1997)][R49]. It supports the entire end-to-end quality assessment process, from requirements analysis to implementation. It also identifies IQ problems through measurement of quality using information quality criteria.

Web Information Quality Assessment Framework (WIQA) [7]: This method uses a set of software components to employ a wide range of different information quality assessment policies to filter information from the Web [Bizer (2007)] [R19]. It provide flexible representation of information together with quality-related meta-information.

Hybrid Approach: In [R16] author present a approach to assessing DQ, which dynamically configure an assessment technique as needed while leveraging the best practices from existing assessment techniques. This

[5] `http://www.istat.it/en/tools/methods-and-it-tools`
[6] `http://web.mit.edu/tdqm/`
[7] `http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/`

approach highly depends on the activity of data processing to select and apply quality assessment technique. This technique identify DQ assessment based on activities from existing techniques in a way that meets differing requirements.

Test-driven Evaluation of linked data quality: In [R25] the authors present a method for assessing the quality of linked data resources, based on a formalization of bad smells and data quality problems. Support data quality integrity constraints that are represented in SPRQL query templates. Use data quality test patterns (DQTP) to reveal a substantial amount of data quality issues.

Using domain ontologies: This approach uses the knowledge modeled in ontologies. Basically the ontology structure is used to provide correction suggestions for invalid data, identify duplicates, and to store data quality annotations at schema and instance level [R45]. Provide support for consistency checking, duplicate detection, and the seamless possibility of metadata annotation.

4.2.2. Analysis and Discussion

One data quality approach can be adopted in multiple study areas. Here most of the studies belong to the second category. Also In table 6 we present each methodology advantages and disadvantages from selected papers.

By exploring all possible ODQ study areas, we obtain various quality assessment methodologies. However these data quality methodologies either standalone or implemented as modules in ODQ. To understand the best practice for ODQ at first we try to analyze each of the methodologies advantages and disadvantages 6. Then we explored tools implemented for ODQ. In table 7 we present ODQ evaluation based on study areas, theme, automation and tool support. Study areas column map quality methodologies to specific areas present in the paper. Researchers used traditional data quality(DQ) techniques in open data. Considering this aspect, in the table we added another area DQ which present traditional data quality.

Table 6: Data quality methodologies

Methodology	Advantages	Disadvantages
Using questionnaire	A practical approach to collect a large amount of data. The result of questionnaires easily quantifies using expert or software tools.	Lack of validation. Hard to predict completeness and accuracy of survey data.
Crowd sourcing data quality assessment	Use of expert crowd to discover and classify quality, Human interpretation for particular quality issues.	Appropriate validation methodology requires for quality control.
Data quality rating	Define acceptable parameters and tolerances, high-level view of the risks associated with data quality issues.	Definition of metrics is separated from their contextual use; Inflexible Design; Scorecard relevancy is based on a hierarchical roll-up of metrics.
Sieve linked data Quality	User-configured quality assessment metrics and fusion functions provide flexibility. It is agnostic to provenance and quality vocabularies.	Assess to the quality of the integrated data rather than original data source. Lack of data diversity in the implication.
Statistical Distributions	Use statistical distributions of properties and types for enhancing the quality of incomplete and noisy dataset, scale to large datasets.	Need Validation for assesses the correctness of statements, purely knowledge base.
Metrics-Driven Framework	Automatically processed using metrics and evaluated based on predictive statistical models for measuring its quality before release [R54]. Adopted theoretical validation based on Property-based measurement framework.	This approach limited with fixed quality dimensions. Need for subjective perception regarding adopted inherent dimension together with validating metrics and perceived quality.
ISTAT	Detect quality issues from a data integration perspective. It provides a variety of simple but effective statistical techniques for quality measurement. It also provides tools for the most relevant data cleaning activities.	It strongly focused on formal norms, most common types of data which cause integrity issues.
CDQM	It provide completeness through considering existing techniques and tools and integrating them in a framework. It also flexible since it supports the user in the selection of the most suitable techniques and tools within each phase and in any context [R30].	It focuses only on obtaining a quantitative assessment of the extent to which business processes are affected by the wrong information.

