Strategies for Semantic Integration of Energy Data in Distributed Knowledge Bases

Álvaro Sicilia, Fàtima Galán, Leandro Madrazo
ARC Enginyeria i Arquitectura La Salle
Universitat Ramon Llull
Barcelona, Spain
[asicilia, fatima, madrazo]@salleurl.edu

Abstract — **This paper reports on the work that is currently being developed in RÉPENER, a research project co-financed by the Spanish National RDI plan 2009-2012. The objective of the project is to apply Semantic Web technologies to create an energy information system which puts together data from different sources, both private and public. To create such a system, it is necessary to integrate different data sources from different domains. Different strategies might be adopted, depending on the contents of the data sources involved. One of them is about adding new external data sources to create brand-new links between the existing ones. This is the strategy thus devised and implemented in this project. In this paper, a description of the process by which two databases with energy information have been linked using a SILK framework is provided.**

Keywords-semantic integration; ontology design; object matching; energy data.

I. INTRODUCTION

Energy-related information is dispersed in proprietary databases and open data sources. It is heterogeneous, since it is generated by different applications and for various purposes (modelling and simulation programs, monitoring systems), and it is compartmentalized by reference to the various stages of the building lifecycle; from design, to construction and operation. For this reason, energy information cannot be properly processed and analyzed because there is a lack of interoperability between applications and databases. Consequently, decision-making actors cannot exploit the benefits of correlative data from different stages and sources due to the lack of a common vocabulary and to the difficulties of accessing the data.

The application of Semantic Web technologies can help to overcome all of these limitations through the application of semantic data-integration processes. In recent years, studies on data integration using ontologies have delivered substantial results. The prime example, which proves the feasibility of tasks solutions, is the Linked Open Data project [1]. Therefore, semantic integration is a core issue in interoperability, particularly in a heterogeneous setting such as the World Wide Web, where different ontologies are used. Semantic integration inevitably leads to inter-ontological mapping, or ontology integration. As stated by Zhdanova [2] and Euzenat [3], ontology matching is a plausible solution to the problem of semantic heterogeneity in many applications. Once the matching is done, the conjunction of ontologies and their interconnections facilitate an integrated access to heterogeneous energy data by providing: (i) a common vocabulary to unify different areas of knowledge or expertise today separated, (ii) an integrated way to explore energy information and its related data; and (iii) a compound bulk data with which to perform data analysis using data-mining techniques. The third feature can retrofit the information system by adding new data relations, which in turn enhance the exploration experience.

The purpose of the RÉPENER [4] project is to develop an ontology-based information system which supports decision-making processes and knowledge discovery by actors who deal in energy management with respect to buildings. The semantic information system which is being developed addresses the interoperability issues between different data sources using semantic technologies. Ontologies are designed using the OWL standard language, and data is exposed on the Internet using the RDF [5] following the Linked Open Data initiative. In this way, the interoperability problem is solved, because all data sources are described by means of a common language – which can be processed by humans and/or machines – using standard protocols. A comprehensive project description can be found in "in press" [6].

The feasibility of the data-integration process and the quality of the interrelationships amongst different data sources is a key issue. In the ideal scenario, data sources of different domains overlap in some concepts, and this allows one to create links between them. For example, on the one hand, a building repository can contain building instances which have a property location naming the city in which the buildings are located. On the other hand, a spatial data repository can contain landmarks with property names. Therefore, both data sources can be connected through the properties' locations and names. However, in an actual scenario, where data sources cannot be modified the process of connecting them is not that simple because there might be elements which do not overlap.

This paper presents the semantic integration process which has been carried out in the RÉPENER project with the objective of integrating data sources having non-overlapping elements.

The content of the following sections of the paper is summarized next. Some of the current strategies, procedures and tools used to perform the integration of data sources are discussed in Section II. In Section III it is described the work done to connect energy related data from different sources

 ISBN: 978-1-61208-240-0

and domains. Finally, the conclusions which can be drawn from the application of the procedures are summarized in Section IV.

