Interpretation and automatic integration of geospatial data into the Semantic Web: Towards a process of automatic geospatial data interpretation, classification and integration usi...

Article *in* Computing · February 2019

DOI: 10.1007/s00607-019-00701-y

CITATIONS
0

READS
105

6 authors, including:

Claire Prudhomme
University of Burgundy
16 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Timo Homburg
Hochschule Mainz
28 PUBLICATIONS **30** CITATIONS
SEE PROFILE

Jean-Jacques Ponciano
Hochschule Mainz
12 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Frank Boochs
Hochschule Mainz
135 PUBLICATIONS **585** CITATIONS
SEE PROFILE

Computing
https://doi.org/10.1007/s00607-019-00701-y

Interpretation and automatic integration of geospatial data into the Semantic Web

Towards a process of automatic geospatial data interpretation, classification and integration using semantic technologies

Claire Prudhomme[1] · Timo Homburg[1] · Jean-Jacques Ponciano[1] · Frank Boochs[1] · Christophe Cruz[2] · Ana-Maria Roxin[2]

Received: 6 September 2017 / Accepted: 10 January 2019

Abstract
In the context of disaster management, geospatial information plays a crucial role in the decision-making process to protect and save the population. Gathering a maximum of information from different sources to oversee the current situation is a complex task due to the diversity of data formats and structures. Although several approaches have been designed to integrate data from different sources into an ontology, they mainly require background knowledge of the data. However, non-standard data set schema (NSDS) of relational geospatial data retrieved from e.g. web feature services are not always documented. This lack of background knowledge is a major challenge for automatic semantic data integration. Focusing on this problem, this article presents an automatic approach for geospatial data integration in NSDS. This approach does a schema mapping according to the result of an ontology matching corresponding to a semantic interpretation process. This process is based on geocoding and natural language processing. This article extends work done in a previous publication by an improved unit detection algorithm, data quality and provenance enrichments, the detection of feature clusters. It also presents an improved evaluation process to better assess the performance of this approach compared to a manually created ontology. These experiments have shown the automatic approach obtains an error of semantic interpretation around 10% according to a manual approach.

Keywords Semantic interpretation · Data quality · Natural language processing · Ontologies · Spatial fusion · Semantic Web

✉ Claire Prudhomme
claire.prudhomme@hs-mainz.de

[1] Mainz University of Applied Sciences, Lucy-Hillebrand-Str. 2, 55128 Mainz, Germany

[2] Laboratoire d'Informatique de Bourgogne (LIB) - EA 7534, University of Bourgogne Franche-Comté, Bâtiment i3M rue Sully, 21000 Dijon, France

Published online: 13 February 2019

1 Introduction

In the domain of disaster management, when a disaster happens, a crisis unit gathers all the representatives of the stakeholders (firemen, police, managers of energy, etc). All of these stakeholders must work in a collaborative way by exchanging information about the situation and cooperating with one another. The crisis unit has to make critical decisions in a short time to rescue and protect the population and maintain a minimal network infrastructure. The efficiency of a decision depends on the skills of the crisis unit, but also on the quantity and the quality of information that it has. Information comes from the different stakeholders or from other sources such as web feature services [66]. However, each of these data sources may originate from different information systems, which are not interoperable and which provide different types of information. This is why a crisis unit is often composed of a team in charge of gathering all of the information to produce a map, providing an overview of the situation. This collection of information during a disaster takes up time that is crucial. To support this community in terms of data and information management, we are working on a semantic-based geographic information system allowing for the integration of heterogeneous data (e.g. information about hospitals, school, shelter, etc.) to support the decision-making about the evacuation process. The goal of this information system is to integrate information from heterogeneous data to create a knowledge base. The data with relevant information includes, a lot of shapefiles [17] from different sources. These shapefiles contain geospatial information, but they also contain other information that can be important for decision-making. The type of information contained in these shapefiles can be very different according to their origins. The research problem of this paper is how to automatically integrate and interpret a set of geospatial data with unknown content so as to retrieve the maximum information. The proposed approach aims at identifying the meaning of the data content by extracting it and representing it in an ontology. Most approaches interpreting and including geospatial data sets in an ontology are manual and/or semi-automatic. These manual and semi-automatic approaches still require user input, making them time-consuming. This input is generally a mapping [18] of the data sets' contents to corresponding classes in the Semantic Web. Providing a mapping for each file would require a lot of effort. Indeed, in addition to being time-consuming which is really not suitable in our context, providing a mapping for data without background knowledge is also a challenge. These types of data are called further non-standard data set schema (NSDS) data sets. This term is used to refer to a data set that is structured in a relational way (i.e. a database table), but is not interconnected to another database table or schema. Shapefiles are a typical example of such a data set structured as a database table. But in contrast to a GML [10] file, for example, they usually do not conform to a standard (i.e. represented by an OGC-[61] endorsed and well-described XSD schema [19]). Although these data might be poorly documented, limiting their usage by creating difficulties in extracting the content meaning, they do provide potentially relevant information to support disaster management. The approach presented in this paper focuses on this type of data and attempts to extract the maximum information automatically. It is based on natural language processing algorithms, geospatial tools and methods to define an ontological schema representing the content of the file and allowing its integration. The natural

language processing techniques used include, named entity recognition and respective language-related ontologies such as WordNet or BabelNet [3,37,38,65]. These elements are combined with geospatial data fusion and traditional reverse geocoding methods [22] to reliably generate a semantic interpretation of the data content. An enrichment of this information from Linked Open Data is then processed to increase the quantity of information. Finally, information about data provenance and quality is added to support a future assessment of the most relevant pieces of information according to their quality. Related works about the different aspects of this approach (semantic mapping, semantic integration, and enrichment with geospatial data quality and provenance) are presented in Sect. 2. Section 3 highlights the methodology behind our approach of automatic data uplift which, is discussed in detail in Sect. 4. We investigate the semantic and automatic integrational quality of our approach by comparing its results with the results obtained by two different manual annotators in Sect. 5. Finally, Sect. 6 presents the advantages and limits of this approach and discusses the perspectives of this research work.

2 Related work

In the related work of enriching linked data with a geospatial dimension, a methodology of processes is presented by [13], who apply it to CSV files [52]. The first process, uplift, transforms the geospatial data into RDF triples [45]. The second one enriches data and consists of link discovery and their incorporation into RDF data. The last one, downlift, transforms the newly enriched RDF data into an enriched CSV file. These different steps of methodology can be found through the goal of the different tools integrated in the GeoKnow project [32] to geographically enrich data with linked data web. This project uses TripleGeo [46] and Sparqlify [58] as tools to do the uplift. The enrichment is done using LIMES [40] for link discovery and Geolift for enriching and data cleaning. The end-user of GeoKnow aims at managing linked data on the web, so no downlift step is included. Instead, the authors visualized the data in a user interface.

Although there are approaches to integrate heterogeneous data, exemplary shown and discussed in [25], a majority of approaches are mainly focused on CSV files [13] and geographic databases [32] whereas these two types of data have similar representations of tabular schemas. The use of semi-automatic schema matching, means the transformation in RDF triples is not fully automated. However, some research like the Datalift project [55], has led to a fully automated RDF data transformation. This project allows for taking many heterogeneous input formats (e.g. databases, CSV [52], GML [10], shapefile [17]) in order to convert them into RDF and interlink them. Their approach is a two step approach. The first step is to convert the input format into RDF triples (subject-predicate-object), where the subject corresponds to an element of a row, the predicate is based on the column name, and the object is the content of a cell corresponding to the intersection between the row of the subject and the column of the predicate. The second step converts these RDF triples into a "well-formed" RDF according to chosen vocabularies. This second step is done thanks to the use of SPARQL [50] CONSTRUCT queries. We tested this approach on one of our shapefiles to compare their "well-formed" RDF to the result of our approach. The

result of this test has provided us with a set of triples based on annotation properties without concepts. Although this approach is similar to ours, it is difficult to compare the two, because our approach aims at extracting a maximum of semantic information by representing the content of a file through concepts, instances, object- and data properties. Similar approaches to ours exist, but are mainly based on semi-automatic schema matching to efficiently extract the semantic of a file. The few approaches using automatic schema matching are less efficient when extracting semantic information. Although, an automatic approach often underperforms when it comes to the accuracy of a semi-automatic approach, which benefits from the expert view of the file, an automatic approach is useful for interpreting data which are provided without expert view. Consequently, our approach provides an automatic approach for semantic data interpretation and its enrichment. As highlighted by the previous related work, the methodology to achieve our goals is composed of essential steps, whose existing works are given in the following subparts of this section. The uplift step [13] is discussed in works related to schema matching, whereas the data enrichment is discussed in works related to link discovery and ontology matching. Interlinked RDF data are often enriched by further information such as data provenance [14,28], which is why a subpart presents the related work of this domain. The downlift step [13] is not discussed in this related work, as it is an end-user specific task depending on the adaptation of the schema of the data file that you want to provide as the end-user.