AIMQ	It is objective and domain dependent technique for quality evaluation by using gab analysis techniques as benchmarking.	The literature does not provide any description of the benchmarking database that is required for the application of the methodology.
DaQuinCIS	Provide support to data trustworthiness. It also uses the concept of data quality certification to manage corresponding quality measures that are exchanged among organizations.	Domain-specific and depends on aggregation on the database before applying quality assessment.
Using Semantic Web Resources	Handle data quality issues related to the growing amount of data available on the semantic web, namely using SPARQL on RDF representations. Also, use of semantic web reference data to spot incorrect literal values and functional dependency violations.	Need of pre-processing dataset or use of knowledge base. Also uncertainty about the quality of the used knowledge bases available in the Semantic Web.
Meta data quality	Automatic approach and can apply to a diverse set of dataset.	Metrics needs validation to be considered sufficiently viable and reliable.
TDQM	It supports the entire end-to-end quality improvement process, from requirements analysis to implementation. It identifies information quality(IQ) problems through measurement of quality using IQ criteria.	Domain-specific and applicable to only technical, economic and organizational phases of business operations. It relates quality issues to improvement techniques.
WIQA - Information Quality Assessment Framework	Flexible representation of information together with quality-related meta-information. Support for different Information filtering policies and explain filtering Decisions [R19].	Based on quality assessment according to task-specific criteria.
Hybrid Approach	Identify DQ assessment based on activities from existing techniques in a way that meets differing requirements. For the DQ assessment, this affords savings in costs, time, and resources that organizations are always striving to contain.	Needed of validation activities through expert and organization trials. Lack of standards methods to ensure the integrity of activities.
Test-driven Evaluation of linked data quality	Support data quality integrity constraints that are represented in SPRQL query templates. Use data quality test patterns (DQTP) to reveal a substantial amount of data quality issues.	Lack of validator for web service format. Missing the use of templates and bindings to fix problems efficiently.
Using domain ontologies	Provide support for consistency checking, duplicate detection, and the seamless possibility of metadata annotation. Define a shared vocabulary for improved interoperability, and data quality management.	Domain-specific and lack of verification of applicability for large-scale datasets.

Table 7: Open Data quality methodologies evaluation

Methodology	Study Areas	Theme	Degree of automation	Tool support
Using questionnaire	OD,OGD	User Driven	Manual	No
Crowd sourcing data quality assessment	OD,LOD	User Driven	Semi-automated	Yes
Data quality rating	OD,OGD	User Driven	Semi-automated	No
Sieve linked data Quality	LOD	User Driven	Semi-automated	Yes
Statistical Distributions	LOD	Data Driven	Semi-automated	Yes
Metrics-Driven Framework	LOD	Data Driven	Semi-automated	No
ISTAT	DQ	Data Driven	Semi-automated	No
CDQM	DQ	Data Driven	Semi-automated	No
AIMQ	DQ	Data Driven	Semi-automated	No
DaQuinCIS	DQ	Data Driven	Semi-automated	No
Using Semantic Web Resources	LOD	Data Driven	Semi-automated	Yes

Meta data quality	LOD	Data Driven	Automated	Yes
TDQM	DQ	Process Driven	Semi-automated	No
WIQA - Information Quality Assessment Framework	LOD	Process Driven	Semi-automated	Yes
Hybrid Approach	DQ	Process Driven	Semi-automated	Yes
Test-driven Evaluation of linked data quality	LOD	Process Driven	automated	Yes
Using domain ontologies	LOD	Process Driven	Semi-automated	Yes

4.3. RQ3: Quality Measurement

4.3.1. Results

The goal of RQ3 is to identify how data quality was assessed. Typically the definition of quality measures implies the identification of the general characteristic –abstract constructs– and then the actual definition of the metrics.

First of all, we focused on the standards and framework, which are used in the selected studies, and on the quality characteristics they propose.

ISO/IEC 25012 SQuaRE is a set of International Standards which consist of different divisions [ISO/IEC 25012 (2008)]. In our review, four papers used the data quality model based on the ISO/IEC 25012 which proposes a model that defines fifteen characteristics considered from two points of view: inherent and system dependent.

Inherent data quality refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions. The characteristics in this set are specifically: Accuracy, Completeness, Consistency, Credibility and Currentness.

System dependent data quality refers to the degree to which data quality is attained and preserved within a computer system when data is used under specified conditions. The characteristics in this set are specifically: Availability, Portability and Recoverability.

