LOD4STAT: a scenario and requirements

Pavel Shvaiko[1], Michele Mostarda[2], Marco Amadori[2], and Claudio Giuliano[2]

[1] TasLab, Informatica Trentina S.p.A., Trento, Italy
[2] Fondazione Bruno Kessler - IRST, Trento, Italy

Abstract. In this short paper we present a scenario and requirements for ontology matching posed by a statistical eGovernment application, which aims at publishing its data (also) as linked open data.

Introduction. Our application domain is *eGovernment*. By eGovernment we mean an area of application for information technologies to modernize public administration by optimizing work of various public institutions and by providing citizens and businesses with better and new services. More specifically, we focus on statistical applications for eGovernment. The driving idea is to capitalize on the statistical information in order to increase knowledge of the Trentino region. Releasing statistical data (with disclosure control) as linked open data aims at simplifying access to resources in digital formats, at increasing transparency and efficiency of eGovernment services, etc. The main challenge is the realization of a knowledge base, which is natively enabled to work with RDBMS tables. Despite this approach has been tailored specifically to the statistical database domain, there is substantial room for generalization. In this view, there was a number of initiatives aiming at releasing governmental data as linked open data to be taken into account: in GovWILD [1] links were established automatically with specifically developed similarity measures, while in [2], the alignment was done semi-automatically with Google Refine. The currently available matching techniques can be well used for automating this process [3].

Scenario. Figure 1 shows the key component, called Statistical Knowledge Base (SKB), of the LOD4STAT system-to-be. The SKB aims at enabling its users to query statistical data, metadata and relations across them without requiring specific knowledge of the underlying database. Users can issue queries, such as *find all data related to population age and employment for the municipality of Trento*. Specifically, user query is analyzed in order to extract concepts out of labels. Then, these are matched at run time against the SKB. For the query example, the term *population age* is connected to *Registry Office*, while *employment* is connected to *Social Security*. The system returns a set of tables, metadata and entities from the Registry Office (with information about population and age) and from the Social Security (with information about employment) containing data for the city of Trento and will suggest possible joins between columns.

The SKB is an interconnected aggregation of ontologies (interpreted in a loose sense), such as WordNet, DBpedia, ESMS[1] what allows both multi-classification and multiple views on data. These ontologies have to be matched among them to enable navigation across them through the respective correspondences. The SKB is also able to export query results in several formats, such as RDF Data Cube and JSON-Stat. The SKB is represented by three (horizontal) layers. The upper layer is a collection of ontologies specific to the statistics domain, e.g., ESMS. The middle layer is composed

[1] http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/metadata

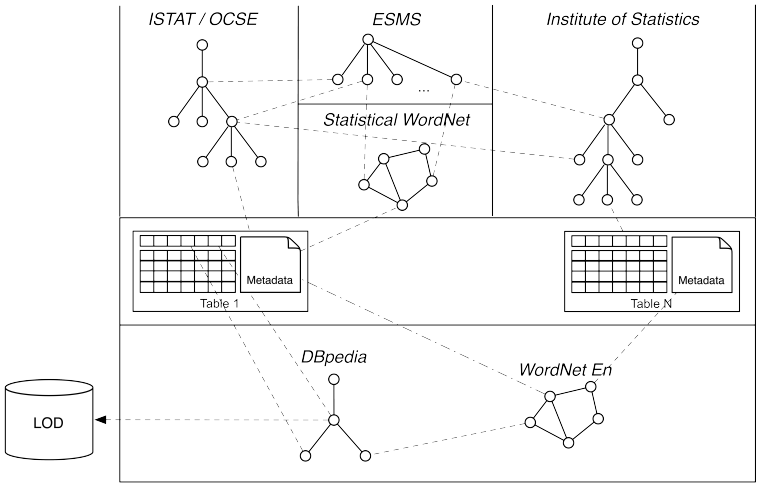

Fig. 1: LOD4STAT: the statistical knowledge base component.

of relational tables associated with metadata. The lower layer is composed of a collection of general purpose ontologies, e.g., WordNet, DBpedia. Table columns are to be matched to the entities of the involved ontologies. Notice that SKB allows for the definition of explicit connections between columns of different tables that can be joined together. Every time an alignment is updated the respective data is updated accordingly.

Requirements. There are several key requirements posed by this application, such as: *performance* - answer queries within 3s., as suggested by the UI interaction practice; *availability* - service up by 99% of time, no more than 15mins. downtime per day during working hours. These requirements put only constraints on run time matching needed between the user query and the SKB ontologies. Matching results can be approximate, though their correctness is preferred over completeness. For what concerns design time matching inside the SKB, it can be performed at design time semi-automatically with sound and complete alignment when any of these knowledge sources evolve. Notice that statistical disclosure control methods use weights associated to the table columns, and it should be possible to inherit them through the respective alignments.

Conclusions and future work. In this short paper we have presented a scenario and requirements for ontology matching within a statistical eGovernment application. We note that such requirements in part have a transversal character, such as for run time query matching and design time matching between the ontologies of the system. However, there are also peculiarities related to the use of alignments, such as support for statistical disclosure control. Future work includes formalization, implementation and evaluation of the system in order to be brought to production.

Acknowledgments. The work has been supported by the Autonomous Province of Trento, Italy.

References

1. C. Böhm, M. Freitag, A. Heise, C. Lehmann, A. Mascher, F. Naumann, V. Ercegovac, M. Hernández, P. Haase, and M.l Schmidt. GovWILD: integrating open government data for transparency. In *Proceedings of WWW*, pages 321–324, 2012.
2. F. Maali, R. Cyganiak, and V. Peristeras. A publishing pipeline for linked government data. In *Proceedings of ESWC*, pages 778–792, 2012.
3. P. Shvaiko and J. Euzenat. Ontology matching: state of the art and future challenges. *TKDE*, 25(1):158–176, 2013.