Dealing with Open Data at ISTAT: First Steps Towards a Perfect Data Portal

Stefano De Francisci

Italian National Institute of Statistics

stefano.defrancisci@istat.it

Abstract. The paper identifies a possible path from a traditional way of dealing with statistical data to an innovative approach based on a concept of assigning a “value” to data in relation to the possibility of sharing and reusing them as well as exploiting their meaning.

In this perspective, we present the Linked Open Data Portal of the Italian National Institute of Statistics – Istat. The portal has been designed and implemented in order to fully enable users to access and browse data in open formats, exploiting the Semantic Web technologies and standards. The Istat’s LOD portal has a relevant role to increase the value, in the above meaning, of Official Statistics data. In addition, the portal represents the first integration and dissemination system published by Istat with a full compliance to the National guidelines for the enhancement of public information.

1) Introduction

The ways to access, handle, use and archive data have been characterized from several stages of evolution along the history of Information Technologies and have undergone many transformations. From simple lists of data, accessible and usable by means of sequential access technologies, to the last frontier of unstructured data repositories, e.g. NoSql databases, each era of new technologies has proposed solutions more and more modern and advanced for storing and accessing data and information. Concepts like bank, warehouse, mart, and, more recently, pool or lake have been used as metaphoric terms to address the different ways for dealing with data. However, even if the last solution is presented by IT vendors as the “perfect” one, we know indeed that this is not actually the case.

The experience teaches that only through the combination of appropriate technologies, suitable reference models, standards shared at global scale and, above all, clarity of objectives and definition of proper contexts and boundaries, it is really possible to implement environments for processing data really fitting the specific requirements at hand. With respect to increasing the value of statistical data, the main features of such an environment should be:

- Degree of openness (referring to the Berners Lee 5 stars model [1])
- Meaning associated to data (by means of the use of semantics layers)
- Levels of granularity (from elementary data to complex indicators)
- Connection among data (i.e. interoperability)
- Way of use/reuse from both humans and machines

To manage all or some of the characteristics above and implement IT solutions effective for sharing data among Statistical Organizations and making them available to the users communities, we can define three main steps:

1) Give each class of users (human or not) the most appropriate way to use the data;
2) Make data in open format, whatever level of openness;
3) Enrich the data with a semantic layer, regardless of the data sharing on public Web sites.

The Linked Data paradigm [2] can answer efficiently and effectively to most of the features mentioned above. In this paper, we will trace a possible path to move from a traditional way for managing data to a cutting edge one with the features above described and that is concretely implemented by Linked data solutions.

2) National Context

In Italy, the Agency for Digital Italy (AgID) annually releases a number of strategic document for Public Administrations (PAs), in its role of enabler of national Public Sector Information (PSI) publication and sharing. Among such documents, AgID publishes national guidelines[1] that paves the way to the use of Open Data to publish and share data among PAs. The guidelines identifies the Linked Data paradigm as the one better enabling semantic interoperability in the collaboration between PAs.

So far there exists a National open data portal, managed by AgID. The portal can be accessed at http://www.dati.gov.it/ and contains more than 10000 datasets provided by 76 PAs. Istat provides around 700 datasets accessible through the AgID portal. Datasets are not stored locally to it, instead, they make reference to the Istat Web warehouse I.stat (available at: http://dati.istat.it).

3) Open Data in Istat: Current Scenario

Istat currently publishes open data according to some main data models and formats (levels are referred to the cited five stars model):

- CSV, mainly tabular, with an header that contains metadata (level 3).
- Excel, metadata described in a dedicated sheet (level 2).
- SDMX, data and metadata expressed according to the same model (level 3).
- Linked Open Data/RDF, data and metadata expressed according to the same model that is RDF data model, also linked to other (external) data (levels 4 and 5).

With respect to the Web systems that makes available the afore-mentioned open data, there are:

- I.stat: Web warehouse (accessible at dati.istat.it) that centralizes most of the Istat open data current provision, where it is possible to download dataset according to CSV, Excel and SDMX formats.

