Publishing Census Data as Linked Open Data

Monica Scannapieco, R. M. Aracri, S. De Francisci, A. Pagano, L. Tosco, L. Valentino

Istituto Nazionale di Statistica – ISTAT
Official Statistics & Data Dissemination

• “Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation.”

  [UN Statistical Division - Fundamental Principles of Official Statistics, Principle 1]

• Data dissemination is a fundamental phase of statistical production processes
Data Dissemination: Models

Data and metadata standardization in the statistical domain:

- **Neuchâtel model**: 10-years work on “a common language and a common perception of the structure of classifications and the links between them”

- **GSIM (Generic Statistical Information Model)**: reference framework of internationally agreed definitions, attributes and relationships that describe the pieces of information that are used in the production of official statistics (information objects)

- **SDMX (Statistical Data and Metadata Exchange)**: ISO international standard, based on XML, available since 2001

- **DDI (Document Data Initiative)**, based on XML, supports the entire research data life cycle (SDMX is mainly oriented to data dissemination)
Istat Data Dissemination

• Istat dissemination architecture based on SDMX:
  – Compliant to Eurostat SDMX Reference Infrastructure
  – SDMX download of data available on Web Warehouse I.stat (http://dati.istat.it)
  – SEP (Single Exit Point) for SDMX-based machine-to-machine communication

• Need to broaden the dissemination to non-statistical/non-SDMX users

• In 2012, the IS-LOD (Istat LOD) project started!
  – ICT Directorate
The IS-LOD Project

Experimental Projects

Production Projects Design

Production Projects Implementation

[2012] [Jan-June 2013] [July 2013-On-going]

• Production projects:
  – SDMX-to-DataCubeVocabulary Translator to be integrated with SEP under a Eurostat grant
  – Official Classifications in LOD, jointly with the Italian Agency for IT (Agenzia per l’Italia Digitale)
  – Census LOD: Population Census Data in LOD
Census-LOD: Data Description

- **Censpop dataset**: describing the population Census indicators, at the territorial level of Census section
  - Published in the past as CSV files or as XLS files
    (http://www.istat.it/it/archivio/104317)

- **Territory dataset**: describing the Italian territorial features from both administrative and geographical perspectives

- **Street dataset**: describing streets with their denominations, civic numbers, etc.
<table>
<thead>
<tr>
<th>COD_REG</th>
<th>COD_PRO</th>
<th>COD_ISTAT</th>
<th>PRO_COM</th>
<th>NOME</th>
<th>ALTITUDINE MINIMA</th>
<th>ALTITUDINE MASSIMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>1005005</td>
<td>5005</td>
<td>Asti</td>
<td>110</td>
<td>295</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>3013004</td>
<td>13004</td>
<td>Albes con Cassano</td>
<td>370</td>
<td>1270</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>5026052</td>
<td>26052</td>
<td>Ormelle</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>97</td>
<td>3097001</td>
<td>97001</td>
<td>Abbadia Lariana</td>
<td>199</td>
<td>1700</td>
</tr>
<tr>
<td>8</td>
<td>99</td>
<td>8099019</td>
<td>99019</td>
<td>Torriana</td>
<td>78</td>
<td>455</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COD_PRO</th>
<th>COD_COM</th>
<th>PRO_COM</th>
<th>SEZ2001</th>
<th>SEZIONE</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td>5001</td>
<td>500100000005</td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5005</td>
<td>500500000343</td>
<td>343</td>
<td>34</td>
<td>17</td>
<td>17</td>
<td>12</td>
<td>15</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>118</td>
<td>5118</td>
<td>51180000013</td>
<td>13</td>
<td>13</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>120</td>
<td>5120</td>
<td>51200000001</td>
<td>1</td>
<td>292</td>
<td>141</td>
<td>151</td>
<td>104</td>
<td>133</td>
<td>7</td>
<td>45</td>
</tr>
<tr>
<td>5</td>
<td>121</td>
<td>5121</td>
<td>51210000037</td>
<td>37</td>
<td>23</td>
<td>11</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Street territory census pop

Monica Scannapieco, LOD, Rome, 20-21/02/2014
Census-LOD: Data Size

• How many data are involved?
  • 402,903 Census Sections
  • 74,482 Localities
  • 2,200 Census Areas
  • 3,631 Geomorphological entities
  • And others classes …

• 43 indicators for each entity:
  • Resident Population – Males
  • Resident Population – age > 74 years
  • Foreigners and stateless persons resident in Italy – Males
  • …
Census-LOD: Test Workflow

• Test project as a first step
• Implemented in Datalift (http://datalift.org/), platform including several tools supporting the whole datasets publication process
• The workflow produced as a result of this phase followed (part of) the process expected by the usage of this platform, namely:

1. Loading the datasets from CSV files into the platform
2. Loading the ontologies modeled as OWL ontologies into the platform
3. Direct mapping
4. URI Policy Design
5. RDF triples generation
6. Linking among datasets
7. Publishing
8. Applications and Visualization
Census LOD: Implementation Issues

• Issues:
  • Large amount of data
  • Complex Ontology
  • Annotations required for all variables (Dissemination Database)

• Activities in progress:
  • New platform definition with RDF graph store that can scale up to billions of triples, supporting bulk and incremental load
  • Use of a «general purpose mapping language»: R2RML (RDB to RDF Mapping Language)
Census-LOD: Production Workflow

