The Italian Integrated System of Statistical Registers: Design and Implementation of an Ontology-based Data Integration Architecture

Roberta Radini, Monica Scannapieco, Laura Tosco

1. INTRODUCTION

Istat has engaged a modernization programme that includes a significant revision of the statistical production. The principal concept underlying such an important change is the usage of a system of integrated statistical registers as a base for all the production surveys; this system will be in the following referred to as the Italian Integrated System of Statistical Registers (ISSR). In [1] a detailed description of the register-based statistics approach as adopted by Statistics Sweden is presented. Such an approach is surely an important reference for the design and implementation of the ISSR, though, Istat (like other non-Nordic countries) does have some peculiarities that have to be taken into account. As an example the system of integrated registers adopted by Statistics Sweden is entirely based on administrative data due to the high quality of such data; this is an assumption that turns to be too much strong for Italy where, instead the quality of administrative data is lower and less controlled and hence dedicated surveys should be taken into account to improve the quality.

2. BASIC CONCEPTS

A system of statistical registers consists of a number of registers that can be linked to each other. The ISSR has been conceptualized as consisting of:

- Base Statistical Registers (BSRs) are composed by a collection of statistics units belonging to populations relevant for official statistics. The variables characterizing such units are "core" variables, meaning that they (i) have a high identification power and (ii) are quite stable in time. In particular, they are: (i) BSR of Individuals, Families and cohabitations; (ii) BSR of Economic Units; (iii) BSR of Places; (iv) BSR of Activities.
- Extended Statistical Registers (ESR), which extend the information available for a population of a specific BSR with other variables.
- Thematic Statistical Registers (TSR), which identifies are not bound to the specification of populations, but rather they have the objective of supporting statistics referred to more than one statistical population.

Figure **1** illustrates such a definition.

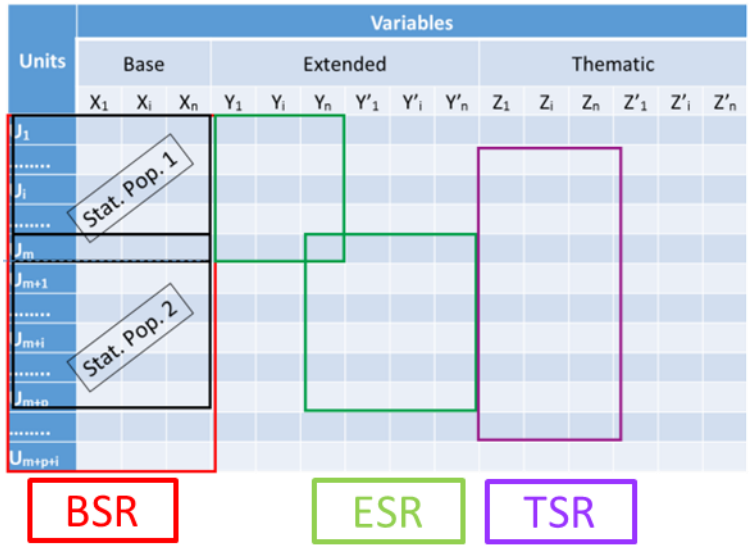

Figure 1: The notions of Base Statistical Registers, Extended Statistical Registers and Thematic Statistical Registers

An example of the reference relationships among such registers is shown in Figure 2. In particular, let us notice that (i) all BSRs but the BSR of Activity are "responsible" for the specific units and provide their identification (ID attributes); (ii) the Work Register (LEED) is a thematic register of the Activity Register.

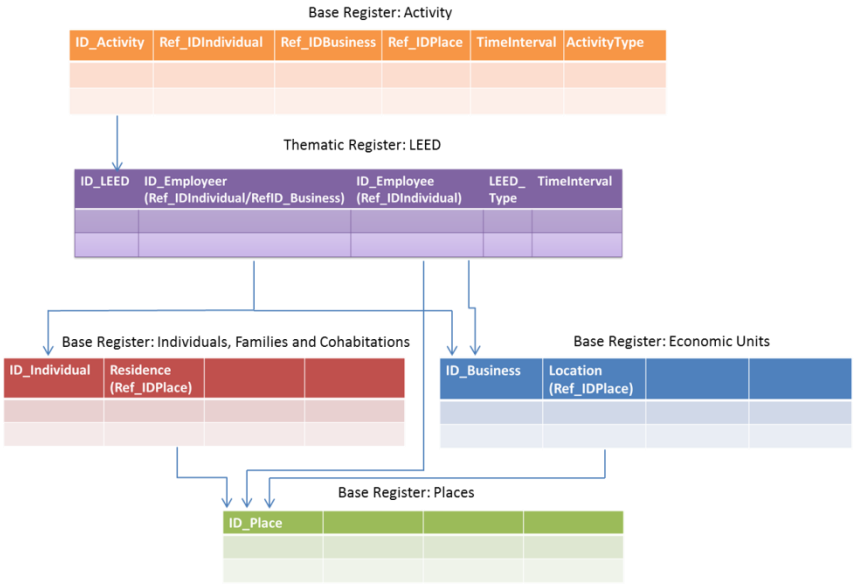

Figure 2: Overall picture of the currently identified registers of the ISSR

3. ISSR AS AN ONTOLOGY-BASED DATA INTEGRATION SYSTEM

In order to design and implement the information architecture of the ISSR, we think that the Ontology Based Data Management (OBDM) paradigm [2] is particularly fitting. This is an approach for accessing, integrating and managing different sources and consists of three layers:

