Knowledge Graph Completeness: A Systematic Literature Review

Subhi Issa, Onaopepo Adekunle, Fayçal Hamdi, Samira Sisaid-Cherfi, Michel Dumontier, Amrapali Zaveri

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 29, 2020, accepted January 17, 2021, date of publication February 2, 2021, date of current version March 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3056622

Knowledge Graph Completeness: A Systematic Literature Review

SUBHI ISSA[1], ONAOPEPO ADEKUNLE[2], FAYÇAL HAMDI[1], SAMIRA SI-SAID CHERFI[1], MICHEL DUMONTIER[2], AND AMRAPALI ZAVERI[2]

[1]Center for Studies and Research in Computer Science and Communication (CEDRIC), Conservatoire National des Arts et Métiers, 75003 Paris, France
[2]Institute of Data Science, Maastricht University, 6229 GT Maastricht, The Netherlands

Corresponding author: Subhi Issa (subhi.issa@cnam.fr)

ABSTRACT The quality of a Knowledge Graph (also known as Linked Data) is an important aspect to indicate its fitness for use in an application. Several quality dimensions are identified, such as accuracy, completeness, timeliness, provenance, and accessibility, which are used to assess the quality. While many prior studies offer a landscape view of data quality dimensions, here we focus on presenting a systematic literature review for assessing the completeness of Knowledge Graph. We gather existing approaches from the literature and analyze them qualitatively and quantitatively. In particular, we unify and formalize commonly used terminologies across 56 articles related to the completeness dimension of data quality and provide a comprehensive list of methodologies and metrics used to evaluate the different types of completeness. We identify seven types of completeness, including three types that were not previously identified in previous surveys. We also analyze nine different tools capable of assessing Knowledge Graph completeness. The aim of this Systematic Literature Review is to provide researchers and data curators a comprehensive and deeper understanding of existing works on completeness and its properties, thereby encouraging further experimentation and development of new approaches focused on completeness as a data quality dimension of Knowledge Graph.

INDEX TERMS Assessment, completeness, data quality, KG, knowledge graph, linked data, LOD, metrics, survey, systematic literature review.

I. INTRODUCTION

The development of Semantic Web technologies like the Resource Description Framework (RDF)[1] has led to unprecedented volumes of data published on the internet as Linked Open Data (LOD)[2] [1]. The collection and publication of such vast amounts of data into a Knowledge Base (KB) is certainly a progression in the right direction towards the *Web of Data*. However, the evolution of KBs exposed as Linked Data such as in the LOD Cloud[3] is generally unrestrained [2], which leads to a variety of quality issues, at various levels; for example, at the schema or at the instance level. An empirical study carried out by Debattista *et al.* [2] shows that datasets published in the LOD cloud have a reasonable overall quality, but significant issues remain concerning some quality dimensions, such as data provenance and completeness. Therefore, by studying only one dimension such as completeness, we have the ability to explore completeness quality issues more thoroughly. For instance, we can detect whether the completeness problem is better dealt with during data collection or integration process.

The Semantic Web promotes the reuse and sharing of this data, as well as its automatic processing by computer agents. The data represented in this way make sense and allow, in theory, for consensual interpretation by all actors (producers and consumers). Linked Data is sometimes called *Knowledge Graph* as referenced by Google in 2012.[4]

Data quality is also a challenge for traditional information systems, hence, rigorous research on ensuring adequate quality of data in relational databases have been carried out, even before the onset of Knowledge Graph. This development has led to a positive impact on the data quality organizational

The associate editor coordinating the review of this manuscript and approving it for publication was Wajahat Ali Khan.

[1]https://www.w3.org/RDF/
[2]http://linkeddata.org/
[3]http://lod-cloud.net/
[4]https://blog.google/products/search/introducing-knowledge-graph-things-not/

processes for relational databases [3], [4]. Thus, the applicability of this approach in the context of Web of Data provides an avenue to leverage the experience gained from traditional information systems. Since high quality of data ensures its fitness for use [5] in a wide range of applications, having the right metrics to assess and improve the quality of Knowledge Graph is of great importance. Several frameworks and approaches have been proposed to evaluate varying dimensions of Linked Data quality; Zaveri *et al.* [6] conducted a comprehensive Systematic Literature Review and identified 18 different quality dimensions that can be applied to assess the quality of a Knowledge Graph. In our article, we focus on how to assess the completeness of Linked Data. Our objective is to qualitatively and quantitatively analyze the existing articles that propose methods to assess several types of completeness dimensions. We also classify the selected articles into various types of completeness.

This article presents a Systematic Literature Review (SLR) on completeness that is one of the most essential dimension in data quality dimensions as stated in [7]. This is because completeness affects other dimensions of data quality such as accuracy, timeliness and consistency. Different comprehensive surveys which focus on data quality methodologies for structured and Linked Data [6], [8], [9] exist in the literature. However, to the best of our knowledge, this SLR is the first one that focuses solely on the completeness of Knowledge Graph. We believe that this SLR will be helpful for Semantic Web researchers to improve the existing approaches or propose new ones with regards to completeness in Knowledge Graphs. This paper is a part of a PhD thesis by Subhi Issa [10] defended on December 13th.

A. OVERVIEW OF KNOWLEDGE GRAPH COMPLETENESS

Completeness is a data quality measure that refers to the amount of information present in a particular dataset [6]. For example, the instance *Albert Einstein* might suffer from a data completeness problem when his birth place is missing in the dataset.

Assessing data quality is one of the challenges that data consumers and providers are facing [11]. It is a multi-faceted challenge, for instance, the term quality is commonly described as *fitness for use* [5] which encompasses several dimensions such as accuracy, timeliness, consistency, correctness, completeness, etc. Nevertheless, these dimensions maybe subjective as completeness implies that the amount of data is sufficient for the consumer's needs which can vary substantially. It can be measured as the percentage of data available divided by the data required, where 100% is the best value. The question, however, is whether we can consider 70% complete data to be of high quality? This amount of information could be sufficient, for example, for the description of a film but not enough for a medical use case. In real-world use cases, incomplete data can lead to missing out on important information and, thus, to inaccurate analysis [12]. Additionally, in terms of timeliness, incompleteness affects the ability to have all the information required at the suitable moment.

In terms of Linked Data, existing literature [6] identified four types of completeness to measure the degree of completeness of data sources. Pipino *et al.* [13] divided completeness into: (i) schema completeness that is the degree to which classes and properties are presented in a schema, (ii) property completeness which is the extent of the missing property values of a specific kind of property, and (iii) population completeness that refers to the ratio of number of represented objects to total number of real-world objects. Later, a new type of completeness is introduced for Linked Data called (iv) interlinking completeness, which checks the existence of links between datasets via their linksets [14]. Thus, we classified the selected articles into one of these types of completeness as illustrated in Section II. As a result of our SLR, we identified three new types of completeness, namely, (v) currency, (vi) metadata and (vii) labelling completeness making it a total of seven types of completeness. It should be noted that assessment requires a reference or gold standard with which to assess against and such a gold standard should ideally operate on Closed World Assumption. However, the gold standards in Linked Data still operate on Open World Assumption [15], but, for the purposes of measuring completeness, we assume that the gold standard is complete.

B. TYPES OF COMPLETENESS

Data completeness is the proportion of existing data to the total required data. Therefore, suppose we were interested in data about *Albert Einstein*, and the information we have is incomplete. This will have different ramifications depending on the type of completeness issue in the data. We explain the types of completeness that have been extracted from the selected papers using the example on *Albert Einstein* as shown in Figure 1.

1) SCHEMA COMPLETENESS

This involves observing if all required properties of the scientist *Albert Einstein* are included, such as ''birthPlace'' and ''birthDate''.

2) PROPERTY COMPLETENESS

The existence of a missing value for a specific property is validated. For example, *Albert Einstein* was married twice and only one value is provided in the given dataset.

3) POPULATION COMPLETENESS

This checks how well values provided in the dataset cover the real-world object. For instance, there are only three states of Germany in the given dataset but in reality there are 16 states.

4) INTERLINKING COMPLETENESS

It is observed if instances used in the dataset are linked to equivalent instances in another existing dataset. For example, the instance *Albert Einstein* is linked to the equivalent one in

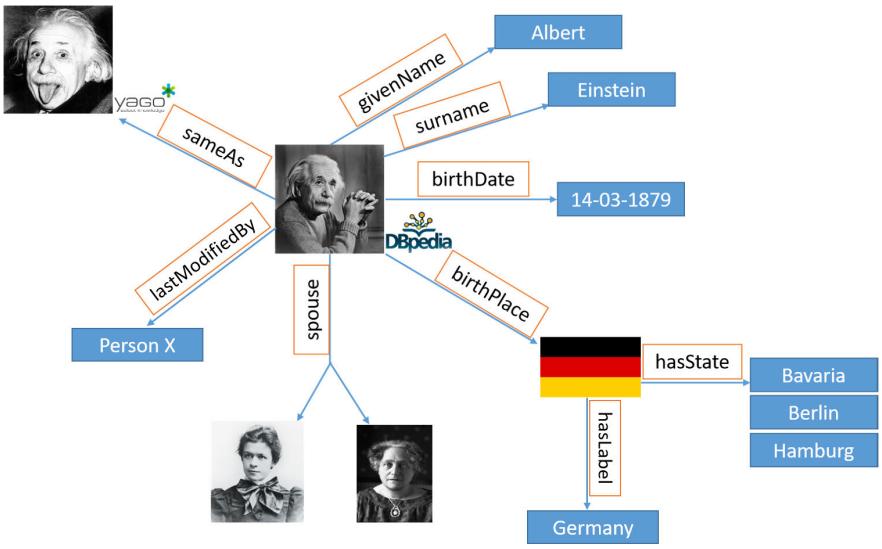

FIGURE 1. Example of Knowledge Graph instance illustrating various types of completeness.

YAGO dataset but does not link to the equivalent instance in Wikidata.

5) CURRENCY COMPLETENESS

This examines how the property values evolve over time. For example, if the dataset captures where *Albert Einstein* lives, it only captured the last location in the USA where he lived. However, he also lived in Germany and no existing versions of the dataset captures this fact.

6) METADATA COMPLETENESS

This involves observing if a sufficient metadata about the dataset is available. For instance, the presence of the name of the last individual who modified the dataset on *Albert Einstein* represented by *lastModifiedBy* relation in Figure 1.

7) LABELLING COMPLETENESS

It is checked whether all entities in the dataset have human and machine readable labels. For instance, if the dataset has a birthplace for *Albert Einstein* depicted by the German flag identified a URI such as *http://rdf.freebase.com/ns/germany.png*. This may not be sufficiently clear to a user that cannot identify flags or the computer; hence, the inclusion of *hasLabel* provides a clear label for such a resource depicted in Figure 1.

This article is structured as follows: In Section I-C we describe our SLR methodology. In Section II, we define and classify the completeness of Knowledge Graph, provide quantitative and qualitative analyses of the selected studies and describe the tools used for evaluating Knowledge Graph completeness. In Section III, we discuss the current challenges and future directions. Finally, we conclude this article in Section IV.

C. SYSTEMATIC LITERATURE REVIEW METHODOLOGY

In this section, we explain our SLR methodology to identify all articles related to Knowledge Graph completeness and we summarize the proposed solutions in terms of (i) the problem addressed, (ii) approaches and metrics proposed and (iii) tools developed to assess the issue of completeness.