[1] National Guidelines for the of the Valorization of the Public Sector Information (in Italian) http://www.agid.gov.it/sites/default/files/linee_guida/patrimoniopubblicolg2014_v0.7finale.pdf. English summary available at: http://www.w3.org/2013/share-psi/workshop/krems/papers/ItalianNationalGuidelines

- SEP (Single Exit Point): Web service SOAP that allows machine-to-machine SDMX data access.
- Thematic Web Sites: where it is possible to download datasets in CSV and Excel formats.
- mIcro.STAT: Web system that gives the possibility to download microdata for public use. Microdata datasets are provided as CSV files.
- LOD-Istat Portal: SPARQL Endpoint enriched with dedicated access GUIs.

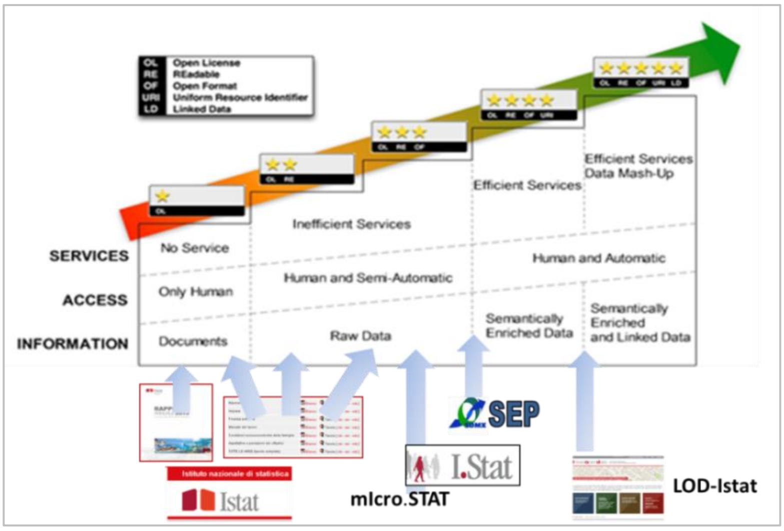


Fig. 1: Open Data at Istat Current scenario

4) Istat’s Linked Open Data Portal

The Linked Open Data Portal of Istat, available at the URL http://datiopen.istat.it, enables users to access and browse data on the basis of the Semantic Web technologies and standards.

The portal, made available since May 2015, represents the first dissemination system published by Istat with a full compliance to the guidelines for the enhancement of public information. In the first year of publication the site has had over 23,000 different visitors, with about 400,000 page views and nearly one million individual accesses.

The main goals of the platform are:

- Providing a single access point for accessing Istat’s open data;
- Similar to dati.gov.it, enabling advanced searching mechanisms, being an open data catalogue managed as a Web site (via a CMS);
- Machine-to-machine access via dedicated Open Data Rest APIs;
- Flexible querying via LOD channel by extending current LOD provision to other domains.
- Advanced navigation mechanisms (faceted browsing, graph browsing, etc.).

The platform has already taken the steps one and two defined in section 1 and is moving along the third one: the system is provided with a wide range of user functionalities and make available data in many formats (e.g.: CSV, JSON, XML, RDF) related, at the moment, to two domains of data: Territory and Population Census.

The publication of the data format in LOD is based on the definition of ontologies (formal representations, shared and explicit conceptualization of a domain of interest). The first group of Linked Open Data published by the Institute consists of data from the Census of Population and Housing in 2011. Two specific ontologies have been defined to represent such data in LOD format: (a) the ontology of territorial bases and (b) the ontology of the census data. Territorial ontology formalizes and describes the Italian territorial features from both administrative and geographical perspectives, while Census data ontology is focused on the metadata of census variables.

From the administrative point of view the territory is divided into State, Geographical Area, Region, Province, Municipality and Sub-municipal area (for the 34 municipalities of greater population size and population of not less than 100,000 inhabitants). From a statistical and geographical point of view the territory is instead partitioned into census sections, enumeration areas and localities. Special areas are also made available, such as geomorphological entities (ponds, fishing valleys, river, lagoon), administrative islands, areas in dispute and special nucleus (hospitals, abbeys, shelters, etc.).

