

# From Data Quality to Big Data Quality: A Data Integration Scenario

Carlo Batini<sup>1</sup>, Anisa Rula<sup>2</sup>

<sup>1</sup>University of Milano-Bicocca

<sup>2</sup>University of Brescia

## Abstract

Big data has made its appearance in many fields, including scientific research, business, public administration and so on. Although, it is acknowledged that there exist different aspects (e.g., acquisition of data, extraction, pre-processing, analysis modelling and functionality, interpretation, etc.) that might affect the benefit of such data, several authors identify *data quality* as the most decisive one. More recently, a variety of data types have arisen from linguistic and visual information, used and diffused through social networks, Internet of things, enterprise and public sector information systems as well as the Web. The big data phenomenon has deeply impacted on the diversity of types of data. In our previous work, we provided a deep investigation on how data quality concepts can be extended to such vast set of data types, encompassing, e.g., semi-structured texts, maps, images and linked data. In this work, we focus on Linked Data, a type of data that can be viewed as big data and study the effect of data quality in a data integration scenario.

## Keywords

Data Quality, Semantic Annotations, Tabular Data

## 1. Introduction

Big data has made its appearance in many fields such as scientific research, business and public administration. There are different definitions provided for the concept of big data [1]. According to Gartner, "Big data is high-volume, high-velocity and/or high-variety information assets that demands cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."<sup>1</sup> In other words, big data goes beyond large amount of data by exceeding the capabilities of traditional computing environments [2]. Although it is acknowledged that there exist different aspects (e.g., acquisition of data, extraction, pre-processing, analysis modelling and functionality, interpretation, etc.) that might affect the benefit of such data [3], several authors identify *data quality* as the most decisive one [2]. Consequently, using data with adequate levels of data quality is paramount to obtain the maximum benefit.

There exist many methodologies proposed in the information system and database communities to identify and manage the quality of the published data, all addressing different aspects of


---

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy

✉ carlo.batini@unimib.it (C. Batini); anisa.rula@unibs.it (A. Rula)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

<sup>1</sup><http://www.gartner.com/it-glossary/big-data/>

quality assessment by proposing appropriate dimensions, measures and tools [4, 5, 6, 7]. More recently, a variety of data types have arisen from linguistic and visual information, used and diffused through social networks, Internet of things, enterprise and public sector information systems as well as the Web. The big data phenomenon has deeply impacted on the diversity of types of data. In this work, we focus on linked data, a type of data that can be viewed as big data [8]. In recent years, the linked data paradigm has emerged as a simple mechanism for employing the Web as a medium for data and knowledge integration where both documents and data are linked. Linked data essentially refers to a set of best practices for publishing and connecting structured data on the Web, which allows publishing and exchanging information in an interoperable and reusable fashion [9]. Many different communities on the Internet such as geographic, media, life sciences and government communities have already adopted these linked data principles. This is confirmed by the dramatically growing linked data Web, where currently more than 150 billion facts are represented<sup>2</sup>.

Modelling and constructing linked data poses different challenges with respect to the three coordinates: *volume*, *variety* and *velocity*. First, the increasing diffusion of the linked data paradigm as a standard way to share knowledge on the Web allows consumers to fully exploit vast amount of open structured data that were not available in the past (high *volume* of data). Second, linked data refers to a Web-scale knowledge base consisting of interlinked published data from a multitude of *autonomous* information providers<sup>3</sup>. Although it is suggested to reuse as much as possible existing vocabularies, still there exist a lot of heterogeneous datasets (*variety* of data). Third, linked data can be considered as a dynamic environment where information can change rapidly and cannot be assumed to be static (*velocity* of data) [10]. In addition to the aforementioned three characteristics, there exists a fourth characteristic named *veracity*. Transforming data into linked data does not necessarily require the definition of a schema thus following a schema-last approach that first publishes the data and subsequently (and optionally) creates the schema. Therefore, the veracity of data should be checked since linked data does not intrinsically prevent to add incorrect or inconsistent information to the datasets.

