On the Experimental Usage of Ontology-based Data Management for the Italian Integrated System of Statistical Registers: Quality Issues

Raffaella Maria Aracri, Italian National Institute of Statistics, aracri@istat.it

Adele Maria Bianco, Italian National Institute of Statistics, bianco@istat.it

Roberta Radini, Italian National Institute of Statistics, radini@istat.it

Monica Scannapieco, Italian National Institute of Statistics, scannapi@istat.it

Laura Tosco, Italian National Institute of Statistics, tosco@istat.it

Federico Croce, Sapienza University of Rome, croce@diag.uniroma1.it

Domenico Fabio Savo, Sapienza University of Rome, savo@diag.uniroma1.it

Maurizio Lenzerini, Sapienza University of Rome, lenzerini@diag.uniroma1.it

Abstract

Ontology-based data management (OBDM) is a recent paradigm for addressing data management based on a conceptualization of the domain of interest, called ontology. A system realizing the vision of OBDM is constituted by three layers: the ontology, that provides a high level, formal, logic-based representation of the above mentioned conceptualization; the data source layer, representing the existing data in the various assets of the system; the mapping between the two layers, which is an explicit representation of the relationship between the data sources and the ontology. Although most works on OBDM focus on querying data through the ontology, recent papers argue that OBDM is a promising tool for assessing the quality of data, especially in the presence of multiple, possibly mutually incoherent data source. We have experimented the OBDM paradigm for data quality assessment in a current project of Istat, namely the Italian Integrated System of Statistical Registers. We have focused on the domain of population data, and we have built an ontology for modeling basic concepts and relationships of this domain, including persons, families, parental relations, citizenship, locations, etc. Then, we have considered a core set of population data and we have specified the mappings from such data sets and the ontology. With such a specification at hand, we have used the MASTRO system for OBDM for carrying out several data quality checks. The preliminary results are extremely encouraging, both in terms of effectiveness of the method and in terms of efficiency of the checking procedures, in the sense that the performance of the quality check is not affected by the (usually expensive) task of reasoning over the ontology.

Keywords: ontology, data integration, statistical register, data quality

1. Introduction

Data analysis is one of the most important IT (Information Technology) tasks in a data-driven society. However, when dealing with Big Data this is far from easy: indeed, as pointed out in (De Giacomo et al. 2018) loading a big data platform with quality data with enough structure to deliver value is a lot of work and requires sophisticated techniques. Thus, it is not surprising that data scientists spend an estimated 50%-80% of their time on accessing, integrating and preparing data for analysis (see CrowdFlower 2016).

In this paper, we follow the idea of using semantics for making data integration, preparation, and governance more powerful. As illustrated in (Lenzerini 2011), using semantics means conceiving information systems where the semantics of data is explicitly specified and is taken into account for devising all the functionalities of the system. Over the past two decades, this idea has become increasingly crucial for a wide variety of information-processing applications and has received much attention in the Artificial Intelligence, Database, Web, and Data Mining communities (Noy, Doan and Halevy 2005). In particular, we concentrate on a specific paradigm, called Ontology-Based Data Management (OBDM), introduced about a decade ago as a new way for modeling and interacting with a collection of data sources (Calvanese et al. 2007; Poggi et al. 2008; Lenzerini 2011). According to such paradigm, the client of the information system is freed from being aware of how data are structured in concrete resources (databases, software programs, services, etc.), and interacts with the system by expressing her queries and goals in terms of a conceptual representation of the domain of interest, called ontology.

More precisely, an OBDM system is an information management system maintained and used by a given organization (or, a community of users), whose architecture has the same structure of a typical data integration system, with the following components: an *ontology*, a set of *data sources*, and the *mapping* between the two.

- The ontology is a conceptual, formal description of the domain of interest of the organization, expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions formally describing the domain knowledge.
- The data sources are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories

are numerous, heterogeneous, each one managed and maintained independently from the others.

- The mapping is a precise specification of the correspondence between the data contained in the data sources and the elements of the ontology. Here element means concept, attribute, or relationship.