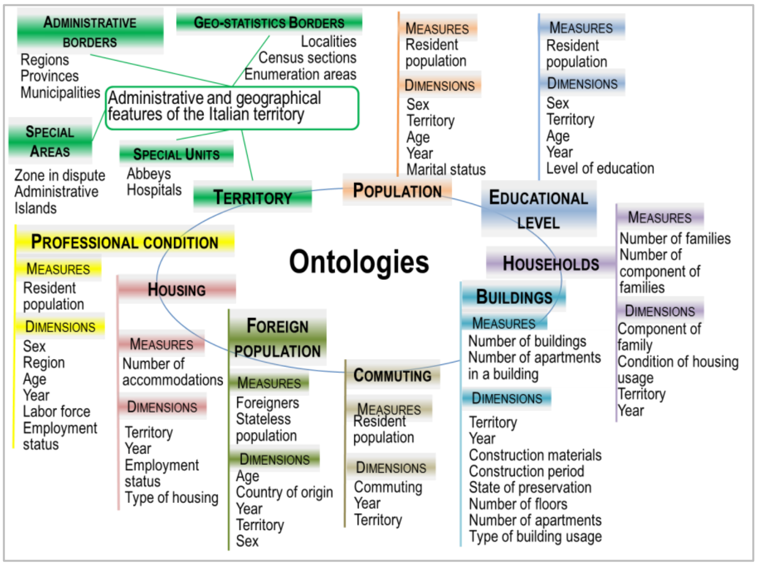


Fig. 2: Ontologies of LOD Portal

With respect to the ontology of census data a specific meta-ontology (Data Cube Vocabulary) was used, useful for the representation of multi-dimensional data. The census variables (over 150) refer to the following subject areas: population, foreign population, households, education level, employment status, commuting, dwellings, buildings. The new frontier of Linked Open Data implies the need to pay particular attention to the quality of the data. For this reason, the LOD portal allows for provenance certification, by using a specific meta-ontology (PROV-O, [3]) that enables a detailed description of the origin of the data. The published data are accompanied in particular by a set of metadata including entities, activities involved in producing a piece of data or thing, responsible for the data, the holder of the rights on the data, last modification date, published title, the reference period, license of publication, description of data, spatial data reference, etc.

5) Linked Open Data Platform: user interactions

From the point of view of the functionality to access and use data, the system provides mechanisms for selection, search, query and retrieval of data (see Figure 3).

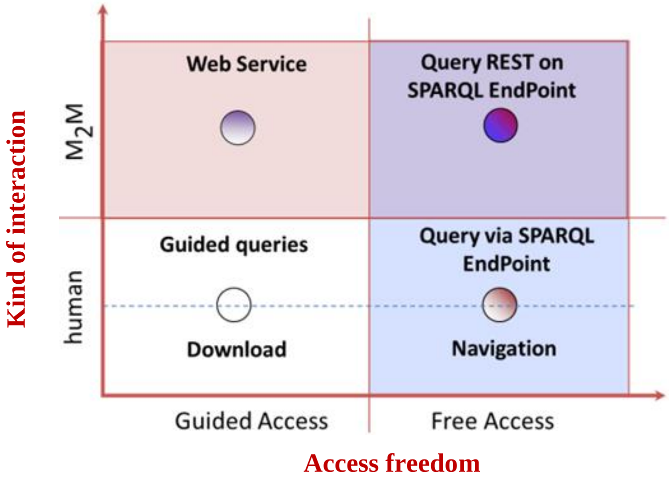
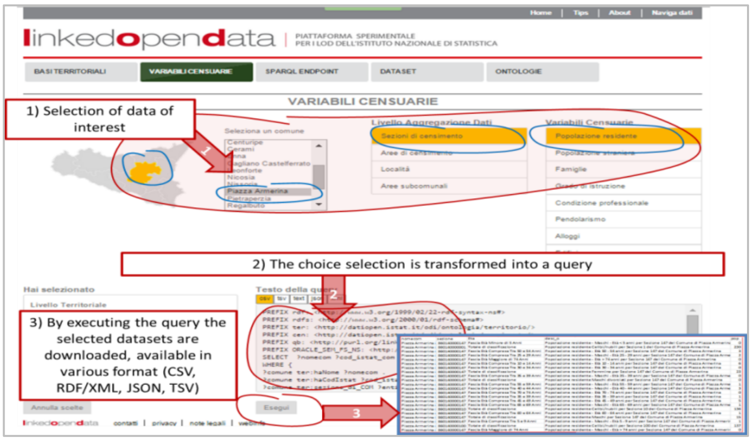


