FAGI-gis: A tool for fusing geospatial RDF data

Giorgos Giannopoulos[1], Nick Vitsas[1], Nikos Karagiannakis[1],
Dimitrios Skoutas[1], and Spiros Athanasiou[1]

IMIS Institute, "Athena" Research Center

Abstract. In this demonstration, we present FAGI-gis, a tool for fusing geospatial RDF data. FAGI-gis is the core component of the FAGI framework, which handles all the steps of the fusion process of two interlinked RDF datasets in order to produce an integrated, aligned and richer dataset that combines data and metadata from both initial datasets. In the demonstation, we showcase how a user can use FAGI-gis's map based UI to perform several fusion actions on linked geospatial entities, considering both spatial and non-spatial properties of them.

1 Introduction

Languages and standards for organizing and querying semantic information, such as RDF(S) and SPARQL, are increasingly being adopted not only within academic communities but also by corporate vendors, which turn to semantic technologies to more effectively organize, expose and exchange their data as Linked Data. However, it is often the case that different data sources, although describing the same real world entities, provide different views of them, either by containing information on different subsets of attributes or even by providing different values on the same attributes. Typical reasons for this is that some sources may be outdated or may serve different purposes. For example, different maps of a city's roads and buildings, obtained from different sources (e.g., governmental, commercial, crowdsourced), may differ in the geometries and coordinates of the depicted geospatial features, as well as on the type, richness and correctness of the metadata associated with them (e.g. names and categories of buildings). As a result, information for the same real world entities is often spread across several heterogeneous datasets, each one providing partial and/or contradicting views of it. These need to be fused in order to acquire a unified, cleaner and richer dataset.

Fusion handles the merging of the linked entities, i.e., for each set of linked entities it produces a richer, more correct and more complete description w.r.t. to the properties describing it. It involves recognizing which properties of the entities correspond to each other and resolving potential conflicts or irregularities, such as different values for the same property, lack of values or properties, differences in metadata quality, etc.

In this demonstration, we present a tool for fusing geospatial RDF data. Based on our findings on the shortcomings of previous works [1], we propose a fusion framework called FAGI (Fusion and Aggregation for Geospatial Information) that includes all the aspects of the process of fusing geospatial RDF data. Specifically, we present the implementation of the core component of the framework: *FAGI-gis*, a tool for performing geospatial processing transformations on RDF geometry features, so that they can be used on complex fusion actions, involving both spatial and non-spatial properties of the interlinked entities.

2 Related Work

Although several works exist on schema integration and interlinking of RDF data, fusion has received less attention and is still a field of ongoing research. Below, we provide a brief overview of the main existing tools. Also, none of the following tools deal with geospatial RDF data. FAGI-gis fills this gap as shown in this demonstration.

Sieve [3] focuses on quality assessment and fusion of Linked Data, being part of a larger framework for Linked Data integration [4] that provides state-of-the-art techniques for data fusion. Fusion takes into account factors such as timeliness of data, provenance, as well as user configurable preference lists on features of the dataset. ODCleanStore [5] is another framework that supports linking, cleaning, transformation and quality assessment operations on Linked Data. The fusion component supports several user configurable fusion strategies, that also consider provenance and quality metadata of the datasets. KnoFuss [6] is a framework for interlinking, conflict detection, and fusion, with main focus on interlinking. It implements several variations of the Jaro-Winkler string similarity metric and an adaptive learning clustering algorithm, which is applied in a configurable way.

3 Geospatial Fusion in FAGI-gis

The central part of the fusion process is the combination of different geometries into richer and more accurate ones. FAGI-gis is the component of our framework that provides the infrastructure for this task. The tool is implemented in Java and Javascript and can be operated either via a command line utility or via a web-based graphical user interface. Since complex geospatial operations are typically time consuming, a PostgreSQL/PostGIS database is used for efficient geospatial indexing and the support of a wide range of efficient calculation and transformation functions. FAGI-gis is publicly available on GitHub[1].

The input of FAGI-gis is two separate RDF datasets and a set of links that interlink entities from one dataset to the other. The output of the tool is a unified dataset, where the geometries of the linked entities, along with the rest, non-spatial properties, are fused according to selected fusion actions. Input and output data are read from SPARQL endpoints and written in Virtuoso RDF Store respectively. Also, the supported vocabularies for representing geospatial features include GeoSPARQL with WKT serialization of geospatial features and Basic Geo.

FAGI-gis supports a set of 15 fusion actions handling both spatial and non-spatial properties. Some of them are of general use and can be applied to both types of properties, while others apply on only one type. Indicatively, FAGI-gis allows the concatenation of strings and geometries, shifting and re-scaling of geometries and mutual handling of semantically related properties (e.g. separate properties that contain different elements of an address can be handled together). Table 1 presents each fusion action, along with the type of property it applies on and a short description of its functionality.

3.1 Tool demonstration

Next, we demonstrate the usage of the software through its graphical user interface[2]. First, the user needs to input the connection information regarding the SPARQL end-

[1] https://github.com/GeoKnow/FAGI-gis

[2] A screencast video is also available at http://vimeo.com/117606305

points containing the two datasets, the local Virtuoso and PostGIS databases and the file containing the pairs of interlinked entities.