Two reviewers, from different institutions (the first two authors of this article), conducted this systematic review by following the systematic review procedures described in [16].[5] According to [16], a systematic review is useful for several reasons, such as: (i) summarize and compare the various methodologies in a domain, (ii) identify open problems, (iii) contribute a hybrid concept comprising of various methodologies developed in a domain, or/and (iv) synthesize new ideas to address open problems. This systematic review tackles, in particular, problems (i), (ii) and (iii). It summarizes and compares various LD completeness data quality assessment methodologies as well as identifying open problems related to LD Completeness. An overview of our search methodology including the number of retrieved articles at each step is shown in Figure 2 and described in detail below.

D. RESEARCH QUESTIONS

In this SLR, we aim to answer the following general research question:

How can we assess the completeness of Knowledge Graph, which includes different types of completeness considering several approaches? We divide this general research question into sub-questions:

- what *types* of completeness currently exist for Knowledge Graphs?

[5]https://www.inf.ufsc.br/ aldo.vw/kitchenham.pdf

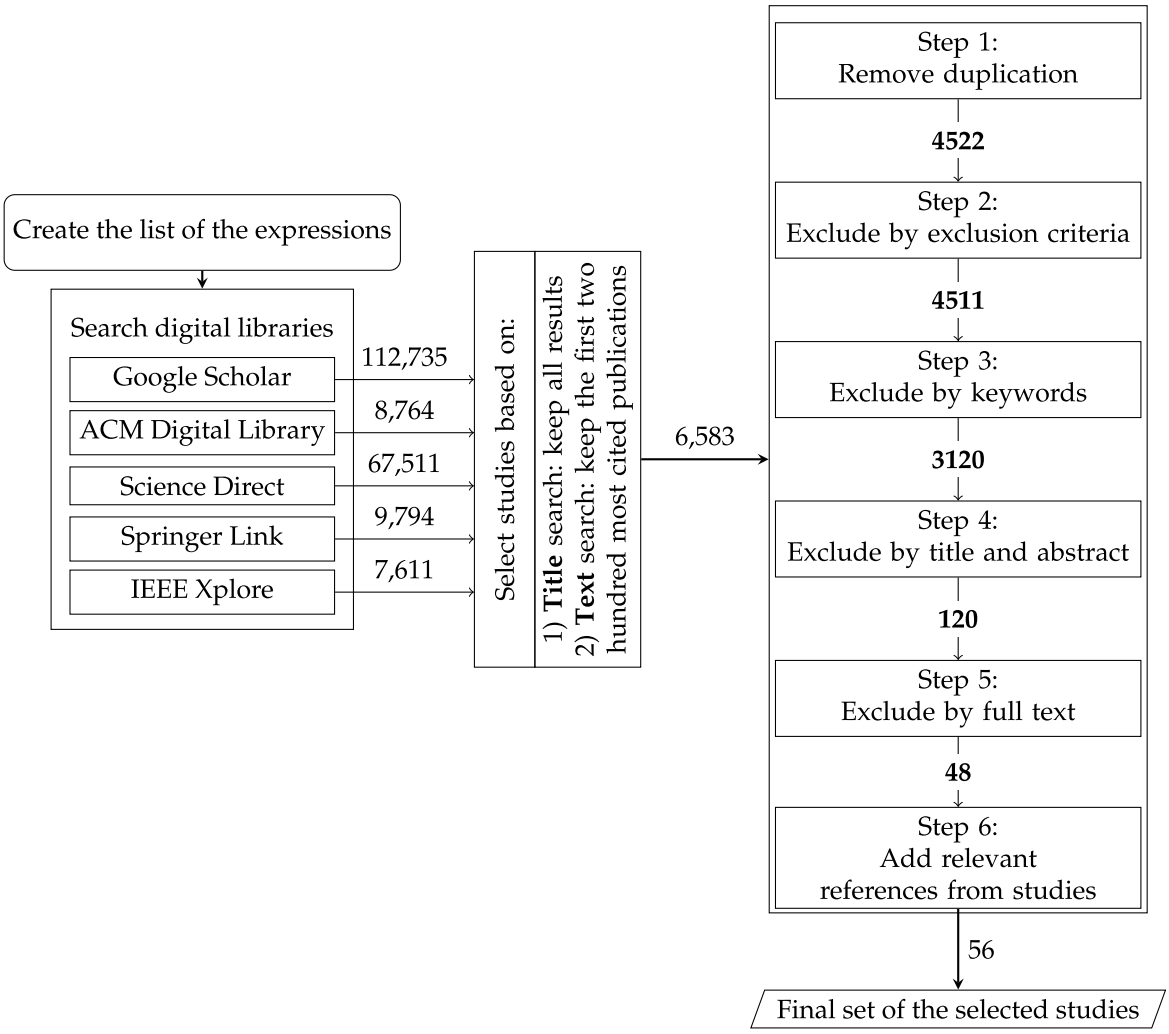

FIGURE 2. Overview of the systematic literature review methodology.

- what are the proposed *approaches and metrics* to identify and measure the completeness of Knowledge Graphs?
- what are the data completeness *problems* being discussed by researchers?
- what *tools* are available to detect completeness of Knowledge Graphs?

E. INCLUSION CRITERIA

- articles published in English
- articles published between 2006-2019[6]
- articles that:
  - studied or measured completeness of Knowledge Graph
  - proposed data completeness methodology or framework
  - proposed and applied metrics for completeness of Knowledge Graph

[6]As the term of Linked Data first appeared in 2006 [1] and Knowledge Graph in 2012

F. EXCLUSION CRITERIA

- articles that have not been peer-reviewed
- articles published in other languages
- master or doctorate thesis, poster, PowerPoint presentation or books
- articles that focused neither on Knowledge Graph nor on semantic web technologies

G. GENERATING A SEARCH STRATEGY

Search strategies in a systematic review are usually iterative and are run separately by two or more reviewers to avoid bias and to maximize coverage of all related articles. We performed a search on Google Scholar as a search engine and the following Databases: IEEE Xplore, ACM Digital Library, Science Direct and Springer Link. Because it is impractical to accept all the returned results when we search for the keywords in the full articles, we limit our research to the most cited 200 articles from each source.

From our perspective, searching only on the title is not efficient and does not always provide all the relevant

articles. This is because authors are often inclined to use agile titles which do not express the real content of the article. Thus, we divide our search strategy into three steps:

- scan article titles based on inclusion/exclusion criteria
- search within text and determine fit based on inclusion/exclusion criteria, abstract and in some cases the full article
- search relevant references in some core articles

Figure 2 provides more details on the exact numbers of articles searched and obtained over the aforementioned steps.

One of the most important parts is defining the search terms. These expressions that that are used to find the articles related to Knowledge Graph completeness should be based on a defined search strategy. This strategy aims to find as many relevant articles as possible. Based on our discussions and testing to obtain the most related articles as possible in our domain, the search string that was proposed contains synonyms of the concept term. As our concept term is ''Knowledge Graph completeness'', we added the alternative spellings, synonyms, as well as terms related to quality. Finally, we connected them using the boolean operators **OR** and **AND**. The expressions that were used to extract the interested studies are:

- Exp. 1: ((''Knowledge Graph'') **OR** (''Linked Data'')) **AND** (quality **OR** assessment **OR** evaluation **OR** methodology **OR** measuring **OR** completeness)
- Exp. 2: (''Linked Open Data'') **AND** (quality **OR** assessment **OR** evaluation **OR** methodology **OR** measuring **OR** completeness)
- Exp. 3: (KG **OR** LOD) **AND** (quality **OR** assessment **OR** evaluation **OR** methodology **OR** measuring **OR** completeness)

After removing about 2000 duplicated papers from the overall search results, we excluded the papers based on exclusion criteria. After that, we excluded them based on the abstract of each paper then the full text of the paper in case that the abstract was not sufficiently clear to take a decision. Finally, we added the relevant papers from the references of the selected papers from the last step.

II. KNOWLEDGE GRAPH COMPLETENESS ANALYSIS

Through our methodology, we identified 56 core articles related to Knowledge Graph or Linked Data completeness. We recognize the challenge of trying to explicitly elaborate on all core articles, hence, we categorized and summarized the main ideas in all the articles with different subsections focusing on different ideas. Section II-A discusses simple statistical analysis and trends from the core articles. Section II-B provides a summary of the core literature categorized according to the type of completeness found. The categories are grouped according to broad problems and approaches recognized. Consequently, Table 5 focuses on the core ideas of the metrics for the core articles. The purpose is to provide a full overview of metrics over all the types of completeness succinctly and avoid clogging. Furthermore, in order to minimize redundancy, we elaborate on some core articles focusing on proposed tools in Section II-C while the main idea is already covered in Section II-B. We have provided appropriate referrals for those who want to further expatiate on how the articles represent the ideas. Also, we selected a few influential and representative literature for discussion in the overview of Section II-B to provide a more comprehensive idea on how LD completeness is tackled.

A. QUANTITATIVE ANALYSIS

The core articles that make up our final list of the selected articles is shown in Table 3. On the other hand, Table 4 shows the list with the types of completeness that each article addresses. We further clustered the selected articles with respect to the type of completeness, as shown in Figure 3. The number labelled nodes represent publications with their reference numbers and the edges link to the type of completeness that the publication covers. We observed that a publication can address multiple types of completeness. Property completeness is the most addressed by the studies, with 22 publications. From the 56 core articles, 32 were published from 2016 till now (2019); hence about 57% of the studies are quite recent and the trend shows that more researchers are getting involved in this domain as the years go by as illustrated in Figure 4. Table 1 shows where researchers are publishing their work, where the top journal is Semantic Web journal and the top conference for publishing being International Semantic Web Conference. We observed that researchers are now publishing more in conferences with 34 articles (61%) published at a conference. 15 articles (24%) were published in journals and seven articles (13%) were published in workshop proceedings. Also, for every article published in a journal there are approximately four articles published in conferences.

B. QUALITATIVE ANALYSIS

In this section, we analyzed the 56 articles qualitatively to extract relevant information regarding Knowledge Graph completeness. After analyzing the selected articles in detail, we identified and extracted 23 ubiquitous metrics which are presented in Table 5 that can be applied to assess the completeness of Knowledge Graph, categorizing them based on the type of completeness covered. As mentioned previously, (i) schema, (ii) property, (iii) population and (iv) interlinking completeness were already identified by [6]. As part of our SLR we identified three more types, namely, (v) currency, (vi) metadata and (vii) labelling completeness.

In the following, we describe each of the seven types by providing a definition and discussing the problems and approaches that they address. We summarize the problems and approaches found for each type of completeness and provide a few examples. The full list of metrics is reported in Table 5.