II. STRATEGIES FOR SEMANTIC INTEGRATION

Historically, data has been stored in relational databases, usually available in offline environments and published on the Internet through web applications which interconnect web documents instead of data instances. The Semantic Web concept coined by Tim Berners-Lee [7] was subsequently undertaken by the Linked Data movement, which has called for the creation of a web of data using Uniform Resource Identifiers (URIs) as the resource identification, Hypertext Transfer Protocol (HTTP) as the universal data retrieval mechanism, and the resource description framework (RDF) as a data model describing things in the world [1].

The web of data is comprised of several heterogeneous data sources which describe different domains using a vocabulary handled by domain experts. The interconnection between data sources is possible thanks to the addition of semantics to data, as achieved by means of metadata descriptions, thereby guaranteeing the interoperability between data. Furthermore, these semantic layers, jointly with the links between the data sources, enable applications to perform smart data analysis which can actually enrich the data. For example, an application could retrieve energy certifications of buildings as built from a repository, in a specific area, and then complement them with regional socioeconomic data from a statistical database. All of these would be for the purpose of applying the data-mining process to compare the energy rating with the economic level of the selected area. Finally, a user can use this improved information based on this comparison to make better decisions.

Data sources are incorporated in the web of data through semantic integration processes in two steps: 1) publishing semantic data which is translating relational data to RDF format, with the objective of exposing data through an SPARQL end point, and of releasing the ontology which models the data, and 2) interlinking data sources between them to create a network, which can subsequently be exploited.

A. Data transformation strategies

To perform the integration of data sources, the source data must first be transformed from a relational database to the RDF language following the Linked Data principles. Data transformation may be implemented as a static ETL (Extract, Transform and Load) process or as a query-driven dynamic process. The static transformation process creates an RDF dump following the application of certain mapping rules. The most significant drawback of the static process is that the most recent data might not be considered. Contrastingly, the dynamic transformation process uses simple queries to access the latest data. A survey published by the W3C RDB2RDF incubator group has identified several tools which can carry out both transformations – static and dynamic – such as Virtuoso RDF View, D2RQ, R2O, or Triplify [8]. The survey concludes that there is not a standard method for the representation of mappings between RDB and RDF and recommends, whenever possible, to implement on-demand mapping to access the latest version of the data. In February 2012, the RDB2RDF group published a R2RML (Relational database to RDF Mapping Language) [9], a recommendation which is currently being implemented in various projects.

B. Linking strategies

The integration process involves interlinking objects of one or several data sources. Each data source usually uses different URIs to identify objects based on domain criteria, even if they describe the same real-world object. Therefore, links between these objects cannot be obtained in a straightforward way. For this reason, finding out that two data objects refer to the same real-world object is a key issue for data integration. Object matching methods (also known as instance consolidation, record linkage, entity resolution or link discovery) are focused on identifying semantic correspondences between objects of different data sources.

Object matches can be set manually or automatically. Typically, manual matching is carried out with small data sources, where it is important to ensure the high quality of the correspondences. If the data source is large, however, it is better to apply automated or semi-automated proposals [1]. The creation of links between objects can be handled with a domain-specific approach or with a universal approach, as stated by Ngonga Ngomo [10]. The second type of approach is not depending of the domain of the data sources, and, therefore, it can be applied to different scenarios. To perform this task, Ngonga Ngomo has identified different tools, among which the SILK framework stands out [11], and proposes an outperformed framework. Both frameworks – SILK and LIMES – generate links between RDF objects based on several similarity metrics (e.g., Levenshtein [12], Euclidean [13], or Jaro-Winkler [14]), which can be chosen and tailored by the user. From a technological point of view, both frameworks gather data from a SPARQL [15] end point, which is described in a configuration file containing the metrics applied and the object selection restrictions.

The aforementioned tools are designed for data sources which contain overlapped objects. In this case, the linking process can be carried out with these tools by setting up the configuration file to match them with the proper similarity metric. As has been stated previously, overlapped objects are not the usual case, and for this reason it is necessary to apply elaborated procedures. Accordingly, this research work proposes a data integration strategy which is based on complementing data sources with external data in order to enable a successful object matching process (Fig. 1).