Fig. 3: User interaction quadrant

In particular, the portal is composed of:

i) Users interfaces for guided queries: they allow downloading dataset to the municipal territorial level through mechanisms that support the selection of sub-municipal areas and subject matter areas of interest, guided queries and downloads in multiple formats (see Figure 4).



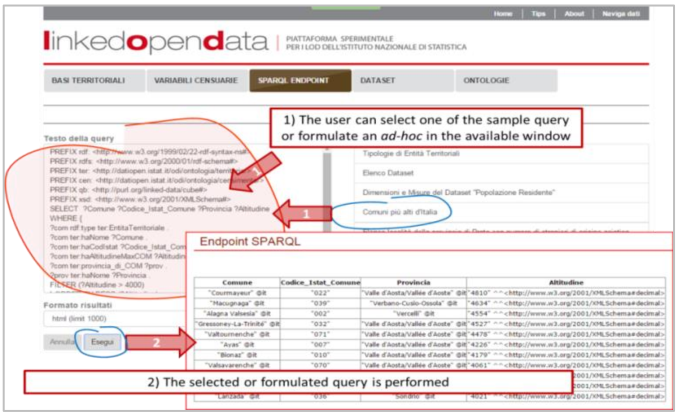
1) Selection of data of interest

2) The choice selection is transformed into a query

3) By executing the query the selected datasets are downloaded, available in various format (CSV, RDF/XML, JSON, TSV)

Fig. 4: Guided interaction flow

ii) Users interfaces for free queries: the system, through a SPARQL endpoint (practical access interface that works like a door through which the system communicates with the outside world) offers the ability to perform free queries that give the user a maximum freedom for data browsing (see Figure 5).



linkedopendata | PIATTAFORMA SPERIMENTALE PER I LOD DELL'ISTITUTO NAZIONALE DI STATISTICA
Home | Tips | About | Naviga dati
BASI TERRITORIALI
VARIABILI CENSUARIE
SPARQL ENDPOINT
DATASET
ONTOLOGIE
1
2
Annulla
Esegui

1) The user can select one of the sample query or formulate an *ad-hoc* in the available window

