Publishing the 15th Italian Population and Housing Census in Linked Open Data

Raffaella Aracri, Stefano De Francisci, Andrea Pagano,

Monica Scannapieco, Laura Tosco, Luca Valentino

Istat – Istituto Nazionale di Statistica

{name.surname}@istat.it

Abstract. The paper describes the first project that Istat, the Italian National Institute of Statistics, has set up for publishing data in LOD on its own SPARQL endpoint. Both the publication process and the underlying technical architecture are described with a focus on design choices (e.g. the adoption of RDF Data Cube Vocabulary for multidimensional data representation and the usage of R2RML for mapping rules specification) and on the adopted technological platform.

1 Introduction

National Statistical Institutes (NSIs) play an important role as data producers, by publishing Official Statistics in the service of citizens and policy-makers. Statistical production processes are indeed intended to produce "data" as their final output. An important phase of such processes is the dissemination phase, dedicated to the design and development of publication tools aimed to reach the widest possible range of users. In this respect, the Linked Data paradigm [1] appears to be extremely promising as part of the dissemination strategy of NSIs.

The increasing Internet penetration and the subsequent proliferation of electronic data exchanges, resulted in the study of new protocols, models and formats specific for the statistical data domain. Ten years work on searching and defining "a common language and a common perception of the structure of classifications and the links between them" has originated the Neuchâtel model, the first relevant standard for data and metadata representation [2].

In recent years, further standard statistical data models have emerged, namely GSIM and SDMX. GSIM (Generic Statistical Information Model) [3] is a reference framework of internationally agreed concepts, attributes and relationships that describes the conceptual view of information relevant to Official Statistics production. SDMX (Statistical Data and Metadata Exchange) [4] is an ISO international standard, based on

XML, available since 2004. It provides a framework (i.e., models, formats, guidelines, software tools, etc.) with the purpose of supporting the exchange of data and metadata. In addition to GSIM and SDMX, DDI (Document Data Initiative) [5] is also emerging as an XML based standard that can be possibly used for data and metadata representation in the statistical domain.
With respect to data standardization activities, NSIs are moving towards two directions; on the one hand they are promoting the development of models that are "ad-hoc" to the statistical domain, like SDMX; on the other hand, they are instead working for the purpose of extending the base of data users, going also beyond the established statistical users. The approach to achieve this latter goal involves the use of worldwide standard models and formats for data sharing, mainly Semantic Web standards. Indeed, several NSIs have set up a SPARQL endpoint for publishing their data in Linked Open Data (LOD) format, including INSEE (France) [20], EL.STAT (Greece) [19] and CSO (Ireland) [21].
In this paper, we describe a project by Istat (the Italian National Institute of Statistics) aimed to publish indicators of the 2011 Population Census as Linked Open Data.
The project, named Census-LOD, is a flagship project that also has the ambitious goal to pave the way for the introduction of LOD as a stable channel for Official Statistics dissemination.
The remaining of the paper is structured as follows. After providing an overview of the dissemination activities related to the latest Italian Population and Housing Census in Section 2, we present the main contribution of the paper in Section 3 by describing the process followed to publish Census data in LOD. In Section 4, we briefly discuss the certified publication of Istat data, while in Section 5 we conclude with final remarks.

2 The 15th Italian Population and Housing Census: an Overview

The data collection phase of the 15th Italian General Population and Housing Census begun in October 2011 and ended in early 2012. The first outputs were disseminated during 2012 (namely, provisional data and legal population). After the dissemination of the legal population, the correction and validation phase for data relating to individuals, households and buildings started. A large amount of statistical analyses are being available since May 2014 through the I.Stat system, the Istat Web warehouse [6].
The reports disseminated via I.Stat have municipalities as the lowest detail of the territorial dimension. As for the previous Census editions, Istat plans to publish a defined subset of indicators also at the territorial level of Census section, which is a sub-municipality level.
Currently, the Italian territory is divided into 400000 Census sections, so this level of publication concerns a huge amount of data that, in the past Census editions, were distributed as CSV or MS Excel files. A provisional version of these files is already available on the Web with a list of 43 indicators related to the population measures.