In addition the standard defines a set of characteristics for both points of view that is composed of Accessibility, Compliance, Confidentiality, Efficiency, Precision, Traceability, and Understandability.

The data quality assessment study by Behkamal et al.[R54] uses ISO/IEC 25012 standard to propose 5 inherent data quality attributes for linked open data. They used a metrics driven quality assessment apporach. They proposed inherent quality characteristics of LOD are: Syntactic Accuracy, Semantic Accuracy, Completeness, Consistency and Uniqueness

OGD 8 principles In our study we considered OGD as a use case for OD process. In the review 11 paper use OGD 8 principles as the base for defining data quality criteria. Which mainly related to use of data quality theme. On December 7-8, 2007, thirty open government advocates gathered in Sebastopol, California and wrote a set of eight principles of open government data[8]. So government data shall be considered open if it is made public in a way that complies with the principles of (1) Complete, (2) Primary, (3) Timely, (4) Accessible, (5) Machine processable, (6) Non-discriminatory, (7) Non-proprietary, and (8) License-free.

Web Information Quality assessments[9] this framework uses quality-based information filtering policies together with quality-related meta-information and set of software component. Using quality-based information filtering policies this framework evaluate multiple information quality dimension [Bizer (2007)] [R19]: Accuracy, Timeliness, Relevancy, Accessible, Interpretability and Believability.

Other Quality of data characteristics according to International Software Testing Qualification Board (ISTQB) [10] include aspects of Currency, Relevance, Consistency, Correctness and Completeness.

To identify related data quality attributes one of the main study is Portal Data quality model (PDQM) [R52], which focuses on data quality in Web portals and presents a total of 42 data quality criteria.

[8] `http://opengovdata.org/`

[10] `http://www.istqb.org`

In addition, we focused on the specific quality characteristics that are present in the selected papers we report a list of the data quality dimensions we found in an online appendix[11]. Moreover, Based on application domain studies reported various set of data attributes. We compiled a list of metrics from selected studies and presented as an online appendix [12] .

4.3.2. Analysis and Discussion

To apply suitable data quality methodology we need to identify data attributes. Based on the quality dimensions list from the selected papers we tried to identify those ceriteras used in ODQ. We classified the data attributes according to ISO/IEC 25012 as being Inherent and System Dependent.

Following we present data quality criteria which are considered by most efforts in the literature.

Inherent data quality Data quality measurement approaches presented in studies mostly focused on inherent data quality with intrinsic potential. Intrinsic, which denotes that data have quality in their own right [R1].

Accuracy It means the extent to which a data or metadata record correctly describes the respective information [R54]. On metadata, this quality dimension directly affects the discoverability of datasets, as good quality metadata enables the dataset to be easily discovered by data consumers [R43].

Completeness This quality dimension deals with the number of completed fields in a data or metadata record [R13][R15]. Thus, a record is considered complete only when the record contains all the information required to have the ideal representation of the described data. The completeness of the metadata, like accuracy, also directly affects the discoverability of datasets.

Consistency The consistency of record fields depends on whether they follow a consistent syntactical format, without contradiction or discrepancy within the entire catalogue of metadata. Apart from the syntactical form, a field is considered to be compatible if the respective values are selected from a fixed set of options [R52][R49].

System dependent It can be refers to domain specific criteria for an application. It emphasizes the importance of the role of systems; that is, the system must be accessible but secure [R1].

Usability This is the most ”generic” quality criterion [R13]. By usability, it means how easily the published data can be used [R54]. It is the most generic as it depends on other quality dimensions whether the published data is usable. It is directly related to what degree the data is: Accessible, Open, Interoperable, Complete and Discoverable.

Timeliness It means the extent to which the data or metadata is up to date[R49]. The organizational approach affects the timeliness of the published data, which depends on whether the data provider directly or indirectly provides the data [R18].

Accessibility This quality dimension is affected by the format in which the data is published, the search tool used, and the discoverability of the dataset [R52]. The quality accessibility dimension has two measures. The cognitive accessibility defines how easy it is for a data consumer to understand the published information. Several aspects of the data affect the cognitive accessibility, such as the ambiguity of the data[R18]. The second measure is the psychological or logical accessibility, which can be defined as the ease with which the relevant dataset is discovered through a data catalog or repository [R46].

Relevancy This dimension focuses the extent to which data are applicable and helpful for users needs [R52]. Relevancy relates to the usefulness of data to generate values.