Testo della query

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ter: <http://datiopen.istat.it/odi/ontologia/territorio/>
PREFIX cen: <http://datiopen.istat.it/odi/ontologia/censimento/>
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?Comune ?Codice_Istat_Comune ?Provincia ?Altitudine
WHERE {
?com rdf:type ter:EntitaTerritoriale .
?com ter:haNome ?Comune .
?com ter:haCodIstat ?Codice_Istat_Com
?com ter:haAltitudineMaxCOM ?Altitudin
?com ter:provincia_di_COM ?prov .
?prov ter:haNome ?Provincia .
FILTER (?Altitudine > 4000)
```

Formato risultati

html (limit 1000)

Tipologie di Entità Territoriali

Elenco Dataset

Dimensioni e Misure del Dataset "Popolazione Residente"

Comuni più alti d'Italia

Endpoint SPARQL

Comune	Codice_Istat_Comune	Provincia	Altitudine
"Courmayeur" @it	"022"	"Valle d'Aosta/Vallée d'Aoste" @it	"4810" ^^<http://www.w3.org/2001/XMLSchema#decimal>
"Macugnaga" @it	"039"	"Verbano-Cusio-Ossola" @it	"4634" ^^<http://www.w3.org/2001/XMLSchema#decimal>
"Alagna Valsesia" @it	"002"	"Vercelli" @it	"4554" ^^<http://www.w3.org/2001/XMLSchema#decimal>
"Gressoney-La-Trinité" @it	"032"	"Valle d'Aosta/Vallée d'Aoste" @it	"4527" ^^<http://www.w3.org/2001/XMLSchema#decimal>
"Valtournenche" @it	"071"	"Valle d'Aosta/Vallée d'Aoste" @it	"4478" ^^<http://www.w3.org/2001/XMLSchema#decimal>
"Ayas" @it	"007"	"Valle d'Aosta/Vallée d'Aoste" @it	"4226" ^^<http://www.w3.org/2001/XMLSchema#decimal>
"Bionaz" @it	"010"	"Valle d'Aosta/Vallée d'Aoste" @it	"4179" ^^<http://www.w3.org/2001/XMLSchema#decimal>
"Valsavarenche" @it	"070"	"Valle d'Aosta/Vallée d'Aoste" @it	"4061" ^^<http://www.w3.org/2001/XMLSchema#decimal>
[illegible]	[illegible]	[illegible]	[illegible]2001/XMLSchema#decimal>
[illegible]	[illegible]	[illegible]	[illegible]2001/XMLSchema#decimal>
"Lanzada" @it	"036"	"Sondrio" @it	"4021" ^^<http://www.w3.org/2001/XMLSchema#decimal>

