Towards a Unified Geospatial Data Reconciliation Platform for Data Analytics

Blerina Spahiu
University of Milano-Bicocca
Department of Informatics, Systems and Communication
Viale Sarca, 336
Milan, Italy
spahiu@disco.unimib.it

Vincenzo Cutrona
University of Milano-Bicocca
Department of Informatics, Systems and Communication
Viale Sarca, 336
Milan, Italy
cutrona@disco.unimib.it

Matteo Palmonari
University of Milano-Bicocca
Department of Informatics, Systems and Communication
Viale Sarca, 336
Milan, Italy
palmonari@disco.unimib.it

ABSTRACT

Many applications rely on the integration of geospatial data but very little effort has been made to consolidate the outputs, in particular to share the resulting mappings with the community and make them reusable. Mappings are extremely valuable assets as they can serve as guide to enrich data with other information residing in other datasets. Geographic data are characterized by different formats and different representation. In this paper we present an ongoing project addressing the need for a flexible and scalable platform of services handling the reconciliation of geospatial data. The platform offers utilities to load, create and maintain mappings between several most know geospatial datasets. On the one hand, this platform allows developers to easily create and manipulate all the mappings required for their application. On the other hand the sets of mappings are also exposed by services, to facilitate integration into applications working with heterogeneous data. Moreover in this paper we discuss several types of real world representation of geospatial data and present a literature review along with a discussion in favor of some interpretations with the aim of defining a shared vision about the problem of geospatial reconciliation that will support further advancements in development of the Semantic Web applications.

PVLDB Reference Format:
Blerina Spahiu, Vincenzo Cutrona, Matteo Palmonari. Towards a Unified Geospatial Data Reconciliation Platform for Data Analytics. *Self-published* under CC BY 4.0 Licence by INSID&S Lab - University of Milan-Bicocca, as a result of the EW-Shopp Project (`www.EW-Shopp.eu`) - GA 732590: 2018.

1. INTRODUCTION

The Semantic Web proposed by Tim Berners-Lee [1] in 2001 has the aim of creating a Web where data are semantically annotated which would enable agents to access the data more intelligently as well as perform tasks on behalf of users in order to discover the desired knowledge. The best practices for publishing and connecting such data on the Web are represented under the term of Linked Data [2]. The number of knowledge bases (datasets) published as Linked Data is constantly increasing with more than 1 184 datasets as of April 2018[1]. Geographical knowledge bases are among the largest in the LOD cloud and have a high impact in everyday applications. The United Nations Initiative on Global Geospatial Information Management (UN-GGIM) assessed 2.5 quintillion bytes of data being generated every day, and a large portion of the data is location-aware [12].

Geospatial data are crucial for many data analytics projects, because of the role that the spatial and temporal variables (and dimensions) play in a very large number of data analyses. For example, a Digital Advertising Company (referred to as DAC in the following), which is analyzing the performance of its campaigns using reports from the Google AdWords platform, needs to aggregate performance data (e.g., impressions) by city, region or country. As another example, an eCommerce Platform owner (referred to as ECP in the following), wants to aggregate sales data by user location (position, city, region, etc.). However, geospatial data play a second, not less important, role in data analytics: they provide one of the main dimensions used to integrate different datasets in such a way that analytics can be computed taking into consideration more variables. In many business analytics scenarios, data integration processes are polarized around a main dataset that contains the principal business variables (named *business data* in the following), which needs to be *enriched* by integrating additional information that is available in a different dataset. For example, the DAC mentioned here above, wants to enrich the Google AdWords data with information about population or weather information associated with cities and regions that occur in their business data. The ECP wants to enrich his business data with sales data acquired by third parties, with weather information, or with data of nearby events that may have had impact on customers' behavior. These kind of analyses are targeted by several business cases where event and weather-based analytics are used to develop data-driven innovative services, like the ones addressed in the EW-Shopp[2] project. In these analysis, data need to be enriched by joining business data and external data sources us-

Peer-reviewed paper accepted and presented at the 2nd Workshop on Mobility Analytics for Spatio-temporal and Social Data, MATES 08, co-located with VLDB 2018, Rio de Janerio, Brasil, 2018. Proceedings of the workshop have not been published by an official publisher.
Paper by INSID&S Lab - University of Milan-Bicocca

DOI:

[1] `http://lod-cloud.net/`
[2] `www.ew-shopp.eu`

ing also spatial data values. Observe that the LOD cloud offers a large amount of data valuable for data enrichment, but these data are still underexploited because the enrichment process requires the reconciliation of heterogenous geospatial references across different datasets. As a result, the enrichment processes described above are still very difficult and time consuming, making it difficult to scale the number of analyses that companies can undertake, and requiring cross-domain expertise that is not easy to acquire.