We observe that the above three layers constitute a sophisticated knowledge representation system that can be managed and reasoned upon with the help of automated reasoning techniques. For example, suitable algorithms allow queries expressed over the ontology to be answered by automatically translating the query in terms of the data sources using the mapping (Calvanese et al. 2007). Although the problem of answering queries over the ontology has been the main focus in the last years, there are several other services that an OBDM system should provide. Data quality assessment (Batini and Scannapieco 2016) is one notable example.

It is often claimed that data quality is one of the most important factors in delivering high-value information services (Fan and Geerts 2012). However, the heterogeneity of the data sources, and the fact that their structure is often dependent of the applications they serve, pose several obstacles to the goal of even checking data quality, let alone achieving a good level of quality in information delivery. How can we possibly specify data quality requirements, if we do not have a clear understanding of the semantics that data should bring? The problem is sharpened by the need of connecting to external data, originating, for example, from business partners, suppliers, clients, or even public sources. Again, judging about the quality of external data, and deciding whether to reconcile possible inconsistencies or simply adding such data as different views, cannot be done without a deep understanding of their meaning. This is the main reason why it is relevant and promising to apply the OBDM paradigm to the problem of data quality assessment (Wand and Wang 1996, Console and Lenzerini 2014a; 2014b). Basing this task on a formal conceptualization of the domain of interest allows us to easily blur out all the meaningless details of the single data source, and focus on real data quality issues. Moreover, different data sources can be analyzed using the same yardstick, i.e., the ontology, and hence analyzed and compared in terms of their quality. Finally, the use of a conceptualization shared among the different assets of an organization allows for data quality assessments that are easy to present and potentially used in many different contexts.

The goal of this paper is to discuss an experience of using OBDM for data quality assessment in the context of the Italian Integrated System of Statistical Registers. We will show how this paradigm and the associated tools can have a significant role in addressing quality issues from the perspective of the usual dimensions studied in the context of data quality, namely, consistency, accuracy, and completeness. In the experimentation, we have focused on the domain of population data, and we have built an ontology for modeling basic concepts and relationships of this domain, including persons, families, parental relations, citizenship, etc. Then, we have specified the mappings linking a core set of population data to the ontology. With such a specification at hand, we have used the MASTRO system (Calvanese et al. 2011) for OBDM for carrying out several data quality checks. The preliminary results are extremely encouraging, in terms of effectiveness of the method and of efficiency of the checking procedures, in the sense that the performance of the quality check is not affected by the (usually expensive) task of reasoning over the ontology. The structure of the paper is as follows. In Section 2 we describe the reference scenario, namely, the Italian Integrated System of Statistical Registers, in Section 3 we briefly present some technical aspects of the OBDM paradigm, and in Section 4 we illustrate the main aspects of the experimentation, mainly through examples. Section 5 concludes the paper by highlighting possible developments of our work.

2. The Scenario

Istat is undergoing a significant revision of the statistical production by investing in the implementation of a system of integrated statistical registers as a base for all the production surveys. This system, named as the Italian Integrated System of Statistical Registers (ISSR), is a logically centralized source where all the data supporting production processes can be accessed. Some registers of the ISSR are defined as base, i.e. containing the core units' variables; these are: (i) Register of Individuals, Families and Cohabitations; (ii) Register of Production Units; (iii) Register of Places; (iv) Register of Activities. Some other registers are extended or thematic, i.e. they add to the units in the base registers some thematic variables (e.g. Women Reproductive Stories) or combine units from different registers (i.e. Labour).

From an architectural perspective, there are three major data building blocks involved in the construction of the ISSR. These are:

- Raw data: input data of the various production processes, coming from administrative archives, or surveys, or sources other than traditional ones (for example, Big Data or GIS spatial data).
- Working data: data processed in the various correction, preparation and integration steps, up to the validation of the results.
- Validated data: these are data contained in the ISSR that have indeed passed the validation checks.

In this scenario, the OBDM approach has been used for quality checking on Working data. For instance, as it will be clarified in the remainder of the paper, several consistency checks through OBDM have been implemented. Let us remark that the complexity of this step is increased by the need of checking cross-registers consistency. Such an effort was aimed to prove the feasibility of realizing at least part of the consistency rules that are typically implemented by using traditional (i.e. not ontology-based approaches) checking processes.

