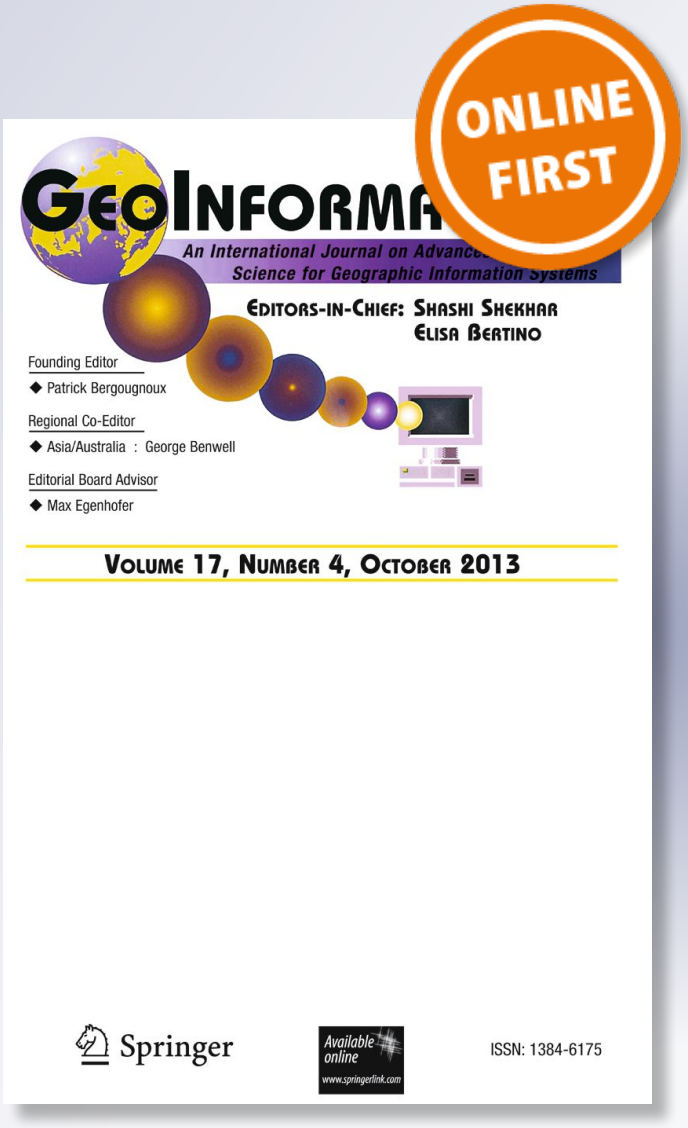
# *Consistency assessment for open geodata integration: an ontology-based approach*

**Linfang Ding, Guohui Xiao, Diego Calvanese & Liqiu Meng**

**GeoInformatica**
An International Journal on Advances of Computer Science for Geographic Information Systems

ISSN 1384-6175

Geoinformatica
DOI 10.1007/s10707-019-00384-9




Springer

Geoinformatica
https://doi.org/10.1007/s10707-019-00384-9

# Consistency assessment for open geodata integration: an ontology-based approach

**Linfang Ding**[1,2] · **Guohui Xiao**[1] · **Diego Calvanese**[1] · **Liqiu Meng**[2]

Received: 9 December 2018 / Revised: 8 August 2019 / Accepted: 1 November 2019 /


**Abstract**
Integrating heterogeneous geospatial data sources is important in various domains like smart cities, urban planning and governance, but remains a challenging research problem. In particular, the production of high-quality integrated data from multiple sources requires an understanding of their respective characteristics and a systematic assessment of the consistency within and between the data sources. In order to perform the assessment, the data has to be placed on a common ground. However, in practice, heterogeneous geodata are often provided in diverse formats and organized in significantly different structures. In this work, we propose a framework that uses an ontology-based approach to overcome the heterogeneity by means of a domain ontology, so that consistency rules can be evaluated at the unified ontological representation of the data sources. In our case study, we use open governmental data from Open Data Portals (ODPs) and volunteered geographic information from OpenStreetMap (OSM) as two test data sources in the area of the province of South Tyrol, Italy. Our preliminary experiment shows that the approach is effective in detecting inconsistencies within and between ODP and OSM data. These findings provide valuable insights for a better combined usage of these datasets.

**Keywords** Open geodata integration · Data quality · Data consistency · Ontology · Semantic technologies

## 1 Introduction

The increasing public availability of big geospatial datasets and the synthesis of such information provide great opportunities for discovering useful knowledge and supporting the decision-making in various domains like smart cities, urban planning and governance [1, 2]. For instance, spatial data infrastructures can fuse different types of open geodata sources, e.g., *Open Government Data* (OGD) and *Volunteered Geographic Information* (VGI), to

✉ Guohui Xiao
xiao@inf.unibz.it

[1] KRDB Research Centre, Faculty of Computer Science, Free-University of Bozen-Bolzano, Bolzano 39100, Italy

[2] Chair of Cartography, Technical University of Munich, 80333 Munich, Germany

enable cross-data analysis [3]. A prerequisite for making sense of these multiple open geodata sources is to perform geodata integration under a clear semantics. However, geodata integration of multiple data sources is in general not an easy task because of their heterogeneity concerning their types, structures, and qualities [4].

Given a set of heterogeneous geodata sources, the starting point of the integration task is usually to understand their respective characteristics and qualities. For instance, OGD are mainly produced by public sectors at different administrative levels and are thus normally confined to certain administrative boundaries. They are mostly published at *Open Data Portals* (ODPs) and are expected to be authoritative and of high quality [5]. It is common that OGD data contain significant amount of geospatial information (e.g., addresses). In contrast, VGI, e.g., *OpenStreetMap* (OSM), is crowdsourced geospatial data resulting from the wide engagement of large numbers of public citizens [6]. Due to the bottom-up manner of data collection, the quality of VGI vary strongly. The geographic coverage of VGI can range from a local event spot to a global coverage. When integrating such heterogeneous geodata sources, a key challenge is a lack of a proper understanding of the value and contributions provided by each single source, and of the relationships between the datasets. In particular, it is crucial to inspect the data quality of each single source and how the mutual consistency of the data sources affects the overall quality of the integrated data. Hence, there is a necessity to assess systematically the consistency within and between the data sources.

In recent years, the Semantic Web and Linked Data communities have been working on adding the semantics to the data and organize them in a simple yet powerful graph-based RDF model [7]. In the GIScience domain, geospatial datasets have been increasingly published as linked data, in which ontologies are used as a means to improve access to and sharing of geographical information [8] and applied in studies like urban development [9]. Popular linked geodata projects include European Open Data Portal and LinkedGeoData. These linked geodata sources provide a vast of advantages to address the challenges for geodata integration, reuse, semantic interoperation, and knowledge formalization. However, best practices of publishing (geo)-data are still missing, which leads to various errors and data quality issues in the released data sets [10]. Hence, consistency assessment of linked datasets remains a critical prerequisite.

From the perspective of database integration, the issue of identifying inconsistencies between datasets can be largely identified at the schema level and at the instance level [11]. Schema-level inconsistency, also called heterogeneity, refers to different ways of organizing the same kind of data. It includes e.g., diverse file formats, different data types, and attribute names. For instance, popular geodata formats include (Geo)JSON, Excel, CSV, Shapefiles, and RDF triples. The locations of geofeature can be provided as geometries in certain geographic coordinate system, but can also be referenced by addresses in natural languages. Even worse, in multilingual areas, these addresses are recorded in different languages following different naming conventions. Another common phenomenon is that data sources are not sufficiently self-described. In particular, one needs to invest major efforts in understanding the possibly confusing and even misleading attribute names. Consequently, the information about the inter-relationship across multiple datasets are largely missing. These problems are also strongly related to the systematic errors in publishing geodata sources as linked data [10]. Without a good understanding of the underlying sources and an agile workflow, such systematic errors can be easily introduced, which are difficult to be detected and fixed later.

Instance-level inconsistency means that data contain logical contradictions [12, 13]. Geodata inconsistencies at the instance level are normally regarded as violating specific

geographic or cartographic constraints or rules. For instance, the topological and the thematic relation of two geofeatures can be contradictory: according to one source, an address is registered in a city; while according to other sources, the geometry of the address is outside the city. Such situation suggests that at least one data source contains some errors and one should be careful when integrating such sources.

The goal of this paper is to develop a framework of data inconsistency assessment for geodata integration. In order to assess the consistency of multiple data sources, we need to place the data on a common ground, for which we resort to ontology-based approaches. An ontology conceptualizes a domain of interest and provides a coherent view of underlying data and thus greatly simplifies the process of consistency assessment. Among the first works, Frank [14] considered consistency constraints on spatial databases from an ontological perspective. Ontologies can also help to improve the accuracy of integration by making the semantic differences of geospatial data explicit. Aracri et al. [15] experimented with the ontology-based data management paradigm for data quality assessment, including consistency, accuracy, and completeness, in the context of the Italian Integrated System of Statistical Registers. Despite the existing extensive theoretical works on geo-ontologies and data quality (see also Section 2), there seems still to be a gap on how to apply these results to the practice of quality assessment for geodata integration apart from some initial works.