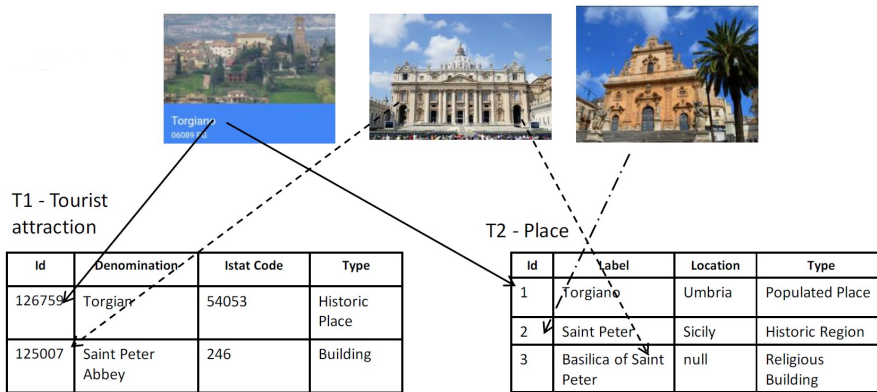
Quality dimensions and metrics are important for measuring quality of linked data. However, an important question is "How do data quality dimensions and metrics evolve in the transformation process from tabular to linked data?" To answer this question, we need to study the evolution of quality dimensions from traditional databases to linked data based on the differences of the structural characteristics of both data models. A study of the evolution between quality dimensions of relational data and linked data according to the differences in their structural characteristics, enables us managing the huge variability of methods and techniques needed to manage data quality in linked data. In this paper, we provide a data integration use case to show that data integration over semantically enriched datasets provides a higher quality result than integrating two relational datasets. In other words, we would like to answer to the following question: *How data quality drives the data integration process?*

**Running Example** Throughout this manuscript, we consider a data integration use case study having two relation tables (see Figure 1). The two tables contain information about Tourist

---

<sup>2</sup><http://lod-cloud.net/state/>

<sup>3</sup>linked data is different from centralized databases that are designed and built by a (single) database administrator.



**Figure 1:** Two relational datasets to be used in the running example.

Attraction (hereafter called T1) and Place (hereafter called T2). T2 contains three instances of place of interest, consisting of the Torgiano locality, the Basilica of Saint Peter, seen as a religious building and the Saint Peter in Modica in Sicily, seen as a place of interest in the historic region (Historic Region). Figure 1 shows the correspondences between the objects of the real world (the observables) and the records of the two tables. We consider the integration process by keeping in mind the quality of the sources and how it will influence the final integration.

**Organizational Structure** This paper is organized as follows. Section 2 presents related work on the mapping of relational databases to linked data and the quality assessment of the RDF mappings. Section 3 presents the differences and similarities of structural characteristics as well as other important differences between the two data models. In Section 4, we analyse the evolution of quality metrics according to the classification model defined in the previous section. Section 5 presents a use case implementing our methodology. Finally, we conclude in Section 6.

## 2. Related Work

In this section we provide details about the state of the art regarding the relational model to linked data mapping problem and the quality of the mappings. This section is divided into two subsections. Section 2.1 compares the characteristics of the relational and linked data and describes the mapping approaches. Section 2.2 describes the quality assessment approaches.

### 2.1. Relational model and linked data

The survey in [11] analyses approaches that tackle the problem from database to ontology mappings. The authors use different criteria for analysing the approaches such as mapping languages or software availability. Before proposing the classification of the approaches, the authors first analyse the characteristics of the data models from Extended Entity Relationship (EER) model to relational model and finally to RDF model.

In addition, the authors in [11] propose other characteristics related to the RDF data model and RDFS as follows: i) classes of individual resources, ii) properties, connecting two resources, iii) hierarchies of classes, iv) hierarchies of properties, and v) domain and range constraints on properties. In this work, we provide both characteristics at instance and schema level. In addition, at instance level we include the position of the elements in a triple e.g. subjects vs objects. The work in [12] identifies 25 structural characteristics separately for the ontology and the relational model. These characteristics were collected as a summary of answers to the following questions (used to understand the differences and similarities): What is it for?, What does it look like? How do you build one? How is it implemented and used? and Where are the semantics? The author in [13] studies the differences between the Semantic Web and the Relational Databases. In particular, he studies the mapping languages. There exists a group on W3C named RDB2RDF<sup>4</sup> that worked on two language mappings standards. The two proposed standards are named Direct Mapping<sup>5</sup> and R2RML<sup>6</sup>.

