Data quality assessment of comma separated values using linked data approach

Aparna Nayak, Bojan Božić, and Luca Longo

SFI Centre for Research Training in Machine Learning,
School of Computer Science,
Technological University Dublin, Dublin, Republic of Ireland
{aparna.nayak, bojan.bozic, luca.longo}@tudublin.ie

Business Information Systems Workshops. BIS 2021. Lecture Notes in Business Information Processing, vol 444. Springer, Cham. https://doi.org/10.1007/978-3-031-04216-4_22

Abstract. With an increasing amount of structured data on the web, the need to understand and convert it into linked data is growing. One of the most frequent data formats is Comma Separated Value (CSV). However, it is not easy to describe metadata such as the datatype, data quality and data provenance along with it. Therefore, to publish CSV on the web, it is required to convert CSV into linked data format. Many approaches exist to facilitate the conversion process from structured data to linked data. However, all methods require additional domain knowledge for the conversion process. The goal of this research is to assist publishers in converting CSV files into linked data without human intervention whilst understanding its quality and root causes of data quality violations. The proposed framework consists of two modules. The first module converts the given CSV file into a knowledge graph based on a proposed ontology which is appended with data quality information. In the second module, triples that have violated the data quality constraints are identified. The results show that it is possible to convert a CSV to a knowledge graph by adding its quality information without the help of external mappings.

Keywords: CSV · Data quality · Knowledge graphs · Linked data · Quality assessment · Root cause analysis

1 Introduction

The Semantic Web aims to publish data on the web that can be interpreted by humans and machines. It also helps to reuse, share, and publish data across the web, making it easy to integrate data from different sources. However, significant amounts of data on the web still reside in various formats other than Resource Description Framework (RDF). The rapid growth of data catalogues on the web has led to publishing data in multiple formats. However, the quality of such published data is not known to its users [2,15]. The comma-separated file (tabular data; irrespective of the delimiter used) format is one of the predominant ways to share information that stores data in a two-dimensional array. The resource

format of 37% of the total available dataset in the European data portal [1] is comma-separated (CSV). The remaining 38% of the dataset is in zip format, which includes CSV files. These files contain data in the form of rows and columns as a two-dimensional array. A row represents numerous properties that belong to one instance. A column represents a set of values that belongs to a specific attribute. Converting delimited files to RDF triples that conform to the principles of Semantic Web is an added advantage to the community for reuse and sharing purposes.

Existing approaches require an external mapping to convert any given dataset into RDF triples. This process requires users to have domain knowledge, as they need to be aware of all existing classes in the available ontology to map each column of the CSV file. The proposed method makes use of direct mapping to uplift data to RDF. Direct mapping [1] defines a simple transformation, used to materialize RDF graphs. This method neither requires users to have domain knowledge, nor the dataset should adhere to any ontology/schema. Furthermore, automatic conversion helps to generate RDF triples from datasets that do not adhere to a specific schema or an ontology. On the other hand, mapping languages require the dataset to adhere to an ontology or schema.

CSV to RDF conversion process extracts additional semantic information from the dataset that is added to the triples generated from the dataset. These additional triples later used to assess the quality and it helps to locate the erroneous triples along with the type of violation. It explains users of the dataset to understand which triples have violated a constraint with a precise location in the knowledge graph (KG). The research project aims to determine "To what extent can CSV files be converted to RDF triples with or without human intervention, thus enabling the user to identify root causes of the data quality problems?"

The remainder of this article is structured as follows. Section 2 covers the related work. Section 3 discusses design of the proposed methodology along with use cases. Section 4 summarizes the details of the dataset and results. Finally, section 5 concludes the paper with future work.