The number of published indicators is expected to grow to around 200 by December 2014 with the inclusion of population, households, dwellings and buildings.

3 The Census-LOD Project

The Census-LOD project aims to make available the Census data at the Census section level in LOD.
The Istat dissemination architecture is so far based mainly on SDMX; in particular, the machine-to-machine data exchange is implemented by a set of Web services that return SDMX data (the system is collectively called SEP - Single Exit Point).
Following the guidelines for the enhancement of the quality of public information [7] distributed by the Agency for Digital Italy, Istat realized the need to broaden the dissemination to non-statistical/non-SDMX users and, in 2012, it started a project implementing a translator from SDMX to RDF Data Cube Vocabulary (RDF-QB) [8]. The realized translator was validated on real Istat datasets, which were selected in order to maximize their diversity with respect to both dataset number of observations and number of dimensions [9]. We are working on the inclusion of the translator into the SEP architecture with the goal of reusing the huge work on metadata already carried out for SDMX-based dissemination.
Later, the possibility of publishing data directly according to LOD paradigm, was taken into consideration independently from the SDMX-based publishing process. In particular, with the Census-LOD project, we decided to develop a platform for the LOD dissemination of: (i) the Population Census indicators at the territorial level of Census sections, linked to (ii) the "Territory" dataset describing the Italian territorial structure including regions, provinces, municipalities and Census sections and to (iii) the "Geonames" ontology that is an international ontology for the description of physical and other territorial characteristics.
In more details, the Census-LOD project consisted of three main phases, specifically:

1. Domain analysis and ontology definition;
2. Triples generation;
3. LOD publishing.

In the following sections, we describe the above phases in details.

3.1 Domain Analysis and Ontology Definition Phase

As a first step of the project, we made an in-depth domain analysis, resulting in a conceptual model of the domain of interest.
We started analyzing two datasets, named *Censpop* and *Territory*; such datasets were used to be published in previous Census editions as CSV and XLS files.
The *Censpop* dataset describes the population Census indicators at the territorial level of Census section. **Table 1** shows an excerpt of one of the *Censpop* files; only three of the total amount of about 200 indicators are shown, namely: (i) total population, (ii) male population and (iii) female population.

Population						
Province Code	Municipality Code	Census Section	Total Population	Population Male	Population Female	...
5	1	50010000005	9	6	3	...
5	5	50050000343	34	17	17	...
5	118	51180000013	13	7	6	...
5	120	51200000001	292	141	151	...
5	121	51210000037	23	11	12	...