## 2.2. Quality Assessment of the Mappings

The database community has deeply investigated the problem of data quality by proposing different methodologies and frameworks [4, 5, 6, 7]. Most of these works have focused on data in relational models, traditionally adopted in Data Base Management Systems [4]. Nevertheless, a variety of data types such as semi-structured texts and linked data, produced over social networks and the Web resulted in an investigation on how data quality concepts can be extended. In [14], we started to investigate the evolution and the adoption of quality metrics according to the structural characteristics of linked data, while in this work we study how the adoption of quality metrics support the data integration process.

Different approaches have been proposed to assess the quality of linked data. These approaches can be distinguished in those applied to the quality assessment of datasets [15, 16] and mapping definitions [17] which can be classified as i) manual, ii) semi-automatic and iii) automatic.

In particular, the work in [15] focuses on the definitions and formalization of quality assessment metrics for linked data. In a more recent work, [16] proposes the formalization of quality metrics from the practical and implementation point of view. In [18], the authors evaluate the quality assessment of crawled datasets containing around 12M RDF triples. The aim was to discuss common problems found in RDF datasets, and possible solutions. More specifically, the work aimed at uncovering errors related to accessibility, reasoning, syntactical and non-authoritative contributions. The authors also provided suggestions on how publishers can improve their data, so that the consumers can find "high quality" datasets. However, all these approaches focus on the quality assessment of the dataset and do not study the evolution of quality metrics; furthermore, they do not provide any evidence about the model adopted for the generation of new quality metrics for linked data based on existing metrics for relational databases.

Whereas [19] and [20] derive actual quality checks from the dataset under assessment, a

---

<sup>4</sup><https://www.w3.org/2001/sw/rdb2rdf/>

<sup>5</sup><https://www.w3.org/TR/rdb-direct-mapping/>

<sup>6</sup><https://www.w3.org/TR/r2rml/>

mapping quality assessment method should take the transformation process into account. Only a few works provide attempts for the quality assessment of RDF mappings [21, 17, 22]. Dimou, et. al. [22] demonstrate that assessing an RDF dataset requires a considerable measure of time, therefore it cannot be often executed, and when that happens, the violations' root is not detected. Despite what might be expected, specifically assessing RDF mappings requires essentially less effort, while the violations' root is detected. In [17], they assess mappings from semi-structured data to RDF by proposing incremental, iterative and uniform validation workflow where violation might derive from incorrect usage of schemas; in addition, they suggest mapping refinements based on the results of these quality assessments. Similar to our work, they focus on three quality dimensions that are completeness, accuracy and consistency. A more recent work in [21] proposes a predicting algorithm to detect incorrect mappings which is applied to the DBpedia dataset. The corresponding discussions do not include any quality metric preservation. So, while there are metrics dedicated to other domains that could be re-used for the RDF quality assessment mapping, there is (to the best of our knowledge) no study to support guiding the generation of RDF mappings based on a quality evaluation of these transformations. Moreover, we could not find an existing method or methodology whose target is the improvement of the quality of RDF integration results through the quality assessment and improvements of mappings.

### 3. Structural Characteristics of Linked Data

A common classification framework characterized by several quality dimensions, allows us to compare dimensions across different data types. The framework is based on a classification in clusters of dimensions proposed in Batini et al. (2012) where dimensions are included in the same cluster according to their similarity.

We consider three main types of data that can be viewed as BD (i) maps, (ii) semi-structured texts and (iii) linked data. Each one of these data types has inherently associated several structural characteristics, which are relevant for the investigation of quality dimensions defined in the literature. In this section we discuss in detail the structural characteristics for linked data.

**Linked Data.** The Web has been in the last years an extraordinary vehicle of production, diffusion, and exchange of information. Data as the lowest level of abstraction, from which information is derived, can be provided on the Web as open data under the open data initiative. Open data is mainly provided in different domains including economy, science, employment, environment and education. Open data gain popularity with the rise of the Internet and World Wide Web especially, with the launch of open-data government initiatives. The philosophy behind open data has been long established in public bodies, while the term "open data" itself is recent. Open data is data that can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike. Open data become open linked data when, according to Tim Berners-Lee (2006):

- Information is available on the Web (any format) under an open license.
- Information is available as structured data (e.g. an Excel sheet instead of an image scan of a table).
- Non-proprietary formats are used (e.g. CSV instead of MS Excel).