The main contribution of this work is an ontology-based data quality assessment framework for geodata integration. The framework leverages the *ontology-based data access* (OBDA) paradigm, which provides a *virtual* unified ontological RDF graph view over the data sources to be integrated. This virtual view is defined by declarative mapping assertions from the data sources to the concepts in the ontology. The mapping and ontology are defined at the schema-level, i.e., data instance independent. During the construction of the ontology and mapping, the schema-level inconsistencies are assessed and possibly fixed. Users can then formulate consistency checking rules over the ontological view as queries to assess data inconsistencies at the instance level, and visually check the results on maps. Query answering over the ontological view is handled by the query-rewriting technique, i.e., ontological queries are translated automatically by an OBDA engine to queries over the data sources. In contrast to the classical costly materialization-based method, our virtualization-based approach avoids explicitly generating RDF triples, and thus allows for a quick iteration of data integration, and enables quality assessment on the fly.

The second contribution is an experiment of evaluating our proposed framework on ODP and OSM datasets, which leads to several insights into data inconsistency issues classified in different categories. The evaluation is illustrated by real-world data in the province of South Tyrol, Italy. We believe that these findings reveal general issues in many geodata integration scenarios.

The following of the paper is structured as follows: Section 2 introduces the related work on (linked) open geodata, the consistency assessment of geodata and OBDA techniques for geospatial data. Section 3 presents a framework for assessing data inconsistency. Test data are described in Section 4. The preliminary experiment is conducted and the results are analyzed in Section 5. We conclude this paper in Section 6.

## 2 Related work

In this section, we survey related work on (linked) open geodata, the consistency assessment of geodata, and ontology-based data access for geodata.

### 2.1 (Linked) open geodata

The open data movement and the booming citizen science have been making large amount of geodata publicly available. On the one hand, governmental or public agencies are collecting and publishing more and more *Open Government Data* (OGD), which are significant resources for increased public transparency [16]. Numerous *Open Data Portals* (ODP) have been built worldwide, allowing anyone to easily search, download and reuse the data for commercial or non-commercial purposes. ODP data are normally confined to certain administrative boundaries, e.g., at provincial, national, or continental levels, and contain significant amount of geospatial information in domains like transportation and environment. ODP data are expected to be authoritative, of high quality, complete, and timely published on the Web [17]. Releasing ODP data without proper quality control may negatively affect dataset reuse and civic participation [5]. However, to prepare high quality ODP data, enormous money and time efforts need to be invested to employ and train employees, understand new legislation, adjust data to new standards. One of the key issues with ODP data quality is the lack of missing standards that provide a common understanding of the datasets. Ontology-based approaches are becoming increasingly popular for addressing this issue. For instance, Austrilian[1] and Italian[2] governments have designed ontologies to describe the characteristics of their published datasets.

On the other hand, large numbers of citizens are engaged to contribute crowdsourced data, especially *Volunteered Geographic Information* (VGI) [6]. VGI has been widely applied in various application domains e.g., crisis response and disaster management [18–20]. OpenStreetMap (OSM) is one of the most popular VGI projects. Compared with ODPs, OSM has a global coverage with all data georeferenced and is updated in a more timely manner. Given the project's volunteering nature, one of the most significant issues is data quality, which has been extensively studied, for example by comparing OSM data with reference datasets [21–23] or by using OSM historical datasets [24]. OSM data have been also modeled by ontologies, e.g., OSMonto [25] and OSM Semantic Network [26].

With the development of Semantic Web technologies geospatial data from OGD and VGI have been considerably released as Linked Data in the last decade [7]. For instance, the European Open Data Portal[3] can be searched via an interactive search engine and through SPARQL queries, and the LinkedGeoData project [27] uses comprehensive geospatial data sources, including OSM, to create a large RDF dataset. The linked data paradigm allows easier data discovery, integration with other data sources, and the development of applications [28].

### 2.2 Consistency assessment of geodata

Consistency is one of the most important geospatial data quality issues along with spatial, thematic, temporal accuracy and resolution, completeness [29]. According to Share-PSI 2.0 [13], data consistency means that data do not contain contradictions. In GIS domain, Egenhofer et al. [12] defined consistency as the lack of any logical contradictions within

[1] https://data.gov.au/dataset/data-gov-au-dataset-ontology
[2] https://github.com/italia/daf-ontologie-vocabolari-controllati
[3] http://data.europa.eu/euodp/en/linked-data

a model of reality. For the evaluation of inconsistency among multiple representations in spatial databases reported by Sheeren et al. [11], two representations of a given geographic phenomenon are inconsistent if and only if the differences between these representations cannot be explained by their respective database specifications; otherwise, the representations are consistent. Brisaboa et al. [30] defined spatial inconsistency by referring to a contradiction between stored data and spatial integrity constraints. Senaratne et al. [31] regarded VGI data consistency as the coherence in the data structures of the digitized spatial data, including conceptual, domain, format, and topological consistency. Other definitions of consistency are related to certain types of datasets and specific applications, e.g., land cover data [32].

The consistency of spatial information needs to consider a variety of issues ranging from the ontological level concerning physical reality, to appropriate conceptual frameworks for analyzing spatial consistency (e.g., models for consistency at multiple representational levels or granularities), and to the specification language of integrity constraints and the design of computational-geometry algorithms to implement consistency checkers [33]. From the perspective of database integration, the issue of identifying inconsistencies between datasets can be largely identified at the schema level and at the data level [11].

At the schema level, different approaches have been developed related to geospatial data consistency assessment. The first approach aims to generate a single unified schema of the originally independent schemata and resolve conflicts between semantic concepts [34]. This centralized approach does not require any mediators. Balley et al. [35] built a global schema and used consistency rules to handle inter-database consistency under a priori consideration that each source database to be consistent with regard to its specifications. The second approach relies on solutions based on mediation or the use of ontologies by taking advantage of their semantic information. Individual schemas can be matched to one ontology [36], or each individual schema is firstly matched to a local ontology whose concepts correspond to database tables and concept properties and relations to class attributes and associations, and then ontology alignment is applied to align these ontologies [37]. For instance, Comber et al. [32] used expert opinions to overcome ontological incompatabilities between time series land-cover datasets and identified land-cover changes from the inconsistency. Yu et al. [38] used Semantic Web technologies to automate the geospatial data conflation using three sets of Points of Interest (POI) data. The third approach uses data-mining methods to infer the schema-level structure necessary for information fusion from instance-level information [39]. For instance, viewing the inconsistency detection as a knowledge-acquisition problem, Sheeren et al. [11] proposed a data-mining approach to partially automate the acquisition of the consistency rules, which were then used in a knowledge-based system for evaluating consistency.

At the instance level, most of the efforts have been devoted to developing geometric feature matching algorithms, which establish explicit links between objects in different representations and resolve geometrical inconsistencies. Among them, road network matching and integration are mostly studied for producing digital navigation map and providing further location-based services [40]. Advanced road-network matching algorithms have been proposed for conflating commercial and administrative digital road network data [41, 42] and more recently for matching crowdsourced spatial trajectories with OSM road network [43]. These works also exploited the topological structures of road intersections and semantic attributes of road segments. Beside geometric matching, topological inconsistency among multiple representations have been extensively studied by imposing topological relation constrains [12, 44].

### 2.3 OBDA for geospatial data

Ontology-Based Data Access (OBDA) is a popular paradigm that enables end users to access data sources through an ontology [45]. The architecture of OBDA has the following components: an ontology, a set of data sources, and the mapping between the two. The ontology provides a high-level description of the domain of interest and is semantically linked to the data sources by means of a mapping consisting of a set of mapping assertions [46]. The standard mapping language is R2RML [47]. The ontology and mapping together, called an OBDA specification, exposes the underlying data sources as a virtual RDF graph, in which the nodes are IRIs (Internationalized Resource Identifiers) and literals representing objects and values respectively. The (virtual) RDF graphs [48] are accessible at query time using the W3C standard SPARQL language [49]. Such SPARQL queries are translated by an OBDA system, e.g., *Ontop* [50], into queries that are directly evaluated by the underlying database engine, without converting and materializing original data as RDF and then storing them in a triple store. The OBDA paradigm has been implemented in several systems, and adopted in many academic and industrial projects [51].

In GIScience domain, OWL and RDF have been successfully applied to model geospatial information. The GeoSPARQL language [52], standard by Open Geospatial Consortium (OGC), is specially designed as a geographic query language by extending SPARQL.

More recently, the Spatial Data on the Web Interest Group, comprised of both W3C and OGC, is working specifically on sharing of spatial data on the Web using Semantic Web technologies. OBDA has been investigated to support geospatial relational databases [53], performing on-the-fly GeoSPARQL-to-SQL translation. It has been used in several use cases, e.g., urban accountant, land management, and crisis mapping [54] and maritime security [55].

## 3 A framework for inconsistency assessment

In this work, we propose an OBDA-based framework to assess the inconsistency of open geodata sources at both schema and instance levels. Our framework, illustrated in Fig. 1, consists of four layers: *(1)* data collection and preprocessing, *(2)* ontology-based data access, *(3)* query-based consistency assessment, and *(4)* visualization. The first two layers are responsible for preparing the data, setting up the OBDA scenario, and assessing the schema-level inconsistencies; the last two layers assess the instance-level inconsistencies using the query-based or visualization-based approaches. In what follows, we first introduce different types of inconsistencies considered in our framework, and then explain the framework in detail.