2.1 Uplift

The target input of our work comprises tabular-structured geospatial data sets, without description of the data content and meaning. The related work of the uplift step concerns the approaches allowing the transformation of data set contents into local ontologies comprised of RDF triples. Concerning geospatial information, related work like [13,14,32] used the GeoSPARQL vocabulary [43] to represent the knowledge base. This OGC standard introduces a vocabulary and SPARQL functions for dealing with geospatial features and gometries. Related approaches to local ontology creation to represent a data set are either semi-automatic or specific to a certain datatype. Bizid et al. [8] convert GML data sets to local ontologies using GML schemas and provide automated interlinking strategies for similarly structured database resources. A more general approach to different geospatial file formats and DBMS is presented by the tool TripleGeo [46], which is based on an Extract-Transform-Load (ETL) process and uses GeoSPARQL to represent geospatial information. Other related work that is more RDBMS-specific is the tool Sparqlify included in the GeoKnow project [32]. Sparqlify provides a RDF view through a SPARQL query, using SPARQL to SQL translation mechanisms. Some tools like BOOTOX [28] have been developed to facilitate the mapping from given relational databases to extract a corresponding ontology from the database schema. Some standards have also emerged, such as R2RML [12] (W3C), which allows for expressing a customized mapping from relational databases to RDF data sets. Debruyne et al. [13] used R2RML to also express a mapping from a CSV file as presented in the paper. However, some automatic process to create a local ontology from a relational database as the direct mapping presented by the W3C

[2] or different processes of schema mapping like the diverse approaches are additionally presented in [51]. This survey classified the different individual matchers as schema-only based and instance/contents-based, which can be both a linguistic or a constraint-based approach at an element level. Among these approaches, Pinkel et al. [48] presented a schema matching based on intermediate graphs obtained by transforming the two inputs corresponding to a relational database and an ontology. These two intermediate graphs are then matched. Their approach uses two matchers based on the graph structure and a lexical one. First it creates a matching using a pairwise connectivity graph to gather pair by pair potential nodes. It then applies a Jaccard similarity matcher [42] and finally, applies a structure matcher using an adaptation of similarity flooding algorithms [36]. The work of Do and Rahm [15] aims at comparing the efficiency of the different types of matchers and assesses their combinations. Their benchmark highlights the efficiency of a combination of matchers and the reusability of matcher results to simplify future mapping. Contrary to the aforementioned approaches, our approach is geodata centric and therefore our automatic approach applies the data integration on data sets with common properties, while at the same time applying an automated interlinking for semantic integration.

2.2 Link discovery and ontology matching

Nentwig et al. [39] conducted a survey of current link discovery frameworks and discussed eleven frameworks by highlighting their specificities. All of the presented frameworks support the relation *owl:sameAs*, but only the Silk Framework [64] and LIMES [24] allow the user to specify other relations. The majority of them require manual configuration. However, four of them have a semi-automatic and adaptive linking specification based on the data set analysis and the identification of the most discriminative properties. Among the four learning-based frameworks, three use supervised learning, whereas only two utilize unsupervised learning. Concerning similarity measures, the eleven frameworks all utilize them, but only five have a structure matcher. Two of them are interesting when it comes to the geospatial domain: Zhishi.links [41] and LIMES (GeoKnow project), because they use geographical coordinates as a similarity measure. These similarity measures intervene in the main step of link discovery, which is the ontology matching. Different techniques of ontology matching are presented in [44]. Their classification is divided between the element and structure-levels, but also between semantic and syntactic techniques. Another method of classification is to use the kind of input rather than the granularity interpretation. In this case, the classification of techniques is divided between context-based techniques, which can be semantic or syntactic, and content-based techniques, which can be terminological, structural, extensional or semantic. As shown by the study of the link discovery frameworks, these techniques are generally combined to obtain better results. However, ontology matching still faces some challenges, which are presented in [57]. In this publication we address the challenge related to the matching of background knowledge by searching to identify concepts related to data content due to the lack of information about the background knowledge of this data.

2.3 Data quality and provenance

In this publication we use the definition of data quality referred by [16] in [21]: Quality is the "degree to which a set of inherent characteristics fulfils requirements". In this meaning, data quality can be defined in various data quality dimensions, two of which we use and highlight in this publication.

2.3.1 Data quality of imported geospatial data

Data quality of geospatial data has been broadly discussed in the geospatial community and can be broken down into the categories of Completeness [9], Semantic Similarity Measurements [56] and Positional Accuracy [20]. While all of the aforementioned methods of quality assessment require a gold standard for comparison [5,29], also developed intrinsic data quality measures like Logical Consistency checks, which can evaluate a geometry with no need for a gold standard. In our article, we refer to those data quality assessment methods to enrich our mapping using data quality annotations in the uplift process.

2.3.2 Semantic integrational quality

Semantic Integrational Quality is the quality of a mapping of a data set to its Semantic Web representation. Zaveri et al. [68] gives an overview on this topic. Tarasowa et al. [60] proposed semantic integrational quality metrics to measure the quality of RDB2RDF mappings in the categories of faithfulness of the output, quality of the mapping implementation, quality of the output and interoperability. Pinkel et al. [47] proposed a benchmarking tool to evaluate relational database to ontology mapping. All of the previous approaches therefore can provide a foundation of evaluating whether a mapping is good or not. Our use case is different in the sense that we are working with previously non-interconnected, i.e. NSDS, data which provide new challenges for the mapping and have very little information about the usage context, whether the data set can be trusted and many other parameters mentioned in the previous publications. We therefore introduce a tailor-made approach for measuring semantic integrational quality in Sect. 5.2.

2.3.3 Provenance

Provenance or Lineage [31] in the GIS community of data can be another data quality indicator, as the history of the authors and their contributions to the current data set can be tracked and used for further analysis. As shown in [13,23] provenance is commonly used in the Semantic Web community for reasoning tasks about the origin of data sets. We build upon the concept of provenance in order to firstly enrich the knowledge base for further analysis and secondly to provide a basis for a potential data quality analysis over several data set revisions in a Semantic Web context.

3 Method

The goal of this approach is to determine what the background knowledge contained in the data set is by determining concepts from the Semantic Web to represent the components of a data set. The geospatial information extraction (detailed in Sect. 4.1) allows for representing the geometry of a feature as presented by GeoSPARQL vocabulary [6]. The schema matching is defined between a relational data set and the Web Ontology Language [63] by creating *owl:Class*, *owl:Individual*, *owl:ObjectProperty* and *owl:DataProperty*. This schema matching is obtained by an analysis of the data set content detailled in the next paragraph. The ontology matching is applied to discover a URI from the Semantic Web which can represent the content of the data set. The process of data interpretation aims at identifying potential concepts through the schema matching, searching for identified concepts in the Semantic Web through the ontology matching, and creating the appropriate representation according to the ontology matching result. The combination of schema and ontology matching produces a local ontology linked to concepts in the Semantic Web, which is populated by the content of the data set (see Fig. 1). The last step of the approach is to annotate the population (set of individuals and their properties) from the data set content with their provenance and quality information. This last step allows for following updates of information and linking them without ignoring their provenance information to provide the possibility to recreate or extend the data set.

3.1 Schema mapping

The schema defines a concept (*owl:Class*) to represent the type of the data set. In addition, it defines an instance for each row of the data set table and a geometry. Each column represents a piece of related information of an instance, therefore it is represented by an *rdf:Property*. However, the specification of the column as *owl:DataProperty* or *owl:ObjectProperty* depends on the ontology matching result performed during the feature value and the descriptor analysis. A specification of the general concept representing the data set can be identified by the combination of the feature value analysis, which identifies reoccurring values and the feature descriptor analysis, which determines a classification for a column name.