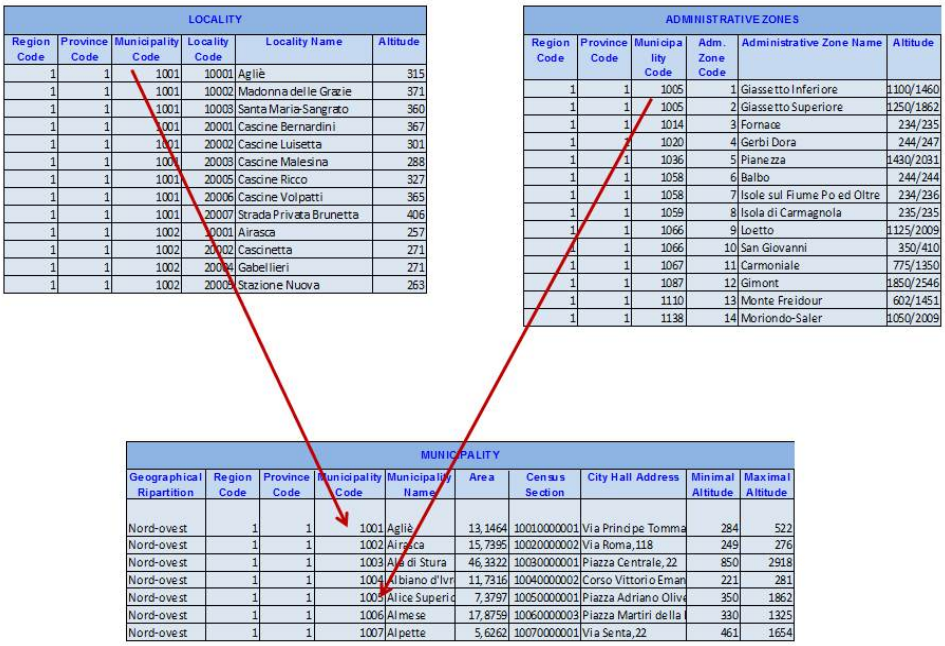
Table 1. Excerpt of *Censpop* Dataset

The *Territory* dataset describes the Italian territorial features from both administrative and geographical perspectives. **Fig. 1** shows an excerpt of one of the *Territory* file composed by several sheets, describing the territory in its different aspects as localities, municipalities, administrative zones etc.
The amount of involved data is huge: there are about 402903 Census sections, 74482 localities, 2200 Census areas, 3631 geomorphological entities and 43 indicators for each entity (e.g., "Resident Population – Males", "Resident Population – age > 74 years", "Foreigners and stateless persons resident in Italy – Males" and so on).
In order to design the ontologies necessary for the data publication, we met periodically domain experts and interviewed them. In addition, we followed an abstraction process starting from the concrete data representations that we had at hand (i.e. CSV and Excel files).
We designed two distinct ontologies: (i) the Territorial Ontology, and (ii) the Census Data Ontology.
The Territorial Ontology is an OWL [10] ontology and describes the administrative and the geographical organization of the territory. It is composed by:

- 95 entities, describing regions, provinces, municipalities, locations, census sections, special areas, special units (e.g., abbeys or hospitals), and so on.
- About 200 roles, describing relationships between entities, e.g. *appartieneACDASC* links a municipality with its sub-municipalities components. Moreover, the ontology describes for each role its cardinalities and its inverse roles. For the entities having a corresponding definition in the Geonames ontology a relation of *EquivalentTo* has been defined with the relative Geonames entity. Finally, subClassOf roles have been defined to descript the hierarchies between entities.

The Census Data Ontology has been written using the Data Cube Vocabulary [8], which is in turn based on OWL: it describes the Census data in terms of measures and dimensions for a total of:

- 6 dimensions, including sex, age classes, citizenship, territory which is the territory defined in the Territorial Ontology;
- 2 measures: number of residents, foreigners and stateless resident in Italy.



LOCALITY

Region Code	Province Code	Municipality Code	Locality Code	Locality Name	Altitude
1	1	1001	10001	Agliè	315
1	1	1001	10002	Madonna delle Grazie	371
1	1	1001	10003	Santa Maria-Sangrato	360
1	1	1001	20001	Cascine Bernardini	367
1	1	1001	20002	Cascine Luisetta	301
1	1	1001	20003	Cascine Malesina	288
1	1	1001	20005	Cascine Ricco	327
1	1	1001	20006	Cascine Volpatti	365
1	1	1001	20007	Strada Privata Brunetta	406
1	1	1002	10001	Airasca	257
1	1	1002	20002	Cascinetta	271
1	1	1002	20004	Gabellieri	271
1	1	1002	20005	Stazione Nuova	263

ADMINISTRATIVE ZONES

Region Code	Province Code	Municipality Code	Adm. Zone Code	Administrative Zone Name	Altitude
1	1	1005	1	Giassetto Inferiore	1100/1460
1	1	1005	2	Giassetto Superiore	1250/1862
1	1	1014	3	Fornace	234/235
1	1	1020	4	Gerbi Dora	244/247
1	1	1036	5	Pianezza	1430/2031
1	1	1058	6	Balbo	244/244
1	1	1058	7	Isole sul Fiume Po ed Oltre	234/236
1	1	1059	8	Isola di Carmagnola	235/235
1	1	1066	9	Loetto	1125/2009
1	1	1066	10	San Giovanni	350/410
1	1	1067	11	Carmoniale	775/1350
1	1	1087	12	Gimont	1850/2546
1	1	1110	13	Monte Freidour	602/1451
1	1	1138	14	Moriondo-Saler	1050/2009