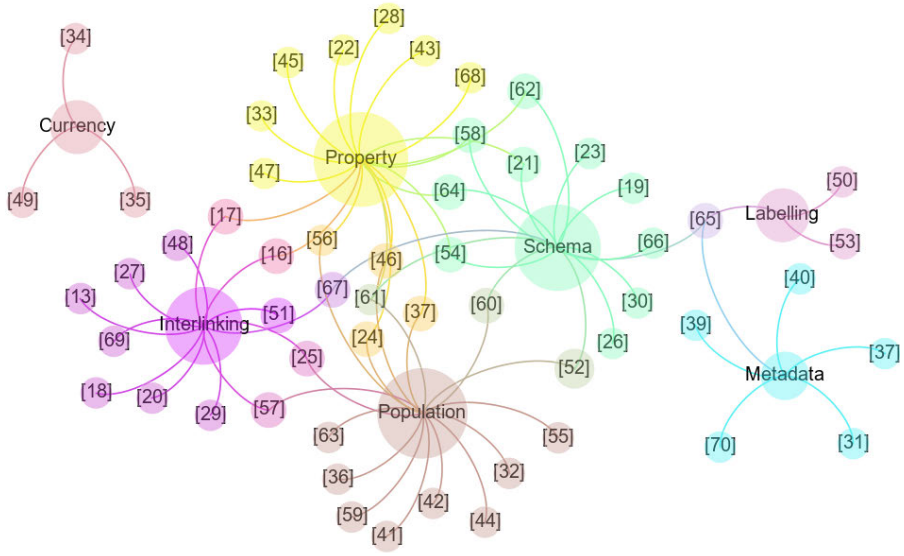

FIGURE 3. Classification of the 56 core articles by type of completeness.

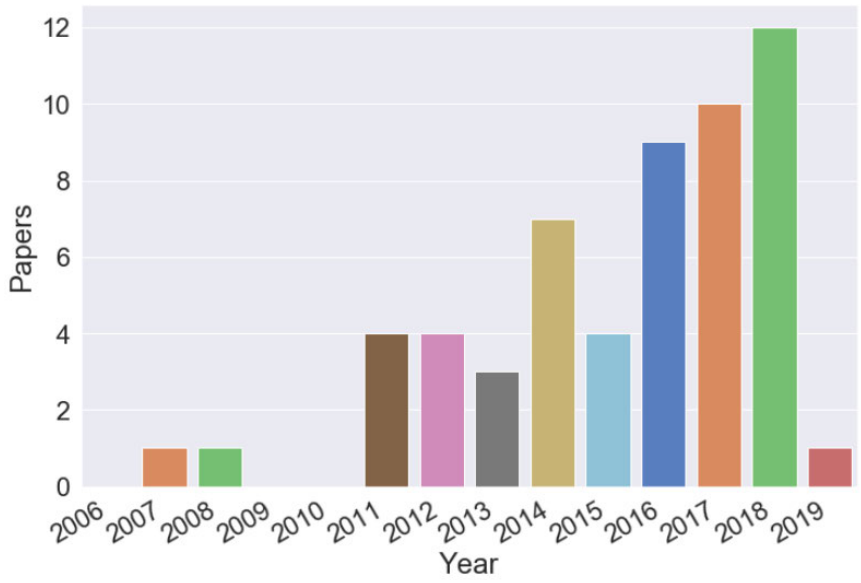

FIGURE 4. Number of core articles by year.

1) SCHEMA COMPLETENESS

The schema of a dataset is considered complete, if it contains all the classes and properties needed for a given task. It is also called ontology completeness [62]. Fürber and Hepp [62] defined schema completeness as the degree to which classes and properties are represented in a schema. In a similar sense but under a different name, Mendes *et al.* [61] defined *intensional completeness* which is the existence of all the attributes in a dataset for a given task. For example, the dataset suffers from a schema completeness problem when the property *capital* is missed from the instance *France*.

Definition 1 (Schema Completeness): Schema completeness is the degree to which the classes and properties of an ontology are represented in a LD dataset.

a: OVERVIEW

One of the most recent researches is the work of Lajus and Suchanek [27] to determine mandatory properties for a class in order to discover the missing facts in a class. The authors presented the incompleteness of KB statistically. They detected the mandatory properties using data from the KB through studying the abundance of properties for a class. For a given class, a mandatory property denotes a relation that every instance of the class should be involved in, such as every city has a population, then considering "population" is a mandatory property for the class "city". Finally, by calculating the ratio of instances that actually have the properties in the data, we get an approximate of the schema completeness.

TABLE 1. List of the core articles by conferences and journals.

Publication Venue	Frequency
Semantic Web journal (SWJ)	6
International Semantic Web Conference (ISWC)	4
International Conference on Web Engineering (ICWE)	3
European Semantic Web Conference (ESWC)	3
International conference on World wide web (WWW)	3
International Journal on Semantic Web and Information Systems (IJSWIS)	3
Workshop on Linked Data Quality (LDQ)	2
International Conference on Theory and Practice of Digital Libraries (TPDL)	2
International Conference on Database and Expert Systems Applications (DEXA)	2
International Conference on Database Theory (ICDT)	2
Government Information Quarterly (GIQ)	1
Open Journal of Semantic Web (OJSW)	1
Linked Data on the Web (LDOW)	1
International conference on Conceptual Modeling (ER)	1
IEEE Conference on Internet of Things, Green Computing and Communications, Smart Data, Cyber, Physical and Social Computing (IEEE SmartData)	1
Joint International Semantic Technology Conference (JIST)	1
Studies in Computational Intelligence (SCI)	1
International Conference on Web Research (ICWR)	1
Journal of Theoretical and Applied Information Technology (JATIT)	1
International Computer Software and Applications Conference (COMPSAC)	1
International Conference on Knowledge Capture (K-CAP)	1
International Symposium on Computer Science and Software Engineering (ICSE)	1
Journal of Data and Information Quality (JDIQ)	1
Journal of Industrial Information Integration	1
OTM Confederated International Conferences On the Move to Meaningful Internet Systems (OTM)	1
International Conference on Semantic Computing (IEEE ICSC)	1
International Workshop on Linked Web Data Management (LWDM)	1
International Workshop on Completing and Debugging the Semantic Web (CoDeS)	1
International Workshop on Semantic Web Enterprise Adoption and Best Practice (WaSABi)	1
European Conference on Information Systems (ECIS)	1
IEEE Transactions on Fuzzy Systems (TFS)	1
International Conference on Conceptual Structures (ICCS)	1
ACM International Conference on Web Search and Data Mining (WSDM)	1
International Conference on Web Intelligence, Mining and Semantics (WIMS)	1
International conference on Distributed event-based systems (DEBS)	1
International Workshop on Big Data and Information Security (IWBIS)	1

TABLE 2. Number of the articles retrieved in each search engine.

		GS	SD	ACM	SL	IEEE
Title	Exp. 1	173	7	557	3	221
	Exp. 2	35	1	103	-	77
	Exp. 3	27	1	41	1	191
Anywhere	Exp. 1	31,100	5,487	6,903	5,365	2,686
	Exp. 2	14,600	573	928	2,270	937
	Exp. 3	66,800	61,442	232	2,155	3,499

Likewise, the authors in [31] proposed a mining-based approach that includes two steps. The first step aims to find the properties patterns that are most shared by the subset of instances extracted from the triple store related to the same category Maximal Frequent Patterns (*MFP*). This set, that is called ''transaction'', will be then used to calculate a completeness value regarding these patterns. The second step carries out for each transaction a comparison between its corresponding properties and each pattern of the *MFP* set regarding the presence or the absence of the pattern. An average is, therefore, calculated to obtain the completeness of each transaction *t* and, hence, the completeness of the whole dataset.

b: PROBLEMS

Several articles address the challenge of development of new tools and frameworks to assess and improve completeness and other data quality dimensions [55], [63], [66]. The authors in [20] were interested in how to apply first order logic predicates and developed a capacity function (i.e., a fuzzy measure) to express completeness. Reference [59] investigated how to employ the similarity between entities in a dataset to determine completeness. In [31], the authors built transaction vectors constituted of sequence of properties that deduced from instances to use them as an input to generate frequent patterns in order to compute the completeness.

c: APPROACHES AND METRICS

There are 13 articles that propose some approaches of metrics about schema completeness [20], [22], [24], [27], [31], [53], [55], [59], [61]–[63], [66], [67]. These existing approaches defined a set of metrics to assess schema completeness such as applying fusion methods or defining quality indicators, or assessing completeness based on extracting a set of frequent/required predicates. Several metrics measure completeness as the ratio of the number of classes/properties presented in a dataset to the total number of classes/properties [22], [53], [61], [63], [66], [67]. Other metrics take into account only the mandatory properties to assess the completeness [24], [31], [72]. [59] measured ratio of similar instances/subjects missing same properties.

2) PROPERTY COMPLETENESS

Property completeness as defined by [6] is the measure of the missing values for a specific property. This is similar to the definition of [13] which referred to it as column completeness. Property completeness is measured by determining if a specific property has missing values. For example, the dataset suffers from a property completeness problem when the property *capital* of the instance *France* does not have a value, namely, *Paris*.

Definition 2 (Property Completeness): Property completeness is the degree to which values for a specific property are available for a given task.

TABLE 3. List of the 56 core articles related to Knowledge Graph completeness.

Article	Citation
A comprehensive quality model for Linked Data	Radulovic et al. [17]
A framework for Linked Data fusion and quality assessment	Nahari et al. [18]
A linkset quality metric measuring multilingual gain in SKOS Thesauri	Albertoni et al. [19]
A Measure-Theoretic Foundation for Data Quality	Bronselaer et al. [20]
A metric-driven approach for interlinking assessment of RDF graphs	Yaghouti et al. [21]
A metrics-driven approach for quality assessment of Linked Open Data	Behkamal et al. [22]
A Model for Linked Open Data Acquisition and SPARQL Query Generation	Alec et al. [23]
A Quality Model for Linked Data Exploration	Cappiello et al. [24]
A Two-Fold Quality Assurance Approach for Dynamic Knowledge Bases: The 3cixty Use Case	Mihindukulasooriya et al. [25]
Analyzing Linked Data Quality with LiQuate	Ruckhaus et al. [26]
Are all people married? Determining obligatory attributes in knowledge bases	Lajus and Suchanek [27]
Are Linked Datasets fit for Open-domain Question Answering? A Quality Assessment	Thakkar et al. [28]
Assessing and Improving Domain Knowledge Representation in DBpedia	Font et al. [29]
Assessing Linked Data Mappings Using Network Measures	Guéret et al. [30]
Assessing linkset quality for complementing third-party datasets	Albertoni et al. [14]
Assessing the Completeness Evolution of DBpedia: A Case Study	Issa et al. [31]
Automated quality assessment of metadata across open data portals	Neumaier et al. [32]
Automatically Generating Data Linkages Using a Domain-Independent Candidate Selection Approach	Song and Heflin [33]
BOUNCER: Privacy-aware Query Processing Over Federations of RDF Datasets	Endris et al. [34]
Capturing the age of Linked Open Data: Towards a dataset-independent framework	Rula et al. [35]
Co-evolution of RDF Datasets	Faisal et al. [36]
Comparing Index Structures for Completeness Reasoning	Darari et al. [37]
Comparison of metadata quality in open data portals using the Analytic Hierarchy Process	Kubler et al. [38]
CROCUS: Cluster-based Ontology Data Cleansing	Cherix et al. [39]
Data Quality Assessment in Europeana: Metrics for Multilinguality	Charles et al. [40]
Dataset Profiling - a Guide to Features, Methods, Applications and Vocabularies	Ellefi et al. [41]
Enabling Fine-Grained RDF Data Completeness Assessment	Darari et al. [42]
Enhancing answer completeness of SPARQL queries via crowdsourcing	Acosta et al. [43]
Enhancing Dbpedia Quality Using Markov Logic Networks	Ali and Alchaita [44]
Ensuring the Completeness and Soundness of SPARQL Queries Using Completeness Statements about RDF Data Sources	Darari et al. [45]
How Linked Data can aid machine learning-based tasks	Mountantonakis and Tzitzikas [46]
Improving Curated Web-Data Quality with Structured Harvesting and Assessment	Feeney et al. [47]
Improving the Quality of Linked Data Using Statistical Distributions	Paulheim and Bizer [48]
Interlinking Linked Data Sources Using a Domain-Independent System	Nguyen et al. [49]
KBQ - A Tool for Knowledge Base Quality Assessment Using Evolution Analysis	Rizzo et al. [50]
Labels in the Web of Data	Ell et al. [51]
Linked Open Data Aggregation: Conflict Resolution and Aggregate Quality	Knap et al. [52]
Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO	Färber et al. [53]
Methodology for Linked Enterprise Data Quality Assessment Through Information Visualizations	Gürdür et al. [54]
Metrics-driven framework for LOD quality assessment	Behkamal [55]
Non-Parametric Class Completeness Estimators for Collaborative Knowledge Graphs-The Case of Wikidata	Luggen et al. [56]
Predicting completeness in knowledge bases	Galárraga et al. [57]
Quality and Complexity Measures for Data Linkage and Deduplication	Christen and Goiser [58]
Recoin: Relative Completeness in Wikidata	Balaraman et al. [59]
Representativeness of knowledge bases with the generalized Benford's law	Soulet et al. [60]
Sieve: Linked Data Quality Assessment and Fusion	Mendes et al. [61]
SWIQA - a Semantic Web Information Quality Assessment Framework	Fürber and Hepp [62]
Test-driven evaluation of Linked Data quality	Kontokostas et al. [63]
Towards a Data Quality Framework for Heterogeneous Data	Micic et al. [64]
Towards a vocabulary for data quality management in semantic web architectures	Fürber and Hepp [65]
Towards an objective assessment framework for Linked Data quality	Assaf et al. [66]
Towards Ontology Quality Assessment	McGurk et al. [67]
Towards the Russian Linked Culture Cloud: Data Enrichment and Publishing	Mouromtsev et al. [68]
Towards unified and native enrichment in event processing systems	Hasan et al. [69]
URI Disambiguation in the Context of Linked Data	Jaffri et al. [70]
What's up LOD Cloud? Observing The State of Linked Open Data Cloud Metadata	Assaf et al. [71]