4.4. RQ4: Validation Methodologies

One of the main issues in open data quality is the lack of method validation techniques. Validation is used to determine if the data quality method satisfies the quality criteria and usage requirements. RQ4 explores the various validation metrics, frameworks and methods used in the selected studies. We divided validation methodologies in two

[11] `http://softeng.polito.it/rifat/AppendixB.pdf`
[12] `http://softeng.polito.it/rifat/AppendixC.pdf`

theme and selected papers based on this theme. We found only six papers that presented studies related to validation methodologies in details. Following we present a short discussion on each validation approaches from selected papers.

In [R43] the authors present a metadata quality assessment using quality metrics on public government data repositories. Though the authors didn't directly implement method validation, they pointed out the need for metrics validation. In the paper they present three possible validation technique: (1) measure the correlation between the value of the quality metrics and the quality as assessed by a human expert, (2) apply the quality metadata on two different sets of metadata to show their discriminatory character, and (3) filter metadata records based on poor quality results.

In [R38] the authors present the quality and validity of a selected OGD through relating it to another dataset and the use of knowledge data discovery (KDD). They use the dataset from Austrian OGD initiative, Vienna stock of trees. They correlate the average annual growth of the stem perimeter of trees with the weather conditions, namely the average. They use the KDD process consisting of the phases: (a) selection of data, (b) pre-processing, (c) data transformation, (d) Data Mining (i.e., statistics and correlation analysis), and (e) interpretation and evaluation of results. The last two phases of the KDD process comprise the application of data mining techniques to analyze the data and interpret the results. They applied statistical methods of correlation analysis. They calculated the Pearson correlation coefficients and created scatter plots for good and significant correlations. Using the correlation data from the dataset and using the domain specific study they show the data as valid.

In [R54] the authors adopted a theoretical validation based on a property-based measurement framework. They considered that measures were theoretically analyzed within the context of measurement theory. They reported two main groups of frameworks for the theoretical validation of metrics in the literature. The first group consists of frameworks directly based on measurement theory principles. The second group expresses the desirable properties of the numerical relational system that need to be satisfied by the metrics. In this work, they have examined the properties of metrics following the second group. i.e. using the property-based measurement framework [Briand et al. (1996)]. This framework provides five types of metrics including size, length, complexity, coupling and cohesion and offers a set of desirable properties for each of these types.

In [R15] a validation technique is reported for open government data benchmarking. It reports a theoretical validation based on a benchmarking conceptualization for e-government using Activity Theory. The conceptualization is based on the mapping of eight generic concepts of the Activity Theory: Activity, Subject, Artifact, Object, Outcome, Community, Roles and Rules into the corresponding concepts in the e-government benchmarking domain.

In [R57] the authors focus on assessing open government data using maturity model. They used expert opinions for the validation process. In particular they sent a OD-MM questionnaire to various set of experts and processed their comments through online tools. The analysis of the comments lead to an improvement proposal of the model design.

Paper [R47] presents a study related to open government data catalogs and their quality perspective. The authors provide an approach to validation based on the OGD catalog records. They identified a process driven by manual review techniques and automated validation techniques. The manual review focused on using an expert for the validation process. Whereas the automatic review procedure was provided through the data-cataloging tool with features that can help to automate the validation, e.g. checking if the required attributes are non-empty or if the provided links are not corrupted.

5. Conclusions

In this paper, we gave an overview of the methods related to open data quality, using the Systematic Literature Review instrument. Overall 63 papers were selected, based on the protocol designed for this study. Among the selected papers, the majority of studies (31 papers) focus on Open Governement Data. Concerning the application domains, public administration – open data published by the government – and population are the most common domains.

Initially, we focused on identifying the most relevant research problems reported by the selected papers. We applied the thematic synthesis approach and presented the research problems based on internal (IQ) and external (EQ). The internal quality theme summarizes the issues based on the technical perspective. The majority of studies mentioned unstructured metadata, lack of accuracy and incompleteness as central issues regarding the internal open data quality.

We were able to identify 16 distinct data quality methodologies. The methodologies in distinct studies cover different quality issues and support diverse sets of features. We divided these methodologies into three main categories: user, data, and process driven. Most of the selected methodologies fall into the category of data-driven approach.