### 3.1 Schema- and instance-level inconsistencies

**Schema-level inconsistencies** In our framework, we are dealing with the following kinds of schema-level inconsistencies:

– Formats of the data sources. Open data are often distributed in diverse formats, among which, Excel, CSV, (Geo)JSON XML, Shapefiles, and RDF triples are popular ones. They correspond to different models, like relational (Excel, CSV), documents (JSON), trees (XML), and graphs (RDF). Dealing with such sources often requires dedicated software tools and customized scripts, especially when sources with different formats need to be accessed in a combined way.

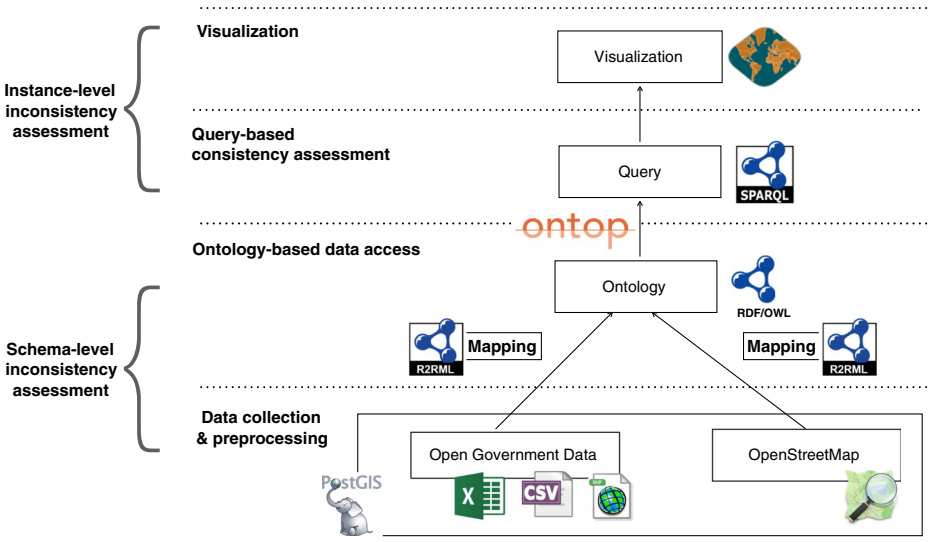



**Fig. 1** An OBDA-based framework for assessing the consistency of OGD and OSM

– Data structure of entities. Even when using the same model, entities (and their attributes) can be structured in significantly different ways. For instance, in the relational model, an object and its related data might be put into one table, or organized in several tables. In particular, data in CSV files are often de-normalized, which means, e.g., that a single row corresponds to several objects, or that the same data about a single object is repeated in several rows. For JSON files, such redundant information often results from structuring data using nesting. Understanding how such data is organized needs expert knowledge on specific sources and on the domain itself.
– Schema element names. Different table or attribute names are often used to represent the same kind of information. For example, organizations may name the same information according to internal conventions or regulations; inside the same organization different table or attribute names can be used when serving specific purposes. A particular case is that of multilingual areas, where schema element names are recorded in different languages according to naming conventions.
– Data types. Some piece of information can also be recorded using different data types. For instance, an identifier might be stored as a string or as a number. A geometry object might be represented as a polygon or as a point, depending on the level of detail; it might also be specified using different coordinate systems, which should be treated analogous to different data types.

**Instance-level inconsistencies** Geodata inconsistencies at the instance level are regarded as violating specific geographic or cartographic constraints. A geospatial object may have thematic (or non-spatial) attributes and geometric (or spatial) attributes; two geospatial objects may be in a topological relation with each other. In this paper we consider geodata inconsistency at the instance level according to their thematic, geometric, and topological aspects.

– Thematic attributes. Thematic attributes describe the non-spatial information of a geospatial object. For example, a hospital should have a unique legal address. If two

different data sources provide two different legal addresses for the same hospital, an inconsistency occurs.
- Geometric attributes. Geometric attributes like location and extent are metric properties of a geospatial object under a spatial reference system. For instance, considering two data sources listing schools, if they cover a common area, and the first source contains a school in that area, then also the second source should contain that school at the same location. If the locations differ, we have a geometric inconsistency.
- Topological properties. Topological properties are expressed by relations between geographic objects, e.g., "overlaps" or "within". Geospatial data conforms to certain topological rules. For example, for two adjacent counties, the polygons representing them should share a common boundary, with no gaps and no overlaps.
- Inconsistencies may also involve multiple types of properties. For example, an address registered in a municipality (thematic information) should be within the municipality (topological property).

### 3.2 The framework

We now describe in detail, layer by layer, the framework shown in Fig. 1. Each of the four layers, corresponds to a specific step in the data processing workflow.

**Data collection and preprocessing** This layer collects and preprocesses the relevant geospatial data by relying on a unique data format. After collecting the relevant datasets, users need to pre-process them, since the data are often provided in different formats, possibly with errors. Typical steps of preprocessing involve necessary conversion of data formats, and data cleaning to remove errors. The datasets are then imported to a specific database (e.g., PostgreSQL with PostGIS extension) to be further processed and queried. During this step, the schema-level inconsistencies due to different data source formats are assessed and resolved.

**Ontology-based data access** The OBDA layer is responsible for providing an ontological view of the underlying heterogeneous datasets to be analyzed. The major challenge here is to overcome the structural heterogeneity of the multiple data sources. The main task in this layer is to build the ontology and the mappings in order to identify the semantics of geodata items across different datasets, and provide consistent ontological vocabularies capturing the contents of underlying datasets. During the construction of ontology and mapping, metadata containing the description of each dataset are often helpful.

The ontology should capture the geographic phenomena to be analyzed. The construction of the ontology can benefit from the classical literature on spatial-temporal ontologies [56] and the identified core concepts in GIScience [57]. Reusing standardized ontologies is also a good practice. For example, the OGC GeoSPARQL ontology [52] is designed for features and geometries; the W3C Semantic Sensor Network (SSN) Ontology [58] is used for sensors and observations. Existing domain ontologies can also be employed when they are available.

The mappings from geodata items to the ontology have been intensively researched. Mappings can be constructed semi-automatically with systems, e.g., Ultrawrap-mapper [59], BootOX [60] and Map-on [61]. Sequeda and Miranker [62] proposed a "pay-as-you-go" methodology for ontology and mapping engineering. However, since the understanding of the semantics of geodata demands the domain knowledge, the process of building OBDA specification with high-quality ontology and mapping cannot be fully automatized and normally takes some manual efforts.

During this step, the schema-level inconsistencies due to different data structures of entities, attribute names, and data types are assessed and resolved by mapping such elements to the proper ontological terms.

**Query-based consistency assessment** The OBDA layer exposes the underlying heterogeneous geodata sources as one integrated (virtual) RDF graph. In this graph, heterogeneous data are described with the same vocabulary (i.e., classes and properties). Thus, queries can be posed over the RDF graph, without knowing how the underlying data is structured. This abstraction allows the analysts to express their information needs through SPARQL query language using high-level concepts that they are interested in. An OBDA engine, e.g. Ontop, is taking care of the SPARQL-to-SQL translation, and the retrieval of the relevant information (e.g., provenance, thematic information, and geometries) from multiple data sources.

Consistency assessment is done by declarative rules over the graph. Those consistency rules can be basic cartographic principles of geodata consistency constraints or a set of predefined criteria, e.g., using geometric measures. The knowledge acquisition of consistency rules can be also an iterative process and collected empirically. Simple rules can be implemented as one SPARQL query. Some consistency conditions maybe complex and require several steps of computations, thus they cannot be directly encoded into one query. Post-processing of the query results is often required, which may involve comparison, matching, and further analysis.

Most instance-level inconsistencies are assessed in this step.

**Visualization** Visualization provides an intuitive and effective way for the understanding of geodata distribution and facilitates the discovery of potential patterns. For the task of consistency assessment, the visualization layer supports the interpretation of the ontology, the queries and the retrieved geodata.

In this framework, multiple visualization technologies can be adopted to represent these information items in this framework. For instance, network visualization techniques are well-suited to represent the involved geospatial concepts in a query, their relations and the related consistency constraints. Cartographic representations, e.g., maps, can not only show the distribution of the retrieved geo-features with their spatial context, but also may immediately reveal the geometrical consistency problems by the mashup representation of data from different sources. For instance, after mapping and visually comparing the surface and point representations of water features in two databases, Sheeren et al. [11] were stimulated to extract rules for assessing the consistency of homologous objects. There have been extensive work on the visualization techniques for representing ontology (see [63] for a survey). Several systems have been developed for visualizing SPARQL queries and results over spatial data, including OptiqueVQS [64], GeoYASGUI [65], and Sextant [66].

This step is often effective in stimulating users to visually discover and assess potential data consistency problems.

### 3.3 Discussion on the framework

We now discuss several aspects of the framework.