3. The Ontology-based data management paradigm and its use for data quality

An OBDM specification I is a triple ⟨O, S, M⟩, where O is an ontology, S is a relational data schema, called source schema, and M is a mapping from S to O. As already stated, O represents the general knowledge about the domain expressed in some logical language. Typically, in OBDM systems, O is expressed in a Description Logics of the DL-Lite family (Calvanese et al. 2007). These languages are characterized by an optimal trade-off between expressive power and efficiency of reasoning, in particular of query answering. The mapping M is a set of mapping assertions, each one linking a query over the source schema to a query over the ontology. Intuitively, a mapping assertion specifies that the presence of a certain pattern in the data source implies that some objects (resp., pairs) are instances of a class (resp., a relation). An OBDM system is a pair (I, D) where I is an OBDM specification, and D is a database for the source schema S, called source database for I. The semantics of (I, D) is given in terms of the logical interpretations that are models of I with respect to D (i.e., satisfy all axioms of O, and satisfy M with respect to D). Notice that OBDM allows dealing with incomplete information, reflected by the fact that a system may have many models, each one corresponding to a mean to complete the partial knowledge possessed by the data sources.

As already said, in OBDM systems, the main service of interest is query answering, i.e., computing the answers to user queries posed over the ontology. Such service amounts to return the so-called certain answers, i.e., the tuples that satisfy the user query in all the models of (I, D). However, in this paper we are interested in using OBDM for data quality assessment, by referring in particular to the three main quality dimensions, namely, consistency, accuracy, and completeness (Console 2016).

Consistency is the quality dimension dealing with the coherence of data. Counterexamples to consistency show that data suffers from integrity problems, thus providing crucial information about the assets owning such data. In the literature, it is often advocated that consistency can be assessed by checking whether data follows specific rules for integrity. However, in traditional approaches, such rules are either implicit, or specified depending on the single data source under analysis. On the contrary, OBDM promotes a new method, where the rules to be checked are derived directly from the ontology, and have been validated by the process of building the conceptual model of the domain.

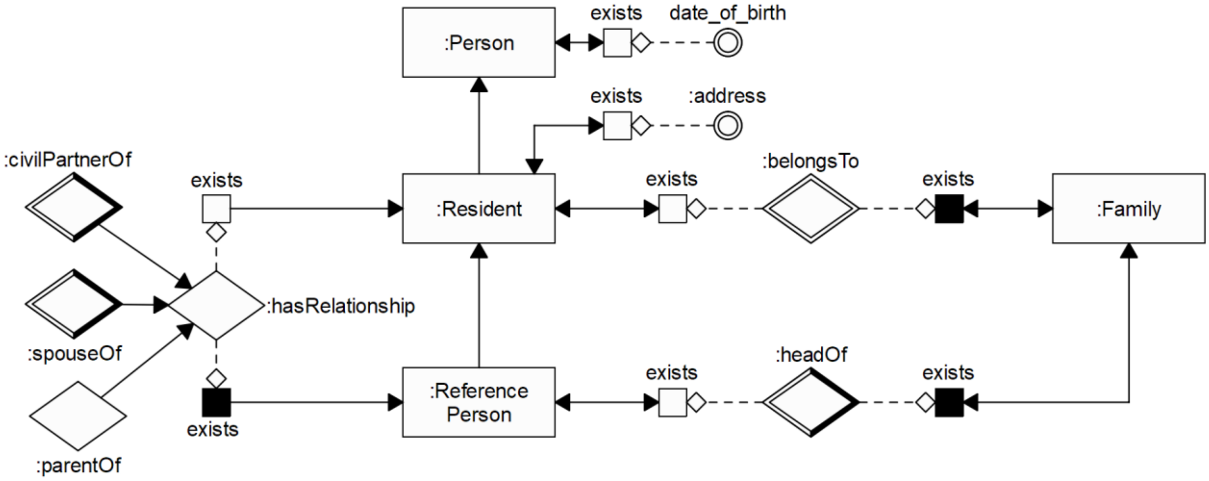
Accuracy deals with questioning the extent to which data accurately represent the real world. In other words, poor data accuracy shows that the information system, although perhaps consistent, represents a wrong state of the real-world. The logical foundation of the OBDM paradigm allows us to develop a new approach to data accuracy. Indeed, OBDA can nicely distinguish between the knowledge about how the world is supposed to be shaped (the ontology), and the knowledge that the current data possess about the world (the sources). From the language point of view, while the sentence α in classical logic specifies that the property asserted by α is true, the sentence $\mathcal{K}$(α) asserts that the system knows that is true, i.e., that holds in every model of the system. Using these sentences, we can impose sophisticated quality requirements. For example, we can express sentences comparing the ontology and the knowledge possessed by the system, and this can give us important insights on the accuracy of the data sources at hand.