a: OVERVIEW

One of the earliest and influential research work which addresses property completeness is the work of Mendes *et al.* [61]. They developed a versatile tool for quality assessment (Sieve) as a part of a Linked Data Integration Framework which deals with data access, schema mapping and identity resolution. This work provided one of the foremost easily extensible framework for Linked Data integration and data fusion. While the tool covers various data quality dimensions for assessment, it adequately provides one of the more widely used implementation for assessing property completeness. Sieve provides a customizable module based on a conceptual model of assessment metrics and scoring functions. The assessment metric is a procedure for measuring the score with regards to a data quality dimension such as property completeness using selected schema elements as a data quality indicator. The definition of how to calculate the score and which indicators to use are user defined. In this work, property completeness (called *extensional completeness*) scoring functions are defined as:

$$\frac{|| \; uniq.\; instance\; values\; in\; dataset \; ||}{|| \; all\; expected\; uniq.\; values\; in\; dataset \; ||} \tag{1}$$

$$\frac{|| \; obj.\; with\; property\; p\; in\; dataset \; ||}{|| \; all\; uniq.\; obj.\; in\; dataset \; ||} \tag{2}$$

These have been generalized to the main idea of percentage of instance values for which a given property exist in Table 5.

Sieve is one of the first readily available tool that is agnostic to provenance and quality vocabularies, allowing users to configure which metadata to read, and which functions to apply via a declarative specification language programmatically.

Likewise, Fürber and Hepp [62] also developed a generic framework for assessing data quality in LD called SWIQA, where predefined rules on the syntax of the literals for the values of a property and other attributes relative to the ideal values of literals are used to define a data quality rule for the property. The ratio of instances that violate a data quality rule and the total number of relevant instances indicates the level of completeness for the property.

b: PROBLEMS

The common research challenges that addresses property completeness is the development of data quality models, metrics and tools upon which benchmarking and evaluation may be carried out. These articles [17], [18], [22], [44], [47],

TABLE 4. List of the 56 core articles classified according to the seven types.

Article/Type of completeness	Schema	Property	Population	Interlinking	Currency	Metadata	Labelling
Bronselaer et al. [20]	✓						
Cappiello et al. [24]	✓						
Issa et al. [31]	✓						
Balaraman et al. [59]	✓						
Lajus and Suchanek [27]	✓						
Assaf et al. [66]	✓					✓	✓
McGurk et al. [67]	✓						
Behkamal et al. [22]	✓	✓					
Behkamal [55]	✓	✓					
Mendes et al. [61]	✓	✓	✓	✓			
Kontokostas et al. [63]	✓	✓					
Färber et al. [53]	✓	✓	✓				
Fürber and Hepp [62]	✓	✓	✓				
Alec et al. [23]		✓					
Font et al. [29]		✓					
Endris et al. [34]		✓					
Ali and Alchaita [44]		✓					
Mountantonakis and Tzitzikas [46]		✓					
Paulheim and Bizer [48]		✓					
Christen and Goiser [58]		✓					
Fürber and Hepp [65]		✓					
Hasan et al. [69]		✓					
Radulovic et al. [17]		✓		✓			
Nahari et al. [18]		✓		✓			
Mihindukulasooriya et al. [25]		✓	✓				
Darari et al. [42]		✓	✓				
Feeney et al. [47]		✓	✓				
Micic et al. [64]		✓	✓				
Galárraga et al. [57]		✓	✓				
Song and Heflin [33]			✓				
Darari et al. [37]			✓				
Cherix et al. [39]			✓				
Acosta et al. [43]			✓				
Darari et al. [45]			✓				
Soulet et al. [60]			✓				
Luggen et al. [56]			✓				
Ruckhaus et al. [26]			✓	✓			
Mouromtsev et al. [68]			✓	✓			
Nguyen et al. [49]				✓			
Jaffri et al. [70]				✓			
Knap et al. [52]				✓			
Albertoni et al. [19]				✓			
Yaghouti et al. [21]				✓			
Thakkar et al. [28]				✓			
Guéret et al. [30]				✓			
Albertoni et al. [14]				✓			
Rizzo et al. [50]					✓		
Rula et al. [35]					✓		
Faisal et al. [36]					✓		
Neumaier et al. [32]						✓	
Kubler et al. [38]						✓	
Charles et al. [40]						✓	
Ellefi et al. [41]						✓	
Assaf et al. [71]						✓	
Ell et al. [51]							✓
Gürdür et al. [54]							✓
Total	13	22	17	13	3	6	3

[55], [61], [63] approached the measurement of property completeness as part of general data quality assessment using novel models, methodology and/or metrics. Reference [34] explored the challenge of applying the evaluation of query answer completeness in a privacy preserving manner and [24] investigated how to develop models for evaluating completeness employing automatic query generation. Reference [57] also studied how to predict the completeness of a knowledge base in the absence of ground truth by deriving completeness assertions from the knowledge base and measuring how many objects are accompanied by a completeness assertion.

c: APPROACHES AND METRICS

There are 22 articles that addressed property completeness either by proposing an approach or metric to measure completeness [17], [18], [22], [23], [25], [29], [34], [42], [44], [46]–[48], [53], [55], [57], [58], [61]–[65], [69]. A prominent methodology for assessing property completeness focuses on the development of novel frameworks towards measuring the level of completeness of a knowledge base such as [22] where the authors use the Goal Question Metric (GQM) method to define metrics for inherent qualities of a dataset. A set of metrics based on measurement-theory have been proposed

TABLE 5. List of completeness metrics.

	Metric	Citation
Schema completeness	* Ratio of number of classes/properties presented in dataset to total number of classes/properties	[22], [53], [55], [61]–[63], [66], [67]
	* Ratio of number of properties of an instance to total number of mandatory properties	[24], [27], [31]
	* Capacity function to validate predicate of completeness at schema level	[20]
	* Ratio of similar subjects/instances missing properties	[59]
Property completeness	* Percentage of values for which a given property exists	[17], [22], [25], [46], [57], [61], [63], [65]
	* Ratio of number of values presented for a specific property to total number of values for a specific property	[44], [58], [62], [69]
	* Completeness measurement based on statistical distributions of properties	[48]
	* Count of property values	[47], [55]
	* Ratio of concepts/predicate pairs	[53]
Population completeness	* Ratio of unique objects on the dataset to all available unique objects in the universe	[25], [26], [33], [37], [39], [47], [53], [62], [64], [68]
	* Multiplicity of the resource and the aggregated multiplicity of all classes where the resource belongs to	[43]
	* Identify a fragment of completeness information to check completeness	[42], [57]
	* Missing objects to add to a dataset to make it representative	
	* Convergence in the amount of estimated missing objects	[60]
		[56]
Interlinking completeness	* Ratio of instances that are interlinked in the dataset to total number of instances in the dataset	[17], [18], [21], [26], [28], [49], [68]
	* Linkset importing	[19]
	* owl:sameAs frequency relative to co-reference	[52], [70]
	* Ratio between the number of triples that are "in-links" and the total number of triples in the RDF graph served as a description of each resource	[28]
	* Extent of connectivity between the dataset under assessment and external sources	[28]
Currency completeness	* Ratio of unique triples in new version of KB to total unique triples over all versions of the KB	[35], [36]
	* Difference between frequency of properties for a class between two KB releases	[50]
Metadata completeness	* Aggregate function on predicates on metadata	[32]
	* Existence , Count of metadata values and ratio of missing metadata values to total metadata properties	[38], [40], [66], [71]
Labelling completeness	* Percentage of URIs with label	[51]
	* Existence, count of labels	[54], [66]

for evaluating the inherent quality characteristics of a dataset where property completeness is evaluated the ratio of the sum of the number of presented properties per instance to the total number of instances in the dataset. Furthermore, [61] proposed a framework for flexibly expressing quality assessment methods as well as data fusion methods where property completeness was explored using the proportion of unique non-missing objects in the dataset. Other approaches include the application of aggregate functions [34], [42] and statistical distributions [18], [44], [48].

Note that, the general assumption is that incorrect data values do not adversely affect the assessment of completeness.

3) POPULATION COMPLETENESS

A dataset is complete if it contains all of real-world objects for a given task, which is also called the completeness at data (instance) level [62]. Population completeness is also termed *extensional completeness* [61]. For example, the dataset suffers from a population completeness problem if it does not have all the French cities.