**Target user** The target of our framework are the users who need to integrate multiple data sources, but don't have a proper understanding of the consistency of the data. With the framework we propose, the users can assess the data quality in a declarative and agile way. In

particular, it is more effective when dealing with inconsistency involving multiple sources, which cannot be easily observed by simply opening files, e.g., in Excel.

**Workflow** We note that the way we use OBDA/I in this paper is different from the classical usage where the aim is to provide a single access point for the integrated data. Rather, our framework focuses on the initial phase of an integration workflow, where the data quality of the datasets needs to be understood. Once the inconsistencies are assessed and fixed, the resulting OBDA specification can be used to publish the integrated data sources as a SPARQL endpoint.

**Advantage of the virtual approach** Our framework relies on the virtual OBDA approach, since materialization is, in general, a costly operation. In particular, when the amount of data is large and/or the data are updated frequently, materialization is not efficient. For inconsistency assessment, the users of our framework might need to revise the mapping and/or the data several times. The virtual approach allows them to experiment with such revisions on the fly, thus improving the efficiency significantly.

**Tooling** The adoption of the proposed framework requires expert knowledge on OBDA technologies, and would strongly benefit from the availability of tools that support the underlying workflow. Although we have a proof-of-concept implementation that shows the feasibility of our proposal, the whole workflow still cannot be fully automated. As future work, we plan to implement the whole framework through proper software components that will make the workflow semi-automatic.

**Correction of inconsistencies** Schema-level inconsistencies are assessed and corrected in the framework. However, correction of instance-level inconsistencies requires an extension of the framework. It also often requires the involvement of data providers to check and correct instance-level inconsistencies.


**Fig. 2** The province of South Tyrol, Italy

**Table 1** ODP datasets from open data portal of South Tyrol

| Dataset | Description | Format | # Entries |
|---|---|---|---|
| Municipality | polygons of municipalities | .shp | 116 |
| Street | street network | .shp | 4324 |
| Address | addresses with street names and house numbers | .shp | 131633 |
| Pharmacy | pharmacies with names, addresses, contact info, etc | .csv | 120 |
| Organization | organizations, (e.g., schools, museums, and offices), with names, addresses, and contact | .csv | 1675 (school:1053) |
| Healthcare | contact of sanitary offices | .csv | 980 |
| Filling station | filling stations | .shp | 163 |

## 4 Test data

We use the province of South Tyrol (German: Südtirol; Italian: Alto Adige) in Italy as our test area. It is an autonomous province in northern Italy and has two official languages German and Italian. Figure 2 shows the geographic location of South Tyrol.

In this study, we collect data from two data sources: *(1)* the open data portal (ODP) of South Tyrol,[4] and *(2)* OpenStreetMap.[5]

**Open data portal of South Tyrol** The portal contains datasets collected from local authorities, companies and relevant stakeholders. At time of writing the agency of the portal has successfully published 463 datasets covering 17 categories on climate, urban planning, health, environment, etc. Users can download datasets in JSON, XML, CSV, or PDF formats. Metadata are available on the web pages as pdf files. The portal also has a Geocatalog platform[6] providing extensive geodatasets like administrative boundaries, satellite images, transportation networks. Vector data with their exact geolocations can be accessed in formats like ESRI Shapefiles, AutoCAD, Google KML, GPX or GeoJSON files. In this study, we use seven datasets including municipality, street, address, pharmacy, organization, healthcare, and filling stations. Table 1 describes these datasets in detail.

**OpenStreetMap** OSM data can be obtained through its official website or from other websites like Geofabrik[7] that provide already processed and structured OSM data in hierarchical regions. The OSM data model represents point, linear and polygonal features by *nodes*, *ways* and *closed ways*. Features can be semantically specified by key-value pairs, so-called tags (e.g., *amenity* = *restaurant*). There are no restrictions on the usage of tags. However, the OSM community agrees to certain key-value combinations for commonly used features and divides features into 23 primary categories (e.g., amenity, building, highway). Each key can take many different values to further specify the sub-type of the mapped feature. In this work, we export OSM data from its official website into an XML file, and then extract the points and polygons located inside South Tyrol. To compare with ODP data in the following experiment, we further extract four types of POI data (i.e., pharmacy, school, healthcare,

[4] http://daten.buergernetz.bz.it/de/
[5] http://www.openstreetmap.org
[6] http://geokatalog.buergernetz.bz.it/geokatalog
[7] https://www.geofabrik.de/

**Table 2** Four types of POI data extracted from OSM

| Dataset | Amenity | Format | Number of entries |
|---|---|---|---|
| Pharmacy | 'pharmacy' | .shp | 128 (point: 119, polygon: 9) |
| School | 'school'/'kindergarten' | .shp | 547 (point: 239, polygon: 308) |
| Healthcare | 'clinic'/'dentist'/'doctors'/'hospital' | .shp | 90 (point:71, polygon: 19) |
| Filling station | 'fuel' | .shp | 165 (point: 111, polygon: 54) |

filling station) from point and polygon datasets by using the attribute of "amenity". Table 2 shows the description of the POIs.

## 5 Experiment and analysis

We apply the proposed framework to assess schema- and instance-level consistency in a single data source and across ODP and OSM. We start by describing our experimental setup, then deal with schema-level and instance-level inconsistencies, and finally discuss our results.

### 5.1 Experimental setup

We use a PostgreSQL database to store all the datasets. For building the ontology and the mapping, we apply the Protégé ontology editor [67] and the Ontop Protégé plugin [50]. The Javascript libraries Openlayers[8] and vis.js[9] are used to implement the web-based visualization system.

**Ontology** The ontology used in this experiment is based on the Territorial Ontology provided by the Italian National Institute of Statistics (Istat)[10] [68] and the GeoSPARQL ontology. The Territorial Ontology uses the vocabulary from GeoNames[11] and W3C provenance ontology[12] to describe the administrative organizations, e.g., region, province, municipality and geographical-statistical organizations. We do not use the OSMonto ontology or OSM Semantic Network, because OSMonto is designed specifically for OSM and constructed following the key-value model of OSM, and OSM Semantic Network does not follow OWL standard and seems not maintained. Hence they are not suitable for the purpose of data integration.

Figure 3 shows a diagram of the ontology. The nodes represent classes and the arrows represent the "is-a" relations between two classes. The classes prefixed with "geosparql:" and "ter:" are from the GeoSPARQL and the Territory Ontology respectively, and those prefixed with ":" are created by us. All geodata are regarded as geospatial features with geometries and thematic properties. For instance, the class "AdminUnit" has sub-classes like

[8] https://openlayers.org/
[9] http://visjs.org
[10] http://datiopen.istat.it/ontologie.php?language=eng
[11] http://www.geonames.org/ontology/documentation.html
[12] https://www.w3.org/TR/prov-primer/

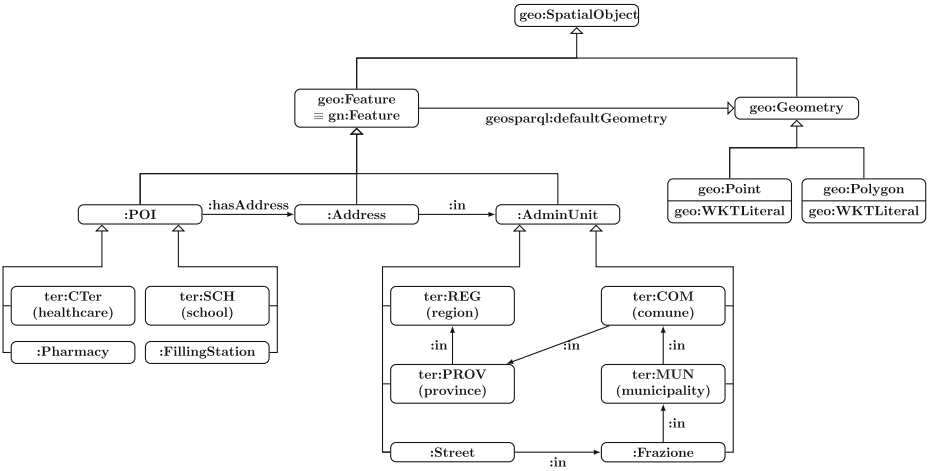



**Fig. 3** A fragment of the ontology

"Province" and "Municipality"; the class "POI" has four sub-classes of "School", "Pharmacy", "Filling Station" and "Healthcare". The underlying geodata from both data sources will be accordingly mapped to the vocabularies in this Ontology.