The Semantic Web community has developed many vocabularies and standards for representing geospatial data on the Web. Despite the effort from the community the semantic reconciliation of geospatial data is still a challenge to build applications that can reconcile among different representation of geospatial data. The difficulty is caused by three aspects; (i) Syntactic heterogeneity (different formats, e.g. RDF, relational, etc.); (ii) Schematic heterogeneity (different generalization hierarchies for the representation of the same real world entity); (iii) Semantic heterogeneity (Disagreement on the meaning and the interpretation of such data). There are many tools available for handling these general challenges with Semantic Web technologies [3, 9]. Some of these approaches addressed the challenge of mapping vocabularies of linked data [7], some of which with a special focus on geospatial information [14]. Several application also used semantic technologies to support the extraction and/or integration of geospatial information [6, 5, 16]. However, to the best of our knowledge there exist no platform that is able to support a seamless reconciliation of geospatial data across datasets that use different identifiers as reference for geospatial entities.

We believe that a significant boost to support data enrichment for data analytics can be achieved if we build a platform such that: it identifies a small number of datasets, which are particularly remarkable in terms of geospatial coverage and content richness, as reference gespatial data sources (e.g., Geonames[3], NUTS, etc.); it bridges the gap between the identifiers used in these different sources (e.g., providing links between Geonames and NUTS); it provides a flexible and effective way to navigate across these interconnected reference geospatial sources (e.g., hosting a service that for an input Geonames identifier returns its equivalent identifier in NUTS, Wikidata, etc.); it provides entity linking functionalities that help finding links from toponyms mentioned in plain text to these reference identifiers (e.g., from "Milan, Italy" to the identifier of Milan in Geonames).

In this paper, we present an ongoing project aimed at building such a platform in a pay-as-you-go fashion, starting from data enrichment use cases that come from the EW-Shopp project. In particular, we first introduce event and weather-based analytics use cases (drawn from the EW-Shopp project) and one use case that clarifies how the data enrichment process is used in a data analyses task (Section 2). Then we introduce the terminology needed to clarify interoperability between different systems of geospatial identifiers (Section 3). We then analyze the state-of-the-art along two dimensions (Section 4): standardization initiatives that address vocabularies and system of identifiers for geospatial data, and projects addressing interoperability across these vocabularies and systems of identifiers. Then, in Section 5, we explain the principles behind our proposal, which include: the selection of reference geospatial data sources, the reconciliation processes that the platform should support, and the architecture of the platform, which should support a variety of reconciliation services. In particular, concerning the selection of reference geospatial data sources, we found that Geonames is the best candidate to work as main reference sources. Other remarkable datasets that need to be interlinked to Geonames are Wikidata, NUTS, DBpedia (for which many links already exists) but also Google's codes, which are used in AdWords and Google Analytics statistics, which serve millions of users worldwide. The architecture of the platform should store links computed between systems of identifiers, services that use these links, as well as reconciliation services for strings input. Finally, the architecture is aimed to be modular, meaning that it should allow to plug new data sources and interlinking services when needed.

We can therefore summarize the main contribution of this paper as follows: (i) we make a detailed analysis of standardization initiatives in geospatial data representation (with particular emphasis on systems of identifiers); (ii) we describe four kinds of reconciliation functionalities that the platform should support; (iii) we describe an architecture to support these kinds of functionalities.

2. DATA ENRICHMENT FOR EVENT AND WEATHER-BASED ANALYTICS

In the following we describe one use case which demonstrates the usefulness of such platform: **Data Enrichment for Digital Marketing Campaigns**.

The business data in Figure 1 describes the performance of digital ads (html links) that are pushed into users' browsers when they submit certain keywords. The data therefore contain: the id of the keyword, the number of clicks collected for the ad for a specific city, and the region of that city. The data can be enriched with two kinds of information:

- **Weather forecasts** (solar radiation, precipitation, temperature, etc.), which in EW-Shopp are provided by the European Centre for Medium-Range Weather Forecasts and are accessible via an API. The API can be queried by specifying a date and a pair of latitude and longitude coordinates. Thus, to collect weather information, we need to associate each city that appears in business data with its coordinates. This information is available in Geonames knowledge base (KB), but we need to reconcile city and region names with Geonames before collecting this information.

- **Demographic information** (population, area and derived density, etc.), which can help determine the potential target of each city. Population is available in Geonames, while city area can be found only in Wikidata. To collect this additional information we first need to reconcile toponyms occurring in the table to Geonames identifiers, then use links (if available) between Geonames and Wikidata to collect the information about the city area.

Other business case addressed in EW-Shopp include event and weather-based analyses of: visitors of locations (e.g., shops); interactions on eCommerce platforms (e.g., numbers of clicks for specific products); sales of specific products; interactions in Customer Relationship Management

[3] http://www.geonames.org/

Keyword	Clicks	City	Region	ID (Geonames	Latitude (Geonames)	Longitude (Geonames)	Population (Geonames)	Area (Wikidata)	Date
194906	64	Altenburg	Thuringia	2822542	50.98763	12.43684	38568	45.6 km^2	11/03/2018
517827	50	Ingolstadt	Bavaria	2951839	48.76508	11.42372	120658	133.35 km^2	12/03/2018
459143	42	Berlin	Berlin	2950157	52.52437	13.41053	3426354	891.68 km^2	12/03/2018
891139	36	Munich	Bavaria	2951839	48.13743	11.57549	1260391	310.71 km^2	11/03/2018
459143	30	Nuremberg	Bavaria	2951839	49.45421	11.07752	499237	186.45 km^2	10/03/2018

Figure 1: Business data that needs to be enriched.

platforms (e.g., inbound and outbound calls from/to certain locations). Examples of first insights gained during the project can be found on the project data blog[4].