Definition 3 (Population Completeness): Population completeness is the degree to which all real-world objects of a particular type are represented in a LD dataset.

a: OVERVIEW

Another purpose of Mendes *et al.* framework [61], that was explained in Section II-B2, is measuring another type of completeness besides to property completeness. It takes into account the instantiations of properties in order to measure population completeness. In this work, the scoring function to compute population completeness is defined as:

$$\frac{||\ obj.\ with\ property\ p\ in\ dataset\ ||}{||\ all\ uniq.\ obj.\ in\ universe\ ||} \tag{3}$$

In this paper, the authors compared Brazilian municipalities in English and Portuguese DBpedia datasets according to all 5565 Brazilian municipalities.

b: PROBLEMS

The popular challenge in this type is to check a KB to see whether it contains all entities of a given type. Reference [25] was interested in how to maintain the quality of KB that evolves and changes frequently. The proposed approach provides an overview of the change of a given KB and a fine-grained analysis. Consequently, the authors in [68] used completeness metrics to assess the quality of newly published Linked Data for cultural heritage. Reference [33] was focused on disambiguation problem, the authors provided a method to scalably resolve entity co-reference in structured datasets. Moreover, efforts have been made to include completeness information in KBs. This is achieved by adding true facts such as *Adele has one brother*. This information is essential to evaluate query completeness and soundness [37], [42], [45].

c: APPROACHES AND METRICS

A set of 17 articles have been proposed to enhance population completeness [25], [26], [33], [37], [39], [42], [43], [45], [47], [53], [56], [57], [60]–[62], [64], [68]. Various metrics that check KBs to see whether they contain all entities of a given type in comparison to real-world data [25] or deal with query completeness via hybrid computation [43], include completeness information as part of the KB that can be used for validation [42]. On the other hand, [26] used a Bayesian Network to model the dependencies among resources that belong to a set of linked datasets and represent the joint probability distributions of relationships among resources. The probability of an individual resource is considered the likelihood of redundancy or indicator of completeness regarding the resource. Soulet *et al.* [60] introduced a method to calculate a lower bound of completeness in KG. The authors discovered the missing facts according to *Benfords Law* to estimate the completeness. In [56], the authors considered non-parametric methods to estimate the class size in order to estimate the completeness of the class.

4) INTERLINKING COMPLETENESS

This type particularly focuses on data integration which is a core tenet of Knowledge Graph. It refers to the instances that are interlinked in the dataset for disambiguation with regards to a reference dataset [30]. For example, the instance *France* linked from French national dataset to another instance *French Republic* in the United Nations dataset.

Definition 4 (Interlinking Completeness): Interlinking completeness is the degree to which instances are interlinked in a LD dataset with respect to some reference dataset(s).

a: OVERVIEW

One of the most influential research work to assess interlinking completeness is the work of Guéret *et al.* [30]. They proposed an automatic tool to evaluate the degree of interlinking using five network measures. Three metrics are from network theory domain to evaluate the variation in the quality with respect to a set of links namely degree, centrality, and the clustering coefficient. The other two measures are developed for LD that are ''open same-as chains'' and ''Description Richness''. They are applied to detect a number of unclosed same-as chains and description enrichment to assess the amount of new properties added through *owl:sameAs* relations. Based on these measures, the proposed approach takes a set of RDF triples from a set of resources and analyzes it to determine whether a set of links can be improved through a quality assessment report.

b: PROBLEMS

In the work of [17], the authors were interested in how to develop a standard data quality model for quality specification and assessment. The authors identified an indicator that are in a specific context of use, how to measure degree of instances associated with all expected and related entities. In the first step, the authors defined base measures for quality evaluation, then combining different base measures to get derived measures and metrics that are obtained by integrating base and/or derived measures. These measures and metrics are used to assess data quality covering various quality dimensions. Albertoni *et al.* [19] explored how to assess the value of interlinks of datasets in terms of information gain via what they refer to as linkset importing. References [33], [70] proposed methods to resolve entity co-reference and completing links in KBs.

c: APPROACHES AND METRICS

We identified 13 articles that focus on the interlinking completeness in LD [14], [17]–[19], [21], [26], [28], [30], [49], [52], [61], [68], [70]. Several approaches have been proposed to assess interlinking completeness. Reference [26] analyzed the quality of data and links in LOD cloud using Bayesian Networks. Additionally, [19] estimated the completeness of a dataset by complementing SKOS thesauri with their *skos:exactMatch* related information. In [49], the authors gathered the essential predicates of data sources using their covering and discriminative abilities. Then, they selected the most suitable alignments based on their confidences and finally, comparing the instances based on the selected alignments.

5) CURRENCY COMPLETENESS

Currency according to [35] is the degree to which data is up-to-date; and in this work, the authors were interested in providing a model for assessing currency and as a result they developed a metric for currency completeness to evaluate the completeness of the currency measurement. Currency completeness is evaluated on the dataset as it is modified and updated over time. For instance, the population in *France* as it varies over the years.

Definition 5 (Currency Completeness): Currency completeness is the degree to which elements of a knowledge base are available as it is updated over time.

a: OVERVIEW

Currency completeness mainly deals with checking for outdated data which are usually inappropriate for most tasks. Rula *et al.* [35] were the first to refer to *currency completeness* as a concept to evaluate the completeness of currency measurement. They developed the first dataset independent framework for assessing the currency of Linked Open Data (LOD) graphs. In order to measure the currency of data, the temporal information about the creation and modification of LOD resources and documents, called versioning metadata is of uttermost importance. Unfortunately, it is difficult to collect versioning metadata in LOD, especially because there are no widely adopted models to represent such metadata, and their characterization. In order to evaluate the approach described by the authors to arbitrarily measure the currency of all resources in a dataset, they defined currency completeness as an evaluation of the number of resources for which currency can be computed over the total number of resources occurring in a dataset.

$$\frac{|\ resources\ set\ with\ currency\ values\ >\ 0\ |}{|\ all\ resources\ set\ in\ dataset\ |} \tag{4}$$

b: PROBLEMS

The issue of how to develop frameworks and metrics for assessing the currency of RDF data is the focus of research such as [35], [73] and currency completeness is the by-product of evaluating the proposed frameworks for assessing currency. Reference [36] also dealt with currency completeness while investigating an approach to deal with the mutual propagation of the changes between a replica and its origin dataset termed as co-evolution.

c: APPROACHES AND METRICS

We found three articles that proposed metrics to measure currency completeness [36], [50] and another article [35] proposed a new framework focusing on LD currency. While timeliness captures the freshness of a specific statement or entity [41], in other words, determines the extent to which data are sufficiently up-to-date for a task; currency completeness measures the completeness of the knowledge base as it is being updated over different versions. As such, currency completeness is the intersection between timeliness and completeness where the degree of completeness is measured as the data becomes more up-to-date. References [35], [73] defined its currency completeness metric as the number of resources for which currency can be computed over the total number of resources occurring in a knowledge base. Furthermore, [36] evaluated currency completeness as the ratio of the number of unique triples in the synchronized dataset to the count of unique triples in the two different versions of the dataset.

6) METADATA COMPLETENESS

Descriptive metadata about datasets enables dataset discovery, and as such [41] provided a comprehensive overview on metadata termed as *dataset profiling* where they also assess metadata completeness. Accordingly, metadata is considered complete if it contains all the fields with values required to properly describe a dataset e.g. a dataset with technical identifiers such as *title or description* without any metadata context is incomplete and reduce the quality of the dataset. Metadata/description of a dataset is expected to be Findable, Accessible, Interoperable and Reusable (FAIR) [74]. For instance, indicating the description of the dataset about *France* that it captures intrinsic properties of the country or all the editors that modified the dataset, thus, making sure the metadata is available and complete.

Definition 6 (Metadata Completeness): Metadata completeness is the degree to which metadata properties and values are not missing in a dataset for a given task.

a: OVERVIEW

One of the most influential research work which addresses metadata completeness is Neumaier *et al.* [32]. They have assessed the metadata of about 260 open data portals towards the searchability, discoverability and usability of their resources. They developed a generic metadata quality assessment framework in which a set of quality metrics for metadata was proposed in line with the Data Catalog Vocabulary (DCAT) metadata standard. Authors map the dataset in various portals to its DCAT instantiation which can be expressed as a tree structure. The assessment metric has been evaluated by selecting an appropriate path in the DCAT instantiation and resolving the path to its value aggregated over the dataset. In particular, metadata completeness was interpreted as existence evaluation of metadata information in the research work. Metadata information such as the existence of access information, dataset provider contact, license for the dataset and timestamps regarding creation and modification of the dataset.

b: PROBLEMS

Open data platforms such as LOD cloud and governmental open data terminal are becoming widespread and important data source for research. Reference [32] mentioned that metadata quality issues in open data portals have been identified as one of the core problems for wider adoption of open data and developed a quality assessment and evolution monitoring framework for web-based data portal platforms, which offers their metadata in different and heterogeneous models. Similarly, [41], [66] investigated the development of frameworks for assessing metadata quality in Knowledge Graph sources and [38], [71] focused on assessing government open data platforms.

c: APPROACHES AND METRICS

There are six articles that focus on metadata completeness [32], [38], [40], [41], [66], [71]. A lot of emphasis is currently

placed on the completeness of the dataset itself only, but the importance of completeness of the metadata cannot be understated. Descriptive metadata about existing datasets are a substantial building block for facilitating entities and datasets linking, entity retrieval, distributed search or query federation. Ellefi *et al.* [41] developed a framework for dataset profiling for the formal representation of a set of features that describes a dataset and allow the comparison of different datasets. They provided a taxonomy, formally represented as an RDF vocabulary of dataset profiling features. Furthermore, [40] presented an approach for capturing multilingualism as part of data quality dimensions, spanning completeness and [38] proposed a framework for comparison of open data portals for metadata quality using analytic hierarchy process.

7) LABELLING COMPLETENESS

Labelling completeness is particular to RDF data, where URIs are used for identification but are not very suitable for indexing purposes and human readability. This warrants that entities have a human readable label and the level for which it is not missing is labelling completeness. For example, the existence of *rdfs:label* for instance of a city *Paris*.

Definition 7 (Labelling Completeness): Labelling completeness refers to the degree to which entities in the dataset have a human readable label.

a: OVERVIEW

According to Ell *et al.* [51], even though it is assumed that labels for resources in a dataset will be made available, among other information, by dereferencing the URI of an entity using the HTTP protocol, following Linked Open Data principles. A whole lot of applications fall back to exposing the URIs of the entities to the user in the absence of more easily understandable representations such as human-readable labels. It is proposed that this is often due to issues such as internationalization, multiple labels for an entity, the computational costs associated with dereferencing, or the use of alternative labeling properties that makes the task of finding a label for a given entity much harder than expected. This work defines a number of metrics that provides a baseline for a quantitative analysis of the state of labeling on the Web.

Completeness of the labels is one of the defined metrics the authors put forward that checks that all resources in the dataset have labels. They defined a function that checks resources in the dataset for *rdfs:label*, or finds a parameter matching a property that assigns a label to a resource either explicitly in the dataset or interlinked from another dataset. The frequency of resource with valid response from the function determines labelling completeness of the dataset.

b: PROBLEMS

Reference [51] was focused on the internationalization of knowledge bases, existence of multiple labels for an entity, the computational costs associated with dereferencing URIs and proposed that all entities in a knowledge base have human readable labels. They also explored development of a metric for labelling completeness. Furthermore, [54] was interested in enterprise data integration and investigated the assessment of data quality for a Linked Enterprise Data in the automotive industry. The authors calculated the average number of resources that did not have a label property defined and stressed the need for it due to the heterogeneous nature of environments that generates different components of the dataset.

c: APPROACHES AND METRICS

Non-information resources are abstract ideas represented in LD dataset that may possess URI but cannot be directly accessed or downloaded via the internet, such as person. The importance of labels for non-information resources in Knowledge Graph cannot be overstated since they provide appropriate context to understand the dataset. It helps indexing and searching the resources and displaying data to end-users that can be easily understood, rather than URIs [51]. Only three articles have examined labelling completeness [51], [54], [66]. According to Ell *et al.* [51], labelling completeness measures the degree to which Linked Data resources have labels, it can be defined as the ratio of URIs with at least one value for a labeling property to all URIs in a given knowledge base. Reference [54] developed a data quality assessment tool in the form of a dashboard to manage data quality of a dataset integrated from various departments of an organization that is part of the automotive industry. Finally, [66] presented a quality measurement tool that helps data providers to rate the quality of their datasets and get recommendations on possible improvements by developing standardized quality indicators to rate datasets.