Action	Type	Functionality
keep target	both	Keeps the value of the first property
keep source	both	Keeps the value of the second property
keep both	both	Keeps both properties separately
concatenate	both	Keeps only one property that contains information from both initial properties
keep complex geometry	spatial	Keeps the geometry that consists of the most points
keep most complete	non-spatial	Keeps the literal of the property, if it contains a large part of the literal of the other property
keep complex geometry and shift it	spatial	Keeps the geometry that consists of the most points and shifts it so that it has as centroid the centroid of the other geometry
keep the average of two points	spatial	Keeps a new point geometry that is calculated by the average of the two initial points
keep one geometry and scale it	spatial	Keeps one of the geometries and rescales it according to some given factor
multi-fusion	non-spatial	Allows the handling of multiple properties describing sub-attributes of a more general attribute of an entity to be handled as a singular property
chain-fusion	non-spatial	Considers for fusion properties that describe an entity but are not directly connected with it

Table 1. FAGI fusion actions

Upon that, two processes are performed: (i) The RDF triples representing the links are loaded in an RDF graph in the local Virtuoso store and (ii) for each dataset, all possible classes that may characterize any of the linked entities are queried from the respective SPARQL endpoints and presented in two distinct lists (Figure 1). The user is able to optionally choose specific classes from both datasets and filter the pairs of linked entities based on them.

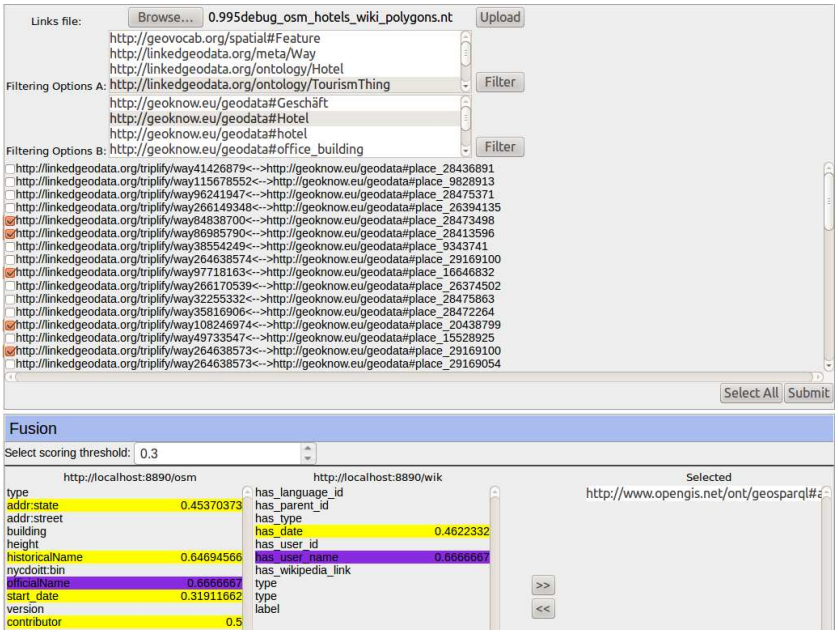

Fig. 1. Class filtering and property matching pane

After this task is performed, the linked entities are visualized on the map of the interface through points or polygons (see polygons in Figure 2). Further, a straight line segment connects each pair of linked entities so that the user can explicitly see on the map the pairs of entities to be fused.

The next step regards property matching. FAGI-gis first selects some sample linked entities pairs and tries to automatically match the properties of the entities for each

pair individually. To this end, it compares the namings of the properties based on their lexical/semantic, textual and literal type similarity. Eventually, the total of the properties for all selected link pairs are presented to the user divided into two lists, one for each of the two input datasets (see “Fusion” panel in Figure 1). When the user selects a property from one list (dataset), the system marks with yellow colour the properties of the other list (dataset) that are found to match. The final selection of the matching is performed by the user who is also able to rename the final, fused property to be kept.

Eventually, the actual fusion task takes place. The user can select a pair of linked entities from the map by clicking on the line segments that represent links. Then, the fusion panel pops-up (Figure 2), that allows to perform different fusion actions, for each pair of properties corresponding to the linked entities. Upon that, the system executes the required transformations and produces RDF triples, either to be output in a new graph, or to replace some of the initial triples of one of the input datasets.

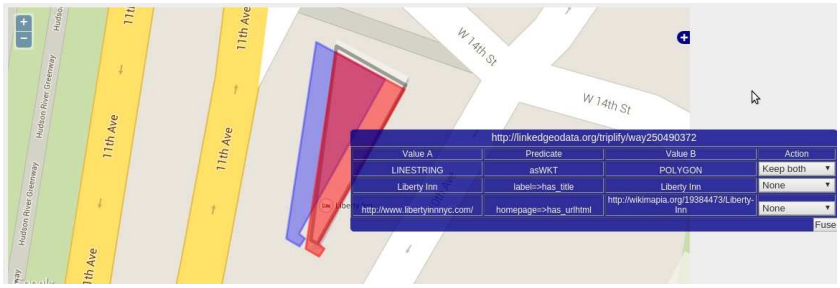

Fig. 2. Fusion pane

4 Evaluation

In the current version of the tool, automatic execution or recommendation of fusion actions is not yet implemented, thus an evaluation of the correctness and quality of the fusion results is not applicable. However, it is interesting to examine whether the tool runs in acceptable times, given large workloads.