- **Ontology**, as a conceptual specification of the domain of interest (represented as a computational ontology).
- **Data sources**, i.e. all the sources that have been selected as input sources to the data integration system to implement. These sources are in general heterogeneous both semantically and technologically.
- **Mapping,** which specifies the relationships among the data sources and the ontology.

The ontology layer drives the access to the system by providing services (e.g. data queries) to the final user. The system enforces a *data source transparency property* for the final user, namely the user does not need to have any notion on where data are, but it is up to the system to "rewrite" ontology queries in terms of queries to data sources.

There are several concrete examples of frameworks on how specifying mappings on the top of database-technology, involving correspondences between ontology classes, data and object properties and database tables, columns and relations respectively (see e.g. [8], [9], [9]).

The main reasons underlying the choice of this paradigm are:

- the complexity of the metadata asset (structural metadata asset or intensional data representation) in terms of hugeness and lack of a direct control (several sources are administrative ones that come with their own semantics). The use of ontologies, which permits a formalization and a machine-actionable representation of such metadata, looks promising in order to deal with such a complexity.
- The need for having an integration layer permitting to virtualize data resources and performing "on-the fly" query answering. We think that OBDM can properly answer to such a requirement of the ISSR as an alternative to rigid and materialized traditional data integration approaches like traditional data warehousing.

National Statistical Institutes have a long experience in dealing with metadata. However, OBDM has a major difference with respect to approaches typically used within NSIs for designing and implementing metadata management systems, namely: ontologies permit to represent metadata "coupled" with data, so they are not only limited to a "documentation" role but they do permit to "govern" the data integration step by ensuring the quality of integrated data.

State of the art technological solutions for data integration seem to indicate Data Virtualization (DV) as a better cost effective data integration strategy if compared to traditional Data Warehousing approaches **Errore. L'origine riferimento non è stata trovata.**.

However, though it is shareable that the time-to-market and operational costs are better in DV approaches, there are other features that should be taken into account, as shown by Table 1.

Features	DV	DW
Storage of Historical Data	NO	YES
Capture Every Change in Production Data	NO	YES (requires integration with CDC)
Multi-Dimensional Data Structures	NO	YES
Data Pre-Aggregation	NO	YES
Query performance on large amounts of data	SLOW (relative to DW)	FAST (relative to DV)
Data Integration on Demand	YES	NO
Operational Cost	LOW (relative to DW)	HIGH (relative to DV)
Time-To-Market	LOW (relative to DW)	HIGH (relative to DV)
Easy to Make Changes	YES (relative to DW)	NO (relative to DV)
Dependence on IT	LOW (relative to DW)	HIGH (relative to DV)

Table 1: Features of Data Virtualization vs. Data Warehousing from **Errore. L'origine riferimento non è stata trovata.**.

In particular, some features related to storage of historical data, multidimensional data structures and performance are better addressed by DW. Though OBDM does not imply any specific architectural approach, surely it could be effectively implemented by platforms adopting DV. However, given that the mentioned features in which the DW approach better performs are important for data management in National Statistical Institutes, we are considering to take into account hybrid solutions; as an example a DV based data architecture could use a DW for historical data storage.

From the statistical user perspective, the advantages of having an ontology-based ISSR are:

- Access to integrated data: for instance the "labour" concept has different definitions according to National Accounts, Structural Business Statistics and Labour Force Survey. Ontologies permit that such different definitions can coexist and underlying data can be accessed consistently.
- Metadata represented and accessible through an IT system: so far statistical metadata models are "not" represented in formal languages; indeed, such models are mainly described in MS Word documents, XLS files, or UML diagrams. Recent efforts have been paid towards the definition of ontologies for such models (e.g. General Statistical Information Model - GSIM ontology [3],General Statistical Business Process Model - GSBPM ontology [4] and Common Statistical Production Architecture - CSPA ontology [5]). In addition, the need for integrating such models among each other and resolve inconsistencies brought to a specific UNECE project "Implementing ModernStats Standards - Linked Open Metadata" ([6], [7]). It is nice to observe how the OWL (Web Ontology Language) representation of GSIM, GSBPM and CSPA highlighted some inconsistencies among them (and even within each model) .