**Mapping** The mapping from the data sets, stored in a PostgreSQL database, to the ontology is correspondingly constructed. In total we have constructed 53 mapping assertions. In Fig. 4, we list five mapping assertion examples related to "Pharmacy" written in the *Ontop* mapping syntax. A mapping assertion takes the form `id: target ← source`, where `id` is an identifier, `source` is an SQL query, and `target` is a triple pattern

```
M_pharmacy_OD_1:
 :pharmacy/{phar_id} a :Pharmacy; :provenance "OD";
    rdfs:label {phar_desc_i}@it, {phar_desc_d}@de;
    geosparql:defaultGeometry :point_geom_pharmacy/{phar_id}.
 ← SELECT phar_id,phar_desc_i,phar_desc_d FROM pharmacies
```

```
M_pharmacy_OD_2:
 :point_geom_pharmacy/{phar_id} a sf:Point; geosparql:asWKT {wkt}.
 ← SELECT phar_id, ST_AsText(geom) as wkt FROM pharmacies, addresses
  WHERE pharmacies.phar_adress_i=addresses.label_it
```

```
M_pharmacy_OSM:
 :osm_point/{osm_id} a :Pharmacy; :provenance "OSM";
  :hasHouseNumber {addr:housenumber}; :hasStreetName {addr:streetit}.
 ← SELECT osm_id, "addr:housenumber", "addr:streetit" FROM osm_points
   WHERE amenity='pharmacy'
```

```
M_point_OSM_1:
 :osm_point/{osm_id} geosparql:defaultGeometry :osm_point_geom/{osm_id} .
 ← SELECT osm_id FROM osm_points
```

```
M_point_OSM_2:
 :osm_point_geom/{osm_id} a sf:Point; geosparql:asWKT {wkt}^^xsd:string .
 ← SELECT osm_id, ST_AsText(geom_3044) AS wkt FROM osm_points
```

**Fig. 4** Example mapping assertions

```
:pharmacy/55 a :Pharmacy.
:pharmacy/55 :provenance "OD".
:pharmacy/55 rdfs:label "Maria delle Grazie"@it .
:pharmacy/55 rdfs:label "Zur Mariahilf"@de .
:pharmacy/55 geosparql:defaultGeometry :point_geom_pharmacy/55.
```

**Fig. 5** Example triples generated by mapping assertion `M_pharmacy_OD_1`

with placeholders like {`column`} such that `column` is an output column in `source`. For instance, the tuple (55, "Maria delle Grazie", "Zur Mariahilf") is an answer to the SQL query in `M_pharmacy_OD_1`, and it generates five RDF triples as shown in Fig. 5. Note that we use the property ":provenance" to describe whether the data is from ODP or OSM. We use langStrings (i.e., strings with language tag) `"Maria delle Grazie"@it` and `"Zur Mariahilf"@de` to distinguish strings in different languages. Finally, a pharmacy has a geometry object, but its WKT has to be retrieved by joining with the table "Address" in another mapping assertion `M_pharmacy_OD_2`. The other three mapping assertion examples are for OSM. Assertion `M_pharmacy_OSM` selects all pharmacies from the `osm_points` table. The last two assertions `M_point_OSM_1` and `M_point_OSM_2` are generic and are used to construct geometries of all OSM points, including pharmacy geometries.

We stress again that the triples derived from the mapping are not needed to be physically generated but they are accessible via SPARQL query using SPARQL-to-SQL rewriting techniques. By avoiding materializing the triples, adding new sources and modifying the OBDA specification becomes rather easy. The virtual approach allows a large flexibility for the experiment. In fact, during our experiment, we often adjust the mapping when we have a better understanding of the data.

**Visual interface** We have developed a web-based interactive visualization system. Figure 6 shows a screenshot of the visual interface. It consists of three views: *(1)* a task view (on the upper left), showing the queries and allowing users to execute specific queries; *(2)* a graph view (on the bottom left), providing a network representation of the executed query with the

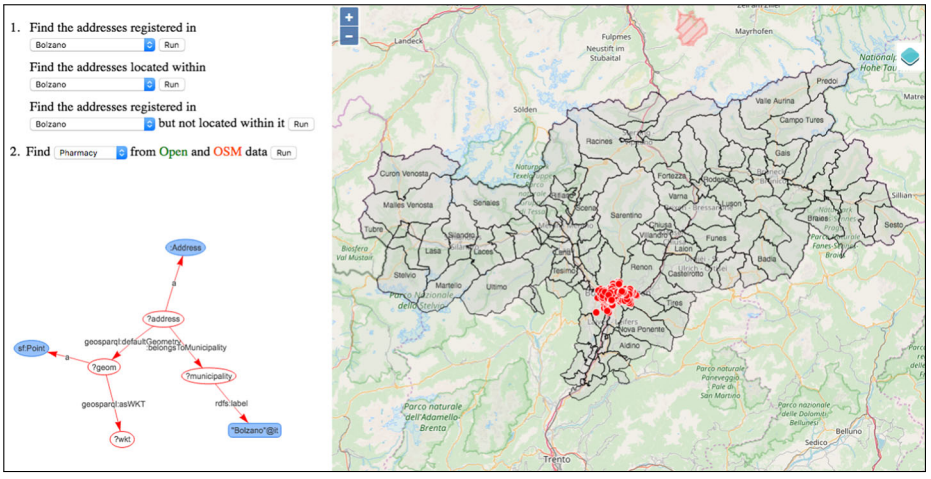



**Fig. 6** The visual interface

objects (represented by nodes) and their relations (represented by edges). *(3)* a map view (on the right), visualizing the query results on a map. In Fig. 6, the first query on finding the addresses registered in Bolzano is executed. The red dots are the queried geo-features.

### 5.2 Schema-level inconsistency

This section introduces the assessment of the schema-level inconsistencies in a single data source and across ODP and OSM.

#### 5.2.1 Inconsistency in a single data source

Schema-level inconsistencies exist within a single source, in particular in the ODP data. Since ODP data are originated from various institutions, the naming of the attributes in the original datasets is rarely kept consistent. Table 3 shows a sample of attributes of three datasets, i.e., "Municipality", "Street", and "Address". For example, the attributes "istat_code", "comistat", and "istat" in these tables (colored in red) are named differently, but they have identical meanings and refer to the unique codes that Istat assigns to each municipality. After finding the correspondences among them, we can use such codes to generate the IRIs of municipalities through mapping. Similarly, "street_code" and "ascot_wege" (in blue), "strt_it" and "desc_i" (in cyan), and "gem_it" and "name_i" (in magenta) refer to the same objects respectively.

We observe that in the OSM datasets the attribute names are consistent across different tables. The main schema-level inconsistency which we have identified has to do with the flexibility of modeling features as points or polygons.

#### 5.2.2 Inconsistency across ODP and OSM

The ODP and OSM data have significantly heterogeneous structures. The ODP data are normally classified according to topics and distributed separately in diverse formats (e.g., pdf, csv, xml, and RDF). In contrast,the OSM data are organized as a large collection of features, each with its geometry and a set of flexible taggings for different information (e.g. addresses and names in possibly multiple languages). When importing OSM data into PostGIS, users even have to specify explicitly which attributes they are interested in.

In our experiment, for instance, health-related data in ODP are already organized in one file including different types of health-related organizations like clinics, dentists, hospitals. This file contains additional information like address, name, telephone, doctor, and opening time. While in OSM, the health-related data are stored as points or polygons, and they can be obtained by filtering the amenity attribute with values "clinic", "dentist", "doctors", or "hospital" (see Table 2).

### 5.3 Instance-level inconsistency

We assess the instance-level inconsistency in a single data source and across ODP and OSM according to the thematic, geometric, and topological constraints of the geo-features. To do so, we have created a suitable set of inconsistency rules, which are formulated over the ontological view as SPARQL queries. For the creation of the rules, we have relied on our expertise in GIS/Cartography and OBDA and on the experience we have obtained by

**Table 3** Inconsistent attribute names in ODP tables

(a) Table: Municipality

| gem_id | istat_code | name_i | name_d | geom |
|---|---|---|---|---|
| g8 | 21008 | Bolzano | Bozen | 01060000E... |

(b) Table: Street

| gcomistat | fraistat | ascot_wege | desc_i | desc_d | geom |
|---|---|---|---|---|---|
| g21008 | 0 | 8280 | VIA LEONARDO DA VINCI | LEONARDO-DA -VINCI-STRASSE | 01050000C... |

(c) Table: Address

| gistat | frac_code | label_it | label_de | street_code | strt_it | num | gem_it | geom |
|---|---|---|---|---|---|---|---|---|
| g21008 | 0 | Via Lenonardo Da Vinci 1/F | Leonardo-Da -Vince Strasse 1/F | 8028 | VIA LEONARDO DA VINCI | 1/F | BOLZANO | 01010C... |

**Table 4** Inconsistent attribute values in multiple ODP tables

| (a) address | | | (b) pharmacy | |
|---|---|---|---|---|
| addr_id | label_it | geom | phar_id | phar_address_i |
| 79543612 | Via Valle Aurina 34 | 01040C.. | 35 | Via Valle Aurina 34 - S. Giorgio |

investigating the data sets described in Section 4. To gain additional insights, we have consulted the data providers and stakeholders familiar with the local territory. The detected inconsistencies are illustrated below using specific examples and visualizations.

### 5.3.1 Data inconsistency in a single data source

We analyze now the inconsistency of ODP and of OSM data separately.