The input used for the evaluation is two datasets extracted from Wikimapia[3], a crowdsourced, open-content collaborative mapping initiative. In particular, we considered a set of cities throughout the world (Athens, London, Leipzig, Berlin, New York) and downloaded the whole content provided by Wikimapia regarding the geospatial entities included in those geographical areas. The aforementioned dumps were transformed into RDF triples in a straightforward way. In order to create a synthetically linked dataset that can be fused, we split the Wikimapia RDF dataset, duplicating the geometries and dividing them into the two datasets in the following way: for each polygon geometry, we created another point geometry located in the centroid of the polygon and then shifted the point by a random (but bounded) factor. The polygon was left in the first dataset, while the point was transferred to the second dataset. The rest of the properties where distributed between the two datasets. This way, every entity that exists in both datasets is considered interlinked among the datasets.

The software is tested in fusing 1000, 10,000 and 100,000 linked entity pairs. We note that these numbers actually correspond to a much higher number of total triples to

[3] http://wikimapia.org/

be fused. The latter is represented by the size of each fused dataset in terms of number of triples that is recorded at each experiment. The second parameter is the size of property chains (sequences) that are considered as metadata of the linked entities. The available dataset had only property chains of depth 2, so our measurements were limited to chains of size 1 (no property chains) and 2. Further, we tested the system on three different fusion actions regarding geospatial properties: keep left (denoted ***KL***) which is a simple fusion action where no special processing is required; shift geometry (denoted ***S***), which requires a spatial transformation to take place; keep both (denoted ***KB***), which requires that a larger number of geometries are kept in the fused dataset.

chain=2						
links	**Total Time (QE)**			**Triples in fused graph**		
	S	***KL***	***KB***	***S***	***KL***	***KB***
1000	20.06	16.13	15.13	21531	21513	23531
10000	96.02	42.46	87.01	222445	222445	242445
100000	276.29	295.47	252.22	4936015	4936015	5136015

chain=1						
links	**Total Time (QE)**			**Triples in fused graph**		
	S	***KL***	***KB***	***S***	***KL***	***KB***
1000	14.86	14.34	13.92	16521	16521	18521
10000	30.27	26.42	27.94	132425	132425	152425
100000	131.87	131.82	152.76	2381137	2381137	2581137

Table 2. Total runtimes and output triples.

Table 2 presents total runtimes of the tool and total number of output triples. All columns present times in seconds except from the last column (triples in fused graph) that counts number of triples. We can see that the tool is able to perform fusion on a dataset of 100000 links in less than 5 minutes for all three tested fusion actions. Note, also that, although the links are only 100000, the final triples of the fused datasets (in the "chain=2" scenario) are 5 Million. This fact, along with the fact that the tool is unavoidably burdened by the RDF store's query execution overhead, which can be further optimized, shows the scalability of the tool and the potential for further improvement.

5 Conclusions

In this demonstration, we demonstrated FAGI-gis, a tool that supports geospatial processing and transformations for fusing spatial and non-spatial properties of interlinked RDF entities. Our future work focuses on enhancing the functionality of the tool, to support more fusion strategies, as well as to increase the efficiency of the underlying operations. Our plans include also the addition of a learning module that will be trained on previous user actions to allow for automatic fusion recommendations.

Acknowledgments This work was supported by a grant from the EU's 7th Framework Programme (2007-2013) provided for the project GeoKnow (GA no. 318159).

References

1. Giannopoulos, G. and Skoutas, D. and Maroulis, T. and Karagiannakis, N. and Athanasiou, S. FAGI: A Framework for Fusing Geospatial RDF Data. In *Proc. of OTM, ODBASE*, 2014.
2. Giannopoulos, G. and Maroulis, T. and Skoutas, D. and Karagiannakis, N. and Athanasiou, S. FAGI-tr: A Tool for Aligning Geospatial RDF Vocabularies. In *ESWC, demo track*, 2014.
3. Mendes, P. N. and Muhleisen, H. and Bizer, C. Sieve: linked data quality assessment and fusion. In *Proc. of the Joint EDBT/ICDT Workshops*, pp. 116-123, 2012.
4. Schultz, A. and Matteini, A. and Isele, R. and Mendes, P. and Bizer, C. and Becker, C. LDIF-A Framework for Large-Scale Linked Data Integration. In *WWW, developer track*, 2012.
5. Michelfeit, J. and Knap, T. Linked Data Fusion in ODCleanStore. In *Proc. of the International Semantic Web Conference (Posters & Demos)*, 2012.
6. Nikolov, A. and Uren, V. S. and Motta, E. and Roeck, A. N. De. Integration of Semantically Annotated Data by the KnoFuss Architecture. In *Proc. of EKAW*,2008.