- Reasoning capability: even if some concepts are not "explicitly" linked, reasoning over ontologies allows to "infer" new knowledge (e.g. new relationships). In this way, statistical users can "discover" implicit patterns that can help in understanding data for their analyses.

4. AN EXAMPLE OF ONTOLOGY MODELLING FOR BSR OF INDIVIDUALS, FAMILIES AND COHABITATIONS

A first initial effort toward the modelling of the ontology for the BSR of Individuals, Families, and Cohabitations is shown in Figure **3**.

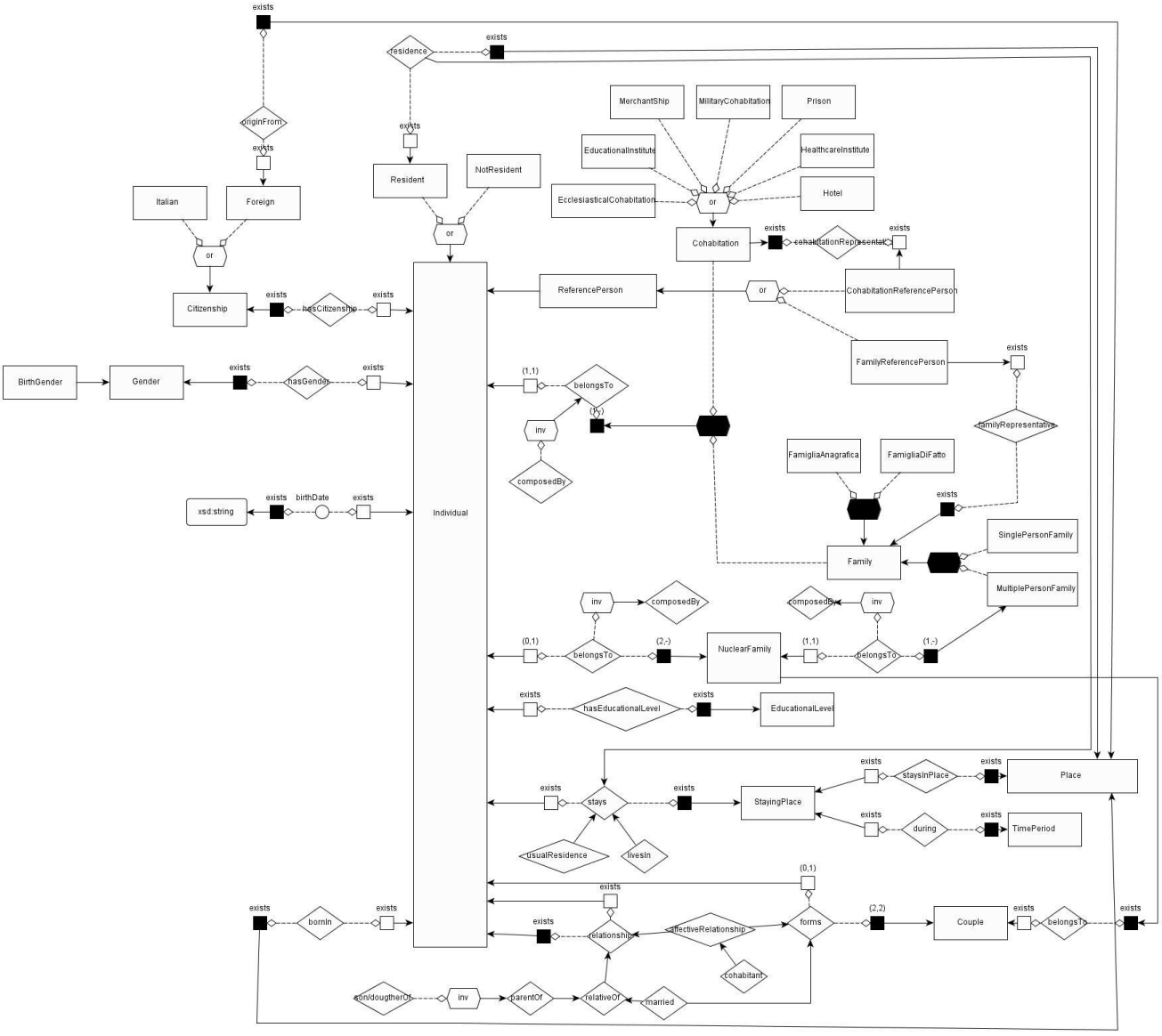

Figure 3 Ontology for the BSR of Individuals, Families and Cohabitations

The notation used to represent the ontology is Graphol ([11],[11]). It is clearly represented that the main concepts are: Individual, Family and Cohabitation.