MUNICIPALITY

Geographical Ripartition	Region Code	Province Code	Municipality Code	Municipality Name	Area	Census Section	City Hall Address	Minimal Altitude	Maximal Altitude
Nord-ovest	1	1	1001	Agliè	13,1464	10010000001	Via Principe Tomma	284	522
Nord-ovest	1	1	1002	Airasca	15,7395	10020000002	Via Roma,118	249	276
Nord-ovest	1	1	1003	Ala di Stura	46,3322	10030000001	Piazza Centrale, 22	850	2918
Nord-ovest	1	1	1004	Albiano d'Ivr	11,7316	10040000002	Corso Vittorio Eman	221	281
Nord-ovest	1	1	1005	Alice Superio	7,3797	10050000001	Piazza Adriano Olive	350	1862
Nord-ovest	1	1	1006	Almese	17,8759	10060000003	Piazza Martiri della	330	1325
Nord-ovest	1	1	1007	Alpette	5,6262	10070000001	Via Senta,22	461	1654

Fig. 1. Excerpt of Territory dataset

Fig. 2 shows an example of an observation expressed with the Data Cube vocabulary. Both ontologies make use of *meta* Ontologies as: (i) SKOS [11] for the description of classifications, (ii) ADMS [12] for the description of interoperability assets, and (iii) PROV ontology [13] for the description of the provenance of the data in terms of information about entities, activities, and people involved in the data production process.

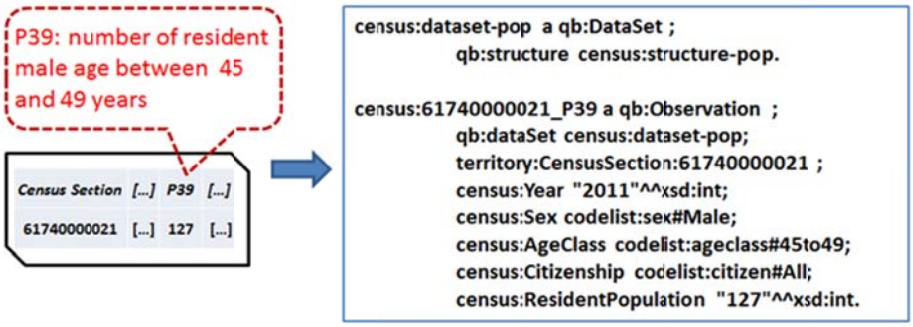


Fig. 2. Example of a Data Cube observation

3.2 Triples Generation Phase

The triples generation phase was performed in order to transform the initial datasets in the LOD format.