It is important to identify what quality methodology is suitable for a specific open data application. By selecting the suitable methodology it is possible to identify the data attributes becomes in a straightforward way. Thus, the next step in our analysis focused on identifying the data quality dimension used in the selected studies. In particular we reported not only the quality dimensions explicitly addressed in the studies but also the mentioned data quality standards and framework.

Another important factor in data quality assessment is the empirical validation of methodologies. We found the there is a lack of research to validate different methodological approaches. Also less focus was given to the development of platforms to make validation feasible. We only found six papers addressing data quality validation techniques.

Based on the research questions and the summaries of results, we present a few conclusing remarks:

RQ1 We identified and explored research problems related to internal and external data quality. Open data initiatives give increasingly focus to external quality related issues, we found but less focus was given to internal quality issues.

RQ2 The majority of open data quality studies are centered on a domain for which they try to address specific issues. This results into a lack of flexibility in the open data quality methodologies; as a consequence, it is hard to apply them to research problems different from the original context. We foresee a set of open data quality guidelines to explore several different research problems, and a platform using generalized data attributes and metrics.

RQ3 Most of the studies focus on few common data attributes such as accuracy and completeness but failed to explore the complete set of the data attributes defined in quality standards.

RQ4 Fewer studies focus on appropriate methodology validation. More focus is needed in this research area.

The results of this review encourage further research on the evaluation of methodology validation, particularly on the conception of structured guidelines which support practitioners in the endeavor of measuring, evaluating, and validating the open data quality measurement.

References

Batini, C. and Scannapieco, M. (2006), *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Bizer, C. (2007), *Quality-Driven Information Filtering- In the Context of Web-Based Information Systems*, VDM Verlag.

Briand, L., Morasca, S. and Basili, V. (1996), 'Property-based software engineering measurement', *Software Engineering, IEEE Transactions on* **22**(1), 68–86.

Cruzes, D., Mendonca, M., Basili, V., Shull, F. and Jino, M. (2007), Extracting information from experimental software engineering papers, *in* 'Chilean Society of Computer Science, 2007. SCCC '07. XXVI International Conference of the', pp. 105–114.

Cruzes, D. S. and Dybå, T. (2011), Recommended steps for thematic synthesis in software engineering, *in* 'Proceedings of the 2011 International Symposium on Empirical Software Engineering and Measurement', ESEM '11, IEEE Computer Society, Washington, DC, USA, pp. 275–284.
URL: *http://dx.doi.org/10.1109/ESEM.2011.36*

ISO/IEC 25012 (2008), 25012:2008 – software engineering – software product quality requirements and evaluation (square) – data quality model, Technical report, ISO/IEC.
URL: *http://iso25000.com/index.php/en/iso-25000-standards/iso-25012*

Kitchenham, B. and Brereton, P. (2013), 'A systematic review of systematic review process research in software engineering', *Information and software technology* **55**(12), 2049–2075.

Kitchenham, B. and Charters, S. (2007), Guidelines for performing systematic literature reviews in software engineering, Technical report, Technical report, EBSE Technical Report EBSE-2007-01.

Kovac, R., Lee, Y. W. and Pipino, L. (1997), Total data quality management: The case of iri., *in* 'IQ', pp. 63–79.

Lee, Y. W., Strong, D. M., Kahn, B. K. and Wang, R. Y. (2002), 'Aimq: A methodology for information quality assessment', *Inf. Manage.* **40**(2), 133–146.
URL: *http://dx.doi.org/10.1016/S0378-7206(02)00043-5*

Pipino, L. L., Lee, Y. W. and Wang, R. Y. (2002), 'Data quality assessment', *Commun. ACM* **45**(4), 211–218.
URL: *http://doi.acm.org/10.1145/505248.506010*

Sayogo, D. and Pardo, T. (2012), Exploring the motive for data publication in open data initiative: Linking intention to action, *in* 'System Science (HICSS), 2012 45th Hawaii International Conference on', pp. 2623–2632.

Appendix A. Selected Papers

References of the SLR where R represents reference no. of selected papers .