2) The selected or formulated query is performed

Fig. 5: Free queries flow

iii) Users interfaces for browsing through the data: the system provides two ways to navigate through the data: a hypertext browser that offers a localized display of data allowing the users to "move" through the data by means the reciprocal links implemented by RDF and a visual navigator which allows to carry out the same navigation through the use of graphs (see Figure 6)

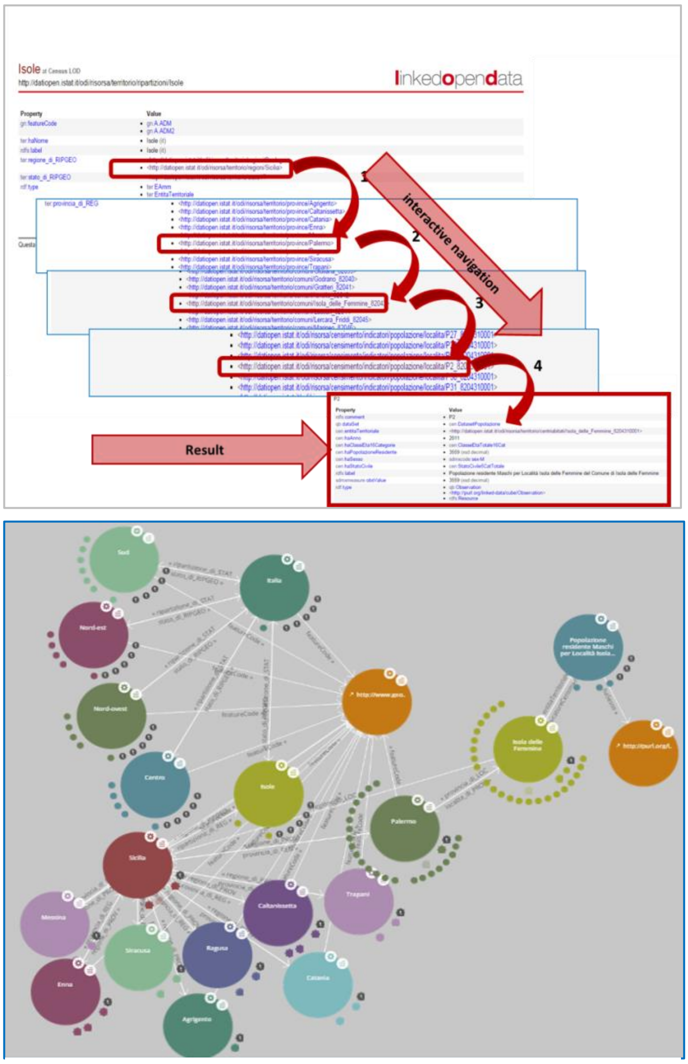
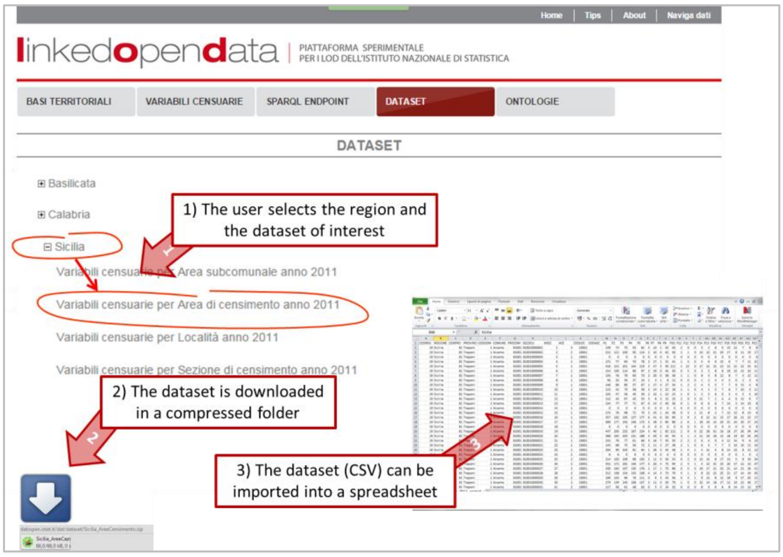


Fig. 6: Interactive and visual navigation

iv) Download Area: It allows users to download in CSV format both data and ontologies. For data, dataset at regional level are available (see Figure 7).


Home | Tips | About | Naviga dati
linkedopendata | PIATTAFORMA SPERIMENTALE PER I LOD DELL'ISTITUTO NAZIONALE DI STATISTICA
BASI TERRITORIALI
VARIABILI CENSUARIE
SPARQL ENDPOINT
DATASET
ONTOLOGIE
DATASET
⊞ Basilicata
⊞ Calabria
⊟ Sicilia
Variabili censuarie per Area subcomunale anno 2011
Variabili censuarie per Area di censimento anno 2011
Variabili censuarie per Località anno 2011
Variabili censuarie per Sezione di censimento anno 2011
1
2
3

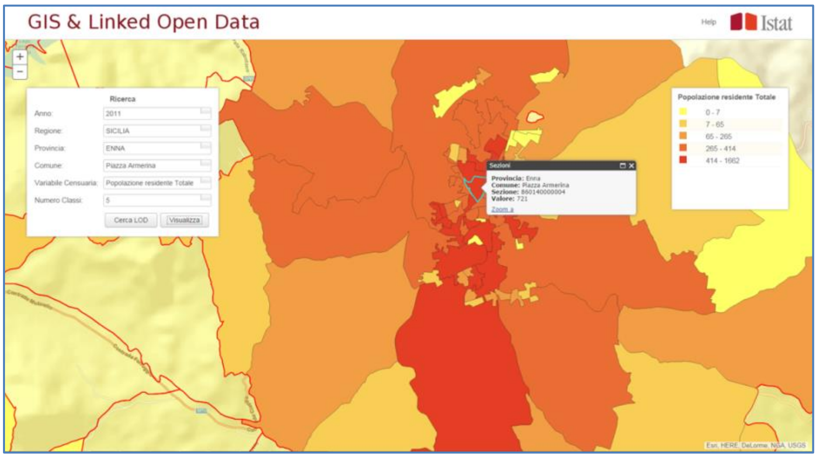
1) The user selects the region and the dataset of interest

2) The dataset is downloaded in a compressed folder

3) The dataset (CSV) can be imported into a spreadsheet

Fig. 7: Download area

Exploiting the potential of the platform, a first experimental experience of integration with geographic data and the Institute's GIS system was also set up. The Linked Geographical Data represent a key point for building up solutions for territorial analysis of data on the Web, such as Linked Open Data. The experiment carried out on Istat portal allows the users to query linked open data via GIS, displaying the contents of maps, as in the example shown in Figure 8.