**Inconsistent thematic attribute values** Inconsistent attribute values can exist across multiple datasets, especially when they are from different providers. Such inconsistency introduces problems when these datasets need to be linked through these attribute values. For instance, Table 4 shows that the addresses in dataset "Address" are standard address names, while in "Pharmacy" they combine standard address names and pharmacy names. The "Address" dataset is associated with geolocations while "Pharmacy" not. For georeferencing pharmacies, it is necessary to match the addresses in "Pharmacy" to those in "Address" dataset. This problem can be reflected at the virtual RDF level. Specifically, using mapping assertion `M_OD_pharmacy_1`, the row in the pharmacy table generates a triple (`:pharmacy/35 a :Pharmacy`). However, due to the mismatch of addresses, mapping `M_OD_pharmacy_2` cannot be triggered and therefore `:pharmacy/35` does not have a geometry in the RDF graph.

The above discussion leads to the following consistency rule: *every pharmacy should connect to an address that in turn connects to a geometry; otherwise, the data is inconsistent*. This can be formulated as the SPARQL query:

```
SELECT * { ?x a :Pharmacy. NOT EXISTS {?x geosparql:hasGeometry
 ?g.}}
```

When the query result is not empty, we can conclude that the data violate the above consistency rule. With this query, we find 32 unmatched addresses out of 120 pharmacies.

**Inconsistent thematic and topological relations** ODP data can contain inconsistent thematic and geometric values. Being collected by different institutions, the thematic values and topological relations of the data are not always consistent. Taking the "Address" dataset in ODP as an example, the feature in the data should satisfy a consistency rule: *if a feature is located within a municipality, its address should be registered in the municipality*, which can be formalized into the SPARQL query in Fig. 7. However, when we query the addresses officially registered in the municipality of Bolzano, we find on the map that some of the queried addresses are located outside of Bolzano (shown in Fig. 8). After checking these addresses, we find that most of them have identical street names of "Via Della Comunale", which refer to the addresses registered for homeless people. In total, there are 289 such addresses.

```
SELECT ?addressName ?wkt
WHERE {
  ?address a :Address ;
    :hasGeometryInWKT ?wkt;
    rdfs:label ?addressName ;
    :belongsTo ?municipality.
  ?municipality a :Municipality ;
    rdfs:label 'Bolzano'@it .
  ?address geosparql:sfDisjoint ?municipality.
  FILTER (lang(?addressName) = 'it')
}
```


**Fig. 7** The SPARQL query for "addresses registered in Bolzano but located outside" and its network representation

### 5.3.2 Data inconsistency across ODP and OSM data

We investigate the data inconsistency of the thematic and geometric properties of POIs in ODP and OSM. Four types of POIs (see Tables 1 and 2) including pharmacy, school, healthcare, and filling station are prepared for the comparison. For each type of POIs, we compare the thematic consistency of address-matched features, and the geometric consistency.

**Thematic inconsistency** For the assessment of the thematic inconsistency, we use thematic attribute "address"based on address matching method. Taking the POI type of "School" as an example, the consistency rule is that: *if there is one school in ODP, then there must be another school in OSM with a matched address; and vice versa*. To evaluate this rule, we first use the following SPARQL query to retrieve all schools in ODP and OSM data.

```
 SELECT ?school ?wkt ?prov ?addressName WHERE {
  ?school a :School ; :provenance ?prov ;
geosparql:defaultGeometry ?geom .
  ?geom geosparql:asWKT ?wkt .
  OPTIONAL { ?school :hasAddress ?address . ?address rdfs:label
  ?addressName.}
 }
```

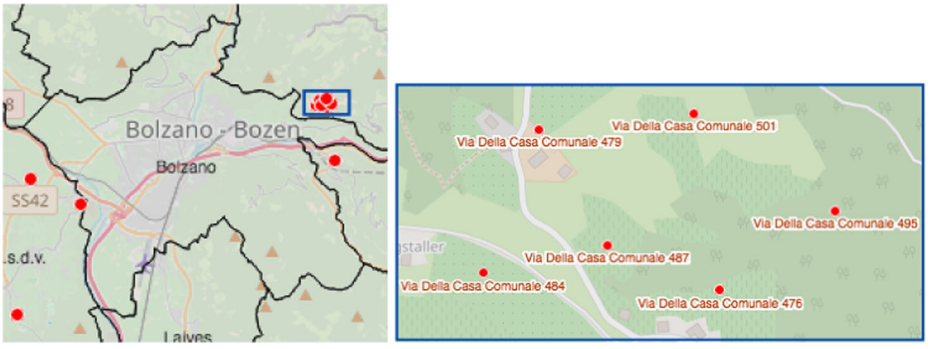
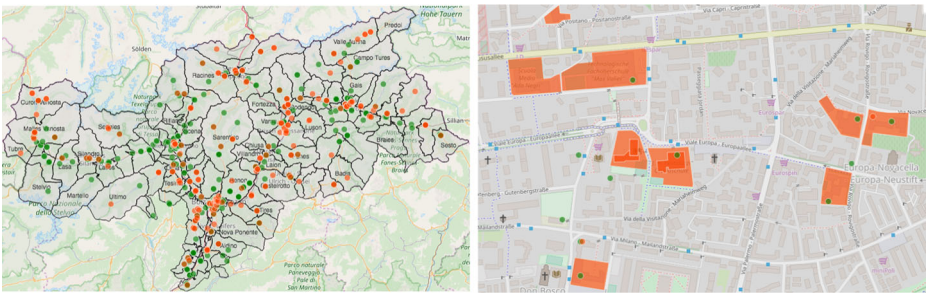



**Fig. 8** Addresses registered in Bolzano but located outside



(a) All schools (b) Schools in a selected area

**Fig. 9** The spatial distribution of **a** all schools from OGD (in green) and OSM (in orange) in the test area, and **b** the schools enlarged in a selected area

Each query result includes the IRI of the school (`?school`), the WKT of its geometry (`?wkt`), its provenance (`?prov`='OSM' or `?prov`='OD'), and optionally its address (`?addressName`). The query results are further processed for address comparison. Here, the address matching method is a simple string comparison.

**Geometric inconsistency** For the assessment of the geometric inconsistency, we use the feature of "geometry" based on geometrical matching method. We continue using the POI type of "School" as an example. The consistency rule is that: *if there is one school in ODP, then there must be another school in OSM with a distance less than a predefined threshold; and vice versa*. To evaluate this rule, in addition to the previous SPARQL query, we use the following SPARQL to retrieve all pairs of neighboring schools from both data sources. Here we choose the threshold of 50m based on empirical knowledge.

```
SELECT ?od_school ?osm_school WHERE {
  ?od_school a :School; :provenance 'OD';
  geosparql:defaultGeometry  ?od_geom.
  ?od_geom geosparql:asWKT ?od_wkt.
  ?osm_school a :School; :provenance 'OSM';
  geosparql:defaultGeometry ?osm_geom.
  ?osm_geom geosparql:asWKT ?osm_wkt.
  FILTER(ogcf:distance(?od_wkt, ?osm_wkt, 'M') < 50)
}
```

Figure 9a maps the spatial distribution of all schools from ODP (in green) and from OSM (in orange). Figure 9b shows schools in an enlarged subarea. From Fig. 9b, we can see that schools from ODP are only points while schools from OSM can be points or polygons. We notice that some schools from ODP do not have corresponding features from OSM (interpreted on the map as: on a specific location there is only a green point but no orange symbol), and vice versa. This kind of geometric inconsistency can be assessed using the previous rule.

Figure 10 summarizes the evaluation results of the thematic and geometric consistency rules of the four types of POIs. Obviously, the thematic inconsistency of the attribute of address is more significant than the geometric inconsistency. One possible reason is that all the OSM data are with geometry information, but only a small portion of them contains

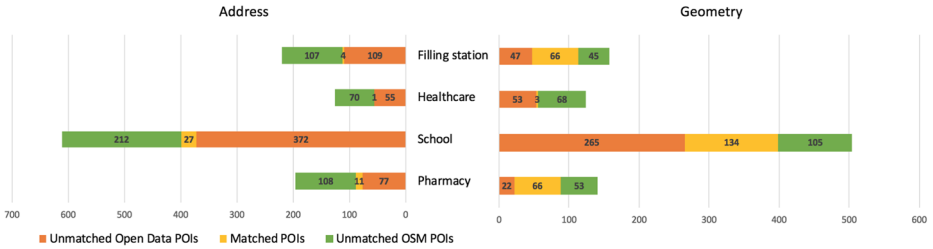



**Fig. 10** The evaluation results of thematic and geometrical consistency rules

address information. Another reason is due to the inconsistent address values in ODP and OSM data.

## 5.4 Discussion

In this experiment, we have assessed the data inconsistency in and across ODP and OSM data at both the schema and individual levels, and have visualzied the results using maps, node-link graphs, and bar charts. We have identified different types of data inconsistency, and summarized them in Table 5. For each type of detected inconsistency, we have used specific examples to illustrate the working principle of our framework. These examples also reflect general geodata inconsistency issues existing in heteregeous open datasets.

**Schema-level inconsistency** Schema-level inconsistencies within and across ODP and OSM can be identified using our framework, more specifically during the construction of the OBDA specification, and can be resolved using suitable mappings. In other words, the expert knowledge about these schema-level inconsistencies is encoded into the mappings so that heterogeneity is hidden at the ontology level. For instance, the inconsistency among different attribute names can be resolved by choosing in the ontology a unique term for each conceptual entity (i.e., concept or property), and by relating such term to the proper attribute(s) in the tables through mappings. Moreover, the heterogeneity of different geometry representations of features can also be abstracted at the ontological level. E.g., both points and polygons of schools in OSM are used to populate the School concept at the ontological level.