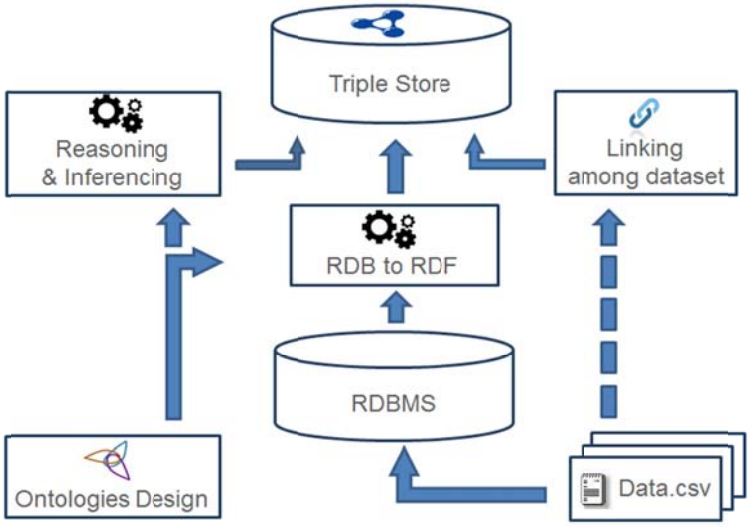


Fig. 3. Triples generation

Fig. 3 shows the designed workflow. As already mentioned, the datasets are originally produced as CSV files; therefore, the first step involves the loading of CSV files into staging tables that reflect the CSV file structure. This step can be easily performed using the direct load utility available in relational databases. Subsequently, we defined a set of mapping rules to generate the RDF triples; such rules permitted to map (i) the concepts stored into the database into (ii) the concepts defined in the ontologies. The rules definition phase should be preceded by a URI policy design for which we followed the best practices descripted in [14].

In order to specify mapping rules we used R2RML [15] that is the language for expressing customized mappings from relational databases to RDF datasets recommended by W3C. In **Fig. 4Fig. 3** an example of R2RML mapping rule is shown. In particular, the rule maps each row in the logical table (base table, view or a valid SQL Query) to a number of RDF triples.
The mapping implementation can be *on-demand* or made by *data materialization*; in the first case the triples are created as a response to a query while, in the second case, they are materialized into the triple store. The triple store loading phase is completed by adding the new triples obtained as result of a reasoning phase; these rules are specified by the defined ontologies and are not coded by explicit mapping rules.

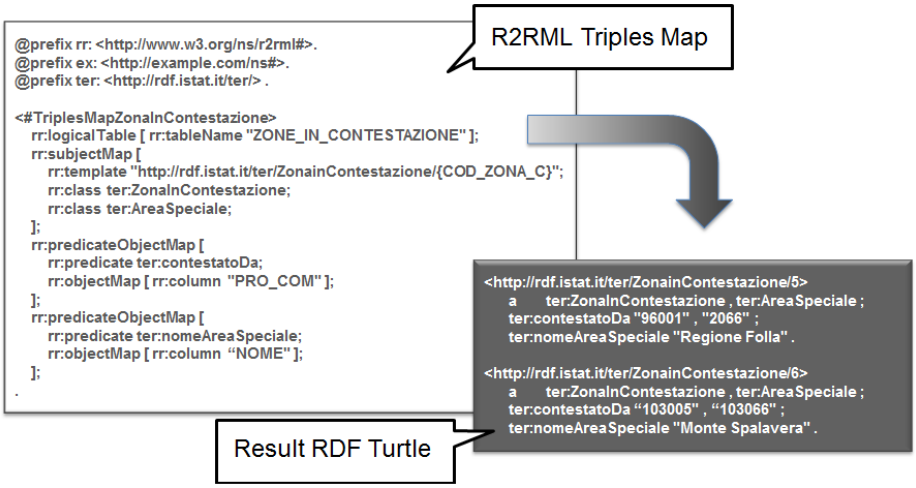


Fig. 4. Mapping of "Area in Dispute" to the corresponding subject with predicate "DisputedBy" and object "Municipaliy"

3.3 LOD Publishing Phase

The publishing phase includes a first step related to the choice of the Web application layer needed to access the triple store: specifically we provide three access points to cover the requirements of the different possible users interested to LOD data.
As shown in Fig. 5, the design of the Web site consists of three different components: (i) a SPARQL endpoint, (ii) a Linked Data Interface (Faceted/Graph browser) and (iii) an ad-hoc GUI for datasets downloading.
Advanced users are supposed to be SPARQL knowledgeable and can directly access the SPARQL endpoint, which in turn can also be used for machine-to-machine communication. Basic users can instead use the Linked Data Interface to browse data or the ad-hoc application that provides a set of predefined queries and functionalities to build customized queries.

The LOD publishing phase results in the publication of the SPARQL endpoint and the related Web containers. This phase is currently in progress; the expected data for Census-LOD Web site publication is December 2014.