Anno: 2011
Regione: SICILIA
Provincia: ENNA
Comune: Piazza Armerina
Variabile Censuaria: Popolazione residente Totale
Numero Classi: 5

Provincia: Enna
Comune: Piazza Armerina
Sezione: 860140000004
Valore: 721

Fig. 8: GIS Portal integration

6) Future Plans and conclusions

As described, the Istat's LOD Portal is quite rich but covers so far only a limited part of its dissemination asset. Indeed, most of Istat datasets reach the level 3 of the five stars model. Given the features of the Linked data paradigm, the plans are to enrich the Istat provision as much as possible towards level 5, especially with respect to microdata.

In addition, the semantic enrichment enabled by the Linked Data paradigm is also planned to be used *within* Istat, in order to support as much as possible the shift of the current statistical production to register-based statistics. Indeed, ontology modelling, as a fundamental step of the Linked Data paradigm, could strongly support the integration of administrative sources and surveys to build statistical registers.

In synthesis, the future plans for the platform are targeted to the implementation of an integrated environment of data, including (see figure 9):

- Open data for public uses (also non linked).
- Linked data for internal integrated uses (through the use of purposefully designed ontologies).
- Elementary and aggregate Linked Open Data for internal and public uses, available for human and machine users, related to the most part of statistics domain and possibly enriched by geographical representation

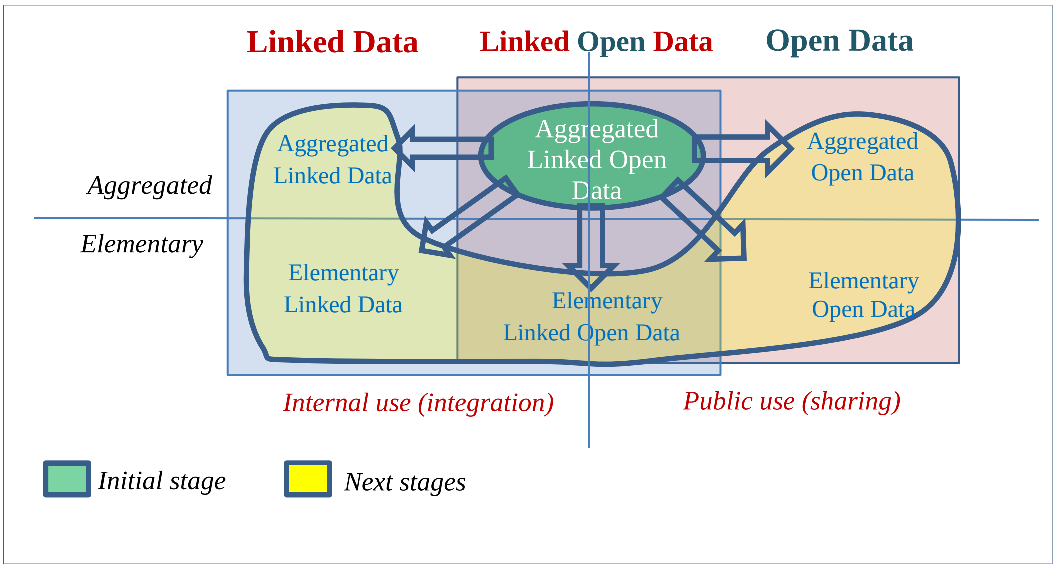


Fig. 9: Open Data at Istat: Evolution scenario

References

1. 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.
2. Linked Data: http://linkeddata.org/
3. PROV Ontology: http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/, 30 April 2013.