R	Publication details
R1	Moraga, C., Moraga, M.., Calero, C. and Caro, A., 2009, August. SQuaRE-aligned data quality model for web portals. In 2009 Ninth International Conference on Quality Software (pp. 117-122). IEEE.
R2	Ruckhaus, E. and Vidal, M.E., 2012, May. LiQuate-estimating the quality of links in the linking open data cloud. In International Workshop on Resource Discovery (pp. 56-82). Springer Berlin Heidelberg.
R3	Al-Khalifa, H.S., 2013, September. A Lightweight Approach to Semantify Saudi Open Government Data. In 2013 16th International Conference on Network-Based Information Systems (pp. 594-596). IEEE.
R4	Behkamal, B., Bagheri, E., Kahani, M. and Sazvar, M., 2014, October. Data accuracy: What does it mean to LOD?. In Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on (pp. 80-85). IEEE.
R5	Corradi, A., Foschini, L. and Ianniello, R., 2014, June. Linked data for Open Government: The case of Bologna. In 2014 IEEE Symposium on Computers and Communications (ISCC) (pp. 1-7). IEEE.
R6	Michelfeit, J. and Necask, M., 2012, July. Linked open data aggregation: conflict resolution and aggregate quality. In Computer Software and Applications Conference Workshops (COMPSACW), 2012 IEEE 36th Annual (pp. 106-111). IEEE.
R7	Segura, A.M., Cuadrado, J.S. and de Lara, J., 2014, September. ODaaS: Towards the Model-Driven Engineering of Open Data Applications as Data Services. In EDOC Workshops (pp. 335-339).
R8	Hoxha, J. and Brahaj, A., 2011, September. Open government data on the web: A semantic approach. In Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference on (pp. 107-113). IEEE.
R9	Hendler, J., Holm, J., Musialek, C. and Thomas, G., 2012. US government linked open data: semantic. data. gov. IEEE Intelligent Systems, 27(3), pp.0025-31.
R10	Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H. and Hall, W., 2012. Linked open government data: Lessons from data. gov. uk. IEEE Intelligent Systems, 27(3), pp.16-24.
R11	Erickson, J.S., Viswanathan, A., Shinavier, J., Shi, Y. and Hendler, J.A., 2013. Open government data: a data analytics approach. IEEE Intelligent Systems, 5(28), pp.19-23.
R12	Heitmann, B., Cyganiak, R., Hayes, C. and Decker, S., 2012. An empirically grounded conceptual architecture for applications on the web of data. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(1), pp.51-60.
R13	Loureno, R.P., 2015. An analysis of open government portals: A perspective of transparency for accountability. Government Information Quarterly, 32(3), pp.323-332.
R14	Lee, G. and Kwak, Y.H., 2012. An open government maturity model for social media-based public engagement. Government Information Quarterly, 29(4), pp.492-503.
R15	Veljkovi, N., Bogdanovi-Dini, S. and Stoimenov, L., 2014. Benchmarking open government: An open data perspective. Government Information Quarterly, 31(2), pp.278-290.
R16	Woodall, P., Borek, A. and Parlikad, A.K., 2013. Data quality assessment: the hybrid approach. Information & management, 50(7), pp.369-382.
R17	Conradie, P. and Choenni, S., 2014. On the barriers for local government releasing open data. Government Information Quarterly, 31, pp.S10-S17.
R18	Detlor, B., Hupfer, M.E., Ruhi, U. and Zhao, L., 2013. Information quality and community municipal portal use. Government Information Quarterly, 30(1), pp.23-32.
R19	Bizer, C. and Cyganiak, R., 2009. Quality-driven information filtering using the WIQA policy framework. Web Semantics: Science, Services and Agents on the World Wide Web, 7(1), pp.1-10.
R20	Kontokostas, D., Zaveri, A., Auer, S. and Lehmann, J., 2013, October. Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. In International Conference on Knowledge Engineering and the Semantic Web (pp. 265-272). Springer Berlin Heidelberg.