- URI identification is used so that people can point at individual data.
- Data is linked to other data to provide context.

Linked data enables publishers to link and publish structured data by generating semantic connections among data sets. Linked data exhibits structural characteristics referring to the issues discussed in [14].

#### 4. Evolution of quality dimensions in Linked Data

We have studied the evolution of quality dimensions and metrics in [14]. As to linked data, we report here the evolution of one quality dimension named accessibility to show how the structural characteristics from one data model to the other influence the way quality is measured. **Accessibility in relational data.** Accessibility measures the ability of the user to access data from his or her own culture, physical status/functions, and technologies available. Several guidelines are provided by international and national bodies to govern the production of data, applications, services, and Web sites in order to guarantee accessibility, with specific concern on accessibility for disabled persons. Guidelines referring to relational tables in Web sites are provided by the World Wide Web Consortium in (WWWC); guidelines identify the characteristics of the HTML representation of tables to be made accessible by means of assistive technologies, for example: for all data tables, identify row and column headers; for data tables that have two or more logical levels of row or column headers, use markup to associate data cells and header cells; for data tables elements, label elements with the "scope", "headers", and "axis" attributes, so that future browsers and assistive technologies will be able to select data from a table by filtering on categories. **Accessibility in linked data.** Public bodies, for reasons of transparency and accessibility, have progressively published public data in order to enable citizens to access data for their own purposes and interests. To make the data accessible in a standard way, the first step is that to release the format of data from proprietary formats to open formats (i.e. RDF), which are not only understood by humans, but also by machines. The format issue is considered in several structural characteristics discussed for linked data in the previous section, corresponding to several possible mechanisms that can be adopted to improve accessibility. In Figure 2 we classify the relevant quality dimensions according to such mechanisms. One mechanism can be the use of HTTP URI, a combination of globally unique identification (through URIs) and a retrieval mechanism (through HTTP), which enables the identification of objects and abstract concepts and their descriptions; in this case the accessibility dimension refers to dereferencability, or resource accessibility. To make datasets available through SPARQL endpoints, the user should indicate the URI of the dataset and the location of the corresponding SPARQL endpoint and should check whether the server responds to a SPARQL query; in this case we refer to dataset accessibility. A further mechanism to access a dataset is by making an RDF dump available for download; in this way the location of the RDF dump can be exploited, and we refer to browsing accessibility.

In order to specify the connection between real world objects, a mechanism of interlinking has been proposed based on the RDF links. Interlinking refers to the degree to which objects are linked to each other, be it within or between two or more data sources. It represents a relevant dimension for accessibility in linked data, since the process of data integration is made



	Structural characteristics				
	<i>Dereferenceable Resource</i>	<i>SPARQL Endpoint</i>	<i>RDF dumps</i>	<i>Interlinking</i>	<i>Licensing</i>
Quality dimension	<i>Quality sub-dimension</i>				
<i>Accessibility</i>	Resource accessibility	Dataset accessibility	Browsing accessibility	Integration accessibility	Reuse accessibility

**Figure 2:** Quality dimensions of linked data classified by linked data structural characteristics.

possible through the links created between various data sets. In this case the accessibility dimension corresponds to integration accessibility, since RDF links describe the relationship between objects and enables discovering new data through integration. Previous approaches to accessibility have evolved to investigate the new juridical licensing aspects of data. Licensing is a new quality dimension not considered for relational databases but mandatory in an open data world. Providing licensing information is an indication of how much data is accessible to be potentially re-used, based on the specification of legal rights and allowances; in this case the accessibility dimension corresponds to reuse accessibility.

As envisioned by this section the evolution of data quality dimensions according to the research coordinates that are relevant in Big Data, such as the variety of data types, is important to be studied to have an understanding of data quality assessment in a new settings of information. In Section 5 we study two additional quality dimensions: the accuracy and the completeness. Our aim is to study the advantage/disadvantage in terms of data quality of the data integration process according to traditional techniques and the "new" techniques applied in linked data.