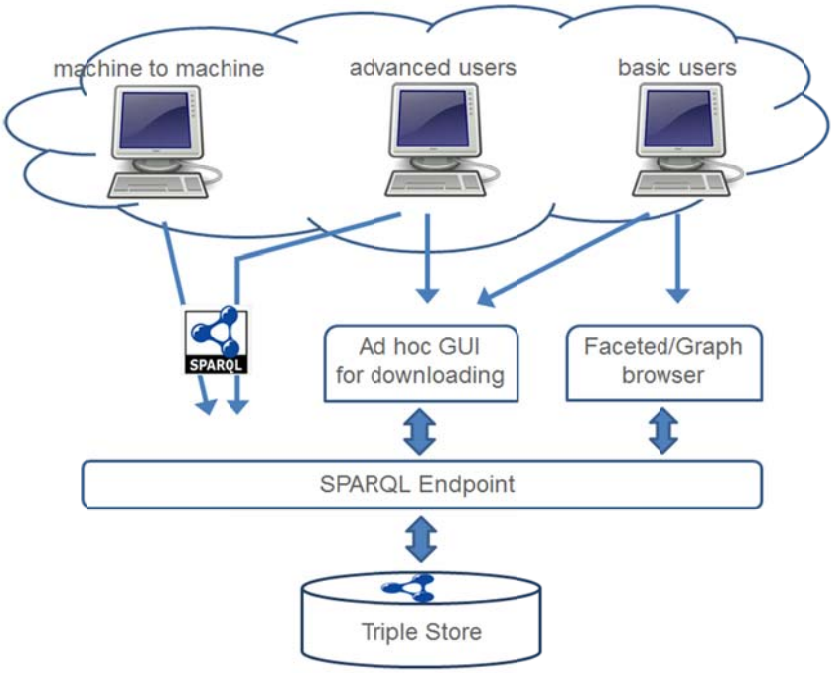


Fig. 5. Web site design

This phase also includes the definition of the technological environment. The core of the platform is the triple store. We analyzed several platforms for publishing LOD, both open and commercial. The result of such an analysis was the choice of Oracle "Spatial & Graph" solution, as triple store and SPARQL query engine, being it fully compliant with the IT infrastructure already existing in Istat. This solution has the major advantage to scale up to billions of triples [9], which is an important requirement for the platform that will support the Istat LOD dissemination channel.
A further benefit of the Oracle platform, is the usage of the R2RML language [15] in order to have a way to specify mapping rules that is generalized, i.e. independent on the specific platform.
The other important architectural choice concerns the dissemination platform; as mentioned in the previous section, we plan to deploy a SPARQL endpoint and two different interfaces for basic and advanced users. The SPARQL endpoint we use is Apache Jena Adapter (Joseki API) an open platform integrated with the Oracle SPARQL query engine. As Graph Browser, we adopted the ELDA framework [17], the open source implementation of the Linked Data API specification released by Epimorphics.

The Linked Data API provides a configurable way to access RDF data using simple RESTful URLs that are translated into queries to a SPARQL endpoint. ELDA customization consists of writing down, using the turtle syntax [18], an API that specifies how to translate URLs into queries.
Fig. 6 shows the complete technological architecture.