C. TOOLS ANALYSIS

From the core articles, we identified nine most common used tools (listed in Figure 5) that automatically or semi-automatically assess completeness of datasets. An overview of different tools and their capabilities along with the type of completeness they focus on, is described below.

1) SIEVE

Sieve[7] [61] is the quality evaluation module within Linked Data Integration Framework (LDIF) [75] which enables automatic data quality assessment by a conceptual model composed of assessment metrics, indicators and scoring functions like (Set Membership, Threshold and Interval Membership) for completeness. It is suitable for assessing schema, property and interlinking completeness.

2) LOUPE

Loupe[8] [17] is a tool that can be used to inspect a dataset to understand which vocabularies (classes and properties) are used with statistics and frequent triple patterns. Starting from

[7] http://sieve.wbsg.de
[8] http://loupe.linkeddata.es/loupe/

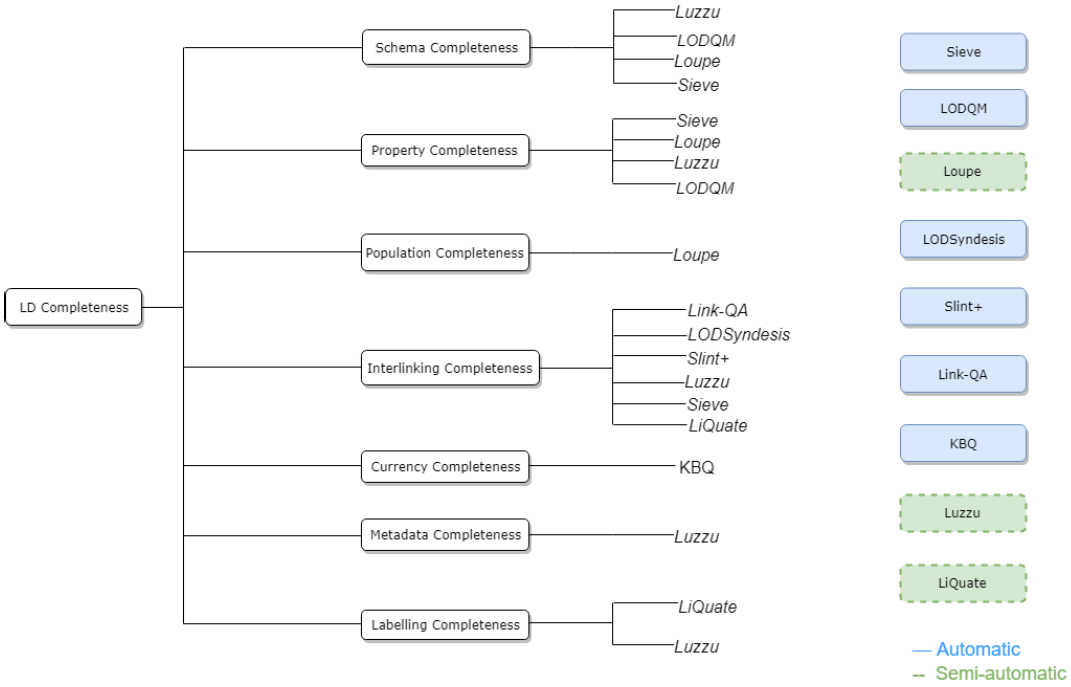

FIGURE 5. Summary of tools based on type of completeness.

the high-level statistics, Loupe allows one to zoom into details down to the corresponding triple with its visual explorer. It is a semi-automatic tool as metrics needs customization by the user. This can be used to assess schema and property completeness. Population completeness can also be assessed with external data.

3) LUZZU

Luzzu[9] [28] is also a Quality Assessment Framework for Linked Open datasets based on the Dataset Quality Ontology (daQ), allowing users to define their own quality metrics. It provides a library of generic quality metrics that users can customize based on domain-specific tasks in a scalable manner, thus it is semi-automatic; it also provides queryable quality metadata on the assessed datasets and assembles detailed quality reports on assessed datasets. It is useful for assessing property, schema, interlinking, metadata and labelling completeness albeit it requires manual configurations by the user.

4) LINK-QA

Link-QA[10] [30] specifies a framework for detection of the quality of linksets using network metrics (degree, clustering coefficient, open sameAs chains, centrality, description richness through sameAs). It is completely automatic and compatible with a set of resources, SPARQL endpoints and/or dereferencable resources and a set of triples as input. It is particularly useful for assessing interlinking completeness.

5) LiQuate

LiQuate [26] is a tool that uses a Bayesian Network to learn the dependencies between properties in RDF data. It is particularly well suited to assess interlinking and population completeness. However, it is semi-automatic and requires configurations from the user.

6) LODsyndesis

LODsyndesis[11] [46] uses novel lattice-based algorithms to find the intersection of datasets in the LOD cloud. The symmetric and transitive closure of the set of *owl:sameAs*, *owl:equivalentProperty* and *owl:equivalentClass* relationships from all datasets was computed for creating semantically enriched indexes. It is a reference for automatically assessing interlinking completeness.

7) Slint+

Slint+[12] [49] (Schema-Independent Linked Data Interlinking) is similar to LODsyndesis, in that, it detects all owl:sameAs links automatically between two given Knowledge Graph sources.

8) LODQM

LODQM[13] [55] is an automatic tool developed around goal-question-metric [76] approach to soliciting metrics used for assessment of datasets. It is suited for assessing schema and property completeness.

[9]https://eis-bonn.github.io/Luzzu/
[10]https://github.com/cgueret/LinkedData-QA
[11]http://83.212.101.188:8081/LODsyndesis/index.jsp
[12]http://ri-www.nii.ac.jp/SLINT/index.html
[13]https://bitbucket.org/behkamal/new-metrics-codes/src

9) KBQ

KBQ[14] [50] is a tool geared towards assessment of quality of datasets based on temporal analysis. It automatically computes the frequency of predicates and the frequency of entities of a given resource type, and compares the frequencies with the ones observed in previous versions of the dataset. It can be specifically used to assess currency completeness.

III. DISCUSSION

A. OVERVIEW

In this SLR, we have analyzed 56 articles that focus on seven types of LD completeness. In total, we identified 23 metrics and nine tools that specifically deal with LD completeness. We observed that some articles examine one type of completeness such as [31], [51], [70] or several types of completeness such as [53], [61]. Furthermore, research is rarely entirely focused on a single aspect of data quality dimension. Among the nine tools analyzed, we discovered six tools that are automatic and the remaining three tools are semi-automatic. Out of the tools, LiQuate does not seem to be online or it is no longer supported while all others are online. Also, there was no formal validation of the methodologies that were implemented as tools. LD completeness challenges tackled in the literature are most often in the development of frameworks for assessment of data quality using various approaches, ranging from the application of network measures [30] to first order logic predicates [20]. In other scenarios, researchers define certain constraints applicable to a LD dataset, such as in the case of a privacy aware assessment framework [24]. Based on our analysis, we have identified several open challenges pertaining to Knowledge Graph quality in general and also specifically for the completeness dimension, which we discuss in the following.

B. OPEN WORLD ASSUMPTION

Typically the Semantic Web follows an Open World Assumption (OWA) [15], which does not allow inferring the truth of a statement only by checking whether the statement is known. OWA assumes that everything we do not know is not yet defined. For data completeness assessment, we often need to define metrics based on Close World Assumption, i.e. assume that everything that is not known can be assumed as false. However, this assumption will most likely not hold in many cases since we often suffer from lack of gold standard and complete data. Consequently, when performing data quality assessment, the metrics have to be evaluated and refined continuously [62].

C. MAINTENANCE OF DATA QUALITY

After the assessment of data quality, the next step is to improve the quality taking into account the results from the assessment. This cycle of assessment and improvement should be done at regular intervals of time and/or when the data is updated. Additionally, a data quality issue in one dataset can ultimately affect the quality of multiple interlinked datasets, thus propagating the errors. Consequently, maintenance of quality becomes challenging in the Web of Data mainly because it is generated from existing data and thus its correction can be even more difficult and time consuming when meta-information provenance is not available anymore [77].

D. QUALITY-BASED QUESTION ANSWERING

When existing linked datasets are published along with their quality information, it can be possible to design a new generation of quality-based question answer systems, which rely on this information to deliver useful and relevant results [78]. In order to provide the answer to user queries in a meaningful way, it is necessary to define what should be in the result, how it can be obtained and how one should represent the query result. In this case, the completeness, consistency (logical/formal), timeliness, etc. of the data affects the results considerably. For example, querying an integrated dataset for a particular flight time, the time from the source with the higher update frequency and more complete information should be chosen. Thus, question answering can be increased in effectiveness and efficiency using data quality criteria as a leverage to filter the most relevant results.

E. STREAM-LINING FUTURE SURVEYS

This SLR took eight months in total to be performed. In order to increase the efficiency and sustainability of such SLRs in the future, we propose (i) future surveys on Knowledge Graph and Linked Data quality, specifically on completeness, tag their articles with the type of completeness (listed in Section II-B) as keywords and (ii) we, as a community, think of combining human and machine effort towards streamlining such SLRs. We resonate with the idea proposed in [79] of *living systematic reviews* combining humans and machines. For some of the repetitive and labor-intensive tasks, machines can assist, such as, for searching relevant articles and eligibility analyzing. Then, humans can assist in extracting relevant information from within the text. Workflows can be developed in which human effort and machine automation can each enables the other to operate in more effective and efficient ways, offering substantial enhancements to the productivity of systematic reviews [79]. In this way, a Systematic Literature Review can be continually updated incorporating new articles as they become available.

IV. CONCLUSION AND FUTURE WORK

The Linked Open Data principles are applied in various domains including life science, media, medicine and e-government. All these areas require high quality of data since human lives are directly impacted; for example, IBM reported that poor quality data costs the US economy $3.1 trillion dollars a year.[15] This raises the need for developing methods to evaluate and improve (Linked) data quality on the Web.

[14] http://datascience.ismb.it/shiny/KBQ/

[15] https://www.ibmbigdatahub.com/infographic/four-vs-big-data

This work focuses on completeness, which is one of the most important dimensions for Knowledge Graph quality assessment [7]. In this article, we surveyed the research topic on completeness of Knowledge Graph. We analyzed 56 studies and classified seven types of completeness. We provided definitions for each type, identified the different kinds of problems that they address, provided approaches and metrics for assessment and analyzed the tools available for assessment of LD completeness. In this SLR, we addressed the research question: *How can we assess the completeness of Knowledge Graphs, which includes different types of completeness considering several approaches?* This is expatiated on in Section I-D where we sub-divided the research question into 4 sub-questions. The first 3 sub-questions are addressed in Section II-B, while the last sub-question is dealt with in Section II-C. There are a number of different reasons why we did this SLR:

- summarize existing approaches concerning Knowledge Graph completeness
- identify problems, approaches, metrics and tools for assessing LD completeness
- realize gaps in existing studies regarding LD completeness, in order to help the researchers find the topic where they should work in
- serve as a starting document for future researchers interested in this topic

One external threat of validity to our work is that we explored state of the art using only the keywords relating to *Knowledge Graph* and *Linked Data* as specified in Section I-G. Therefore, this may lead to the omission of interesting approaches that do not use these keywords. We already discussed the internal threat of validation due to open world assumption of LD in Section III. The future direction of our work entails the plan to expand our research by adding more relative keywords *RDF dataset* and *RDF graph* as RDF is the most common way of representing Knowledge Graph and a lot of quality assessment methodologies were performed relying on the properties of RDF. Moreover, we intend to develop our search strategy in order to cover types of completeness having alternative names such as *schema* and *ontology* completeness and include terms such as *coverage* as a synonym for completeness. Finally, we also intend to select other core dimensions like accuracy or timeliness of knowledge graphs and replicate the review process for in-depth analysis as a data quality dimension.