The enrichment process will be carried out using a tabular data transformation application as a service, which also supports semantic data management (transformation to RDF and graph-based data hosting), that extends the DataGraft and Grafterizer tools [18]. The data enrichment process is performed by users following an approach similar to the one supported in OpenRefine[5], a popular data management tool. In fact, we provide OpenRefine-compliant reconciliation services that can be seamlessly used in both OpenRefine and Grafterizer, with the aim of supporting more scalable and efficient data processing when using the second tool.

3. PRELIMINARIES

In order to support interoperability within the platform we define data formats, vocabularies and systems of identifiers that are the basis for the reconciliation process.

Definition 1. **Data Format.** A data format specifies how to encode data for storage in a computer file.

Some data format specifies complex data structures that are associated with file extensions. The data format can consist in a specific syntax for a full-fledged formal language, for example, XML/RDF is a data format for data represented using the RDF data model, with the XML/RDF syntax. Other formats specify the data structure to represent specific kind of information in different more complex data formats. For example, complete date plus hours and minutes (YYYY-MM-DDThh:mmTZD, e.g. 1997-07-16T19:20+01:00)[6] is a data structure to represent information about time points that can be used in CSV as well as XML files.

Definition 2. **Vocabularies (and ontologies)** A vocabulary defines any specification of the terminology to be used to represent information in a domain of interest.

Most of vocabularies are defined within specific formal languages used by computer programs, such as XML or RDF. In particular, vocabularies usually specify classes of objects as well as properties used to describe and interrelate these objects. The degree of specification of the meaning of the terms defined in a vocabulary may diverge from vocabulary to vocabulary. In this paper we consider vocabularies proposed to support interoperability on the web, which are usually defined for the XML and RDF languages. Vocabularies are also frequently referred to as ontologies, where the latter term is used in particular when the meaning of the terminology is specified by means of logical languages. In this paper we will use the term ontology and vocabulary interchangeably without committing to a specific degree of specification with logical axioms.

Definition 3. **Systems of identifiers. (SIs)** A systems of identifiers is a system that specifies identifiers for objects in any model.

Examples of shared systems of identifiers are: location identifiers provided by the Geonames dataset or SKU (Stock Keeping Unit) provided as identifiers for products. We consider shared systems of identifiers also syntax that support the intentional specification of objects in infinite sets; for example, numbers, pairs of longitude and latitude, which identify specific points in a coordinate system.

It is important to distinguish between formats and SIs, vocabularies and SIs, and the adjectives standard and shared referred to vocabularies and systems of identifiers.

- **Formats vs. SIs:** Sometimes data formats, e.g., format for complete date, hours and minutes specified in ISO 8601, define also systems of identifiers; in this paper we consider time formats as systems of identifiers.
- **Vocabularies vs. SIs:** The distinction between vocabularies and systems of identifiers is not sharp. It may happen that one authoritative data source, e.g., a KG based on Linked Data specification like DBpedia, presents terms to be used to refer to classes, data types and properties, as well as to entities. In addition, systems of classifications, e.g., product category taxonomies, define classes of objects, which are usually specified by means of identifiers and associated with objects at the instance level.
- **Standard vs. shared:** Authoritativeness of vocabularies and SIs may depend on two different processes:
  - Standardization initiatives, which bring a community of stakeholders to agree upon the specification of vocabularies or SIs. These initiatives may be driven by: (1) International or national organizations dedicated to the specification of standards, e.g., the International Organization for Standardization (ISO), GS1 (a not-for-profit

[4] `http://www.ew-shopp.eu/data-blog/`
[5] `http://openrefine.org/`
[6] `https://www.w3.org/TR/NOTE-datetime`

organization that develops and maintains global standards for business communication), local or federal governments; (2) Working groups defining best practices, e.g., W3C working groups.

– Adoption by a large community of stakeholders, which may lead organization resources to be de facto used by a variety of applications, e.g., Wikipedia, which provides descriptions of real world entities used by many data processing tools.

In this paper the expressions Standard or Shared Vocabularies (SSVs) and Standard or Shared Systems of identifiers (SSSI) are used to refer in general to vocabularies/systems of identifiers that are frequently used because of standardization initiatives or that gained authoritativeness because of the use by a large community of stakeholders.

4. RELATED WORK

In this section we make an in depth analysis for the standardization of vocabularies or system of identifiers used for geospatial data and discuss in favor of some interpretations in order to define a shared vision of geospatial data reconciliation. Furthermore we compare our work to approaches explicitly proposed to reconcile geospatial information.

4.1 State-of-the-art on the representation of spatial data

Several efforts have been dedicated to standardize representations of spatial data.