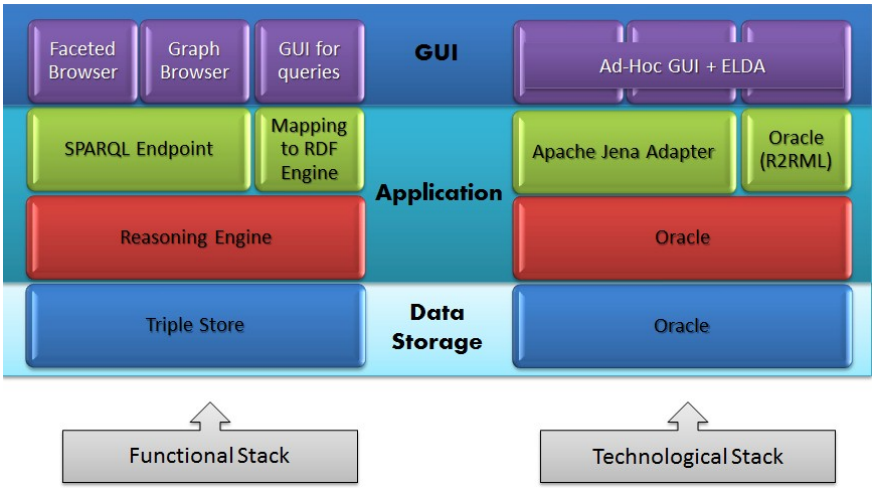


Fig. 6. Functional and technological stacks

4 Certifying Istat Data

People make trust judgments based on data provenance that may or may not be explicitly offered to them. Since the data represented in LOD format, are accessible via machine-to-machine communication, it is necessary to explicitly represent provenance information in an accessible way to machines, so that software agent can make trust judgments.
This issue has been recognized extremely important, and, indeed, the Italian guidelines for the enhancement of public information [7], point out the issue of data provenance suggesting the usage of the W3C PROV Ontology [13]. We used the PROV framework to certify the origin of the published data and the role of Istat as official data producer. In more details, we associated to data provenance metadata that specify who is responsible for the data, which are the entities, activities and agents involved in the generation/manipulation processes, allowing to certify data quality and reliability.

5 Conclusions

Census-LOD is the first project that deploys Istat data on a SPARQL endpoint that is owned by Istat itself. The publication of Census indicators at the sub-municipality

level will be available in December 2014. Published data will be enriched with information related to the administrative and geo-morphological division of the national territory. All published data will be accompanied by their own formal semantics specification through the Territory and Census ontologies.
The LOD-based data dissemination will provide several benefits, including:

- Fostering harmonization of information concepts also within Istat. Indeed, a schema integration step is necessary in order to publish data and this will also provide a positive feedback by forcing statistical processes to share a common data semantics.
- Improving machine-to-machine data provisioning by Istat, by providing an RDF-based channel in addition to the already available SDMX one.
- Providing final users with advanced functionalities like navigational querying and information discovery.

References

1. Linked Data: http://linkeddata.org/
2. Neuchâtel Model: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=14319930
3. GSIM: http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model
4. SDMX: http://sdmx.org/
5. DDI: http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/
6. I.Stat: http://dati.istat.it
7. Linee Guida per la valorizzazione del patrimonio informativo pubblico (in Italian): http://www.agid.gov.it/sites/default/files/linee_guida/patrimoniopubblicolg2014_v0.6.pdf, 2014.
8. RDF Data Cube Vocabulary: http://www.w3.org/TR/2013/CR-vocab-data-cube-20130625/, 25 June 2013
9. Aracri R., De Francisci S., Pagano A., Scannapieco M., Tosco L., Valentino L.. : “Integrating Statistical Data with the Semantic Web: The ISTAT Experience” negli Atti della Congresso Nazionale AICA “Frontiere Digitali: dal Digital Divide alla Smart Society" 8-20 Settembre 2013
10. Ontology Web Language (OWL): http://www.w3.org/TR/owl-ref/, 10 February 2004
11. Simple Knowledge Organization System (SKOS): http://www.w3.org/TR/2009/REC-skos-reference-20090818/, 18 August 2009
12. Asset Description Metadata Schema (ADMS): https://joinup.ec.europa.eu/asset/adms/home
13. PROV Ontology: http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/, 30 April 2013.
14. W3C URI guidelines: http://www.w3.org/2011/gld/wiki/223_Best_Practices_URI_Construction, 2011.
15. RDB to RDF Markup Language (R2RML): http://www.w3.org/TR/r2rml/, 27 September 2012

16. Michel F., Montagnat J., Faron-Zucker C: “A survey of RDB to RDF translation approaches and tools”, Research report - http://hal.archives-ouvertes.fr/hal-00903568 November 2013
17. ELDA: http://www.epimorphics.com/web/tools/elda.html
18. TTL: http://www.w3.org/TeamSubmission/turtle/, 28 March 2011
19. EL.STAT RDF site: http://linked-statistics.gr/
20. INSEE RDF site: http://rdf.insee.fr/
21. CSO RDF site: http://data.cso.ie/