Completeness deals with the question whether a given source contains all the relevant data about a certain phenomenon. Again, we can show that assessing completeness can be done by comparing the ontology and the knowledge possessed by the system (in particular the knowledge deriving from a given source). For example, by using OBDM, we can ask the system if all the data regarding the instances of a given concept in the ontology is stored in a certain data source. Every deviation from this property is a formal indication that completeness is compromised for the data source under considerations.

4. The experience

In the proof of concept carried out in Istat, we focused on the portion of the domain related to persons, including residential data, and their family relationships. In this domain, a family is constituted by a group of persons linked to the reference person (head of the family) by a family relationship and having the same residential address. Each person can belong to at most one family. A family can be constituted by a single person. In this case, the only member of the family is indeed the reference person. The relationship with the head of the family can be of several kinds, including for instance parental relations, marital and civil partnerships. For each person some information are collected, among which the date of birth and place of birth. All residents must have a family to which they belong and a residential address. Figure 1 depicts a simplified version in Graphol (Console et al. 2014) of the ontology.

Figure 1. Graphical representation of an excerpt of the ontology



The data sources connected to the ontology are those containing information about persons and families. To link such data sources to the ontology we specified about 100 mapping assertions which have been validated and optimized by Istat experts.

The concept Person and the attributes date_of_birth and address are linked to the data source table Tab_pers[idp,date_of_birth,address] by means of the following mapping assertions:

$$\begin{aligned}
&\text{Tab\_pers}(x,y,z) && \rightarrow \text{Person}(x)\\
&\text{Tab\_pers}(x,y,z) \wedge y \neq \text{NULL} && \rightarrow \text{date\_of\_birth}(x,y)\\
&\text{Tab\_pers}(x,y,z) \wedge z \neq \text{NULL} && \rightarrow \text{address}(x,z)
\end{aligned}$$

While the data about families and their members are retrieved from the Tab_family[idf,idp,head_flag] through the following mapping assertions:

$$\begin{aligned}
&\text{Tab\_family}(x,y,z) && \rightarrow \text{Family}(x)\\
&\text{Tab\_family}(x,y,z) && \rightarrow \text{belongsTo}(y,x)\\
&\text{Tab\_family}(x,y,z) \wedge z = \text{TRUE} && \rightarrow \text{headOf}(y,x)
\end{aligned}$$

We used the MASTRO system for OBDA for carrying out several data quality checks.

We start presenting an example of how the OBDA paradigm can be used for identifying consistency issues in the underlying data sources. As mentioned before, members belonging to the same family cannot have different residential addresses. This is expressed over the ontology by means of the following consistency constraint:

$$\begin{aligned}
\forall x,y,z,v,w \quad & \text{Family}(x) \wedge \text{belongsTo}(y,x) \wedge \text{address}(y,v)\\
& \wedge \text{belongsTo}(z,x) \wedge \text{address}(z,w) \wedge v \neq w \quad \rightarrow \bot
\end{aligned}$$

The verification of consistency constraints is automatically performed by the MASTRO system and allows to identify the data at the sources violating the constraints. Subsequently, such data return to the GSBPM Process phase to ensure correction and validation.