## 5. Use Case

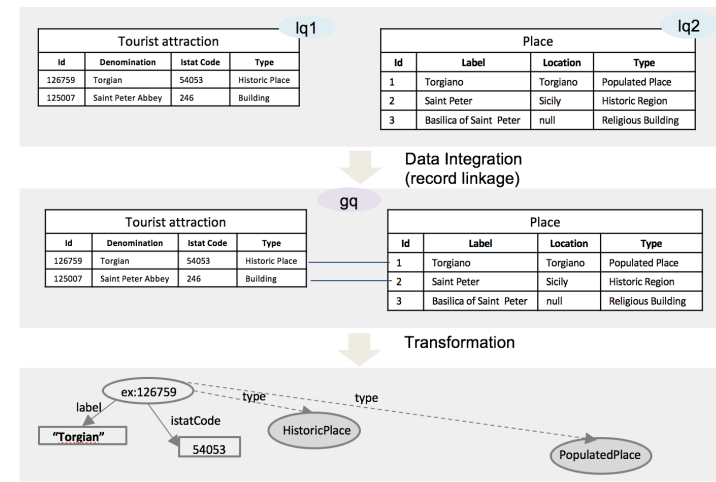
This section presents a practical data integration use case that shows how the transformations can affect the overall quality of single and integrated datasets. We show a data integration process on two datasets performed according to two different workflows. In Section 5.1, the first workflow considers the integration of the two tabular datasets based on traditional techniques and after the transformation of the integrated dataset into linked data. In Section 5.2, the second workflow considers the transformation in linked data of the two datasets and after their integration. In both workflows the two datasets of Tourist attraction (T1) and Place (T2) of the running example introduced in Section 1 are used. We simplified the schema information by using the same names for the same attributes when possible since we are not interested in schema matching but in instance matching.

**Instance Matching.** Given two records  $r_1$  and  $r_2$  which represent two distinct observables in the universe, an integration process must allow us to decide whether:

- $r_1$  and  $r_2$  refer to the same observable, in this case there exists a match between the two objects.
- $r_1$  and  $r_2$  do not refer to the same observable, in this case there is no match between the two objects.

In the first case, we refer to either *true positives*, or to *false positives*. In the second case, we refer either to true negatives or to false negatives. The quality of the distance measurement should be an indication of minimizing false positives or false negatives. In the following, we discuss the two workflows.

## 5.1. Integration vs Transformation



**Figure 3:** First workflow: data integration first and then transformation.

In the first workflow (see Figure 3), the integration of the two datasets is performed first by using traditional techniques such as record linkage followed by the transformation into RDF and enrichment with external knowledge bases.

- We measure the distance between the first record of T1 (*Torgian*) and the first record of T2 (*Torgiano*) on *Denomination* and *Label* attributes respectively, see Figure 3. The edit distance returns 1, thus there is a **match**.
- We measure the distance between the second record of T1 (*Saint Peter Abbey*) and the second record of T2 (*Saint Peter*) *Denomination* and *Label* attributes respectively. Jaccard distance function returns 0.33, thus there is a **match**.
- We measure the distance between the second record of T1 (*Saint Peter Abbey*) and the third record of T2 (*Basilica of Saint Peter*) *Denomination* and *Label* attributes respectively. The Jaccard distance function returns 0.60, thus there is a **non-match**.

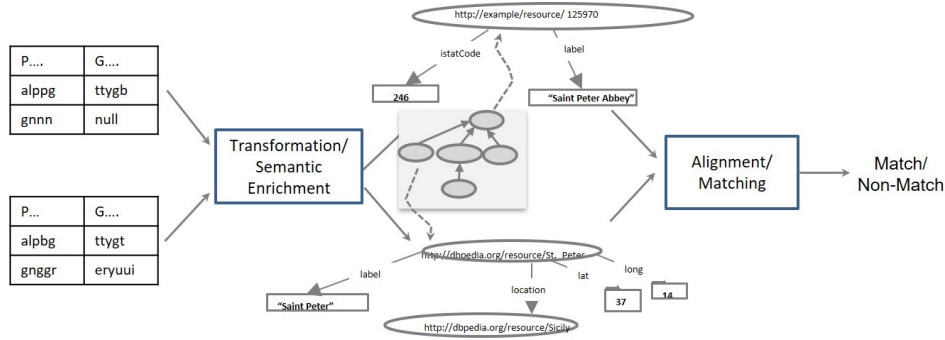
The final result compared with the “true” matches are: a true positive (R11 and R21 records), a false positive (R12 and R22 records), a false negative (R12 records and R23 records).