WGS84[7] is a coordinate system based on Decimal Minutes format to identify points in the coordinate space (SSSI). ISO 3166-1 is an ISO standard for codes to identify countries, dependent territories, and special areas of geographical interest which have also been used in statistics. . Remarkably, Federal Information Processing Standards (FIPS) of US government has defined several standards to identify geographical entities, used, in particular, in statistics (e.g., in the US Census). In Europe, the **Classification of Territorial Units for Statistics (NUTS)**[8] and **Local Administrative Units (LAU)** are recommended as standard identifiers of territorial units by **INSPIRE**[9]. NUTS are organized along three levels of aggregation, while LAU represent districts and municipalities. INSPIRE is a directive that aims to create a European Union spatial data infrastructure to enable the sharing of environmental spatial information among public sector organizations, facilitate public access to spatial information and assist in policy-making across boundaries.

ISO 19107:2003[10] can be considered a vocabulary that specifies conceptual schemas for describing spatial characteristics of geographic features, and a set of spatial operations. It treats vector geometry and topology up to three dimensions. It defines spatial operations for use in access, query, management, processing, and data exchange of geographic information for spatial (geometric and topological) objects of up to three topological dimensions.

WOEID[11] is a SSSI by Yahoo! that identifies spatial entities provided by Yahoo! GeoPlanet. Google use 98227 spatial identifiers named **Geotargets**[12] that are used in AdWords (Google advertising platform). These identifiers cover cities, provinces, but also airports and other points of interest. AdWords is adopted by a large community of developers and the spatial reference include; criteria ID, Google unique identifiers for location; name, the best available name in English; canonical name, a more specific name that includes parent territories (e.g., province and countries for cities); parent ID, the criteria ID of a parent; country code, the ISO-3166-1 alpha-2 country code; target type, the type associated with the location (e.g, City), status, the status of the location (e.g., active). These identifiers are mapped to other SSSIs described above only at the country level.

4.2 SSSVs and SSI to represent geographical information in RDF

Geographic data on the Web use different vocabularies or standards. Below we summarize some state-of-the-art vocabularies or SSSI that are used by these datasets.

Only 44 datasets in the LOD cloud have as primary topic geographic, while many other datasets contain information about geographical data but labeled with other topics, such as DBpedia. The geographic category contains datasets such as Geonames and LinkedGeoData[13] comprising information about geographic entities, geopolitical divisions, and points of interest. Concerning the geographic domain, the **W3C Geo** is the most widely used vocabulary (in 21 datasets), followed by the spatialrelations[14] ontology of Ordnance Survey (OS). At the same time, the analysis reveals that the property geo:geometry is used in 1 322 302 221 triples, exceeded only by the properties rdf:type (6 251 467 091 triples) and rdfs:label (1 586 115 316 triples).

Basic Geo (WGS84 lat/long) is a vocabulary that is used to represent latitude and longitude and other information about spatially-located things, using the WGS84 as datum[15]. The use of RDF as a carrier for latitude and longitude is motivated because of the capability of RDF for integrating data belonging to different domains. Basic Geo is used not only to describe maps, but also entities positioned on the map. The vocabulary defines a class called Points, whose members are points. These points can be described by properties of WGS84 and other properties from other vocabularies. Entities on the maps are called points, which are further described with latitude, longitude and altitude from WGS84 specification. For example we might use an externally defined property such as bornNear or withinFiveMilesFrom or using other properties for latitude, longitude and altitude in non-WGS84 systems. This vocabulary is used by more than 37% of the datasets in the LOD cloud [15]. Another geo vocabulary is the **NeoGeo** Vocabulary [16] which differently from the others makes a distinction between a Feature and a Geometry providing two classes spa-

[7] http://earth-info.nga.mil/GandG/publications/tr8350.2/tr8350_2.html
[8] http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units
[9] https://inspire.ec.europa.eu/
[10] https://www.iso.org/standard/26012.html
[11] https://developer.yahoo.com/geo/geoplanet/guide/concepts.html
[12] https://developers.google.com/adwords/api/docs/appendix/geotargeting
[13] http://linkedgeodata.org
[14] http://www.ordnancesurvey.co.uk/ontology/spatialrelations.owl
[15] https://www.w3.org/2003/01/geo/
[16] http://geovocab.org/doc/neogeo/

tial:Feature and geom:Geometry classes which have the relation geom:geometry among them. This vocabulary allows only WGS84 coordinates. The **SmOD Custom Vocabulary**[17] extends the Inspire Vocabulary[18] with other terms and properties regarding agroforestry management, such as land parcels, chemical characteristics of the soil and climate condition. It has 8 classes, 23 properties and 8 other namespaces. This is vocabulary is not a W3C standard.

The vocabulary designed by OGC **GeoSPARQL** standard does not reuse W3C Geo vocabulary but proposes another class Point instead. Geometries of geographical data represented in RDF with the GeoSPARQL vocabulary are represented by literals encoded consistently with other OGC standards. The OGC GeoSPARQL standard supports representing and querying geospatial data on the Semantic Web. GeoSPARQL defines a vocabulary for representing geospatial data in RDF, and it defines an extension to the SPARQL query language for processing geospatial data. In addition, it is designed to accommodate systems based on qualitative spatial reasoning and spatial computations[19].