For verifying the accuracy of data, the OBDA specification is enriched with ad-hoc constraints. For instance, in our domain, the reference person of each known family and his/her address must be known. The following rule describes this constraint.

$$\forall x \quad \mathcal{K}(\text{Family}(x)) \quad \rightarrow \quad \exists y,z\, \mathcal{K}(\text{headOf}(y,x) \wedge \text{address}(y,z))$$

Completeness constraints compare the knowledge (inferred and not) of the ontology with the information stored in the data source. The Tab_pers table should contain all the known persons of the Italian Integrated System of Statistical Registers. To verify this constraint, the following rule is added to the OBDA specification which asserts that persons known at the ontology level must be stored in the Tab_pers table.

$$\forall x \quad \mathcal{K}(\mathrm{Person}(x)) \rightarrow \exists y,z\ \mathrm{Tab\_pers}(x,y,z)$$

To answer a query **q** posed by the user over the ontology, the system rewrites **q** in a set of queries **Q** encoding the knowledge in the ontology (Calvanese et al. 2007). In particular, when asking for the persons known by the ontology, the query Person(x) is rewritten by the system in the set of queries including the query belongsTo(x,y), asking for all the persons that are also a member of a family. Hence, according to the mapping assertions shown earlier, to retrieve all the persons from the data sources, the system queries both the Tab_pers table and the Tab_family table. Thanks to this procedure, the constraint given above can detect whether the Tab_family table contains persons not stored in the Tab_pers table, hence highlighting an incompleteness issue of the Tab_pers table.

5. Conclusions

In this paper, we presented our experience of using OBDM for data quality assessment in the context of the Italian Integrated System of Statistical Registers and demonstrated how this paradigm can be effectively used for this purpose. In particular, we showed how OBDM supports the designer in checking consistency, accuracy and completeness of data sources. We plan to continue our work along several directions. First, we would like to extend both the ontology and the mapping for capturing the whole ISSR domain and data sources. This will allow us to better evaluate both the approach and the tools from the performance point of view. In case this evaluation shows critical aspects in performance we plan to study suitable specialized optimization techniques. Another interesting direction is to investigate how OBDM can help in addressing other data quality dimensions such as confidentiality.

5. References

Batini. C. and Scannapieco, M. (2016). Data and Information Quality - Dimensions, Principles and Techniques. Data-Centric Systems and Applications, Springer.

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Rosati, R., Ruzzi, M., and Savo, D. F. (2011), The MASTRO system for ontology-based data access. Semantic Web Journal.

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., and Rosati, R. (2005), DL-Lite: Tractable Description Logics for Ontologies. In Proc. of, AAAI 2005: 602-607.

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., and Rosati, R. (2007), Tractable reasoning and efficient query answering in description logics: The DL-Lite family. Journal of Automated reasoning, 39(3), 385-429.

Console, M. (2016), Ontology-based data quality: principles, methods, and algorithms. PhD Thesis, Sapienza University of Rome.

Console, M., Lembo, D., Santarelli, V., and Savo, D. F. (2014), Graphol: Ontology Representation through Diagrams, In Proc. of DL 2014.

Console, M. and Lenzerini, M. (2014a), Data Quality in Ontology-based Data Access: The Case of Consistency. AAAI, 1020-1026.

Console, M. and Lenzerini, M. (2014b), Reducing global consistency to local consistency in Ontology-based Data Access. ECAI, 219-224.

CrowdFlower (2017), 2007 Data Scientist Report. Available at: https://visit.crowdflower.com/WC-2017-Data-Science-Report_LP.html.

De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., and Rosati, R. (2018), Using Ontologies for Semantic Data Integration. A Comprehensive Guide Through the Italian Database Research, 187-202.

Fan, W. and Geerts, F. (2012), Foundations of Data Quality Management. Synthesis Lectures on Data Management, Morgan & Claypool Publishers.

Lenzerini, M. (2011), Ontology-based data management. In Proc. of CIKM 2011. Noy, N. F., Doan, A., and Halevy, A. Y. (2005). Semantic Integration, AI Magazine 26(1): 7-10.

Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., and Rosati, R. (2008). Linking Data to Ontologies. J. Data Semantics 10: 133-173.

Wand, Y. and Wang, R. Y. (1996), Anchoring Data Quality Dimensions in Ontological Foundations. Commun. ACM 39(11): 86-95.