 ISBN: 978-1-61208-240-0

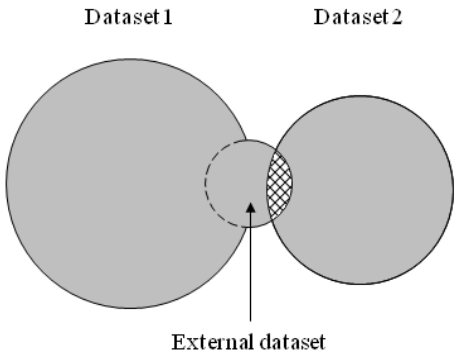

Figure 1. Elaborated data integration strategy.

The procedure of linking databases involves the attachment of external data properties to objects of a data source for, in a second step, applying the link-generation tools mentioned above. The first step is to identify the potential objects which can describe the same object of the world but do not have sufficient properties to be matched. For example, a data source DS1 which contains the monitoring data of a building whose location is described with a string property (e.g., the name of the place) and a data source DS2 that contains weather stations which are geo-located (e.g., longitude and latitude). Both objects might be connected through the property *hasWeatherStation*, where each building is linked to its closest station (Fig. 2). Because a string value and the set of longitude-and-latitude properties cannot be compared, a possible solution is to add *geo-localization* properties to the building objects or *place name* properties attached to weather-station objects.

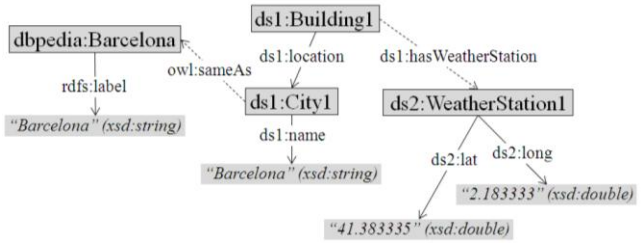

Figure 2. Object matching example.

Once the potential objects have been identified and analyzed, it is necessary to find new data properties in external data sources which can be used for object matching. The search process is carried out following the links between data sources, usually based on the *owl:sameAs* property. For specialized domains, it is important to be supported by a domain expert who can guide the exploration. In the previous example, city objects might have an *owl:sameAs* property linking DS1 with Dbpedia city objects which might have geo localization properties. When the links to external data sources are not available, it is necessary to generate them in order to access to the external data properties using link-generation tools. Finally, the data properties of external data sources are included in the existent data sources which can then be matched to other data sources.

The data sources are enriched by this method, which is based on the addition of new data properties gathered from external data sources instead of inferring new class relationships, taking into account the ontology itself as the enrichment step of the Linked Open Data life cycle.

This procedure has its weakness in the fact that external data must be added to the source-data sources. This is not always feasible. For instance, it occurs when the semantic data is generated through an RDF wrapper. In these cases, a possible solution is to publish RDF links in an intermediary RDF store (e.g. [16] service), which link-generation tools can then use to gather data.

III. IMPLEMENTATION

This section describes the work which has been done to connect energy data sources applying the integration procedures previously described. The proprietary data sources used in the implementation have been provided by ICAEN – an organization of the Catalan government which gathers the energy certificates of newly planned buildings which include their simulated performance – and by Aemet, the Spanish meteorological agency which provides measurements made by a network of meteorological stations throughout Spain. The Ontology Engineering Group from the Universidad Politécnica de Madrid has published the Aemet data source through an SPARQL end point [17]. These data sources have been combined with the ultimate goal of assembling data from different sources and domains, thereby enabling the final user to understand the figures of buildings' energy certifications when applied to the building environment, particularly in the context of climate.

The semantic integration process is divided into two parts: the relational data transformation and the data interconnection.

A. Data transformation

The main purpose of data transformation is to create interconnected ontologies with the ability of integrating the different data sources. The data transformation embraces two actions: 1) the creation of an ontology which fits with the database, and 2) the transformation of data according to the ontology thus designed.