Geo OWL[20] provides an ontology which closely matches the GeoRSS feature model and which utilizes the existing GeoRSS vocabulary for geographic properties and classes. The ontology provides a compatible extension of GeoRSS practice for use in more general RDF contexts.

schema.org and the DBpedia Ontology provide cross domain vocabularies that cover also several spatial concepts and properties. SKOS is also used as language to represent geographical features types in the Geonames KB.

There are also other very common and very used KBs that provide SSSIs such as Geonames and LinkedGeoData which are not standard. Geonames is the most used vocabulary between the two in the datasets of the LOD cloud. It is the main vocabulary of the geonames.org dataset and has 7 classes and 26 properties. LinkedGeoData ontology has been derived from concepts defined by Open Street Map. It is the main ontology of the linkedgeodata.org dataset.

4.3 Similar Approaches

The GeoKnow project focused on integrating geographical information on the Web and semantically processing the information such as efficient browsing and exploration [11]. GeoKnow Generator[21] integrates different tools for; the creation and maintenance of qualitative geospatial information from existing unstructured data; mapping and exposing existing structured geospatial information on the Web; efficient spatial indexing; automatic fusing and aggregation of geospatial data as well as it allows exploring, searching, and curating the spatial data by using machine learning techniques. In this project, different tools have been extended and optimized in order to process geospatial information but it lacks any support for automated workflows with multiple tools and data streams as input.

The GEISER project[22] aims at creating a platform for flexible integration of different geospatial and sensors data. The objective is to design and implement innovative functionality for developing services for transforming, storing, integrating and processing such data. Machine learning approaches will be applied for tasks such as computing topological relations between resources and time-efficient generation of link specifications. The resulting tools will be integrated as microservices in an open cloud-based platform.

GeoLink dataset has leveraged Linked Data principles to create a knowledge graph that allows users to query and reason over some of the largest geoscience data repositories in the United States [4]. The dataset contains more than 45 million RDF triples as well as a collection of ontologies and geo-visualization tools. The GeoLink knowledge graph comprises datasets such as R2R (environmental sensor data collected by the U.S. academic research fleet), BCO-DMO (data and information generated during oceanographic research efforts), IODP (collection of sediments, rocks and fluids from beneath the seafloor), MBLWHOI (marine life and its environment), SESAR (rocks specimens, water samples, and sediment cores), DATAone (earth and environmental sciences), AGU-NSF (annual conferences as well as NSF funded proposals related to geophysics), NGDB (geochemical content of samples from American stream sediments, soils, and waters) and USAP (ice coverage, snowfall totals, and glacial movements). Similarly our platform can be seen as a composition of the most known geospatial datasets but in difference we provide only reconciliation links among geospatial identifiers and do not provide a unique ontology.

5. SYSTEM ARCHITECTURE

In this section we describe the platform for geospatial reconciliation.

5.1 Datasets Selection

The central dataset for our platform is Geonames. We choose Geonames because; (1) it is the biggest dataset describing geo entities, (2) it is a multilingual dataset for entities labels, and (3) the administrative hierarchy for a given entity can be accessible via APIs. The Geonames RDF dump contains information for about 11, 7mio geo entities as of March 2018. The other dataset to be considered is DBpedia-2016-10 as it describes 6, 6mio entities, where 1, 9mio have geo coordinates and 840k are places. Moreover DBpedia being a cross domain dataset, contains also other descriptive data for geospatial entities, thus it is a very important dataset for reconciliation. Another important dataset is Wikidata which contains 3 496 401 triples (3 583 882 distinct instances) that have a link to Geonames. Among links to Geonames, Wikidata has 29 783 links to Yahoo! WOEID and 4 004 links to Facebook Places. There exists also 1 355 direct links to NUTS codes. While in DBpedia instances are linked to Geonames datasets with an owl:sameAs. Except of owl:sameAs links there are many links such as rdfs:seeAlso from DBpedia to local datasets (e.g. Embrune in France is linked through as owl:sameAs to the database of INSEE). The datasets considered in the platform are shown in Figure 2.

5.2 Reconciliation Process

Suppose the owner of a DAC is interested in enriching her data with data from different KBs for geomarketing scoping. She has in her dataset only the toponymy and the country. Without the reconciliation platform she has to query the

[17] https://www.w3.org/2015/03/inspire/smod
[18] https://www.w3.org/2015/03/inspire/
[19] http://www.opengis.net/doc/IS/geosparql/1.0
[20] https://www.w3.org/2005/Incubator/geo/XGR-geo-20071023/W3C_XGR_Geo_files/geo_2007.owl
[21] http://generator.geoknow.eu/
[22] https://www.projekt-geiser.de/en/welcome/