## 5.2. Transformation vs Integration

In the second workflow, the integration of the datasets is performed after the transformation and enrichment steps, see Figure 4. In the semantic transformation and enrichment steps, the

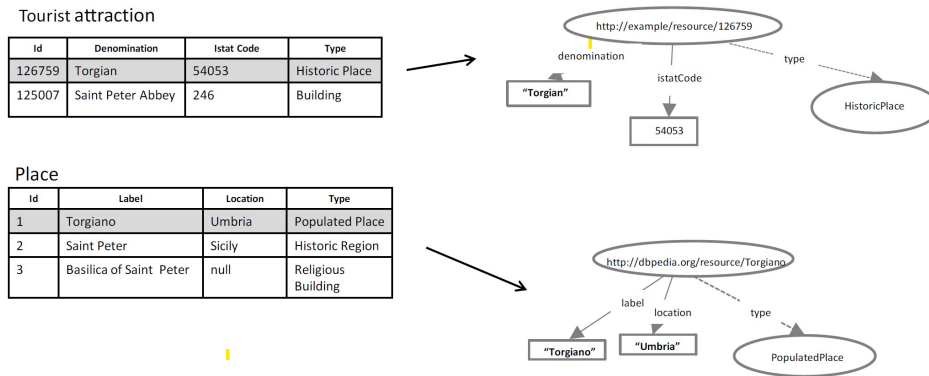


datasets are transformed into a semantically richer model, the RDF model, which, by the virtue of its graph structure, it can exploit semantic resources and techniques available on the Web. For example, we can connect some of the data present in the two graphs to a taxonomy or ontology, so that it is possible to compare the data not only by means of distances measured between strings, but exploiting their meaning provided by the taxonomies.



**Figure 4:** Second workflow: transformation first and then integration.

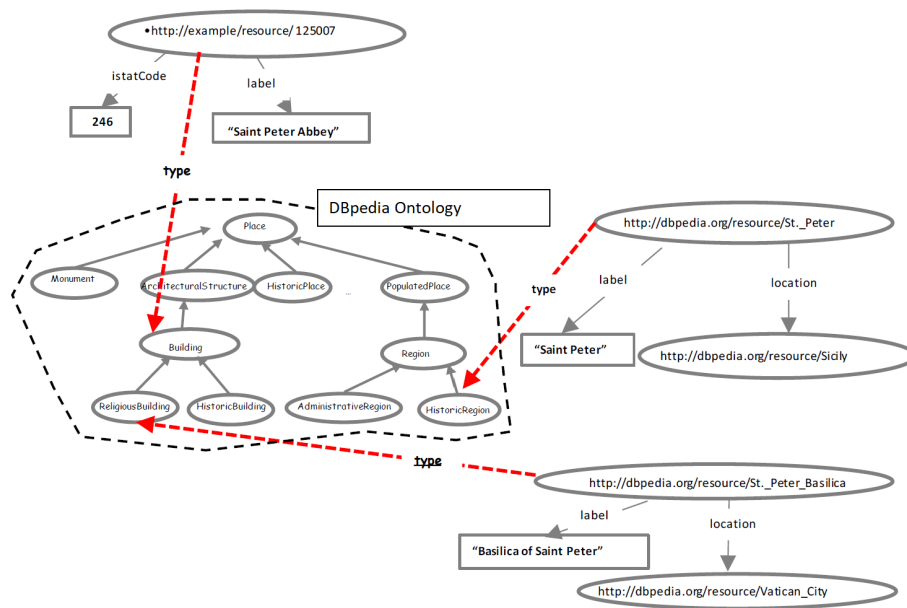
First, we transform the tables from the relational model into the RDF model; Figure 5 shows the RDF graph relating to the first records of the two tables (R11 and R21) and then those relating to the second record of the Tourist attraction table (R21) and to the second and third records of the Place table (R22 and R23).