2 Related work

Several solutions have been developed to map non RDF data into RDF. This section briefly introduces some tools that help to map CSV to RDF, and data quality metrics helps to assess the RDF. An automatic conversion framework is implemented to deploy linked data on the Pan-European website [5]. This framework reads CSV headers to generate mappings. In another similar approach, CSV files that belong to the same domain are clustered before mapping it into RDF format [14]. However, one of the drawbacks of such method is that the system requires column must be declared as either a data property or as an object property in their ontology. Another problem is that a typographic error

[1] `https://www.europeandataportal.eu/data/eu-international-datasets`

Business Information Systems Workshops. BIS 2021. Lecture Notes in Business Information Processing, vol 444. Springer, Cham. https://doi.org/10.1007/978-3-031-04216-4_22

in the column name does not match any properties in the ontology. On the other hand, CSV files are uplifted with the help of associated metadata [11]. The conversion process fails in the absence of metadata. A comprehensive analysis of data mappings helps to uplift any given dataset to RDF mentioned in [8]. Sheet2RDF [6], is a semi-automatic approach to uplift CSV to RDF triples. The graphical user interface of sheet2rdf enables some refinements without the need to explicitly use the underlying mapping specification language. Data uplifting techniques either require an external mapping function or additional software.

Data quality is a multidimensional concept. There are several frameworks for assessing the quality of linked data. A systematic review conducted in [17] identified a number of different data quality dimensions applicable to Linked Data. Luzzu [4] is one of the frameworks that consider most of the quality dimensions defined by [17] to evaluate the linked data. Semquire [10], Databugger [9], LD Sniffer [12] are some other existing data quality assessment tools/frameworks. However, the root causes of the triples that have violated the constraints are not identified in existing data quality assessment frameworks. Luzzu framework is extended to support root causes of quality violations [16] by identifying the erroneous triples based on the failure of the metric. The data quality metrics are associated with corresponding exception class and problematic parts such as subject, object, and predicate. One of the recent works [3], identifies the constraint violated triples using EYE reasoner[2]. However, all the described constraints are domain-dependent. The proposed technique does not require a user to write any external mappings. In addition, it appends data quality information to the RDF triple store along with the identification of root causes of triples that have violated the data quality constraints.

3 Design

This section gives an overview of the system design that includes a data quality assessment framework that assesses the data quality and reports the quality violated triples, if any. One of the best practise suggested by W3C is to publish data on the web along with data quality information [2]. The main goal of the proposed method is to convert a CSV file into RDF triples that also contains data quality information. RDF data is generated with the help of direct mapping which is appended with quality information. Furthermore, these RDF triples are evaluated to identify erroneous triples.

W3C [7] has described two methods to convert any tabular data into RDF triples. Standard mode conversion frames the information gleaned from the cells of the tabular data with details of the rows, tables, and a group of tables within which that information is provided. Minimal mode conversion includes only the information gleaned from the cells of the tabular data. The proposed model follows minimal mode as the dataset considered for the experiment is not an annotated file. Most of the vocabulary terms mentioned by W3C are applicable

[2] `http://eulersharp.sourceforge.net/`

Business Information Systems Workshops. BIS 2021. Lecture Notes in Business Information Processing, vol 444. Springer, Cham. https://doi.org/10.1007/978-3-031-04216-4_22

only for annotated CSV files. However, the standard mode of conversion is used wherever appropriate.

The proposed CSV2RDF ontology conceptually maps all the components of the CSV file to the classes of the ontology as described in figure 1. It connects dataset properties to the data arranged in the form of rows and columns. The contents of the CSV file is converted, as shown in figure 2. Each row of the CSV is considered to be a class, as mentioned in the CSV2RDF ontology. All the columns of the dataset are considered object properties of the row to maintain the order of data. All the cell values are translated into RDF literals for the given CSV file.