R21	Ding, L., Lebo, T., Erickson, J.S., DiFranzo, D., Williams, G.T., Li, X., Michaelis, J., Graves, A., Zheng, J.G., Shangguan, Z. and Flores, J., 2011. TWC LOGD: A portal for linked open government data ecosystems. Web Semantics: Science, Services and Agents on the World Wide Web, 9(3), pp.325-333.
R22	Wijnhoven, F., Ehrenhard, M. and Kuhn, J., 2015. Open government objectives and participation motivations. Government information quarterly, 32(1), pp.30-42.
R23	Abu-Shanab, E.A., 2015. Reengineering the open government concept: An empirical support for a proposed model. Government Information Quarterly, 32(4), pp.453-463.
R24	Galiotou, E. and Fragkou, P., 2013. Applying linked data technologies to Greek open government data: a case study. Procedia-Social and Behavioral Sciences, 73, pp.479-486.
R25	Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R. and Zaveri, A., 2014, April. Test-driven evaluation of linked data quality. In Proceedings of the 23rd international conference on World Wide Web (pp. 747-758). ACM.
R26	Zuiderwijk, A. and Janssen, M., 2014, June. The negative effects of open government data-investigating the dark side of open data. In Proceedings of the 15th Annual International Conference on Digital Government Research (pp. 147-152). ACM.
R27	Verma, N. and Gupta, M.P., 2013, October. Open government data: beyond policy & portal, a study in Indian context. In Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance (pp. 338-341). ACM.
R28	Bizer, C., Heath, T. and Berners-Lee, T., 2009. Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts, pp.205-227.
R29	Kalampokis, E., Tambouris, E. and Tarabanis, K., 2011, August. Open government data: A stage model. In International Conference on Electronic Government (pp. 235-246). Springer Berlin Heidelberg.
R30	Batini, C., Cappiello, C., Francalanci, C. and Maurino, A., 2009. Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3), p.16.
R31	Mendes, P.N., Mhleisen, H. and Bizer, C., 2012, March. Sieve: linked data quality assessment and fusion. In Proceedings of the 2012 Joint EDBT/ICDT Workshops (pp. 116-123). ACM.
R32	Gonzlez, J.C., Garcia, J., Corts, F. and Carpy, D., 2014, June. Government 2.0: a conceptual framework and a case study using Mexican data for assessing the evolution towards open governments. In Proceedings of the 15th Annual International Conference on Digital Government Research (pp. 124-136). ACM.
R33	Frber, C. and Hepp, M., 2010, October. Using semantic web resources for data quality management. In International Conference on Knowledge Engineering and Knowledge Management (pp. 211-225). Springer Berlin Heidelberg.
R34	Alani, H., Dupplaw, D., Sheridan, J., OHara, K., Darlington, J., Shadbolt, N. and Tullo, C., 2007. Unlocking the potential of public sector information with semantic web technology. In The semantic web (pp. 708-721). Springer Berlin Heidelberg.
R35	Stvilia, B., Gasser, L., Twidale, M.B. and Smith, L.C., 2007. A framework for information quality assessment. Journal of the American society for information science and technology, 58(12), pp.1720-1733.
R36	Kalampokis, E., Tambouris, E. and Tarabanis, K., 2011. A classification scheme for open government data: towards linking decentralised data. International Journal of Web Engineering and Technology, 6(3), pp.266-285.
R37	Maali, F., Cyganiak, R. and Peristeras, V., 2012, May. A publishing pipeline for linked government data. In Extended Semantic Web Conference (pp. 778-792). Springer Berlin Heidelberg..
R38	Radl, W., Skopek, J., Komendera, A., Jger, S. and Mdritscher, F., 2013, September. And Data for All: On the Validity and Usefulness of Open Government Data. In Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies (p. 29). ACM.