- Ontologies Design
- Ontologies Publish
- Mapping R2RML
- Reasoning & Inferencing
- GUI Design and Implementation

[Diagram showing the workflow]

Monica Scannapieco, LOD, Rome, 20-21/02/2014
Mapping Examples

Example D2RQ Mapping
@prefix map: <#> .
@prefix ter: <http://rdf.istat.it/ter/> .
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
map:ZonaInContestazione a d2rq:ClassMap;
d2rq:dataStorage map:database;
d2rq:uriPattern "ter/ZonainContestazione/@@ZONE_IN_CONTESTAZIONE.COD_ZONA_C|urlify@@";
d2rq:class ter:ZonaInContestazione;
d2rq:class ter:AreaSpeciale;
d2rq:classDefinitionLabel "Zone in contestazione";
map:contestatoDa a d2rq:PropertyBridge;
d2rq:belongsToClassMap map:ZonaInContestazione;
d2rq:property ter:contestatoDa;
d2rq:propertyDefinitionLabel "Codice Comune contestatario";
d2rq:column "ZONE_IN_CONTESTAZIONE.PRO_COM" .

Example R2RML mapping
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix ex: <http://example.com/ns#>.
@prefix ter: <http://rdf.istat.it/ter/> .

<#TriplesMapZonaInContestazione>
rr:logicalTable [ rr:tableName "ZONE_IN_CONTESTAZIONE" ];
rr:subjectMap [(
  rr:template "http://dati.istat.it/ter/ZonainContestazione/{COD_ZONA_C}";
  rr:class ter:ZonaInContestazione;
  rr:class ter:AreaSpeciale;
);
nr:predicateObjectMap [(
  rr:predicate ter:contestatoDa;
  rr:objectMap [ rr:column "PRO_COM" ];
);].

Result (Turtle)
<http://dati.istat.it/ter/ZonainContestazione/5>
a ter:ZonaInContestazione , ter:AreaSpeciale ;
ter:contestatoDa "96001" , "2066" ;
ter:nomeAreaSpeciale "Regione Folla" .

Mapping of «Area in Dispute»
to the corresponding subject with predicate «DisputedBy»
and object «Municipality»
Ontologies (1)

Two distinct Ontologies (so far):

- Territorial Ontology
- Census Data Ontology

Common features:

- OWL Ontologies
- Use of Meta Ontologies:
  - **SKOS**: skos:Concept, …
  - **ADMS**: adms:AssetRepository, …
  - **Data Cube Vocabulary**: qb:DataSet, qb:Observation, …
  - **PROV**: prov:wasGeneratedBy, …
  - **GeoNames**: gn:name, gn:countryCode, gn:parentCountry, …
Ontologies (2)

Territorial Ontology

Description of principal classes of the domain, as:

- **Administrative**
  - Region
  - Province
  - Municipality

- **Geographical-Statistical**
  - Location
  - Census Section

- **Special Areas**
  - Contested Zone
  - Administrative Island

- **Special Units**
  - Abbey
  - Hospital
  - Climatic Colony
Census Data Ontology

Use of *RDF Data Cube Vocabulary* that allows to publish multi-dimensional data

**DIMENSIONS**
- Sex
- Age
- Marital Status

**MEASURE**
- Resident Population
- Number of dwellings

**DIMENSIONS**
- Construction Period
- Intended Use
- Number of floors
Certifying Istat Data

- Istat data are the results of established methodological procedures: Official Statistics has a precise meaning in terms of quality and trust of the statistical information product

- We used the W3C PROV Ontology as a structured description of the provenance of the data we intend to publish
  - Where data come from
  - Official data sources according to European and National regulation
  - Domain standard conformance (e.g., variant and version of a statistical classification)
  - ...
# Platform Requirements

<table>
<thead>
<tr>
<th></th>
<th>Oracle</th>
<th>D2RQ</th>
<th>Virtuoso Open Source edition</th>
<th>DataLift + Sesame</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ontology Data Mapping</strong></td>
<td>YES (R2RML)</td>
<td>YES (proprietary &amp; R2RML)</td>
<td>YES (proprietary &amp; part of R2RML)</td>
<td>Yes (direct mapping)</td>
</tr>
<tr>
<td><strong>Storing RDF Triples</strong></td>
<td>Yes (billions of triples)</td>
<td>NO (mapping on-demand with relational db)</td>
<td>Yes</td>
<td>Yes (small triplestore)</td>
</tr>
<tr>
<td><strong>Querying/Reasoning</strong></td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td><strong>SPARQL Endpoint</strong></td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td><strong>Scalability</strong></td>
<td>YES</td>
<td>Depends on the used db</td>
<td>?</td>
<td>NO</td>
</tr>
<tr>
<td><strong>Integration with Istat Environment</strong></td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
</tbody>
</table>
Concluding Remarks

• Cens-LOD is the first production process that deploys Istat data on an Istat SPARQL Endpoint
  • 2014: Publication of CensPop and Territory
  • 2015: Addresses

• LOD-based data dissemination will allow:
  • Machine-to-machine data provisioning by Istat (currently only SDMX datasets via SEP)
  • Widening the range of Istat data users
  • Improving efficiency of data exchange flows with Italian administrations
  • …and much more 😊!