3.2 Ontology matching

Our ontology matching is a hybrid matching using string-based and instance-based techniques.
The string-based technique uses the Google translate API[1] to determine the language of the data set. Next, this language is used in a process of string similarity matching with the label of the same language of a chosen ontology from the Semantic Web. The string similarity is computed through the Levenshtein measure [34]. This technique is mainly applied to column heads (see Sect. 4.3). In the case of a non-matched concept

[1] http://code.google.com/p/google-api-translate-java.

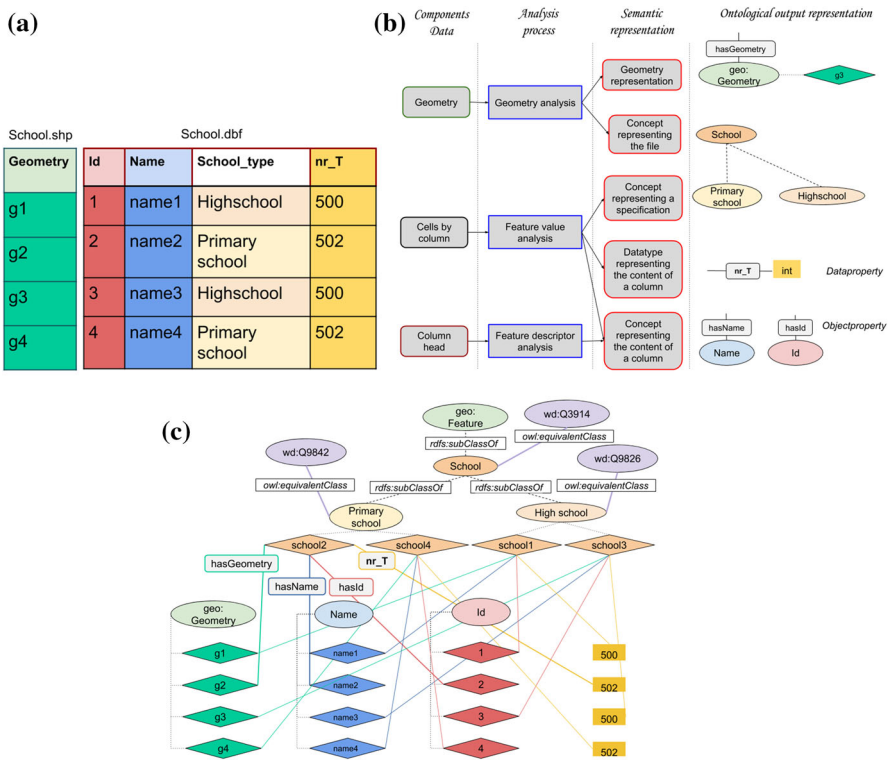

Fig. 1 Methodology overview applied to an example: **a** input example, **b** schema matching example according to the result of the steps of the ontology matching, **c** output example with Wikidata matching

and a non-English label, the column head is translated into English, and the process is repeated. A column head may also be a compound noun, which must first be split in order to do the matching with the Semantic Web.

The instance-based technique introduced in Sect. 4.2 is also applied by an analysis of the cells' content to determine a potential existing matching individual from the Semantic Web. After their identification, their classes are retrieved and the most frequently occuring class is assigned as the concept of the column.

3.3 Combination

Executing the different steps of analysis gives us four sets of concepts, which are used to build the resulting local ontology as follows:

1. *Geometry detection set* If a class has been detected using the geometry detection, this class (or the highest ranked class) will be taken to describe the data set.
2. *File name detection set* If a class has been detected via analysis of the filename, this class will be taken if no appropriate geometry class has been detected.

3. *Property detection set* Properties and their respective ranges as detected by the feature descriptor analysis are created and used for address columns as determined by its respective analysis.
4. *Individual detection set* Individuals are created according to the recognized classes and values that can be resolved to URIs will be created as the corresponding individuals.

Thanks to the schema matching, the data set is considered as a derived graph. Through this derived graph and the concept matching, our process dynamically creates the different identified concepts, their individuals, the objectproperties relating to them, and the data properties for the column without concept, to ultimately link them together according to our derived graph.

4 Interpretation process

This section presents the approach to interpreting heterogeneous data. These heterogeneous data are geodata sets with a tabular structure containing a set of objects with a description of their geometry and other features corresponding to information related to these objects. The challenge of this approach is extracting information from data without background knowledge.

4.1 Geometry representation

We use the GeoSPARQL vocabulary [6] to represent the geospatial data we uplift and extend the vocabulary using the interpreted data sets we import. GeoSPARQL defines a spatial object as a geometry and a linked feature. Representing a geometry in GeoSPARQL requires to retrieve and identify the geometry type of each entity of the geospatial data set. Automatically detecting the correct representation of the feature linked to the geometry is the objective of this publication.

4.1.1 Geometry matching

The geospatial Semantic Web consists of geospatial ontologies such as GeoNames [67] and LinkedGeoData Ontology [4]. These ontologies gather a great number of classified geometries to describe the object corresponding to said geometry. Our first approach was to use a small enough buffer, around the geometry to identify a concept. By buffer, we mean either an encompassing rectangle around a point geometry, or an encompassing rectangle around the centroid of a non-point geometry. The buffer is increased dynamically if no appropriate results have been found in the last iteration. The process of one interation is illustrated in Fig. 2.

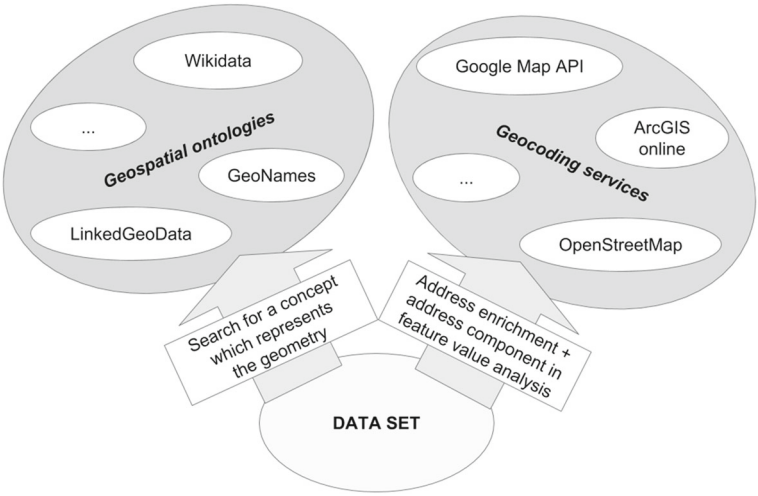

Fig. 2 Geometry and data set specification [49]

4.1.2 Geometry comparison

In many cases, not only the matching of a class is important, the merging of the attributes of two particular geometries and/or the relation of a particular geometry to encompassing geometries is also of interest. In the example of a hospital complex, it is to be expected, that the complex itself consists of various buildings, some of which may share the concept of a hospital, but some of which (e.g. a hospital chapel building) may not include a hospital annotation in the respective ontology. To be able to determine the exact geometry concerned, we employ the following geometry matching and spatial fusion techniques from the geospatial world:

- *Similarity metrics* Haussdorff Distance [26], Fréchet Distance [1], Shape Similarity [62], Overlapping Degree [7]
- *Geometrical features* Diameter, Length, Number Of Points

Similar to comparing label values of the Semantic Web to columns of our data set to be integrated, we use geometry matching algorithms for data sets providing enough geometrical information for verification (Polygons or LineStrings). This approach is extendable to encompass possible further metrics, many of which are introduced in [62]. The more metrics are applied, the more information our algorithm can evaluate as a basis for geometrical similarity assessment.

4.2 Feature value analysis

The first step of feature description analysis is to identify the information that appears frequently in the data sets. In this process, empty and NULL values are ignored. The types of information and their detection are described below and illustrated in Fig. 3.