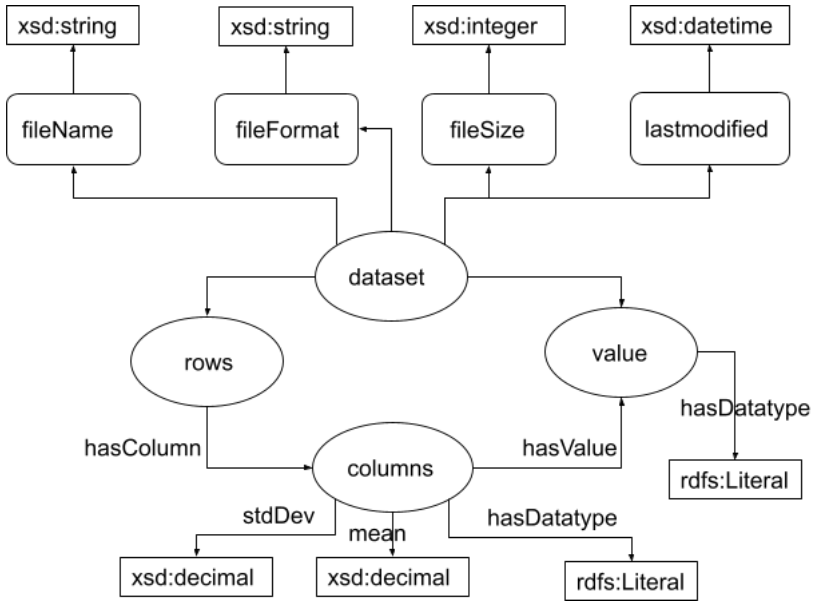

Fig. 1. Ontology describing CSV2RDF data

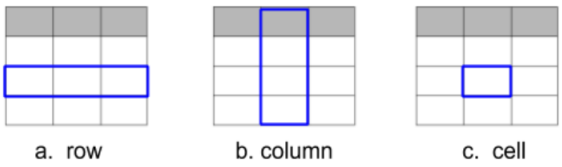

Fig. 2. Row-Column-Cell as RDF triples((a) subject - (b) predicate - (c) object

Business Information Systems Workshops. BIS 2021. Lecture Notes in Business Information Processing, vol 444. Springer, Cham. https://doi.org/10.1007/978-3-031-04216-4_22

Data quality assessment pipeline in case of direct mapping is as shown in figure 3. The triple modeling module is responsible for uplifting CSV into RDF. The proposed ontology is used to convert CSV into RDF. The data module is responsible for collecting all the required RDF triples, axioms, and constraints. Axioms contain CSV2RDF ontology. Constraints represent the rules that are used to identify the quality violated triples. The conversion process generated RDF triples along with meta-information such as file name, file format, last modified, and data quality information. The quality metrics used to evaluate the knowledge graph are obtained from the comprehensive list of data quality metrics [17].

Typed literals are created by comparing the lexical token of each cell with basic datatypes. Datatypes such as int, float, date, string and URI are considered in this experiment. RDF:nil is assigned as the datatype for all the empty cells present in the CSV. RDFS:Literal is assigned to all other remaining cells if they do not match any of the afore-mentioned basic datatypes. The datatype is predicted for each column based on values that are present in the column. The lexical form of a cell value is compared with all considered datatype using a regular expression, and matched datatypes are counted. The datatype which has the highest value is assigned to the column.

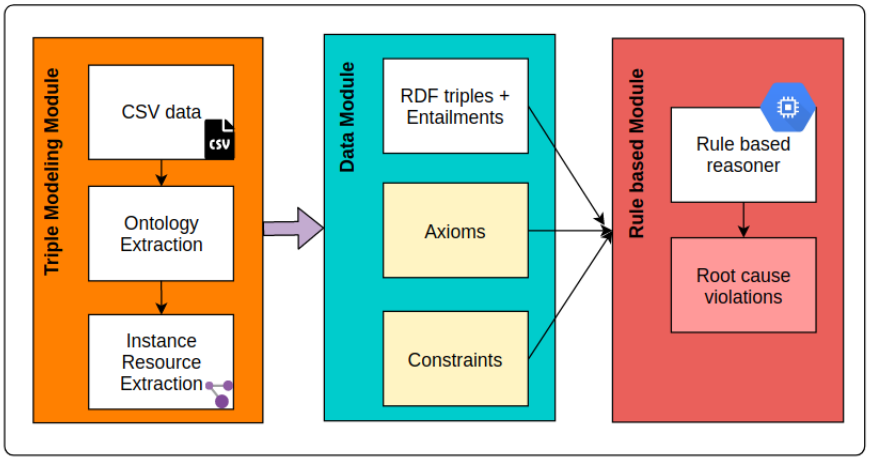

Fig. 3. CSV2RDF direct mapping with data quality assessment

The RDF triples are given as input to the Apache Jena framework. The Jena inference subsystem is designed to allow a range of inference engines or reasoners to be plugged into Jena. Such engines are used to derive additional RDF assertions, which are entailed from some base RDF together with any optional ontology information, the axioms and rules associated with the reasoner. The conventional model makes use of a single reasoner. The proposed model benefits

Business Information Systems Workshops. BIS 2021. Lecture Notes in Business Information Processing, vol 444. Springer, Cham. https://doi.org/10.1007/978-3-031-04216-4_22

from the cascaded reasoner. RDF triples and OWL axioms are given as input to the OWL reasoner, as shown in figure 4. The inferential model generated by the OWL reasoner is given as input to the generic reasoner. The generic reasoner validates the OWL inferential model against the user-defined constraints, and the resulting violations are reported to the user in triple format.