To create an ontology, the ICAEN data source [18] has been cleaned so as to eliminate unnecessary data, and consistency methods have been applied by energy domain experts. With the collaboration of energy domain experts and ontology engineers, an informal data structure has been developed in order to build an ontology which includes all the terms and categories identified. The ontology relies on a foundational ontology created for this project, which encompasses the building energy domain as well as other domains (social, economic). However, not all the data-source content has been contemplated in the ontology. For this reason, it has been necessary to define new concepts and relations to integrate this data in the ontology, as existing ontologies to be reused could not be found.

Once the ICAEN ontology has been created, the data is transformed to RDF. For the data transformation, a dynamic transformation process has been chosen, given the need for

 ISBN: 978-1-61208-240-0

access to the most recent data. The tool selected to carry out this work is D2RQ [19], because it supports database translation with high performance. It dynamically rewrites the SPARQL queries into SQL. This is a stable, lightweight solution. It represents mappings with an easily customizable D2RQ mapping language, enables configuration changes in real time, is independent of the database provider. It is currently being developed to support R2RML and it publishes data in HTML, RDF and through a SPARQL end point. The D2RQ tool has been configured to transform the ICAEN database according to the ontology thus designed.

B. Data linking

As a result of this transformation process, all of the data sources have become accessible through a SPARQL end point. The integration of the ICAEN and Aemet data sources has been accomplished, generating links between buildings' *ProjectData* and *WeatherStation* objects according to the *icaen:hasWeatherStation* property (Fig. 3) using the SILK tool [7].

The first attempt at integration was carried out by matching *ProjectData* and *WeatherStation* objects through the data properties: *icaen:ID_Localitat* (the name of the city in which the building is located) and *aemet:stationName* (the name of the station, which is usually the same as the one of its closest city). The number of valid generated links was low, because nearly all station names were based on a mixture of the city name and internal terms, which interfered the matching. To solve this problem, we looked for external data sources which could provide additional data. The Linked GeoData repository – which provides spatial data such as roads, cities, mountains or points of interest – was selected as a source of external data. Linked GeoData objects are geo-located using latitude and longitude data properties, which are also used by weather stations. Therefore, it constitutes a feasible source of additional data.

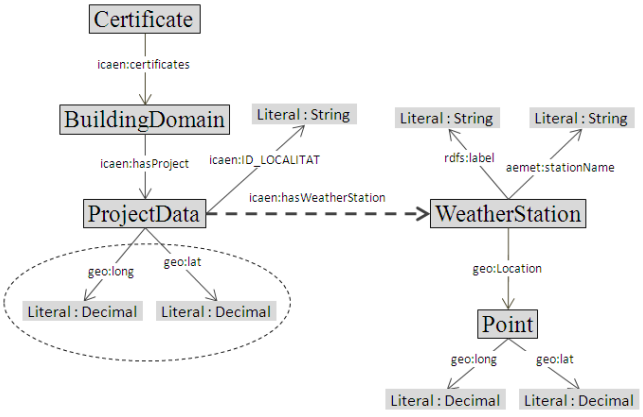

Figure 3. ICAEN and Aemet integration.

The SILK tool has been configured with the Levenshtein similarity function, which is best suited to compare string chains. ICAEN objects have been filtered by the *icaen:ProjectData* class, and Linked GeoData objects have been filtered by the *lgdo:City*, *lgdo:Town*, and *lgdo:Village* classes. The ICAEN data source contains 1,805 objects and the Linked GeoData 720323. The SILK tool has found out 1,398 links between both data sources; thus, 77% of the buildings achieved links to a Linked GeoData place in less than an hour of execution time.

To attach geo-localization properties to ICAEN objects, a script which generates two RDF triples for each ICAEN object was developed: one for the latitude property and the other for the longitude one. The script queries the end points with SPARQL and generates a N-Triples file, which is later uploaded to the ICAEN data source.