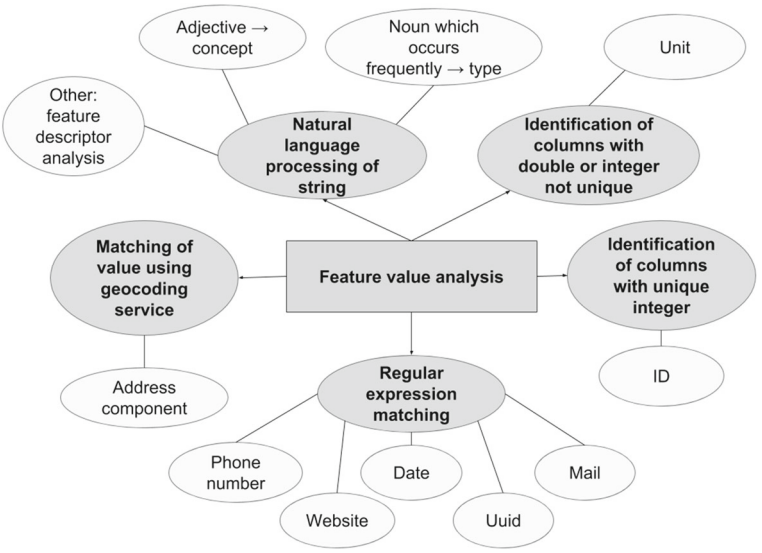

Fig. 3 Identification by value analysis [49]

– *Address components* The specificity of geodata sets is that they contain a geometry for each spatial object. The usage of the spatial object geometry with a geocoding service (in our case, Google Maps API [59]), allows for address enrichment, explained in [49]. The information retrieved is compared with the different value of the cell to determine which column contains information concerning the geographic address of the object.
– *ID* The process of an eventual ID discovery corresponds to an analysis of values in order to identify a column which fulfils the following constraints: the value has to be an integer and has to be unique. If we discover UUIDs for example, they will be categorized using appropriate regular expressions. IDs could be used as individual descriptors in a later process.
– *Unit* A double generally represents a quantification, which is why an analysis of all columns determines that a column could represent a quantity with a unit, if all values are Double or Integer. Something that is usually measured in any unit (e.g. 2.5 °C) or is a description of an amount (2.5 apples). If we can identify the column type from its descriptor, then we may be able to use it to draw conclusions about the unit associated with this type, otherwise it will be associated to Unitless. Work on integrating e.g. DBpedia [3] with unit ontologies has been done by [54] and is extended manually by our projects work for most common units.
– *Regular expression* A set of regular expressions has been defined for: a date, a phone number, an email address, a website url and a UUID. This set of regular expressions is then applied to all strings in order to check whether the string matches with one of those regular expressions. The elements identified as a date are stored thanks to a data property with the name of the column and the type *xsd:date*. Information corresponding to a phone number, an email address, and

a website is stored using FOAF ontology [30] properties foaf:phone, foaf:mbox, foaf:homepage. The UUID is stored as a data property.
– *Remaining string* Natural language processing in the form of named entity recognition, and POSTagging is applied to all strings which have not yet been identified (using the Stanford NLP Toolkit [35]). For the moment, this natural language processing is specific to German and English, and may be extended to further languages in the future. It is aiming to determine whether the string is an adjective or a noun. The values of the column, containing a majority of adjectives will become an instance of the concept linking to the general concept with an object property. When a column contains a set of nouns which occur frequently, we assume the column describes a type of the general object, as stated in [49]. The value of this column is processed to identify a set of nouns without redundancy, and then, the nouns which composed this set are added as a subclass of the general concept which represents the file. When all values have been analysed, the process of feature descriptor analysis (cf. 4.3) begins and is applied to all column names which have not yet been processed by the value analysis, to the adjective column, and to the nouns that become a subclass.

The second step consists in the recognition of a named entity. When doing string value analysis it is important to separate named entities from nouns, because their representations in ontologies usually vary from each other. In a geospatial context it is likely that geographic names are found in a data set in order to link them to other relevant geospatial resources. One example could be an administrative district or a specific administrative building like the town hall or the parliament using its name [11].

4.2.1 Unit assignment

In addition, we want to use domain information to identify missing ranges for non-classified features. For this purpose, we employ the ontology of units of measure and related concepts [54] to relate concepts of the geospatial domain to typical measures (e.g. length, height and width for buildings) and give "reasonable" upper and lower bounds for their values. For example, if we know the tallest and the shortest building in the world as upper and lower bounds for height, we can infer that height is usually measured by a double and can estimate average building heights for countries and regions using a statistical analysis of an elevation model. This information can help us to classify the unit of values of a column (in this case "height") more precisely. As the class of geometric objects on a map is limited and easy to generalize (e.g. a fire station is a building, a motorway is a road), it is feasible to relate measures to classes of objects in a general way, and this can easily be extended if more specific concepts and information are created. Figure 4 presents a usage example of the unit graph.

4.3 Feature description analysis

In the case of our data sets of the column name, the feature description analysis (Fig. 5) can give us valuable information about properties and classes in ontologies

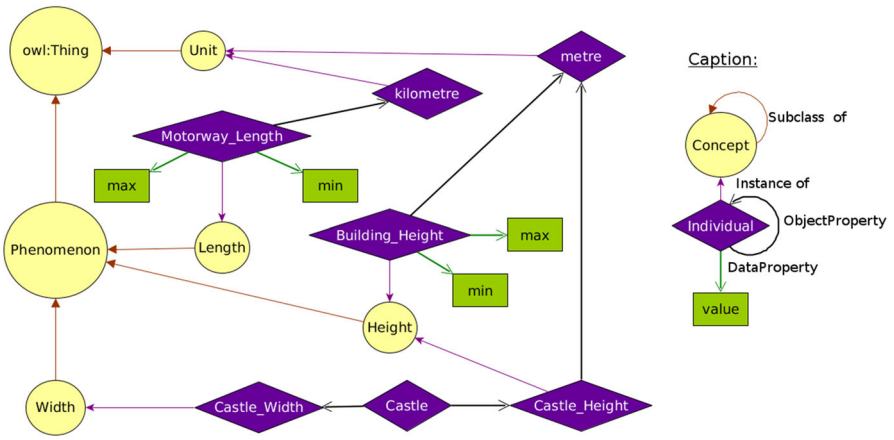

Fig. 4 Unit graph representing the application of [54]

that represent the column's content. However, column names are represented in natural language and with a limited context to parse from, which can limit disambiguation methodologies if needed. In addition, before an analysis of the feature descriptor can be conducted, the following pre-processing steps must be conducted:

- Detection of the language being used in the column's name using, for example, the Google Translate API.
- Recognition of common abbreviations and replacement of those with their long form using abbreviation lists for the particular language.

We conduct the analysis of column names as follows:
For a list of triplestores to be examined, we try to match a concept first using its URI and, if this fails, using a label matching approach. If there is no concept after these two steps, we translate the given column name to English and try the aforementioned steps again. We discovered that using an English translation is not always possible, as the translation of the full term does not necessarily represent a word that can be found in a dictionary or ontology. Compound words often needed to be split and investigated separately. In that regard, we analysed the parts of compound nouns from their ending to their beginning and tried to resolve possible concepts from those noun parts.

Listing 1 Splitting of compound nouns

```
1 Bauarbeiter -> Arbeiter
  primary school -> school
```

If we cannot find a concept for the column's name using all of the aforementioned methods, we declare the column unresolvable. If we have many results for the respective column, we will rank the results using the Levensthein Distance to find out the concepts name which comes closest to the column's name. This concept will be used to describe the column in the local ontology.

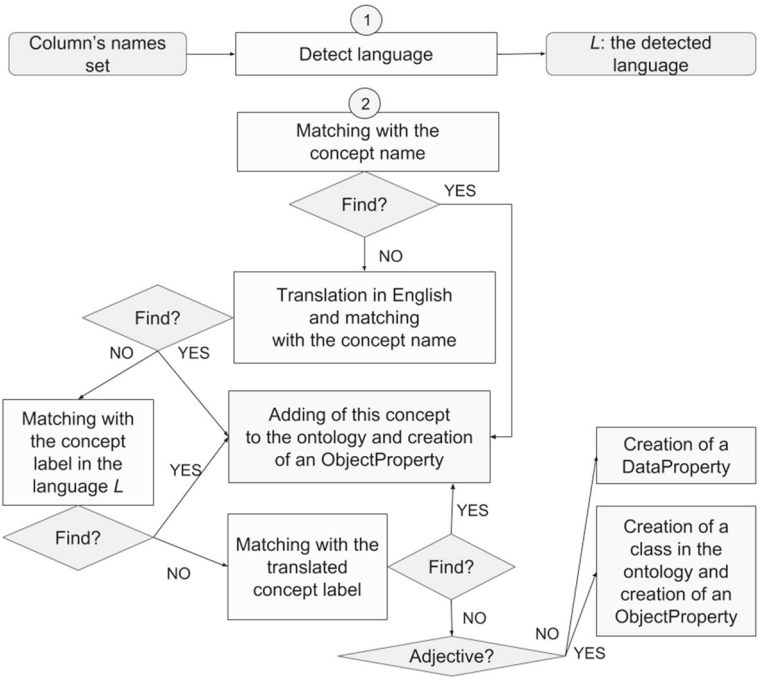

Fig. 5 Process of linking with Semantic Web resource [49]