We hope that researchers find this work as a comprehensive introduction to LD completeness and identify future research challenges to address.

ACKNOWLEDGMENT

(Subhi Issa and Onaopepo Adekunle contributed equally to this work.)

REFERENCES

[1] T. Berners-Lee. (2006). *Linked Data-Design Issues*. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html

[2] J. Debattista, C. Lange, S. Auer, and D. Cortis, "Evaluating the quality of the LOD cloud: An empirical investigation," *Semantic Web*, vol. 9, no. 6, pp. 859–901, Sep. 2018.

[3] M. Scannapieco and C. Batini, "Completeness in the relational model: A comprehensive framework," in *Proc. ICIQ*, 2004, pp. 333–345.

[4] A. Motro and I. Rakov, "Estimating the quality of databases," in *Proc. Int. Conf. Flexible Query Answering Syst.* Roskilde, Denmark: Springer, 1998, pp. 298–307.

[5] J. M. Juran, F. M. Gryna, and R. S. Bingham, *Quality Control Handbook* (McGraw-Hill Handbooks). New York, NY, USA: McGraw-Hill, 1974.

[6] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for linked data: A survey," *Semantic Web*, vol. 7, no. 1, pp. 63–93, Mar. 2015.

[7] M. Margaritopoulos, T. Margaritopoulos, I. Mavridis, and A. Manitsaris, "Quantifying and measuring metadata completeness," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 4, pp. 724–737, Apr. 2012.

[8] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–52, Jul. 2009.

[9] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic Web*, vol. 8, no. 3, pp. 489–508, Dec. 2016.

[10] S. Issa, "Linked data quality: Completeness and conciseness," Ph.D. dissertation, CNAM, Paris, France, 2019.

[11] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, Mar. 1996.

[12] A. Zaveri, J. R. N. Vissoci, C. Daraio, and R. Pietrobon, "Using linked data to evaluate the impact of research and development in Europe: A structural equation model," in *The Semantic Web—ISWC 2013*, H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, and K. Janowicz, Eds. Berlin, Germany: Springer, 2003, pp. 244–259.

[13] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[14] R. Albertoni and A. G. Pérez, "Assessing linkset quality for complementing third-party datasets," in *Proc. Int. Conf. Database Theory*, 2013, p. 52.

[15] N. Drummond and R. Shearer, "The open world assumption," in *Proc. eSI Workshop, Closed World Databases Meets Open World Semantic Web*, vol. 15, 2006, p. 23.

[16] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele Univ., Keele, U.K., Durham Univ., Durham, U.K., Tech. Rep. EBSE 2007-001, 2007.

[17] F. Radulovic, N. Mihindukulasooriya, R. García-Castro, and A. Gómez-Pérez, "A comprehensive quality model for linked data," *Semantic Web*, vol. 9, no. 1, pp. 3–24, Nov. 2017.

[18] M. K. Nahari, N. Ghadiri, Z. Jafarifard, A. B. Dastjerdi, and J. R. Sack, "A framework for linked data fusion and quality assessment," in *Proc. 3th Int. Conf. Web Res. (ICWR)*, Apr. 2017, pp. 67–72.

[19] R. Albertoni, M. Martino, and P. Podestà, "A linkset quality metric measuring multilingual gain in SKOS Thesauri," in *Proc. CEUR Workshop*, vol. 1376, 2015, pp. 1–9.

[20] A. Bronselaer, R. D. Mol, and G. D. Tre, "A measure-theoretic foundation for data quality," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 627–639, Apr. 2018.

[21] N. Yaghouti, M. Kahani, and B. Behkamal, "A metric-driven approach for interlinking assessment of RDF graphs," in *Proc. Int. Symp. Comput. Sci. Softw. Eng. (CSSE)*, Aug. 2015, pp. 1–8.

[22] B. Behkamal, M. Kahani, E. Bagheri, and Z. Jeremic, "A metrics-driven approach for quality assessment of linked open data," in *Proc. Int. Conf. Database Expert Syst. Appl.*, vol. 9, 2014, pp. 64–79.

[23] C. Alec, C. Reynaud-Delaître, and B. Safar, "A model for Linked Open Data acquisition and SPARQL query generation," in *Proc. Int. Conf. Conceptual Struct.* Springer, 2016, pp. 237–251.

[24] C. Cappiello, T. D. Noia, B. A. Marcu, and M. Matera, "A quality model for Linked Data exploration," in *Proc. Int. Conf. Web Eng.*, vol. 9671, 2016, pp. 397–404.

[25] N. Mihindukulasooriya, G. Rizzo, R. Troncy, O. Corcho, and R. García-Castro, "A two-fold quality assurance approach for dynamic knowledge bases: The 3cixty use case," in *Proc. CEUR Workshop*, vol. 1586, 2016, pp. 1–12.

[26] E. Ruckhaus, M. E. Vidal, S. Castillo, O. Burguillos, and O. Baldizan, "Analyzing linked data quality with liquate," in *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.*, vol. 8798, 2014, pp. 488–493.

[27] J. Lajus and F. M. Suchanek, "Are all people married?: Determining obligatory attributes in knowledge bases," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 1115–1124.