**Figure 5:** Second workflow: transformation into RDF of the records R11, R21.

After finishing with the transformation we can compare the values in the triples, i.e., *Torgiano* and *Torgian*. A similarity metric based on the edit distance between the two values is applied. The result obtained is 1 meaning that the two records represent the same observable object. We achieved the same result in the first workflow too. To further enrich our tabular data, we use the DBpedia ontology, available in the Linked Open Data Cloud (<https://wiki.dbpedia.org/>).

At this point, we can compare the similarity of the entities not only on similarity metrics based on the edit distances between two strings but based on the meaning of the two entities



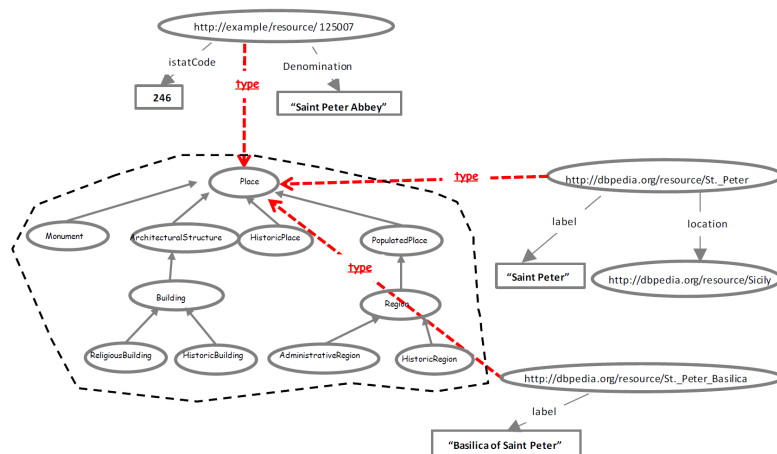
**Figure 6:** Second Workflow: distances in the DBpedia ontology.

which is measured based on the distance between concepts in a hierarchy. By measuring the distances between the DBpedia concepts of the R12 record and the concepts of the R22 and R23 records, we can deduce that *Saint Peter Abbey* actually corresponds to *Basilica of Saint Peter* (distance in the tree equal to 1), and not to *Saint Peter* (distance equal to 5).

Now suppose we produce another transformation and a semantic enrichment in which we are less precise; in this way we introduce the problem of semantic accuracy known as imprecise annotation or classification. Suppose we classify all instances to be of type *Place* instead of classifying them with a more specific type such as *HistoricalPlace* or *Building*. Further, we apply the matching to the last transformation (see Figure 7) and we obtain distances all equal to 0, thus producing a qualitative result worse than before, because we introduced a false positive. This is due to the fact that the classification of the ontology is not exploited, and the attribution of meaning remains at a high level of abstraction which does not help to exploit the ontological classification. This example has shown that during the transformation process there can be quality problems, and a *degradation* of the meaning of the concepts represented can occur. So the overall quality, in our case of integration, can improve if semantic enrichment is exploited, but at the same time, we should be careful during transformation process since noisy data can be inserted and thus worsening the precision with which they represent the reality.

The final outcome of the second workflow is given as follows: 2 pairs true positive and 1 pair true negative. In conclusion, the initial example demonstrated the superiority of the process where first a transformation of the model from relational to RDF is performed and then a semantic enrichment (creation of correspondences with DBpedia) is applied.

Our model shows through this use case that generated the two datasets according to the two different workflows, through the quality metrics we can evaluate which of the two workflows



**Figure 7:** Second Workflow: distances in the DBpedia ontology for records R12, R21, R22.

generated the dataset with the highest quality. As we can see from this results, the best workflow with the highest quality is the one that takes as input the two datasets that are first transformed and then aligned to existing vocabularies when possible and finally enriched with additional data from other datasets.