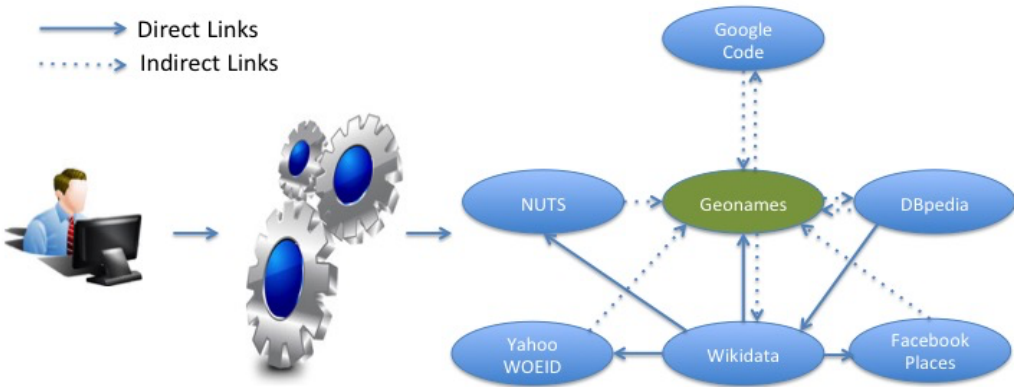

Figure 2: An overview of the reconciliation process.

reference datasets separately. At first she makes a query of the form Milan, Italy (city, country) in each of the most common geospatial dataset such as Geonames, DBpedia and Wikidata. In this situation, she needs to retrieve additional information and the geospatial ID for each dataset. In DBpedia she finds that there are about 45 properties describing the resource of Milan, Italy There are two sameAs links (links that declares two items to be identical) from DBpedia connecting this resource to the ones in Geonames. In this case she extracts two Geonames ID for the resource of Milan. Accessing such links she could verify that between them one is wrong as it refers to the city of Milan in Quebec, Canada. For her is quite simple to decide which one to consider as she can verify the information contained in each link. She makes the same query in Geonames dataset and retrieves among other information also the information about administrative hierarchy and the alternate names in different language. The information collected so far is not enough as for her geomarketing scooping she also needs to know identifiers such as NUTS, Facebook Places, WOEID, Google's Code, etc. She can have some of this identifiers in Wikidata dataset. Thus, she searches for Milan, and obtains a list of candidates for the query. Once she understands which is the right resource she accesses the link containing the descriptive information she needs. Among other information there are also different identifiers that identify equivalent resources in different KBs, such as ISTAT ID, Geonames ID, Facebook Places ID, etc. Finally she was able to find all the information she wanted but such process takes time and a lot of efforts. Such process requires expertise in writing SPARQL queries and competences in different integration platforms.

Our platform can be queried in three different ways; by searching for a toponymy (Rome, Paris, ect.), coordinates (41.89193, 12.51133; 48.85341, 2.3488, ect) or by querying the whole address (via degli Olmetti 5B 00060 Formello (ROMA); 6, rue Arsene Houssaye, 75008, Paris). Moreover a user can have one identifier belonging to one of the datasets and wants to retrieve the other IDs. For this reason the entry point can be any of the reference dataset, which the user can select a priori or make a query without selecting the source dataset. In Table 1 it is shown how the reconciliation process occurs among datasets and for each of them we give details if the link among datasets is direct or indirect (meaning that the dataset can be reached through other datasets). In the case a user searches a toponymy we make an exact match in Geonames labels and after retrieve all the other goespatial IDs from the platform. Reverse geocoding is used when coordinates is the information in input, while in the case a complete address is queried, we parse the address and consider it as part of the toponymy case.

5.3 System Architecture

The main aim of the platform is to support users in an efficient and effective way in the reconciliation process of spatial data. Since the project is still in an early phase, we have only started to implement and validate our design decisions. The logical architecture is given in Figure3. The platform has four functionalities:

- **String matching and disambiguation.** Searching for a toponymy and retrieving for example the Geonames ID.
- **Mapping via unique label.** Searching for a unique label (e.g. a code bar) for example in linking products.
- **Matching with lookup and disambiguation.** Searching for a Google Code and retrieving the Geonames ID.
- **Mapping via ID.** Searching for one of the IDs, e.g. from Geonames ID to DBpedia ID.

The link discovery is a component that implements different matching strategies. This is very important as the matching strategies are different for each of the functionalities described above. The data collector runs in batch the updates which are taken directly from the KBs and stores only the mappings or the relevant information in the private store. After, it interacts with the link discovery module in order to imply a check in the existing mapped links and find new ones if possible. These links are mappings among different KBs and are stored in an indexed database. We intend to use also state-of-the-art tools and techniques such as SILK [10] and LIMES [13] or other state-of-the-art techniques [17] for the generation of such links in order to build correct mappings.

As an example we have created the reconciliation of links between datasets for the Province of Milan (Italy). We retrieved 135 entities, which could be cities, towns, and villages under the Province of Milan in Geonames dataset. For each of them we extracted the name, alternate names, the

Table 1: Geographic information reconciliation among KBs

Name	Geonames	Wikidata	DBpedia	Google Code	Toponymy	Coordinates	Complete Address
Geonames		Direct	Indirect	Direct	Direct	Direct	Parsing
Wikidata	Direct		Direct	Indirect	Direct	Direct	Parsing
DBpedia	Indirect	Direct		Indirect	Direct	Direct	Parsing
Google Code	Direct	Indirect	Indirect		Direct	Geocoding	Direct
Toponymy	Direct	Direct	Direct	Direct		Geocoding	Parsing
Coordinates	Direct	Direct	Direct	Indirect	Reverse Geocoding		Reverse Geocoding
Complete Address	Parsing Direct	Parsing Direct	Parsing Direct	Parsing Direct	Parsing	Geocoding	