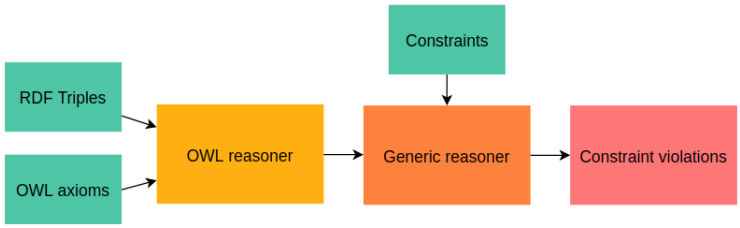

Fig. 4. Cascaded reasoner in Apache Jena

The generic reasoner admits RDF triples as input to validate them against the constraints and axioms. Constraints help to identify erroneous triples in the dataset. Data quality problems such as the presence of outliers and datatype mismatch have been identified using constraints. The explanatory message reported by the reasoner helps the user to understand the exact position of the erroneous triple in the dataset, thereby identifying the root causes of the quality problems. Constraints are written in Datalog as shown in figure 5.

The knowledge graph is constructed by considering metrics that are listed in table 1. Considering the KG is constructed from raw data quality metrics such as deprecated class/property, use of long URIs and some other metrics have been taken care of. Metrics listed under the domain ontology dependent column are assessed with the help of an OWL reasoner. Root cause analysis of ontology dependent metrics is accomplished with the help of the cascaded OWL reasoner. The cascaded reasoner evaluates ontology dependent quality metrics followed by root cause analysis of the metrics mentioned in table 2.

Table 2 denotes the data quality metrics that are implemented for quality assessment purpose. Assessment refers to the evaluation of the entire dataset for the specific metric. Root cause analysis refers to the detection of triples that have violated a constraint that describes the specific quality metric. All the metrics used for the evaluation are considered from [17]. All columns are assumed to have a header while converting CSV to RDF triple. Semantic Web typically follows the open-world assumption(what is not known to be true or false might be true, or the absence of information is interpreted as unknown information, not as negative information). In this research, the quality metric population completeness is determined by considering the closed world assumption.

Datalog constraints are written to identify the triples that have violated the constraints. Rule R1 in figure 5 compares the datatype of the column with the

Business Information Systems Workshops. BIS 2021. Lecture Notes in Business Information Processing, vol 444. Springer, Cham. https://doi.org/10.1007/978-3-031-04216-4_22

Table 1. Data quality metrics and corresponding dependency

Metric	KG from raw data	Domain ontology dependent
No use of deprecated class/property	✓	✗
No misreported content types	✓	✗
Detection of long URI	✓	✗
Syntax error	✓	✗
Prolix RDF features	✓	✗
Inverse-functional properties	✗	✓
No use of entities as members of disjoint classes	✗	✓
Correct domain/range definition	✗	✓
Misplaced classes and properties	✗	✓

Table 2. Implemented data quality metrics

Metric	Assessment	Root cause identification	Dimension
Detection of ill typed literals	✓	✓	Syntactic validity
Population completeness	✓	✗	Completeness
No inaccurate values	✓	✗	Semantic accuracy
No outliers	✓	✓	Semantic accuracy
Language used	✓	✗	Interpretability

data type of the value. Datatype mismatch error is thrown in case of a mismatch between the column and the value. Rules R2 and R3 in the figure 5 helps to locate the outliers based on lower and upper quantiles. Outliers are computed only for columns that have integer datatype.

The users of this framework will be either CSV data consumers who wish to uplift their data to RDF or publishers who wish to keep multiple data source format. The following use cases were listed based on the functionalities that the framework supports.