5. QUESTIONS

We are currently in the very initial phase of designing the ISSR. The main questions we have in this phase are:

1) Ontology of BSR of Individuals, Families, and Cohabitations. Is the modelling effort going in the right direction?
2) Modelling of the Time dimension. Some concepts have duration (start time, end time) as ID attribute. As an example, for the LEED (Linked Employer Employee Data) Register (also called Work register), which is a thematic statistical register of the BSR of Activities, it is defined a basic LEED concept consisting of IDEmployer, IDEmployee, IDLabourType, IDDuration. Some concepts may have attributes (non-ID) with a duration. As an example a LEED has a LabourAddress with a duration (but can have a different one in another period of time). Some instances can be instances of a concept in a period and instances of another concept in another period. For instance a family can no –longer exists and can be split in time in two families: longitudinal analyses should be able to access to these kind of "transformations". Our need is to model Time so that all these situations (and probably further ones) could be properly represented. Is there any consolidated approach for modelling the Time dimension with such requirements?
3) Ontology complexity and architectural mapping: we are aware that in order to use the current Data Base Management Systems (DBMSs) technology we should have ontologies first-order rewriteable, so that first order queries (e.g. SQL queries) can be posed to data sources. However, we are looking for enterprise-scale architectural solutions that could support: (i) mapping specification and (ii) efficient query rewriting according to such mappings. Supposing that OWL is used for ontology modelling, is there any (efficient) SPARQL to SQL query rewriting technological solution ready to be used on an enterprise scale? We have already used OWL for data publication according to the Linked Data paradigm ([15], [16]) and we think that we could proper use it for representing the ISSR's ontologies. However, given the inefficiency of having data represented as RDF triples (which we have directly experienced), we would like to stay with RDBMs data. Is this the right approach to follow?

6. REFERENCES

[1] A. Wallgren, B. Wallgren: Register-Based Statistics. Statistical Methods for Administrative Data. Wiley, 2014.

[2] M. Lenzerini. Ontology-based data management. In Proc. of the 20th Int. Conf. on Information and Knowledge Management (CIKM 2011), pages 5–6, 2011.

[3] M. Scannapieco, L. Tosco, D. Gillman, A. Dreyer, G. Duffes: "An OWL Ontology for the Generic Statistical Information Model (GSIM): Design and Implementation". In the Proceedings of the 4th International Workshop on Semantic Statistics, http://ceur-ws.org/Vol-1654/article-03.pdf..

[4] Cotton F., Gillman D.: "Modeling the Statistical Process with Linked Metadata", SemStats 2015, available at http://ceur-ws.org/Vol-1551/article-06.pdf

[5] A. Dreyer, G. Duffes, F. Cotton: "An OWL Ontology for the Common Statistical Production Architecture". In the Proceedings of the 4th International Workshop on Semantic Statistics, http://ceur-ws.org/Vol-1654/article-06.pdf.

[6] IMS – Implementing ModernStats Standard Project http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=122323917

[7] Implementing ModernStats Standards Linked Open Metadata Design Guidelines http://www1.unece.org/stat/platform/download/attachments/129172661/HLG-MOS%20-%20IMS%20Design%20Guidelines_Jan2017.docx?version=1&modificationDate=1483969944574&api=v2
[8] K. Čerāns, G. Būmans: RDB2OWL: a language and tool for database to ontology mapping. In: Proceedings of the CAiSE 2015 Forum at the 27th International Conference on Advanced Information Systems Engineering (CAiSE 2015), Kista, Sweden, pp. 81–88 (2015)
[9] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, and G. Xiao. Ontop: Answering SPARQL Queries over Relational Databases. Semantic Web Journal. 2016.
[10] C. Civili, M. Console, G. De Giacomo, D. Lembo, M. Lenzerini, L. Lepore, R. Mancini, A. Poggi, R. Rosati, M. Ruzzi, V. Santarelli, D. Savo. MASTRO STUDIO: Managing Ontology-Based Data Access Applications. In: Proc. of the 39rd Int. Conf. on Very Large Data Bases (VLDB 2013). 2013
[11] Graphol: http://www.dis.uniroma1.it/~graphol
[12] Marco Console, Domenico Lembo, Valerio Santarelli, Domenico Fabio Savo: Graphol: Ontology Representation Through Diagrams. In Proc. of the 27th Int. Workshop on Description Logic, 2014.
[13] R. F. van der Lans: Data Virtualization for Business Intelligence Systems, 2012
[14] L. J.Pullokkaran: Analysis of Data Virtualization & Enterprise Data Standardization in Business Intelligence Working Paper Composite Information Systems Laboratory CISL# 2013-10. Sloan School of Management, MIT, Boston 2013.
[15] Christian Bizer, Tom Heath e Tim Berners-Lee: Linked Data—The Story So Far (PDF), in International Journal on Semantic Web and Information Systems, vol. 5, nº 3, 2009, pp. 1–22.
[16] Istat LOD portal: www.datiopen.istat.it