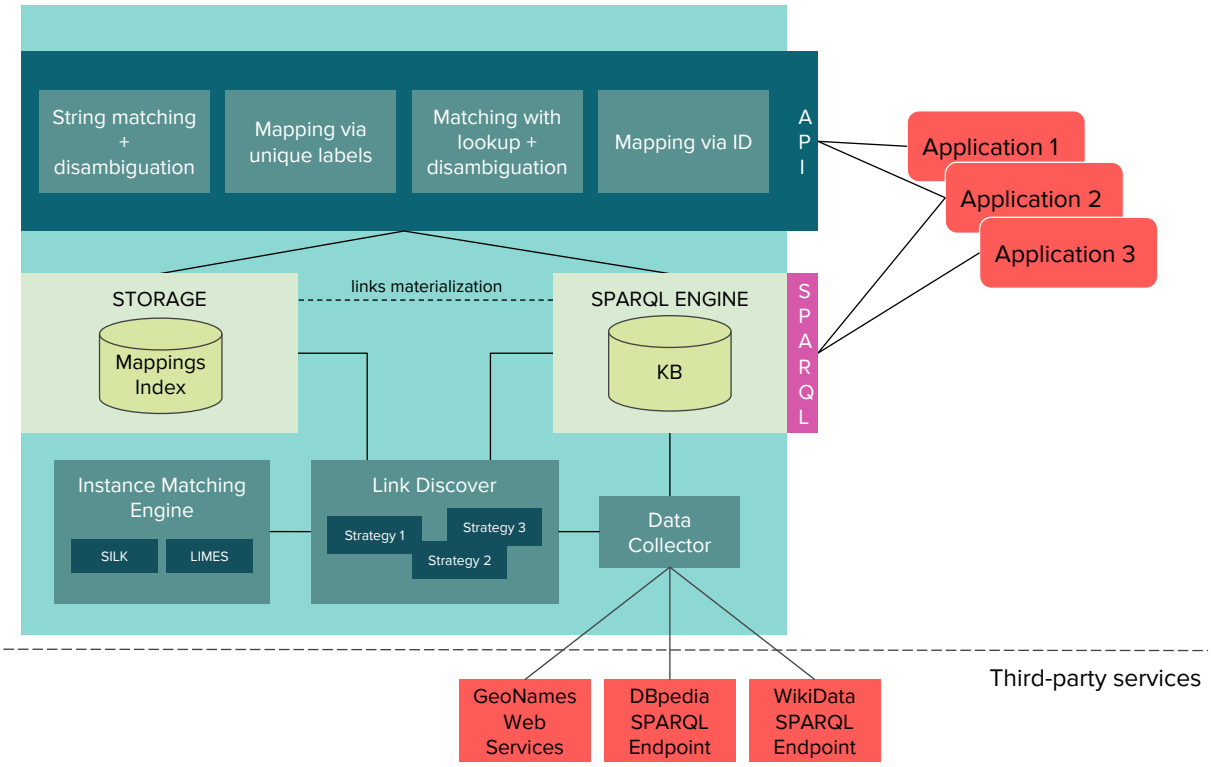

Figure 3: An overview of the logical architecture.

Table 2: Sample of the mappings for some of the cities of Province of Milan

Name	GeonamesID	WikidataID	DBpediaID	Yahoo WOEID	Facebook Places	Google Code	NUTS	Latitute
Metropolitan city of Milan	gn:3173434	wdt:Q15121	dbr:Metropolitan_City_of_Milan	NULL	NULL	08463	ITC4C	45.45186
SestoSan Giovanni	gn:6536522	wdt:Q43005	dbr:Sesto_San_Giovanni	12681991	NULL	1008491	NULL	45.53937
Vimodrone	gn:6541664	wdt:Q42419	dbr:Vimodrone	726242	NULL	1008507	NULL	45.51537
CerroMaggiore	gn:6539674	wdt:Q42507	dbr:Cerro_Maggiore	NULL	109303599096282	1008431	NULL	45.59527

administrative codes (adm1, adm2, adm3), population, latitude, longitude, elevation, country code, and time zone. For each of the geonames ID we query the Wikidata dataset, to obtain the Wikidata ID, Facebook Place ID, Yahoo! WOEID ID and the NUTS identifier. We could not find all the information for every entity but respectively 117, 2, 14, 4 identifiers. Once we have the Wikidata ID, we could query DBpedia to obtain the DBpedia identifier and we obtained 117 IDs (for each resource in Wikidata there exists the relative entity in DBpedia). Finally to find the Google Code ID we use SILK to map the entities of Province of Milan to the Google's Codes. We could generate 77 Google Codes as an exact match for the cities under the province of **Milan**.

So far we have developed the service of fetching the information from Geonames given a toponymy, the coordinates[23]. As every KB points to Geonames is easy to navigate through all of them following mapping links. Although as a first step we are considering only the links which are already mapped between datasets, the main challenge is to create and find all the mappings among the reference KBs such as the case of the NUTS and Yahoo WOEID for the resource of Cerro Maggiore.

