Dealing with Open Data at ISTAT

First Steps Towards a Perfect Data Portal:

a Cutting-edge Approach for Dealing with Open Data

Giovanni A. Barbieri
Stefano De Francisci
Italian National Statistical Institute (Istat)
• Our proposal is simple: [...] the federal government [...] is to provide data that is easy for others to reuse, rather than to help citizens use the data in one particular way or another

• Open infrastructures that enable citizens to make their own uses of the data

• Reverse the current policy, which is to regard government websites themselves as the primary vehicle for the distribution of public data, and open infrastructures for sharing the data as a laudable but secondary objective [Robinson,Yu, Zeller and Felten 2008]
• Government information that is nominally publicly available is in fact difficult to access either because it is not online or, if it is online, because it is not available in useful and flexible formats [Brito 2008]

• “Structured data”
  – Associated structured XML file would allow a user to sort the data by ascending or descending date, alphabetically by headline or author, by number of words, and in many other ways
Why is Linked Data an Opportunity?

Linked Data as a semantically rich paradigm for data representation

- Rich enough for the strict requirements of Official Statistics
- Formal and well-defined data structures, i.e. ontologies

Linked Data as an international standard (W3C)

- Tools availability and independence
- Beyond statistical users
- RDF: Resource Description Framework (W3C) (subject-predicate-object)
## The 5-Star Model

<table>
<thead>
<tr>
<th>Data</th>
<th>Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Data</td>
<td>*</td>
<td>Open License</td>
</tr>
<tr>
<td>Open Data</td>
<td>**</td>
<td>Readable</td>
</tr>
<tr>
<td>Open Data</td>
<td>***</td>
<td>Open Format</td>
</tr>
<tr>
<td>Linked Open Data</td>
<td>****</td>
<td>Uniform Resource Identifier</td>
</tr>
<tr>
<td>Linked Open Data</td>
<td>*****</td>
<td>Linked Data</td>
</tr>
</tbody>
</table>

![Diagram](image)
Italian Strategies: Digital Agenda and Open Data

Publication and sharing of national public-sector information

National guidelines for the development of public-sector information

Linked data paradigm for enabling semantical interoperability within the public sector
The National Open Data Portal

Published and maintained by AgID
More than 10,000 datasets
Provided by 76 public bodies

Istat makes available through the AgID portal more than 700 different datasets

http://www.dati.gov.it/

http://dati.istat.it
A First Step, but Not Enough
In addition to I.Stat, Istat currently publishes open data according to some main data models and formats

**Excel**
Metadata described in a dedicated sheet

**CSV**
Mainly tabular, with a header that contains metadata

**SDMX**
Description of data and metadata according to shared standards

**Linked Open Data/RDF**
Data and metadata semantically described and linked to other (external) data
The Linked Open Data (LOD) web site, allows to accesses and browse data in open format based on technology and Semantic Web standards. The LOD can be inquired directly from any application and responds to the need, expressed by the community of users, to have standardized and interoperable data.

The platform allows:

- Direct access to data via Web Services
- M2M solutions (e.g. GIS-LOD)
- Data conversion
- Export to productivity tools
- Visualization by means of external tools
To increase the value of statistical data, some advanced features can be addressed through the LOD Portal.
Steps to a «perfect» data portal

**STEP 1**
*Give each class of users (human or not) the most appropriate way to use the data*

- Strong connection among data
- Clear and shared meaning associated to data
- Enabling any class of user to use data
- Open Data strategy

**STEP 2**
*Make data in open format, whatever the level of openness*

- Open Data strategy
- Multiple levels of granularity
- Diverse degrees of openness

**STEP 3**
*Enrich the data with a semantic layer, regardless of the release on public web sites*

- Strong connection among data
- Open Data strategy
- Clear and shared meaning associated to data
- Diverse degrees of openness
STEP 1

Give each class of users the most appropriate way to use the data

**Interaction Modes**

- **Guided Access**
  - Web Service
  - Guided Queries
  - Download

- **Free Access**
  - Query REST on SPARQL EndPoint
  - Free Query via SPARQL EndPoint
  - Navigation

**Freedom of access**
Interaction Modes

(Human) User technical skills

Type of interaction

Basic Intermediate Advanced

Predefined Queries (Set of simple and customizable queries)

Guided Queries

Free Queries (SPARQL Queries)

Free Query via SPARQL EndPoint

Download

Navigation

Guided Access Free Access

Freedom of access
STEP 1  Give each class of users the most appropriate way to use the data

Possible uses of the portal

Interoperability

Integration

Graphical User Interface

Interactive browsing
**STEP 2** Make data in open format, whatever level of openness

Datasets in traditional open formats (CSV, Spreadsheet, etc.) can be accessed and downloaded

<table>
<thead>
<tr>
<th>Choice of Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piemonte</td>
</tr>
<tr>
<td>Valle d'Aosta</td>
</tr>
<tr>
<td>Lombardia</td>
</tr>
<tr>
<td>Trentino Alto Adige/Südtirol</td>
</tr>
<tr>
<td>Veneto</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Choice of specific dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sicilia</td>
</tr>
<tr>
<td>Variabili censuarie per Sezione di censimento anno 2011</td>
</tr>
<tr>
<td>Variabili censuarie per Area di censimento anno 2011</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resulting Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Istat</td>
</tr>
</tbody>
</table>
STEP 3  Enrich the data with a semantic layer
Describes the administrative and geographical features of the Italian territory

Describes the measures and dimensions of the indicators w/r/t people

Describes the measures and dimensions of the indicators w/r/t households

Describes the measures and dimensions of the indicators w/r/t dwellings

Describes the measures and dimensions of the indicators w/r/t households

Administrative Boundaries
- Regions
- Provinces
- Municipalities

Statistical-Geographical Boundaries
- Localities
- Census sections

Special Areas
- Contested areas
- Enclaves

Special Units
- Abbeys
- Hospitals

Ontologies

Dimensions
- Sex
- Age
- Marital status

Measures
- Residents

Territory

Population

Dwellings

Households
Census Data Ontology

- Use of RDF Data Cube Vocabulary that allows to publish multi-dimensional data
- Dimensions:
  - Sex
  - Age class
  - Citizenship
  - Territory (defined in the Territorial Ontology)
  - Construction Period
  - Number of floors
  - …
- Measures:
  - Number of residents
  - Foreigners and stateless resident in Italy
  - Number of Housing
  - …
The future plans for the platform are targeted to the implementation of an integrated environment of data, including:

- Open data for public uses (also non-linked)
- Linked data for internal integrated uses (through the use of ad-hoc designed ontologies)

Elementary and aggregate Linked Open Data for internal and public use, available for human and machine users, related to the most part of statistics domain and possibly enriched by geographical representation
Elementary data

Aggregated data

Linked data

Linked Open Data

Open Data

Internal use (integration)

Public use (sharing)

Initial stage

Next stages
• Macroscale vs microscale modeling
  – Pseudo-Einstein (as simple as possible but not simpler)
  – Von Neumann (agent-based modeling)
• Technological constraint → enabling technology
• It widens the space of what is feasible:
  – In production: our experience with SBS.Frame
  – In analysis and research…
• A paradigm shift?
  – Statistical mechanics vs Agent-based modeling
  – Just because you can doesn’t mean you should
• Back to open data
  – From dissemination to release (“data liberation” at StatCan)
  – The regulators have need to new rules in line with the new scenarios