UC1: Uplifting of CSV: Users who have CSV files can uplift their data to RDF. Direct mapping is a simple technique to map CSV to RDF. It also helps the naive users to get the RDF format of data.

UC2: Assessment of data: Users should be able to assess data. Linked data that is uplifted from CSV is assessed to understand the quality of data.

UC3: View data quality problems: Users will get the data quality problems in the form of a report that is generated using data quality vocabulary [13]. This helps the user to understand the exact cause of the problem.

Business Information Systems Workshops. BIS 2021. Lecture Notes in Business Information Processing, vol 444. Springer, Cham. https://doi.org/10.1007/978-3-031-04216-4_22

```
[R1: (?r ?c ?o) (?o ns1:hasValue ?d)
(?c ns1:hasDatatype ?e) notDType(?d, ?e) ->
(?c ns1:constraintType ns1:datatypemismatch)
(?c ns1:expectedDT ?e) (?c ns1:foundDT ?d)
(?c ns1:constraintElement ?r) ]

[R2: (?r ?c ?o) (?o ns1:hasValue ?v)
(?c ns1:stdDev ?std) (?c ns1:mean ?m)
product(?std,3,?threshold) difference(?m ?threshold
?LL)  lessThan(?v, ?LL) -> (?c ns1:constraintElement ?v)
(?c ns1:contraintType ns1:OutlierLowerThreshold)  ]

[R3: (?r ?c ?o) (?o ns1:hasValue ?v)
(?c ns1:stdDev ?std) (?c ns1:mean ?m)
product(?std,3,?threshold) sum(?m ?threshold ?UL)
greaterThan(?v, ?UL) -> (?c ns1:constraintElement ?v)
(?c ns1:contraintType ns1:OutlierUpperThreshold) ]
```

Fig. 5. Datalog constraints for datatype mismatch and outlier detection

4 Results

The experiment is conducted by considering CSV files from various themes such as agriculture, education, health, house, and synthetic data. Table 3 represents dataset information considered for the experiment. The table gives information about the number of rows, number of columns, and different datatype that the file contains. One of the significant outputs of root cause violations of the datatype

Table 3. Dataset information

Dataset	Rows	Columns	Datatype
Agriculture	18	7	String, URL, Decimal, Integer
Education	13	9	String, Integer, URL
Health	18	16	Integer, String
House	21613	16	Integer, Decimal
Synthetic data	10	9	Integer, String, URL, Decimal

mismatch is as shown in the listing 6. The exact location of the quality problem is highlighted in error thrown by the program. It includes row number, column name, expected value and found value. The datatype mismatch is found as a result of additional space in the beginning of the email id which was present in the original dataset. The detailed analysis of the problem helps the user to understand data quality in the dataset. For example, in listing 6, *ns1:R2* refers to the row number and data property is *ns1:Email*. The expected datatype refers to the datatype of the column, and found datatype indicates the datatype associated with the literal. This is one of the significant outputs found for the datatype

Business Information Systems Workshops. BIS 2021. Lecture Notes in Business Information Processing, vol 444. Springer, Cham. https://doi.org/10.1007/978-3-031-04216-4_22

mismatch when the RDF triples are verified against the constraints mentioned in listing 5. Additional datatype mismatch triples are the result of the presence of *NULL* values. The algorithm assigns *RDF: nil* to literal values if it finds any empty cells. When such values are compared against a column, a datatype error is thrown.

```
ns1:Email ns1:constraintElement ns1:R2 ;
        ns1:constraintType ns1:datatypemismatch;
        ns1:expectedDT xsd:string ; ns1:foundDT
        " braylib@wocklowcoco.ie"^^xsd:decimal .
```

Fig. 6. Root cause violation : Datatype mismatch

Outlier constraints (R2 and R3 in figure 5) were not projected due to the fact that no outliers were present in the dataset. The synthetic data is generated to locate all the root causes that the experiment addresses. The experiment is publicly available on Github [3].

5 Conclusions and future work

This research presents an ontology-based framework for constructing the knowledge graph from a CSV file without human intervention. The proposed framework consists of two core modules. The first module deals with converting the representation of knowledge from frames to semantic networks, which are also appended with data quality information. The second part of the work deals with building a knowledge base from production rules to identify triples that have violated the constraints. Creating a knowledge graph from raw data benefits from incorporating multiple quality metrics.