Table 1 Example file "FireS" represented as a DB table with column clustering

ID	the_geom	Height	Type	City	District	Vehicles	GMLID	Financier
FireS.1	POINT(..)	12	Main	Berlin	Mitte	10	mdbcs99	Berlin City
FireS.2	POINT(..)	8	Small	Hamburg	Harburg	2	sdasd69	Hamburg City

4.4 Feature column clustering

In accordance with the nature of a geospatial data set we expect to find features that are directly related to the geometry and/or its most likely concept, and columns which are indirectly or not related to the most likely concept of the data set. In the example of a data set describing fire stations (Table 1), we therefore expect features to describe the geometry better, called geometry attributes (e.g. address components, GML IDs, etc.), and features related the domain of fire-fighting, called concept attributes (e.g. number of fire-fighting vehicles). Remaining features belong to a different domain that is indirectly associated with the main concept domains of the data set (e.g. the financier of the fire station). This principle is illustrated in Fig. 6.
We match every feature which has been identified in the data set with a domain as follows:

– Creation of a list of superclasses and topics for the identified concepts.
– Calculation of the graph distances between superclasses found and topics.

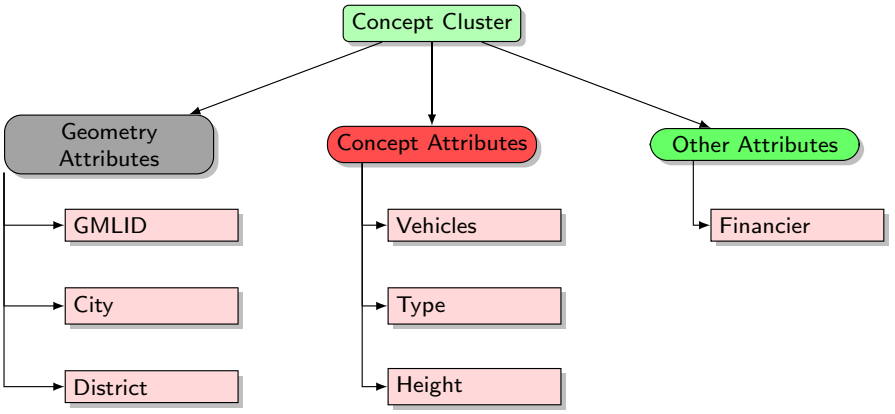

Fig. 6 Concept cluster analysis

– Picking the most specific matching superclasses as feature column clusters for the respective features.

The result is a map of domains to sets of concepts provided by features and columns. We use this map to decide which features are essential for describing the geometry (features associated with the geometry and domain of the feature class).

4.5 Data quality and provenance enrichment

At this point in the algorithm, we receive a mapping of one class to represent the whole data set, at best one interpretation per column of the data set and one interlink per instance to a corresponding individual in LinkedGeoData and GeoNames using an owl:sameAs relation between their representation as feature, if a corresponding geometry is found. However, we can also take advantage of this relationship to enhance the data set using data quality analysis tools and by adding provenance information. Here we import typical provenance information such as the origin of the data set, its creation and/or modification date, format, web service and authors, depending on availability, using the PROV-O vocabulary [33] from both the geospatial data set we import and the LinkedGeoData data set using the OpenStreetMap history. In addition we can add intrinsic data quality metric results such as geometry validity, geometry simplicity, its diameter or the number of points included in the geometry. Intrinsic metrics serve as basic information for the usefulness of a specific geometry in a specific data set to the typical GIS end user. Consequently, we can also include extrinsic data quality metric results which were created in the geometry matching process described in Sect. 4.1.

4.6 Conclusion

Compared to the last version of our algorithm presented in [49], we added four new improvements. Firstly, we improved the geometry matching process by adding shape

similarity measurements. This allows us to create better confidence scores for the concepts we match and to exclude certain concepts which have a contradicting geometry (e.g. we are matching a polygon geometry and can discard a LineString geometry or a non-overlapping polygon geometry). In addition we added the detection of feature column clusters in order to determine the usefulness of columns for the data set. Having a categorization of columns allows us to judge a mapping not only on its overall completeness, but additionally on its potential usefulness to the end user. Thirdly, we improved our unit detection approach by providing an approach to add reasonable boundaries for common observations present in geospatial concepts. Being able to detect whether a detected unit column is consistent with real world observations (e.g. if the height of a building should exceed 10 km) improves the confidence in the unit detection algorithm and the overall mapping result. Lastly, the addition of provenance and basic data quality evaluation results allows the end user to compare the quality of the data sets they have imported either by means of authenticity of the provider or via data quality metrics classifying the data set intrinsically.

5 Evaluation of the interpretation process

5.1 Experimental setup

The experimental setup explains the different data sets and the different approaches which have been used and their goals.

5.1.1 Data sets

To test our approach, we needed to apply it to a set of data which provides a diversity of information with different qualities. With our project SemGIS[2] being in the context of disaster management, we have chosen five files: two files about schools, two files about hospitals and a file about rescue organisations which will serve as buildings to possibly be evacuated or emergency unit buildings respectively. Files about schools and hospitals come from two different regional sources, one each for the city of Cologne[3] and the other for the German *Bundesland* of Saarland[4] in Germany. The two different files have some similarities, but do not provide the same type of information. They allow us to evaluate the integration of a similar data set (with a same subject) from different data sources and with different contents. Thanks to three fields (school, hospital and rescue organisation) and different data sources, we obtain a diversity of information with different qualities, as the column names and their contents which describe the information are built differently. As shown in Table 2, these data sets are composed of geospatial attributes (GA) (e.g. street, postal code, district), concept attributes (CA)

[2] http://i3mainz.hs-mainz.de/de/projekte/semanticgis.

[3] https://offenedaten-koeln.de Open data portal of Cologne to retrieve data that we have converted in shapefiles.

[4] http://geoportal.saarland.de/arcgis/services/Internet/Gesundheit/MapServer/WFSServer Web service allowing for retrieving data from Saarland that we have converted in shapefiles.

Table 2 Number of columns addressing each cluster for each data set

Data set name	Geospatial attributes	Concept attributes	Other attributes
Hospital in Cologne	9	3	0
Hospital in Saarland	9	32	2
School in Cologne	10	4	1
School in Saarland	9	8	6
Fire brigade	14	5	2

(e.g. name, email, phone number, or number of places for each service of a hospital, like in the data set from Saarland), and other attributes (OA) (e.g. stakeholder). Sometimes, these column names are an abbreviation, a complete name, a composed name separated by an underscore or an abbreviation of two words, still representing the same semantic meaning.

5.1.2 Experiments

To assess the relevance and the efficiency of our automatic approach, we applied two other approaches to the data set.

Experiment 1: Manual approach by non-expert This manual approach consists of creating an ontology to represent each file of the data set. There are several ways to create an ontology and to understand the meaning of a file when you are a non-expert of the processed data.[5] To compare our automated approach, we asked two different persons to create their own ontologies to represent each data content of our use case. One of those persons participated in the development of the automated approach, and thus followed the same process of ontology construction. The result provided by this person is called manual ontology 2 (MO2). As the process of ontology creation is similar, we consider it the gold standard for our experiment when it comes to the expected outcome for our automated approach (AO). The second person was free to apply their own method of ontology construction. The result of this second person corresponds to the manual ontology 1 (MO1) and provides another point of view for ontological representation. During the experiment, we evaluated the similarity between these two manual approaches, and then the similarity of each with our automatic approach. The comparison between a manual approach and our approach allows the comparison of a human method with a computational method.

Experiment 2: Other automatic approach with LogMap [27] LogMap allows to match two ontologies according to two automatic approaches: one using string matching and one using a matching repair algorithm. A part of our automatic process is matching our local ontology with the Semantic Web (DBpedia 2016 [3] and Wikidata [65]). In order to compare our automatic approach with LogMap, we created a very simple local ontology with the name of the column as Concept and applied the

[5] A non-expert is someone who knows about Semantic Web technologies but does not know the context and the goal of the data set.

different matchings of LogMap to this simple local ontology with DBpedia and, as a comparison, with Wikidata.