Once the ICAEN objects contain geo localization properties from the Linked GeoData data source (the dotted circle in Figure 3), the SILK tool is called to generate links between the ICAEN and Aemet data sources. In this case, a geographical distance function is selected to use both the *geo:long* and *geo:lat* properties for the purpose of comparing objects. The Aemet data source contains only 260 weather stations, so the execution time is less than 4 seconds for generating 1,305 links, thereby covering 72% of the ICAEN buildings, or 93% of the ICAEN buildings which have links to linked GeoData objects (Table I).

TABLE I. LINK GENERATION COVERAGE

Icaen objects:	1805 (100%)
Linked to Linked GeoData:	1398 (77%)
Linked to Aemet:	1305 (72%)

IV. CONCLUSIONS

The work thus far developed facilitates semantic integration processes in linked data environments, taking advantage of existing links between data sources or by generating intermediary links. The procedure has been validated in a case study which demonstrates the feasibility of using external data sources to integrate semantic data.

It has been suggested that intermediary RDF stores can be used when it is not possible to add external data to a data source. As far as we know, this process is not possible using current link-generation tools, because they cannot integrate external data. Further work to be done in this regard is to have generation tools use federated queries in order to take advantage of the existing links.

Data sources will be released simultaneously with the user interface and project services by the end of the research project.

ACKNOWLEDGEMENT

RÉPENER is being developed with the support of the research program BIA 2009-13365, co-funded by the Spanish National RDI Plan 2009-2012.

REFERENCES

[1] T. Heath and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space," Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 2011, pp. 1-136. Morgan & Claypool.

 ISBN: 978-1-61208-240-0

[2] A. Zhdanova, “Towards a community-driven ontology matching,” Proc. 3rd international conference on knowledge capture (K-Cap’05), ACM, Oct. 2005, pp. 221-222.

[3] J. Euzenat, “Semantic technologies and ontology match-ing for interoperability inside and across buildings,” Proc. 2nd CIB workshop on eeBuildings data models, Oct. 2011, pp. 22-34.

[4] http://arc.housing.salle.url.edu/repener 09.08.2012

[5] http://www.w3.org/RDF/ 09.08.2012

[6] L. Madrazo, A. Sicilia, M. Massetti, and F. Galan, “Semantic modeling of energy-related information throughout the whole building lifecycle,” Proc. 9th European Conference on Product and Process Modelling (ECPPM), Jul. 2012 .

[7] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, vol. 284, May 2001, pp. 29-37, doi:10.1038/scientificamerican0501-34.

[8] S. S. Sahoo, W. Halb, S. Hellmann, K. Idehen, T. Thibodeau, S. Auer, J. Sequeda, and A. Ezzat, “A Survey of Current Approaches for Mapping of Relational Databases to RDF,” W3C publication, 2009. http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf (accessed July 16, 2012).

[9] http://www.w3.org/TR/r2rml/ 09.08.2012

[10] A.-C. Ngonga Ngomo and S. Auer, “LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data,” Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI), IJCAI/AAAI, Jul. 2011, pp. 2312-2317.

[11] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov “Discovering and maintaining links on the web of data,” Proc. 8th International Semantic Web Conference (ISWC), Springer, Oct. 2009, pp. 650-665.

[12] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” Soviet Physics Doklady, vol. 10, 1966, pp. 707-710.

[13] E. Deza and Michel M. Deza, “Encyclopedia of Distances,” Springer Berlin Heidelberg, 2009, pp. 94.

[14] W. Winkler, “The state of record linkage and current research problems,” Technical report, Statistical Research Division, U.S. Bureau of the Census, 1999.

[15] http://www.w3.org/TR/rdf-sparql-query/ 09.08.2012

[16] http://sameas.org/ 09.08.2012

[17] http://aemet.linkeddata.es/sparql_en.html 09.08.2012

[18] http://www20.gencat.cat/portal/site/icaen 09.08.2012

[19] C. Bizer and R. Cyganiak, “D2R Server – Publishing Relational Databases on the Semantic Web,” Poster at the 5th International Semantic Web Conference (ISWC), Springer, Nov. 2006.

 ISBN: 978-1-61208-240-0