**Instance-level inconsistency** The ontological representation allows us to express the instance-level consistency rules at the integrated RDF graph. More specifically, we can formalize consistency rules as SPARQL queries, which is then translated by our framework into queries over the data sources to be assessed. Normally, each query result corresponds to one violation of a rule. The results can be later summarized and used to address the identified issues. In this experiment, we have assessed the inconsistency of 1) thematic attribute values in a single data source, 2) semantic and topological relations in a single data source, and 3) the thematic and geomteric properties across the two different data sources. These examples show a spectrum of common issues happening in practice and also demonstrate how we can identify them with our framework.

**Table 5** Summary of the detected inconsistency issues ("–" indicates "not discussed in this paper")

| Data source | Inconsistency | | | |
|---|---|---|---|---|
| | Schema-level | Instance-level | | |
| | | Thematic | Geometric | Topological |
| ODP data | attribute names, e.g., national code of municipalities, Table 3 | attribute values, e.g., addresses of pharmacies, Table 4 | – | semantic and topological relations, e.g., withIn relation, Fig. 8 |
| OSM data | possibility of modeling features as points and polygons | – | – | – |
| ODP and OSM | ODP data with different topics in separate files; OSM in a file with a large collection of features | addresses of POIs in OSM and ODP, Fig. 10 (left) | geometries of POIs in OSM and ODP, Fig. 10 (right) | – |

## 6 Conclusion

In this paper, we investigate the consistency assessment issues of multiple geodata sources in the context of data integration. Two levels of data inconsistency at schema and instance levels are identified. We propose a general framework using the ontology-based approach, which provides a coherent view of the underlying data sources, and hence enables a lightweight approach to the assessment using high-level queries and visualization. In this framework, the schema-level inconsistencies are mainly assessed by the two bottom layers of data preprocessing and the OBDA layer. The instance-level inconsistencies are mainly assessed by the two top layers of query-based consistency assessment and visualization. Preliminary experiments have been conducted using ODP and OSM data collected in the province of South Tyrol, Italy. We conduct the analysis of consistency assessment at both schema and instance levels. The analysis results show that the approach is feasible to reveal inconsistencies within and across both data sources.

This work forms a basis to solve the identified inconsistencies and to improve the quality of datasets using our framework. In the future, we plan to expand our experiments to more data sources, e.g. GeoNames and LinkedGeoData. We also plan to investigate other data quality issues, e.g., completeness, based on our approach.

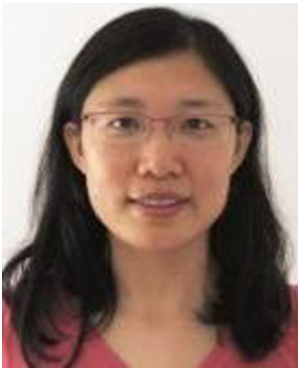
**Acknowledgements** This research has been partially supported by the EU H2020 project INODE, by the Italian PRIN project HOPE, by the European Regional Development Fund (ERDF) Investment for Growth and Jobs Programme 2014-2020 through the project IDEE (FESR1133), by the Free University of Bozen-Bolzano through the projects QUADRO, KGID, and GeoVKG, by the Jiangsu Industrial Technology Research Institute (JITRI), and by the Changshu Fengfan Power Equipment Co., Ltd.

## References

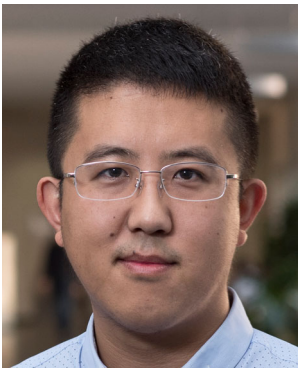
1. Hashem IAT, Chang V, Anuar NB, Adewole K, Yaqoob I, Gani A, Ahmed E, Chiroma H (2016) The role of big data in smart city. Int J Inf Manag 36(5):748–758
2. Hao J, Zhu J, Zhong R (2015) The rise of big data on urban studies and planning practices in China: review and open research issues. J Urban Manag 4(2):92–124. Big/Open Data for Urban Management
3. Wiemann S, Bernard L (2016) Spatial data fusion in spatial data infrastructures using linked data. Int J Geogr Inf Sci 30(4):613–636
4. Schaumberger A (2006) Full integration of geodata in gis. In: Socrates-Erasmus Summer School, Brno
5. Vetrò A, Canova L, Torchiano M, Minotas CO, Iemma R, Morando F (2016) Open data quality measurement framework: definition and application to open government data. Gov Inf Q 33(2):325–337
6. Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221
7. Bizer C, Heath T, Berners-Lee T (2009) Linked data - the story so far. Int J Semantic Web Inf Syst 5(3):1–22
8. Kuhn W (2001) Ontologies in support of activities in geographical space. Int J Geogr Inf Sci 15(7):613–631
9. Teller J, Lee JR, Roussey C (eds) (2007) Ontologies for urban development. Volume 61 of studies in computational intelligence. Springer, Berlin
10. Janowicz K, Hu Y, McKenzie G, Gao S, Regalia B, Mai G, Zhu R, Adams B, Taylor KL (2016) Moon landing or safari? A study of systematic errors and their causes in geographic linked data. In: GIScience. Volume 9927 of LNCS. Springer, pp 275–290
11. Sheeren D, Mustière S, Zucker JD (2009) A data-mining approach for assessing consistency between multiple representations in spatial databases. Int J Geogr Inf Sci 23(8):961–992
12. Egenhofer M, Clementini E, Felice PD (1994) Evaluating inconsistencies among multiple representations. In: Sixth international symposium on spatial data handling, pp 901–920
13. Janev V, Höchtl J (2016) Best practice: enable quality assessment of open data. Technical report W3C
14. Frank AU (2001) Tiers of ontology and consistency constraints in geographical information systems. Int J Geogr Inf Sci 15(7):667–678

15. Aracri RM, Bianco AM, Radini R, Scannapieco M, Tosco L, Croce F, Savo DF, Lenzerini M (2018) On the experimental usage of ontology-based data management for the italian integrated system of statistical registers: quality issues. In: The 9th European conference on quality in official statistics (Q2018)
16. Ubaldi B (2003) Open government data – towards empirical analysis of open government data initiatives. OECD Working Papers on Public Governance
17. Peled A (2013) Re-designing open data 2.0. J eDemocracy Open Govern 5(2):187–199
18. Heinzelman J, Waters C (2010) Crowdsourcing crisis information in disaster-affected Haiti. US Institute of Peace Washington, DC
19. MacEachren AM, Robinson AC, Jaiswal A, Pezanowski S, Savelyev A, Blanford J, Mitra P (2011) Geo-twitter analytics: applications in crisis management. In: 25th International cartographic conference, pp 3–8
20. Latif S, Islam KMR, Khan MMI, Ahmed SI (2011) Openstreetmap for the disaster management in Bangladesh. In: 2011 IEEE conference on open systems, pp 429–433
21. Haklay M (2010) How good is volunteered geographical information? A comparative study of openstreetmap and ordnance survey datasets. Environ Plan B: Plan Des 37(4):682–703
22. Girres JF, Touya G (2010) Quality assessment of the french openstreetmap dataset. Trans GIS 14(4):435–459
23. Fan H, Zipf A, Fu Q, Neis P (2014) Quality assessment for building footprints data on openstreetmap. Int J Geogr Inf Sci 28(4):700–719
24. Barron C, Neis P, Zipf A (2014) A comprehensive framework for intrinsic openstreetmap quality analysis. Trans GIS 18(6):877–895
25. Codescu M, Horsinka G, Kutz O, Mossakowski T, Rau R (2011) Osmonto – an ontology of openstreetmap tags. In: State of the map Europe (SOTM-EU)
26. Ballatore A, Bertolotto M, Wilson DC (2013) Geographic knowledge extraction and semantic similarity in OpenStreetMap. Knowl Inf Syst 37(1):61–81
27. Stadler C, Lehmann J, Höffner K, Auer S (2012) Linkedgeodata: a core for a web of spatial open data. Sem Web J 3(4):333–354
28. Koubarakis M, Bereta K, Papadakis G, Savva D, Stamoulis G (2017) Big, linked geospatial data and its applications in earth observation. IEEE Internet Comput 21(4):87–91
29. Veregin H (1999) Data quality parameters. Geograph Inform Syst 1:177–189
30. Brisaboa NR, Rodríguez MA, Seco D, Troncoso RA (2015) Rank-based strategies for cleaning inconsistent spatial databases. Int J Geogr Inf Sci 29(2):280–304
31. Senaratne H, Mobasheri A, Ali AL, Capineri C, Haklay MM (2017) A review of volunteered geographic information quality assessment methods. Int J Geogr Inf Sci 31(1):139–167
32. Comber A, Fisher P, Wadsworth R (2004) Integrating land-cover data with different ontologies: identifying change from inconsistency. Int J Geogr Inf Sci 18(7):691–708
33. Rodríguez A (2005). In: Inconsistency issues in spatial databases. Springer, Berlin, pp 237–269
34. Devogele T, Parent C, Spaccapietra S (1998) On spatial database integration. Int J Geogr Inf Sci 12(4):335–352
35. Balley S, Parent C, Spaccapietra S (2004) Modelling geographic data with multiple representations. Int J Geogr Inf Sci 18(4):327–352
36. Quix C, Ragia L, Cai L, Gan T (2006) Matching schemas for geographical information systems using semantic information. In: Meersman R, Tari Z, Herrero P (eds) On the move to meaningful internet systems 2006: OTM 2006 workshops. Heidelberg, Berlin, pp 1566–1575
37. Nathalie A (2009) Schema matching based on attribute values and background ontology. In: 12th AGILE International conference on geographic information science, vol 1, pp 1–9
38. Yu F, McMeekin DA, Arnold L, West G (2018) Semantic web technologies automate geospatial data conflation: conflating points of interest data for emergency response services. In: Progress in location based services, vol 2018. Springer International Publishing, Cham, pp 111–131
39. Duckham M, Worboys M (2005) An algebraic approach to automated geospatial information fusion. Int J Geogr Inf Sci 19(5):537–557
40. Meng L (2017) From multiple geodata sources to diverse maps. In: H L, X S (eds) Frontiers in geoinformations. Higher Education Press, pp 191–218
41. Zhang M, Meng L (2008) Delimited stroke oriented algorithm - working principle and implementation for the matching of road networks. J Geogr Inf Sci 14(1):44–53
42. Hackelöeer A, Klasing K, Krisp JM, Meng L (2013) Comparison of point matching techniques for road network matching. In: 8th International symposium on spatial data quality, Hong Kong, pp 87–92
43. Yang J, Meng L (2014) Feature selection in conditional random fields for map matching of gps trajectories. In: Lecture notes in geoinformation and cartography, progress in location-based serivces. Springer, pp 121–135