Content of the ontologies used for the experiment An overview of the content for each of the ontologies belonging to our experiments makes it possible to show the degree of interpretation for each technique. The number of concepts (Table 3) and the object properties (Table 4) represent the number of interpreted elements, whereas the number of data properties (Table 4) represents the amount of non-interpreted information among the column names. Table 3 presents the concepts of each ontology according to the following classification: geospatial attributes (GA), concept attributes (CA), other attributes (OA), concepts from enrichment step (E), and the total number of concepts (T).

5.2 Evaluation

5.2.1 Methodology

To evaluate our results, we compared them with the results of manual ontologies, whose respective methodology of construction and expectations according to the automated methodology are given in Sect. 5.1.2. This evaluation is based on the following point-based scoring system: 1 point is assigned for each match between two same concepts (or properties), meaning same URI or *equivalent class* and 0.5 points are assigned for each match between two similar concepts (or properties). A similar concept can be estimated according to a hierarchical link as *Subclass of*, due to a too specific concept (e.g. a column with city district names classified as "City District of Cologne" instead of "City District"). It can also be estimated as two different concepts which could have a similar meaning (e.g. dbpedia:Usage, which could have a general meaning of usage, but which represents a usage in a specific domain ($dbc : Applied\_Linguistics$) and dbpedia:Use, which has a more general meaning). A percentage is then computed according to the maximum possible similarity between the two sets of comparison (*set(a)* and *set(b)*). This computation corresponds to the following equation:

$$\frac{\sum_{i=0}^{n} sim_i}{Min(set(a), set(b))} \tag{1}$$

where $0 <= sim_i <= 1$ corresponds to a point of similarity between two concepts, and where $|set(a)| <= n$ and $|set(b)| <= n$ represent the total number of concepts for each set respectively.

5.2.2 Results

Our previously presented evaluation [49] was based on the comparison between identified concepts, using the string similarity measure of Resnik [53]. This previous evaluation focused on the concept's identification, but did not provide an evaluation of the ontology structure. That is why this evaluation focuses on both the ontological structure, by considering concepts and the different types of properties, and the

Table 3 Number of concepts by cluster for each ontology

Data set name	Automatic					Manual 1					Manual 2					LogMap
	GA	CA	OA	E	T	GA	CA	OA	E	T	GA	CA	OA	E	T	
H.C. DBpedia	2	3	0	7	12	3	0	0	0	3	4	1	0	4	9	1
H.S. DBpedia	3	3	0	6	12	2	2	0	0	4	3	4	0	4	11	2
S.C. DBpedia	1	5	0	7	13	4	1	0	0	5	3	7	0	3	13	2
S.S. DBpedia	2	2	0	6	10	2	3	0	0	5	5	3	0	4	12	1
F. DBpedia	4	1	0	9	14	4	3	0	0	7	5	4	1	3	13	1
H.C. Wikidata	4	1	0	10	15	5	0	0	0	5	8	3	0	4	15	0
H.S. Wikidata	7	1	0	12	20	3	1	1	0	5	5	5	1	4	15	0
S.C. Wikidata	4	5	0	11	20	5	1	0	0	6	8	7	0	3	18	0
S.S. Wikidata	0	4	0	13	17	3	0	0	0	3	7	4	0	4	15	0
F. Wikidata	3	4	0	11	18	2	7	0	0	9	7	4	1	3	15	0

Table 4 Number of properties for each ontology

Data set name	Automatic		Manual 1		Manual 2	
	Object properties	Data properties	Object properties	Data properties	Object properties	Data properties
H.C. DBpedia	8	9	3	7	4	8
H.S. DBpedia	10	40	3	14	9	16
S.C. DBpedia	9	11	5	8	5	9
S.S. DBpedia	9	16	4	6	8	14
F. DBpedia	13	22	10	13	10	7
H.C. Wikidata	10	11	5	6	12	2
H.S. Wikidata	10	41	5	29	11	13
S.C. Wikidata	11	8	6	8	11	5
S.S. Wikidata	16	20	3	13	10	11
F. Wikidata	17	21	9	11	12	5

Table 5 Concept similarity between ontologies

Data set name	Similar concepts			
	MO1 & MO2	MO1 & AO	MO2 & AO	LogMap & AO
H.C. DBpedia	2 (66.67%)	1.5 (50%)	7 (77.78%)	0 (0%)
H.S. DBpedia	2 (50%)	2.5 (62.50%)	7 (63.64%)	1 (50%)
S.C. DBpedia	4 (80%)	2 (40%)	8 (61.54%)	1 (50%)
S.S. DBpedia	2 (40%)	2 (40%)	6 (50%)	0.5 (50%)
F. DBpedia	5 (71.43%)	2 (28.57%)	6 (46.15%)	1 (100%)
Average similarity	61.62%	44.21%	59.82%	50%
H.C. Wikidata	5 (100%)	1 (20%)	5.5 (36.67%)	0 (0%)
H.S. Wikidata	3 (60%)	3 (60%)	8.5 (56.67%)	0 (0%)
S.C. Wikidata	5 (83.33%)	3 (50%)	10.5 (58.33%)	0 (0%)
S.S. Wikidata	2 (66.67%)	3 (100%)	7.5 (50%)	0 (0%)
F. Wikidata	8 (88.89%)	4 (44.44%)	6 (40%)	0 (0%)
Average similarity	79.78%	54.89%	48.34%	0%
Total average	70.7%	49.55%	54.08%	25%

"well-concepts" identification, by comparing the result of our automated approach with others. Our evaluation results are presented in two tables: one for the concept similarity in Table 5 and one for the property similarity in Table 6. The data sets used for our experiments are about Hospital (H), School (S) and Fire brigade (F). The data set on the fire brigade corresponds to the north of Germany, whereas the data sets for hospitals and schools correspond to the city of Cologne (C) and the German *Bundesland* Saarland (S). These tables present the comparison between our automatic generated ontology (AO), the first manual ontology (MO1), the second manual ontology (MO2) and LogMap (the latter only for concepts' similarity). The process has been applied twice for each data set: once to link concepts to DBpedia and once with Wikidata.

5.2.3 Observation and interpretation

Comparison with LogMap LogMap is based on a repair matching in order to link two ontologies. Table 3 presents the number of concepts successfully matched with concepts from DBpedia and Wikidata. The result of LogMap with DBpedia has obtained a low matching, only few concepts have been detected with the data set. This low level of matching illustrates the difficulty of identifying concepts from the value of data sets. Our automatic approach identified more concepts than LogMap. Although LogMap was not specified in the interpretation of the meaning, but rather in the matching, we use it to compare the part of our approach which allows matching of a column name with a concept. Thus we can say that the step of feature value analysis obtains a better result, which is due to its combination of several steps of matching based on its natural language processing. LogMap found no match in the matching with Wikidata. In comparison with DBpedia, Wikidata has the specific characteristic of having a URI

Table 6 Property similarity between ontologies

Data set name	Similar object properties			Similar data properties		
	MO1 & MO2	MO1 & AO	MO2 & AO	MO1 & MO2	MO1 & AO	MO2 & AO
H.C. DBpedia	1.5 (50%)	1.5 (50%)	3 (75%)	6 (85.71%)	5 (71.43%)	6 (75%)
H.S. DBpedia	1 (33.33%)	1 (33.33%)	2 (22.22%)	7 (50%)	13 (92.86%)	11 (68.75%)
S.C. DBpedia	3 (60%)	2 (40%)	2 (40%)	6 (75%)	6 (75%)	7 (77.78%)
S.S. DBpedia	1 (25%)	1 (25%)	4 (50%)	3 (50%)	3 (50%)	10 (71.43%)
F. DBpedia	4 (40%)	3 (30%)	3 (30%)	4 (57.14%)	9 (69.23%)	2 (28.57%)
Average similarity	41.67%	35.67%	43.44%	63.57%	71.70%	64.31%
H.C. Wikidata	5 (100%)	1 (20%)	2 (20%)	2 (100%)	4 (66.67%)	2 (100%)
H.S. Wikidata	2 (40%)	3 (60%)	4 (40%)	10 (76.92%)	26 (89.66%)	10 (76.92%)
S.C. Wikidata	5 (83.33%)	3 (50%)	4 (36.36%)	4 (80%)	5 (62.5%)	4 (80%)
S.S. Wikidata	2 (66.67%)	3 (100%)	6 (60%)	9 (81.82%)	12 (92.31%)	10 (90.91%)
F. Wikidata	6 (66.67%)	3 (33.33%)	4 (33.33%)	3 (60%)	8 (72.73%)	2 (40%)
Average similarity	71.33%	52.67%	37.94%	79.75%	76.77%	77.57%
Total average	56.15%	44.17%	40.69%	67.75%	74.24%	70.94%

with an identifier that is not a string similar to a label. We assume this can imply some difficulties for LogMap when assessing string similarity on URIs. Moreover, LogMap uses also the hierarchy of the ontology, but there is no hierarchical information with this type of data, which is another problem with a tool for ontology matching. Our advantage is that our approach considers and combines several types of information to identify a concept as, for example, a label or a comment. In the case of Wikidata, taking the label of the concept into account during the analysis allows for improving the result. Table 5 shows the similarity between the concepts identified by Logmap and concepts identified by our automatically generated ontology. The four concepts identified by LogMap are *dbo:District*, *dbo:Place*, *dbo:Hospital*, and *dbo:School*.