6. CONCLUSIONS

In this paper we present a first platform aimed at support data enrichment with different geospatial datasets. The platform supports reconciliation against a set of reference datasets upon different kinds of data input, which include toponyms, coordinates, complete addresses and other identifiers. Links extracted or computed off-line among these reference datasets support then bridging across them in a seamless way. In particular, at the moment, we support reconciliation against and navigation across (Geonames, DBpedia, Wikidata, NUTS, Yahoo WOEID!, Facebook Places, and Google Codes)

In future work, we plan to extend the coverage of the links that are currently established between some of the considered datasets pairs (e.g., between Google Codes and Geonames), as well as considering more datasets . As a complementary task, we also want to correctness of the links used in the platform and extracted from existing sources (e.g., the links between the identifiers of Milan, Italy, in DBpedia and Geonames). We would also like to implement a multi-user feedback loop to validate uncertain links, following a previous work applied to ontology matching [8]. Finally, we plan to implement ranking techniques for toponym queries.

7. ACKNOWLEDGMENTS

This research has been supported in part by EU H2020 projects EW-Shopp - Grant n. 732590, and EuBusiness-Graph - Grant n. 732003.

8. REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific american*, 284(5):34–43, 2001.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22, 2009.

[3] S. Bortoli, P. Bouquet, and B. Bazzanella. Okkam synapsis: connecting vocabularies across systems and users. In *Semantic Web Collaborative Spaces*, pages 181–205. Springer, 2013.

[4] M. Cheatham, A. Krisnadhi, R. Amini, P. Hitzler, K. Janowicz, A. Shepherd, T. Narock, M. Jones, and P. Ji. The geolink knowledge graph. *Big Earth Data*, 0(0):1–13, 2018.

[5] I. F. Cruz, V. R. Ganesh, C. Caletti, and P. Reddy. Giva: a semantic framework for geospatial and temporal data integration, visualization, and analytics. In *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 544–547. ACM, 2013.

[23] `https://github.com/UNIMIBInside/conciliator`

[6] I. F. Cruz, V. R. Ganesh, and S. I. Mirrezaei. Semantic extraction of geographic data from web tables for big data integration. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, pages 19–26. ACM, 2013.

[7] I. F. Cruz, M. Palmonari, F. Caimi, and C. Stroe. Building linked ontologies with high precision using subclass mapping discovery. *Artificial Intelligence Review*, 40(2):127–145, 2013.

[8] I. F. Cruz, M. Palmonari, F. Loprete, C. Stroe, and A. Taheri. Quality-based model for effective and robust multi-user pay-as-you-go ontology matching 1. *Semantic Web*, 7(4):463–479, 2016.

[9] S. Gao, L. Li, W. Li, K. Janowicz, and Y. Zhang. Constructing gazetteers from volunteered big geo-data based on hadoop. *Computers, Environment and Urban Systems*, 61:172–186, 2017.

[10] R. Isele, A. Jentzsch, and C. Bizer. Silk server - adding missing links while consuming linked data. In *Proceedings of the First International Workshop on Consuming Linked Data, Shanghai, China, November 8, 2010*, 2010.

[11] J. J. Le Grange, J. Lehmann, S. Athanasiou, A. Garcia-Rojas, G. Giannopoulos, D. Hladky, R. Isele, A.-C. N. Ngomo, M. A. Sherif, C. Stadler, et al. The geoknow generator: managing geospatial data in the linked data web. *Linking Geospatial Data*, 2014.

[12] J.-G. Lee and M. Kang. Geospatial big data: challenges and opportunities. *Big Data Research*, 2(2):74–81, 2015.

[13] A. N. Ngomo and S. Auer. LIMES - A time-efficient approach for large-scale link discovery on the web of data. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 2312–2317, 2011.

[14] R. Parundekar, C. A. Knoblock, and J. L. Ambite. Discovering concept coverings in ontologies of linked data sources. In *International Semantic Web Conference*, pages 427–443. Springer, 2012.

[15] M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the linked data best practices in different topical domains. In *International Semantic Web Conference*, pages 245–260. Springer, 2014.

[16] V. R. Shivaprabhu, B. S. Balasubramani, and I. F. Cruz. Ontology-based instance matching for geospatial urban data integration. In *Proceedings of the 3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, page 8. ACM, 2017.

[17] B. Spahiu, C. Xie, A. Rula, A. Maurino, and H. Cai. Profiling similarity links in linked open data. In *32nd IEEE International Conference on Data Engineering Workshops, ICDE Workshops 2016, Helsinki, Finland, May 16-20, 2016*, pages 103–108, 2016.

[18] D. Sukhobok, N. Nikolov, A. Pultier, X. Ye, A. J. Berre, R. Moynihan, B. Roberts, B. Elvesæter, M. Nivethika, and D. Roman. Tabular data cleaning and linked data generation with grafterizer. In *The Semantic Web - ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 - June 2, 2016, Revised Selected Papers*, pages 134–139, 2016.