## 6. Conclusions

This work demonstrate the importance of integration in the data life cycle. Data integration is inherent to the data modelling of the datasets. To evaluate whether a data transformation from one data model to the other is appropriate we should ask whether the quality of the transformed dataset has increased meaning that the resources available on the Linked Data cloud provide additional information on the accuracy and the completeness of the initial dataset. Moreover, we should check if we can take advantage of this transformation in the data integration process so that available techniques of matching based on hierarchy of the concepts can be applied. The lesson learnt from this use case is that it is better to do first integration followed by transformation if no domain knowledge is needed. Otherwise, transformation followed by alignment is better when: i) aligning attributes to existing vocabularies; ii) identification of the distance between concepts is defined; iii) enriching the dataset with the missing values.

## References

- [1] C. P. Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, *Information Sciences* 275 (2014) 314 – 347.
- [2] D. Loshin, *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph*, Morgan Kaufmann, Amsterdam, 2013.

- [3] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management* 35 (2015) 137–144.
- [4] C. Batini, M. Scannapieco, *Data and Information Quality: Dimensions, Principles and Techniques*, Springer, 2016.
- [5] Y. W. Lee, D. M. Strong, B. K. Kahn, R. Y. Wang, AIMQ: a methodology for information quality assessment, *Information Management* 40 (2002) 133 – 146.
- [6] L. L. Pipino, Y. W. Lee, R. Y. Wang, Data quality assessment, *Communications of the ACM* 45 (2002) 211–218.
- [7] R. Y. Wang, A product perspective on total data quality management, *Communications of the ACM* 41 (1998) 58 – 65.
- [8] P. Hitzler, K. Janowicz, Linked data, big data, and the 4th paradigm., *Semantic Web* 4 (2013) 233–235.
- [9] T. Heath, C. Bizer, *Linked data: Evolving the web into a global data space*, volume 1, Morgan & Claypool Publishers, 2011.
- [10] T. Käfer, A. Abdelrahman, J. Umbrich, P. O’Byrne, A. Hogan, Observing linked data dynamics, in: *ESWC*, volume 7882, Springer, 2013, pp. 213–227.
- [11] D. Spanos, P. Stavrou, N. Mitrou, Bringing relational databases into the semantic web: A survey, *Semantic Web* 3 (2012) 169–209.
- [12] M. Uschold, Ontology and database schema: What’s the difference?, *Applied Ontology* 10 (2015) 243–258. URL: <https://doi.org/10.3233/AO-150158>. doi:10.3233/AO-150158.
- [13] J. F. Sequeda, Integrating relational databases with the semantic web: A reflection, in: *Reasoning Web. Semantic Interoperability on the Web*, 2017, pp. 68–120.
- [14] C. Batini, A. Rula, M. Scannapieco, G. Viscusi, From data quality to big data quality, *J. Database Manag.* 26 (2015) 60–82.
- [15] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey, *Semantic Web* 7 (2016) 63–93. URL: <http://dx.doi.org/10.3233/SW-150175>. doi:10.3233/SW-150175.
- [16] J. Debattista, C. Lange, S. Auer, D. Cortis, Evaluating the quality of the LOD cloud: An empirical investigation, *Semantic Web* 9 (2018) 859–901.
- [17] A. Dimou, D. Kontokostas, M. Freudenberg, R. Verborgh, J. Lehmann, E. Mannens, S. Hellmann, R. Van de Walle, Assessing and refining mappingsto rdf to improve dataset quality, in: *ISWC*, Springer, 2015, pp. 133–149.
- [18] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, S. Decker, An empirical survey of linked data conformance, *JWS* 14 (2012) 14–44.
- [19] C. Fürber, M. Hepp, Swiqa - a semantic web information quality assessment framework, in: *19th ECIS*, 2011, p. 76.
- [20] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, A. Zaveri, Test-driven evaluation of linked data quality, in: *Proceedings of the 23rd international conference on World Wide Web*, ACM, 2014, pp. 747–758.
- [21] M. Rico, N. Mihindukulasooriya, D. Kontokostas, H. Paulheim, S. Hellmann, A. Gómez-Pérez, Predicting incorrect mappings: A data-driven approach applied to dbpedia, in: *33rd ACM SAC*, 2018, pp. 323–330.
- [22] A. Dimou, D. Kontokostas, M. Freudenberg, R. Verborgh, J. Lehmann, E. Mannens, S. Hellman, R. Van de Walle, Dbpedia mappings quality assessment, in: *ISWC*, 2016, pp. 1–4.