Comparison with manual ontologies Thanks to the comparison between the content of two manual ontologies, we can observe that two different persons can create two very different ontologies from the same data set, since their number of concepts and properties are very different. However, a certain level of a common base exists between the manual approaches, with a similarity average of 64.87%. Generally, we can observe that the first manual ontology has fewer concepts than the second manual ontology and that the number of concepts of the automatically generated ontology without considering the enrichment is between the two manual ontologies. According to the criteria of quantity, that means the automatically generated ontology has a good result in interpreting the heterogeneous file. By comparing the similarity between the manual approaches and the automatic approach, we observe a common base of 55.62%. This similarity score between manual approaches and an automatic approach is 10% less than the similarity between two manual approaches. The similarity of concepts between the automatic approach and the manual approach defined as a gold standard is a bit higher than with the other manual approach. However, we can observe this similarity between concepts at around 50% regardless of the number of concepts that have been determined in a manual approach. A quality aspect can be seen in this similarity comparison. However, the quality cannot be assessed in a global view due to the difference observed when linking with DBpedia and Wikidata. In general, linking with DBpedia obtained fewer concepts than with Wikidata. That means Wikidata allows the obtaining of a better quantity of concepts than DBpedia. DBpedia is very rich in individuals, but has a very flat class hierarchy compared to Wikidata when it comes to our data set. The diversity of concepts in Wikidata also implies, a diversity of concept choice for the file interpretation.

5.3 Conclusion

The comparison of the concept similarity between the two manual approaches has shown that a semantic structure can have an impact of around 40% on the similarity between two ontologies resulting from the interpretation of a file. The similarity comparison between the automatic approach and manual approaches has shown a difference of 10% more compared to the two manual approaches. This difference of 10% provides an estimation of matching error or unadapted matching. Moreover, thanks to the classification of identified concepts between different clusters, we observe that the automatic approach is, to an extent, less efficient in interpreting geospatial attributes

from the feature descriptor and value analysis steps than a manual annotator (which could be expected). However, the geocoding enrichment step compensates for this lack of geospatial interpretation by adding complementary information, which increased the similarity with the manual approaches.

6 Conclusion

We presented a fully automated approach for interpreting geospatial data and its features by creating an ontology linked to the Semantic Web to represent its semantic interpretation. This approach is realized through the combination of different schema and ontology matching techniques. The implicit schema used to create the ontology adapts itself according to the result of the elements matching with ontologies from the Semantic Web. By focusing on the geometry as a new central point for concept detection, we can build a local ontology structured around this main concept which allows for retrieving complementary geospatial information. The benefit of this approach is to allow an automatic semantic interpretation of geospatial data without background knowledge of this data. This benefit offers the possibility to gather geospatial information through a common representation from web feature services of associated open data portals, which often lack background knowledge. In addition, it allows the retrieval and combination of information from different sources thanks to concept and geometry matching, providing a great source of information. Gathering and linking data allows the provision of a set of information which can be compared to determine its quality and identify the most adapted information to the addressed use case. In the context of disaster management, having the right information plays a major role in the efficiency of response, but the quantity and quality of the information matters. That is why automatic and semantic integration of different data with its provenance is suitable for assessing its quality and using the information adapted most to a situation. Although our approach allows for interpreting the content of data without background knowledge, some features corresponding to columns still fail to be interpreted properly. Cryptic names and values that are used to represent features are particular challenging both for our approach and for humans. In terms of extension of the algorithm, we shall highlight some ideas in our future work.

Future work In our future work we face two major challenges. We firstly would like to extend our integration approach to support relational databases. Applying our approach per database table in a relational database should in our opinion be a reliable enough basis for interlinking data sets if all database tables can be reliably imported. We would like to investigate this idea using some examples of geospatial databases in a disaster management context. Secondly, we are working on downlift methods to enrich geospatial data sets that we have imported using our integration approach. Especially in the geospatial community, RDF and ontologies are not the most common data formats for processing geospatial data. Instead, downlifting APIs or mechanisms to support common formats like GeoJSON, SHP or GML are needed for better acceptance of semantic integration in the geospatial community. For NSDS data, this requires the extension of the provenance information to include mappings and source annotations.

For standardized geospatial data, local ontologies can provide the basis for downlifting by preparing data structures for provenance.

Acknowledgements We are funded by the German Federal Ministry of Education and Research (https://www.bmbf.de/en/index.html Project Reference: 03FH032IX4).

References

1. Alt H, Godau M (1995) Computing the Fréchet distance between two polygonal curves. Int J Comput Geom Appl 5(01n02):75–91
2. Arenas M, Bertails A, Prud'hommeaux E, Sequeda J (2012) A direct mapping of relational data to RDF. W3C recommendation. https://www.w3.org/TR/rdb-direct-mapping/
3. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) Dbpedia: a nucleus for a web of open data. In: The semantic web, Springer, pp 722–735
4. Auer S, Lehmann J, Hellmann S (2009) Linkedgeodata: adding a spatial dimension to the web of data. In: International semantic web conference, Springer, pp 731–746
5. Barron C, Neis P, Zipf A (2014) A comprehensive framework for intrinsic openstreetmap quality analysis. Trans GIS 18(6):877–895
6. Battle R, Kolas D (2011) Geosparql: enabling a geospatial semantic web. Semant Web J 3(4):355–370
7. Berretti S, Del Bimbo A, Pala P (2000) Retrieval by shape similarity with perceptual distance and effective indexing. IEEE Trans Multimed 2(4):225–239
8. Bizid I, Faiz S, Boursier Patriceand Yusuf JCM (2014) Integration of heterogeneous spatial databases for disaster management. In: Parsons J, Chiu D (eds) Advances in conceptual modeling: ER 2013 workshops, LSAWM, MoBiD, RIGiM, SeCoGIS, WISM, DaSeM, SCME, and PhD symposium, Hong Kong, China, November, 2013, revised selected papers. Springer, Cham, pp 77–86. https://doi.org/10.1007/978-3-319-14139-8_10
9. Brassel K, Bucher F, Stephan EM, Vckovski A (1995) Completeness. In: Guptill SC, Morrison JL (eds) Elements of spatial data quality. Elsevier, Amsterdam, pp 81–108
10. Burggraf DS (2006) Geography markup language. Data Sci J 5:178–204
11. Buscaldi D, Rosso P (2008) Geo-wordnet: automatic georeferencing of wordnet. In: LREC
12. Das S, Sundara S, Cyganiak R (2012) R2RML: RDB to RDF mapping language, W3C recommendation. World Wide Web Consortium, Cambridge
13. Debruyne C, McGlinn K, McNerney L, O'Sullivan D (2017) A lightweight approach to explore, enrich and use data with a geospatial dimension with semantic web technologies. In: Proceedings of the fourth international ACM workshop on managing and mining enriched geo-spatial data, ACM, p 1
14. Debruyne C, Meehan A, Clinton É, McNerney L, Nautiyal A, Lavin P, O'Sullivan D (2017) Ireland's authoritative geospatial linked data. In: International semantic web conference, Springer, pp 66–74
15. Do HH, Rahm E (2002) Coma: a system for flexible combination of schema matching approaches. In: Proceedings of the 28th international conference on very large data bases, VLDB endowment, pp 610–621
16. Eren H (2016) 8 standards in process control and automation. In: Liptak BG, Eren H (eds) Instrument engineers' handbook, volume 3: process software and digital networks, vol 3. CRC Press, Boca Raton, p 155
17. ESRI E (1998) Shapefile technical description. An ESRI white paper
18. Euzenat J, Shvaiko P (2007) Ontology matching. Springer, Berlin
19. Gao S, Sperberg-McQueen CM, Thompson HS, Mendelsohn N, Beech D, Maloney M (2009) W3C XML schema definition language (XSD) 1.1 part 1: structures. W3C Candidate Recomm 30(7.2):16
20. Goodchild MF, Hunter GJ (1997) A simple positional accuracy measure for linear features. Int J Geogr Inf Sci 11(3):299–306
21. Grantner E (2007) ISO 8000: a standard for data quality. Logist Spectr 41(4):4–6
22. Guo H, Song GF, Ma L, Wang SH (2009) Design and implementation of address geocoding system. Comput Eng 35(1):250–251
23. Hartig O, Zhao J (2009) Using web data provenance for quality assessment. CEUR workshop proceedings