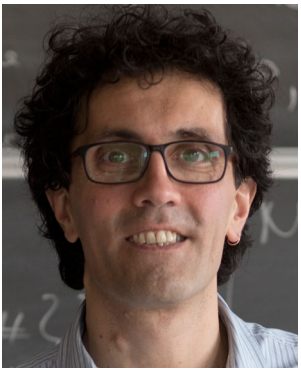
44. Paiva JAdC (1998) Topological equivalence and similarity in multi-representation geographic databases. PhD thesis the University of Maine
45. Xiao G, Calvanese D, Kontchakov R, Lembo D, Poggi A, Rosati R, Zakharyaschev M (2018) Ontology-based data access: a survey. In: Proc.of the 28th int. joint conf. on artificial intelligence (IJCAI), IJCAI/AAAI
46. Poggi A, Lembo D, Calvanese D, De Giacomo G, Lenzerini M, Rosati R (2008) Linking data to ontologies. J Data Semantics 10:133–173
47. Das S, Sundara S, Cyganiak R R2RML: RDB to RDF mapping language. W3C Recommendation, World Wide Web Consortium (September 2012) Available at http://www.w3.org/TR/r2rml/
48. Manola F, Mille E RDF primer. W3C Recommendation, World Wide Web Consortium (February 2004) Available at http://www.w3.org/TR/rdf-primer-20040210/
49. Harris S, Seaborne A SPARQL 1.1 query language. W3C Recommendation, World Wide Web Consortium (March 2013) Available at http://www.w3.org/TR/sparql11-query
50. Calvanese D, Cogrel B, Komla-Ebri S, Kontchakov R, Lanti D, Rezk M, Rodriguez-Muro M, Xiao G (2017) Ontop: answering SPARQL queries over relational databases. Sem Web J 8(3):471–487
51. Xiao G, Ding L, Cogrel B, Calvanese D (2019) Virtual knowledge graphs: an overview of systems and use cases. Data Intell 1:201–223
52. Perry M, Herring J (2011) GeoSPARQL - a geographic query language for RDF data. OGC Candidate Standard OGC 11-052r3 Open Geospatial Consortium
53. Bereta K, Xiao G, Koubarakis M (2019) Ontop-spatial: ontop of geospatial databases. Journal of Web Semantics
54. Bereta K, Xiao G, Koubarakis M, Hodrius M, Bielski C, Zeug G (2016) Ontop-spatial: geospatial data integration using GeoSPARQL-to-SQL translation. In: Proceedings of the ISWC 2016 posters & demonstrations track. Volume 1690 of CEUR workshop proceedings., CEUR-WS.org
55. Brüggemann S, Bereta K, Xiao G, Koubarakis M (2016) Ontology-based data access for maritime security. In: ESWC. Volume 9678 of lecture notes in computer science. Springer, pp 741–757
56. Frank AU (2003) Ontology for spatio-temporal databases. In: Spatio-temporal databases: the CHOROCHRONOS approach. Volume 2520 of lecture notes in computer science. Springer, pp 9–77
57. Kuhn W (2012) Core concepts of spatial information for transdisciplinary research. Int J Geogr Inf Sci 26(12):2267–2276
58. W3C (2017) Semantic sensor network ontology. W3C Recommendation, W3C
59. Sequeda JF, Miranker DP (2015) Ultrawrap mapper: a semi-automatic relational database to RDF (RDB2RDF) mapping tool. In: International semantic web conference (posters & demos). Volume 1486 of CEUR workshop proceedings., CEUR-WS.org
60. Jiménez-Ruiz E, Kharlamov E, Zheleznyakov D, Horrocks I, Pinkel C, Skjæveland MG, Thorstensen E, Mora J (2015) Bootox: Practical mapping of rdbs to OWL 2. In: International semantic web conference (2). Volume 9367 of lecture notes in computer science. Springer, pp 113–132
61. Sicilia Á, Nemirovski G, Nolle A (2017) Map-on: a web-based editor for visual ontology mapping. Sem Web 8(6):969–980
62. Sequeda JF, Miranker DP (2017) A pay-as-you-go methodology for ontology-based data access. IEEE Internet Comput 21(2):92–96
63. Katifori A, Halatsis C, Lepouras G, Vassilakis C, Giannopoulou E (2007) Ontology visualization methods — a survey. ACM Comput Surv (CSUR) 39(4):10
64. Soylu A, Kharlamov E, Zheleznyakov D, Jiménez-Ruiz E, Giese M, Skjæveland MG, Hovland D, Schlatte R, Brandt S, Lie H, Horrocks I (2018) OptiqueVQS: a visual query system over ontologies for industry. Sem Web 9(5):627–660
65. Beek W, Folmer E, Rietveld L, Walker J (2017) Geoyasgui: the GeoSPARQL query editor and result set visualizer. ISPRS - International archives of the photogrammetry, remote sensing and spatial information sciences XLII-4/W2, pp 39–42
66. Nikolaou C, Dogani K, Bereta K, Garbis G, Karpathiotakis M, Kyzirakos K, Koubarakis M (2015) Sextant: visualizing time-evolving linked geospatial data. J Web Sem 35:35–52
67. Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubézy M, Eriksson H, Noy NF, Tu SW (2003) The evolution of protégé: an environment for knowledge-based systems development. Int J Hum-Comput Stud 58(1):89–123
68. Aracri RM, Radini R, Scannapieco M, Tosco LGarrigós I, Wimmer M (eds) (2018) Using ontologies for official statistics: the istat experience. Springer International Publishing, Cham



**Linfang Ding** is a postdoc researcher at the KRDB Research Centre for Knowledge and Data, Free University of Bozen-Bolzano, Italy, and at the Chair of Cartography, Technical University of Munich, Germany. She received her Ph.D. degree in cartography from Technical University of Munich, Germany, in 2016. Her research interests are geodata integration, ontology-based data access, geovisual analytics and mobility data analysis.



**Guohui Xiao** is an assistant professor at the KRDB Research Centre for Knowledge and Data, Free University of Bozen-Bolzano, Italy. He received his Ph.D. degree in computer science from Vienna University of Technology, Austria, in 2014. He is currently leading the development within the Ontop team. His main research interests are geodata integration, ontology-based data access, and geovisual analytics. His main research interests are knowledge representation, description logics, semantic web, database theory, ontology-based data access, and optimization and implementation of reasoning engines.



**Diego Calvanese** is a full professor at the KRDB Research Centre for Knowledge and Data, Free University of Bozen-Bolzano, Italy. His research interests include formalisms for knowledge representation and reasoning, ontology languages, description logics, conceptual data modeling, data integration, graph data management, data-aware process verification, and service modeling and synthesis. He is one of the editors of the Description Logic Handbook. He is a fellow of the European Association for Artificial Intelligence (EurAI) since 2015.

**Liqiu Meng** is a full professor of Cartography at the Technical University of Munich. Her research interests are geodata integration, mobile map services, multimodal navigation algorithms, and geovisual analytics. She is a member of "German National Academy of Sciences" and "Bavaria Academy of Sciences". She is currently serving as the Senate of German Aerospace Center DLR, one of the Series Editors for the Springer Lecture Notes "Geoinformation and Cartography", and vice president of the International Cartographic Association.