[28] H. Thakkar, K. M. Endris, J. M. Gimenez-Garcia, J. Debattista, C. Lange, and S. Auer, "Are linked datasets fit for open-domain question answering? A quality assessment," in *Proc. 6th Int. Conf. Web Intell., Mining Semantics*, Jun. 2016, pp. 1–12.
[29] L. Font, A. Zouaq, and M. Gagnon, "Assessing and improving domain knowledge representation in DBpedia," *Open J. Semantic Web*, vol. 4, no. 1, pp. 1–19, 2017.
[30] C. Guéret, P. Groth, C. Stadler, and J. Lehmann, "Assessing Linked Data mappings using network measures," in *Proc. Eur. Semantic Web Conf.*, vol. 7295, 2012, pp. 87–102.
[31] S. Issa, P. H. Paris, and F. Hamdi, "Assessing the completeness evolution of DBpedia: A case study," in *Proc. Int. Conf. Conceptual Modeling* (Lecture Notes in Computer Science), vol. 10651. Cham, Switzerland: Springer, 2017, pp. 238–247.
[32] S. Neumaier, J. Umbrich, and A. Polleres, "Automated quality assessment of metadata across open data portals," *J. Data Inf. Qual.*, vol. 8, no. 1, pp. 1–29, Nov. 2016.
[33] D. Song and J. Heflin, "Automatically generating data linkages using a domain-independent candidate selection approach," in *Proc. 10th Int. Semantic Web Conf. Semantic Web (ISWC), Part I*, in Lecture Notes in Computer Science, vol. 7031, L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. F. Noy, and E. Blomqvist, Eds. Bonn, Germany: Springer, Oct. 2011, pp. 649–664, doi: 10.1007/978-3-642-25073-6_41.
[34] K. M. Endris, Z. Almhithawi, I. Lytra, M.-E. Vidal, and S. Auer, "BOUNCER: Privacy-aware query processing over federations of RDF datasets," in *Proc. 29th Int. Conf. Database Expert Syst. Appl. (DEXA), Part I*, in Lecture Notes in Computer Science, vol. 11029, S. Hartmann, H. Ma, A. Hameurlain, G. Pernul, and R. R. Wagner, Eds. Regensburg, Germany: Springer, Sep. 2018, pp. 69–84, doi: 10.1007/978-3-319-98809-2_5.
[35] A. Rula, M. Palmonari, and A. Maurino, "Capturing the age of linked open data: Towards a dataset-independent framework," in *Proc. IEEE 6th Int. Conf. Semantic Comput.* Washington, DC, USA: IEEE Computer Society, Sep. 2012, pp. 218–225.
[36] S. Faisal, M. Kemele Endris, S. Shekarpour, S. Auer, and M. E. Vidal, "Co-evolution of RDF datasets," in *Proc. Int. Conf. Web Eng.*, vol. 9671, 2016, pp. 225–243.
[37] F. Darari, W. Nutt, and S. Razniewski, "Comparing index structures for completeness reasoning," in *Proc. Int. Workshop Big Data Inf. Secur. (IWBIS)*, May 2018, pp. 49–56.
[38] S. Kubler, J. Robert, S. Neumaier, J. Umbrich, and Y. Le Traon, "Comparison of metadata quality in open data portals using the analytic hierarchy process," *Government Inf. Quart.*, vol. 35, no. 1, pp. 13–29, Jan. 2018.
[39] D. Cherix, R. Usbeck, A. Both, and J. Lehmann, "CROCUS: Cluster-based ontology data cleansing," in *Proc. CEUR Workshop*, vol. 1240, 2014, pp. 7–14.
[40] V. Charles, J. Stiller, P. Kiraly, and W. Bailer, "Data quality assessment in Europeana: Metrics for multilinguality," in *Proc. Int. Conf. Theory Practice Digit. Libraries*, vol. 2038, 2018, p. 11.
[41] M. B. Ellefi, Z. Bellahsene, J. G. Breslin, E. Demidova, S. Dietze, J. Szymanski, and K. Todorov, "RDF dataset profiling—A survey of features, methods, vocabularies and applications," *Semantic Web*, vol. 9, no. 5, pp. 677–705, 2018, doi: 10.3233/SW-180294.
[42] F. Darari, S. Razniewski, R. E. Prasojo, and W. Nutt, "Enabling fine-grained RDF data completeness assessment," in *Proc. Int. Conf. Web Eng.*, vol. 9671, 2016, pp. 170–187.
[43] M. Acosta, E. Simperl, F. Flöck, and M.-E. Vidal, "Enhancing answer completeness of SPARQL queries via crowdsourcing," *J. Web Semantics*, vol. 45, pp. 41–62, Aug. 2017.
[44] M. Ali and M. Alchaita, "Enhancing DBpedia quality using Markov logic networks," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 12, pp. 3924–3936, 2018.
[45] F. Darari, W. Nutt, S. Razniewski, and S. Rudolph, "Completeness and soundness guarantees for conjunctive SPARQL queries over RDF data sources with completeness statements," *Semantic Web*, vol. 11, no. 3, pp. 441–482, 2020, doi: 10.3233/SW-190344.
[46] M. Mountantonakis and Y. Tzitzikas, "How linked data can aid machine learning-based tasks," in *Proc. Int. Conf. Theory Pract. Digit. Libraries*, in Lecture Notes in Computer Science, vol. 10450. Cham, Switzerland: Springer, 2017, pp. 155–168.
[47] K. C. Feeney, D. O'Sullivan, W. Tai, and R. Brennan, "Improving curated Web-data quality with structured harvesting and assessment," *Int. J. Semantic Web Inf. Syst.*, vol. 10, no. 2, pp. 35–62, Apr. 2014.
[48] H. Paulheim and C. Bizer, "Improving the quality of linked data using statistical distributions," *Int. J. Semantic Web Inf. Syst.*, vol. 10, no. 2, pp. 63–86, Apr. 2014.
[49] K. Nguyen, R. Ichise, and B. Le, "Interlinking linked data sources using a domain-independent system," in *Proc. Joint Int. Semantic Technol. Conf.*, in Lecture Notes in Computer Science, vol. 7774. Berlin, Germany: Springer, 2013, pp. 113–128.
[50] G. Rizzo, M. Torchiano, and P. Torino, "KBQ—A tool for knowledge base quality assessment using evolution analysis," in *Proc. Int. Conf. Knowl. Capture*, 2017, pp. 1–6.
[51] B. Ell, D. Vrandecic, and E. P. B. Simperl, "Labels in the Web of data," in *Proc. 10th Int. Semantic Web Conf. (ISWC), Part I*, in Lecture Notes in Computer Science, vol. 7031, L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. F. Noy, and E. Blomqvist, Eds. Bonn, Germany: Springer, Oct. 2011, pp. 162–176, doi: 10.1007/978-3-642-25073-6_11.
[52] T. Knap, J. Michelfeit, and M. Necasky, "Linked open data aggregation: Conflict resolution and aggregate quality," in *Proc. IEEE 36th Annu. Comput. Softw. Appl. Conf. Workshops*, Jul. 2012, pp. 106–111.
[53] M. Färber, F. Bartscherer, C. Menne, and A. Rettinger, "Linked data quality of DBpedia, freebase, OpenCyc, wikidata, and YAGO," *Semantic Web*, vol. 9, no. 1, pp. 77–129, Nov. 2017.
[54] D. Gürdür, J. El-khoury, and M. Nyberg, "Methodology for linked enterprise data quality assessment through information visualizations," *J. Ind. Inf. Integr.*, vol. 15, pp. 191–200, Sep. 2019.
[55] B. Behkamal, "Metrics-driven framework for LOD quality assessment," in *Proc. 11th Int. Conf. Semantic Web, Trends Challenges (ESWC)*, in Lecture Notes in Computer Science, vol. 8465, V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, and A. Tordai, Eds. Crete, Greece: Springer, May 2014, pp. 806–816, doi: 10.1007/978-3-319-07443-6_54.
[56] M. Luggen, D. Difallah, C. Sarasua, G. Demartini, and P. Cudré-Mauroux, "Non-parametric class completeness estimators for collaborative knowledge graphs—The case of Wikidata," in *Proc. Int. Semantic Web Conf.* Springer, 2019, pp. 453–469.
[57] L. Galárraga, S. Razniewski, A. Amarilli, and F. M. Suchanek, "Predicting completeness in knowledge bases," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, Feb. 2017, pp. 375–383.
[58] P. Christen and K. Goiser, "Quality and complexity measures for data linkage and deduplication," in *Quality Measures in Data Mining* (Studies in Computational Intelligence), vol. 43, F. Guillet and H. J. Hamilton, Eds. Springer, 2007, pp. 127–151, doi: 10.1007/978-3-540-44918-8_6.
[59] V. Balaraman, S. Razniewski, and W. Nutt, "Recoin: Relative completeness in Wikidata," in *Proc. Int. Conf. World Wide Web (WWW)*, 2018, pp. 1787–1792.
[60] A. Soulet, A. Giacometti, B. Markhoff, and F. M. Suchanek, "Representativeness of knowledge bases with the generalized Benford's law," in *Proc. Int. Semantic Web Conf.* Springer, 2018, pp. 374–390.
[61] P. N. Mendes, H. Mühleisen, and C. Bizer, "Sieve: Linked data quality assessment and fusion," in *Proc. Joint EDBT/ICDT Workshops (EDBT-ICDT)*, 2012, pp. 116–123.
[62] C. Fürber and M. Hepp, "SWIQA—A semantic Web information quality assessment framework," in *Proc. Eur. Conf. Inf. Syst. (ECIS)*, 2011, p. 76.
[63] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri, "Test-driven evaluation of linked data quality," in *Proc. 23rd Int. Conf. World Wide Web (WWW)*, 2014, pp. 747–758.
[64] N. Micic, D. Neagu, F. Campean, and E. H. Zadeh, "Towards a data quality framework for heterogeneous data," in *Proc. IEEE Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, Jun. 2017, pp. 155–162.
[65] C. Fürber and M. Hepp, "Towards a vocabulary for data quality management in semantic Web architectures," in *Proc. Int. Workshop Linked Web Data Manage.*, New York, NY, USA, 2011. pp. 1–8.
[66] A. Assaf, A. Senart, and R. Troncy, "Towards an objective assessment framework for linked data quality," *Int. J. Semantic Web Inf. Syst.*, vol. 12, no. 3, pp. 111–133, Jul. 2016.
[67] S. McGurk, C. Abela, and J. Debattista, "Towards ontology quality assessment," in *Proc. CEUR Workshop*, 2017, pp. 94–106.
[68] D. Mouromtsev, P. Haase, E. Cherny, D. Pavlov, A. Andreev, and A. Spiridonova, "Towards the Russian linked culture cloud: Data enrichment and publishing," in *Proc. Eur. Semantic Web Conf.*, vol. 9088, 2015, pp. 637–651.
[69] S. Hasan, S. O'Riain, and E. Curry, "Towards unified and native enrichment in event processing systems," in *Proc. 7th ACM Int. Conf. Distrib. Event-Based Syst. (DEBS)*, New York, NY, USA, 2013, p. 171.
[70] A. Jaffri, H. Glaser, and I. Millard, "URI disambiguation in the context of linked data," in *Proc. CEUR Workshop*, vol. 369, 2008, pp. 1–5.

[71] A. Assaf, R. Troncy, and A. Senart, "What's up LOD cloud? Observing the state of linked open data cloud metadata," in *Proc. Eur. Semantic Web Conf.*, vol. 9341, 2015, pp. 247–254.
[72] S. Issa, P.-H. Paris, F. Hamdi, and S. S.-S. Cherfi, "Revealing the conceptual schemas of RDF datasets," in *Proc. Int. Conf. Adv. Inf. Syst. Eng.* Springer, 2019, pp. 312–327.
[73] A. Rula, L. Panziera, M. Palmonari, and A. Maurino, "Capturing the currency of DBpedia descriptions and get insight into their validity," in *Proc. COLD*, 2014, pp. 1–12.
[74] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. D. S. Santos, P. E. Bourne, and J. Bouwman, "The fair guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.
[75] A. Schultz, A. Matteini, R. Isele, N. Pablo Mendes, C. Bizer, and C. Becker, "LDIF—A framework for large-scale linked data integration," in *Proc. 21st Int. World Wide Web Conf. (WWW)*, Apr. 2012, pp. 1–3.
[76] R. Victor Basili, G. Caldiera, and H. Dieter Rombach, *The Goal Question Metric Approach, Volume I*. Hoboken, NJ, USA: Wiley, 1994.
[77] A. Zaveri, A. Maurino, and L.-B. Equille, "Web data quality: Current state and new challenges," *Int. J. Semantic Web Inf. Syst.*, vol. 10, no. 2, pp. 1–6, Apr. 2014.
[78] F. Naumann, *Quality-Driven Query Answering for Integrated Information Systems*, vol. 2261. Springer, 2003.
[79] J. Thomas, A. Noel-Storr, I. Marshall, B. Wallace, S. McDonald, C. Mavergames, P. Glasziou, I. Shemilt, A. Synnot, and T. Turner, "Living systematic reviews: 2. Combining human and machine effort," *J. Clin. Epidemiol.*, vol. 91, pp. 31–37, Nov. 2017.

SUBHI ISSA received the master's degree in computer science from the University of Paris-Saclay, in 2014, and the Ph.D. degree in semantic web technologies and knowledge graph from Conservatoire National des Arts et Métiers, Paris, France, in 2019. His research is focused on assessment and improving of knowledge graph quality on real-world data. His Ph.D. Thesis was titled "Linked Data Quality: Completeness and Conciseness."

His research interests include data science, machine learning, and artificial intelligence.

ONAOPEPO ADEKUNLE received the master's degree in computer science from the African University of Science and Technology, Abuja.

He is currently a Ph.D. Researcher with Institute of Data Science (IDS), Maastricht University. His research is focused on the application of machine learning to improve pension savings in The Netherlands. Before joining IDS in May 2018, he joined the Graduate School of Computer Science in 2015, for postgraduate studies. His research interests include data curation, machine learning, and artificial intelligence.

FAYÇAL HAMDI received the M.Sc. degree and the Ph.D. degree in computer science from the University of Paris-Sud, in 2008 and 2011, respectively.

He is currently an Associate Professor with the Information and Decision Systems Engineering (ISID) Group, Conservatoire National des Arts et Métiers (CNAM), Paris. His research work is mainly focusing on semantic web technologies, ontology alignment, ontology engineering, data integration, and large-scale ontology matching and linked data. In the last few years, he was involved in different semantic web projects, such as WebContent, GeOnto, DataLift, GioQoso, Huma, and SAFECARE, in collaboration with academic and industrial partners. He was the Principal Actor in the realization of the National Institute of the Geographic and Forest Information (IGN) public data portal.

SAMIRA SI-SAID CHERFI received the Ph.D. degree from the University of Paris 1 Panthéon-Sorbonne.

She obtained her accreditation to supervise research at the University of Paris 1 Panthéon-Sorbonne. She is currently a Full Professor and the Head of the Computer Science Department, Conservatoire National des Arts et Métiers. She supervised several Ph.D. students. She has more than 60 papers published in international conferences and journals. She was involved in several projects on data quality, data privacy, and information systems security. Her research interests include information systems methodologies, information systems quality and security, methods and tools for information systems engineering, method engineering, and quality assessment and improvement. She chaired several international events within conferences of her interest domain, such as RCIS, CAiSE, ER, and so on.

MICHEL DUMONTIER is currently a Distinguished Professor of Data Science with Maastricht University, the Founder and the Director of the Institute of Data Science, Maastricht University, and is a Co-Founder of the Findable, Accessible, Interoperable and Reusable (FAIR) Data Principles. His research aims to unlock the potential of data for scientific research. He is an expert in building and mining knowledge graphs for drug discovery and personalized medicine. He is a Principal Investigator with the Dutch National Research Agenda, the European Open Science Cloud, the NCATS Biomedical Data Translator, and the MCSA ITN KnowGraphs. He is internationally recognized for his contributions in bioinformatics, biomedical informatics, and semantic technologies including ontologies and linked data.

Prof. Dumontier is the Editor-in-Chief of the journal *Data Science* and an Associate Editor for the journal *Semantic Web*.

AMRAPALI ZAVERI received the Ph.D. degree in computer science from the University of Leipzig, Germany.

She is currently a Postdoctoral Researcher with the Institute of Data Science, Maastricht University. Previously, she was a Postdoctoral Researcher with the Biomedical Informatics Department, Stanford University, USA. She conducted a comprehensive survey on existing data quality assessment methodologies. Additionally, she evaluated crowdsourcing methodologies for the assessment and improvement of linked data quality as well as biomedical metadata quality. She worked on finding the optimal balance between machine learning and crowdsourcing with non-experts and experts toward data quality assessment. Her research interests include data quality, knowledge interlinking and fusion, and biomedical and health care research.