24. Hillner S, Ngomo ACN (2011) Parallelizing limes for large-scale link discovery. In: 7th international conference on semantic systems, ACM, pp 9–16
25. Homburg T, Prudhomme C, Würriehausen F, Karmacharya A, Boochs F, Roxin A, Cruz C (2016) Interpreting heterogeneous geospatial data using semantic web technologies. In: International conference on computational science and its applications, Springer, pp 240–255
26. Huttenlocher DP, Klanderman GA, Rucklidge WJ (1993) Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 15(9):850–863
27. Jiménez-Ruiz E, Grau BC (2011) Logmap: logic-based and scalable ontology matching. In: International semantic web conference, Springer, pp 273–288
28. 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, Springer, pp 113–132
29. Kainz W (1995) Logical consistency. Elem Spat Data Qual 202:109–137
30. Kalemi E, Martiri E (2011) FOAF-academic ontology: a vocabulary for the academic community. In: 2011 third international conference on intelligent networking and collaborative systems (INCoS), IEEE, pp 440–445
31. Lanter DP (1990) Lineage in GIS: the problem and a solution, NCGIA National Center for Geographic Information and Analysis. http://infoscience.epfl.ch/record/51713
32. Le Grange JJ, Lehmann J, Athanasiou S, Garcia-Rojas A, Giannopoulos G, Hladky D, Isele R, Ngomo ACN, Sherif MA, Stadler C, et al (2014) The geoknow generator: managing geospatial data in the linked data web. In: Linking geospatial data
33. Lebo T, Sahoo S, McGuinness D, Belhajjame K, Cheney J, Corsar D, Garijo D, Soiland-Reyes S, Zednik S, Zhao J (2013) PROV-O: the PROV ontology. W3C recommendation. https://www.w3.org/TR/prov-o/
34. Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions, and reversals. Soviet Phys Dokl 10:707–710
35. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: 52nd annual meeting of the association for computational linguistics: system demonstrations, pp 55–60
36. Melnik S, Garcia-Molina H, Rahm E (2002) Similarity flooding: a versatile graph matching algorithm and its application to schema matching. In: 18th international conference on data engineering, 2002. Proceedings, IEEE, pp 117–128
37. Miller GA (1995) Wordnet: a lexical database for english. Commun ACM 38(11):39–41
38. Navigli R, Ponzetto SP (2010) BabelNet: building a very large multilingual semantic network. In: Proceedings of the 48th annual meeting of the association for computational linguistics, Association for computational linguistics, pp 216–225
39. Nentwig M, Hartung M, Ngonga Ngomo AC, Rahm E (2017) A survey of current link discovery frameworks. Semant Web 8(3):419–436
40. Ngomo ACN, Auer S (2011) Limes-a time-efficient approach for large-scale link discovery on the web of data. In: IJCAI, pp 2312–2317
41. Niu X, Rong S, Zhang Y, Wang H (2011) Zhishi.links results for OAEI 2011. In: Ontology matching, vol 220
42. Niwattanakul S, Singthongchai J, Naenudorn E, Wanapu S (2013) Using of Jaccard coefficient for keywords similarity. In: Proceedings of the international multiconference of engineers and computer scientists, vol 1
43. OGC (2011) OGC geosparql—a geographic query language for RDF data. Technical report
44. Otero-Cerdeira L, Rodríguez-Martínez FJ, Gómez-Rodríguez A (2015) Ontology matching: a literature review. Expert Syst Appl 42(2):949–971
45. Pan JZ (2009) Resource description framework. In: Staab S, Studer R (eds) Handbook on ontologies. Springer, Berlin, pp 71–90
46. Patroumpas K, Alexakis M, Giannopoulos G, Athanasiou S (2014) Triplegeo: an ETL tool for transforming geospatial data into RDF triples. In: ICDT workshops, pp 275–278
47. Pinkel C, Binnig C, Jiménez-Ruiz E, Kharlamov E, May W, Nikolov A, Sasa Bastinos A, Skjæveland MG, Solimando A, Taheriyan M et al (2016) RODI: benchmarking relational-to-ontology mapping generation quality. Semant Web 9(1):25–52
48. Pinkel C, Binnig C, Jimenez-Ruiz E, Kharlamov E, Nikolov A, Schwarte A, Heupel C, Kraska T (2017) IncMap: a journey towards ontology-based data integration. In: Mitschang B, Nicklas D, Leymann

F, Schöning H, Herschel M, Teubner J, Härder T, Kopp O, Wieland M (eds) Datenbanksysteme für Business, Technologie und Web (BTW 2017). Gesellschaft für Informatik, Bonn
49. Prudhomme C, Homburg T, Ponciano JJ, Boochs F, Roxin A, Cruz C (2017) Automatic integration of spatial data into the semantic web. In: WebIST 2017
50. Prud E, Seaborne A, et al (2008) SPARQL query language for RDF. W3C Recommendation. https://www.w3.org/2001/sw/DataAccess/rq23/
51. Rahm E, Bernstein PA (2001) A survey of approaches to automatic schema matching. VLDB J 10(4):334–350
52. Repici J (2010) The comma separated value (CSV) file format. Creativyst Inc, San Carlos
53. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. arXiv:cmp-lg/9511007
54. Rijgersberg H, van Assem M, Top J (2013) Ontology of units of measure and related concepts. Semant Web 4(1):3–13
55. Scharffe F, Atemezing G, Troncy R, Gandon F, Villata S, Bucher B, Hamdi F, Bihanic L, Képéklian G, Cotton F, et al (2012) Enabling linked-data publication with the datalift platform. In: Proceedings of AAAI workshop on semantic cities
56. Schwering A (2008) Approaches to semantic similarity measurement for geo-spatial data: a survey. Trans GIS 12(1):5–29
57. Shvaiko P, Euzenat J (2013) Ontology matching: state of the art and future challenges. IEEE Trans Knowl Data Eng 25(1):158–176
58. Stadler C, Unbehauen J, Lehmann J, Auer S (2013) Connecting crowdsourced spatial information to the data web with sparqlify. Technical report, University of Leipzig
59. Svennerberg, G (2010) Beginning Google Maps API 3. Apress
60. Tarasowa D, Lange C, Auer S (2015) Measuring the quality of relational-to-RDF mappings. In: International conference on knowledge engineering and the semantic web, Springer, pp 210–224
61. van Rees E (2013) Open geospatial consortium (OGC). Geoinformatics 16(8):28
62. Veltkamp RC (2001) Shape matching: similarity measures and algorithms. In: SMI 2001 international conference on shape modeling and applications, IEEE, pp 188–197
63. Vertan C, Wozu O (2007) Web ontology language (OWL). W3C Recommendation. https://www.w3.org/TR/owl-features/
64. Volz J, Bizer C, Gaedke M, Kobilarov G (2009) Silk-a link discovery framework for the web of data. In: LDOW, vol 538
65. Vrandečić D, Krötzsch M (2014) Wikidata: a free collaborative knowledgebase. Commun ACM 57(10):78–85
66. Vretanos PA (2005) Web feature service implementation specification. Open Geospatial Consort Specif 1325:04–094
67. Wick M, Vatant B, Christophe B (2015) Geonames ontology. http://www.geonames.org/ontology/documentation.html
68. Zaveri A, Rula A, Maurino A, Pietrobon R, Lehmann J, Auer S (2016) Quality assessment for linked data: a survey. Semant Web 7(1):63–93