Some of the limitations of this research are i. The annotated CSV files are not considered ii. Semantic enrichment of the knowledge graph. Annotated CSV files contain additional metadata that can be used to identify similar entities on web using natural language processing. Semantic enrichment of the knowledge graph addresses mapping the contents of CSV file to DBpedia classes thus connecting data silos. The future work of this research tries to overcome the limitations of the system along with incorporating more data quality metrics and automatic refinement based on quality violation explanations.

Acknowledgements This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

[3] `https://github.com/aparnanayakn/csvdataqualityassessment`

Business Information Systems Workshops. BIS 2021. Lecture Notes in Business Information Processing, vol 444. Springer, Cham. https://doi.org/10.1007/978-3-031-04216-4_22

References

1. Arenas, M., Bertails, A., Prud'hommeaux, E., Sequeda, J.: A direct mapping of relational data to rdf. W3C recommendation **27**, 1–11 (2012)
2. Bernadette Farias Lóscio, Caroline Burle, N.C.: Data on the web best practices. W3C recommendation (2017)
3. De Meester, B., Heyvaert, P., Arndt, D., Dimou, A., Verborgh, R.: Rdf graph validation using rule-based reasoning. Semantic Web (Preprint), 1–26 (2020)
4. Debattista, J., Auer, S., Lange, C.: Luzzu—a methodology and framework for linked data quality assessment. Journal of Data and Information Quality (JDIQ) **8**(1), 1–32 (2016)
5. Ermilov, I., Auer, S., Stadler, C.: User-driven semantic mapping of tabular data. In: Proceedings of the 9th International Conference on Semantic Systems. p. 105–112. Association for Computing Machinery, New York, NY, USA (2013)
6. Fiorelli, M., Lorenzetti, T., Pazienza, M., Stellato, A., Turbati, A.: Sheet2rdf: A flexible and dynamic spreadsheet importlifting framework for rdf. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **9101**, 131–140 (2015)
7. Jeremy Tandy, Ivan Herman, G.K.: Generating rdf from tabular data on the web. W3C recommendation (2015)
8. Junior, A.C., Debruyne, C., Brennan, R., O'Sullivan, D.: Funul: A method to incorporate functions into uplift mapping languages. In: Proceedings of the 18th International Conference on Information Integration and Web-Based Applications and Services. p. 267–275. iiWAS '16, Association for Computing Machinery, New York, NY, USA (2016)
9. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R.: Databugger: A test-driven framework for debugging the web of data. pp. 115–118. Association for Computing Machinery, Inc (2014)
10. Langer, A., Siegert, V., Göpfert, C., Gaedke, M.: Semquire - assessing the data quality of linked open data sources based on dqv. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11153**, 163–175 (2018)
11. Mahmud, S., Hossin, M., Hasan, M., Jahan, H., Noori, S., Ahmed, M.: Publishing csv data as linked data on the web. Lecture Notes in Electrical Engineering **605**, 805–817 (2020)
12. Mihindukulasooriya, N., García-Castro, R., Gómez-Pérez, A.: Ld sniffer: A quality assessment tool for measuring the accessibility of linked data. In: European Knowledge Acquisition Workshop. pp. 149–152. Springer (2016)
13. Riccardo Albertoni, A.I.: Data on the web best practices: Data quality vocabulary. W3C recommendation (2016)
14. Sharma, K., Marjit, U., Biswas, U.: Automatically converting tabular data to rdf: an ontological approach. International journal of Web Semantic Technology **6**, 71–86 (2015)
15. Umbrich, J., Neumaier, S., Polleres, A.: Quality assessment and evolution of open data portals. In: 2015 3rd international conference on future internet of things and cloud. pp. 404–411. IEEE (2015)
16. Vaidyambath, R., Debattista, J., Srivatsa, N., Brennan, R.: An Intelligent Linked Data Quality Dashboard. In: AICS 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science. pp. 1–12
17. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web **7**(1), 63–93 (2016)

Business Information Systems Workshops. BIS 2021. Lecture Notes in Business Information Processing, vol 444. Springer, Cham. https://doi.org/10.1007/978-3-031-04216-4_22