R39 Prieto, L.M., Rodrguez, A.C. and Pimiento, J., 2012, October. Implementation framework for open data in Colombia. In Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance (pp. 14-17). ACM.

R40 de Mendona, R.R., da Cruz, S.M.S., De La Cerda, J.F., Cavalcanti, M.C., Cordeiro, K.F. and Campos, M.L.M., 2013, June. LOP: capturing and linking open provenance on LOD cycle. In Proceedings of the Fifth Workshop on Semantic Web Information Management (p. 3). ACM.

R41 Hartig, O. and Zhao, J., 2009, October. Using web data provenance for quality assessment. In Proceedings of the First International Conference on Semantic Web in Provenance Management-Volume 526 (pp. 29-34). CEUR-WS. org.

R42 Zuiderwijk, A., Janssen, M. and Parnia, A., 2013, June. The complementarity of open data infrastructures: an analysis of functionalities. In Proceedings of the 14th Annual International Conference on Digital Government Research (pp. 166-171). ACM.

R43 Reiche, K.J. and Hfig, E., 2013, July. Implementation of metadata quality metrics and application on public government data. In Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual (pp. 236-241). IEEE.

R44 dos Santos Brito, K., da Silva Costa, M.A., Garcia, V.C. and de Lemos Meira, S.R., 2014, June. Brazilian government open data: implementation, challenges, and potential opportunities. In Proceedings of the 15th Annual International Conference on Digital Government Research (pp. 11-16). ACM.

R45 Brggemann, S. and Grning, F., 2009. Using ontologies providing domain knowledge for data quality management. In Networked Knowledge-Networked Media (pp. 187-203). Springer Berlin Heidelberg.

R46 Paulheim, H. and Bizer, C., 2014. Improving the quality of linked data using statistical distributions. International Journal on Semantic Web and Information Systems (IJSWIS), 10(2), pp.63-86.

R47 Kuera, J., Chlapek, D. and Neask, M., 2013, August. Open government data catalogs: Current approaches and quality perspective. In International Conference on Electronic Government and the Information Systems Perspective (pp. 152-166). Springer Berlin Heidelberg.

R48 Sandoval-Almazan, R. and Gil-Garcia, J.R., 2014, September. Towards an evaluation model for open government: A preliminary proposal. In International Conference on Electronic Government (pp. 47-58). Springer Berlin Heidelberg.

R49 Floridi, L. and Illari, P. eds., 2014. The philosophy of information quality (Vol. 358). Springer.

R50 Villazn-Terrazas, B., Vilches-Blzquez, L.M., Corcho, O. and Gmez-Prez, A., 2011. Methodological guidelines for publishing government linked data. In Linking government data (pp. 27-49). Springer New York.

R51 Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S. and Lehmann, J., 2013, October. Crowdsourcing linked data quality assessment. In International Semantic Web Conference (pp. 260-276). Springer Berlin Heidelberg.

R52 Calero, C., Caro, A. and Piattini, M., 2008. An applicable data quality model for web portal data consumers. World Wide Web, 11(4), pp.465-484.

R53 Knap, T., Neask, M. and Svoboda, M., 2012, September. A framework for storing and providing aggregated governmental linked open data. In International Conference on Electronic Government and the Information Systems Perspective (pp. 264-270). Springer Berlin Heidelberg.

R54 Behkamal, B., 2014, May. Metrics-driven framework for lod quality assessment. In European Semantic Web Conference (pp. 806-816). Springer International Publishing.

R55 Shvaiko, P., Farazi, F., Maltese, V., Ivanyukovich, A., Rizzi, V., Ferrari, D. and Ucelli, G., 2012, November. Trentino government linked open geo-data: A case study. In International Semantic Web Conference (pp. 196-211). Springer Berlin Heidelberg.

R56 van der Waal, S., Wcel, K., Ermilov, I., Janev, V., MiloÅevi, U. and Wainwright, M., 2014. Lifting open data portals to the data web. In Linked Open Data–Creating Knowledge Out of Interlinked Data (pp. 175-195). Springer International Publishing.

R57 Solar, M., Concha, G. and Meijueiro, L., 2012, September. A model to assess open government data in public agencies. In International Conference on Electronic Government (pp. 210-221). Springer Berlin Heidelberg.

R58 Debattista, J., Lange, C. and Auer, S., 2014, April. daQ, an Ontology for Dataset Quality Information. In LDOW.

R59 Kalampokis, E., Tambouris, E. and Tarabanis, K., 2013, September. Linked open government data analytics. In International Conference on Electronic Government (pp. 99-110). Springer Berlin Heidelberg.

R60 Hchtl, J. and Reichstdter, P., 2011, August. Linked open data-a means for public sector information management. In International Conference on Electronic Government and the Information Systems Perspective (pp. 330-343). Springer Berlin Heidelberg.

R61 Alexopoulos, C., Zuiderwijk, A., Charapabidis, Y., Loukis, E. and Janssen, M., 2014, September. Designing a second generation of open data platforms: Integrating open data and social media. In International Conference on Electronic Government (pp. 230-241). Springer Berlin Heidelberg.

R62 Piprani, B. and Ernst, D., 2008, November. A model for data quality assessment. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 750-759). Springer Berlin Heidelberg.

R63 Martin, S., Foulonneau, M., Turki, S. and Ihadjadene, M., 2013, June. Open data: Barriers, risks and opportunities. In Proceedings of the 13th European Conference on eGovernment: ECEG (pp. 301-309).