Publishing Linked Statistical Data: Aragón, a case study.

Oscar Corcho[1], Idafen Santana-Pérez[1], Hugo Lafuente[2], David Portolés[3], César Cano[4], Alfredo Peris[4], and José María Subero[4]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid
{ocorcho, isantana}@fi.upm.es
[2] Localidata
hlafuente@localidata.com
[3] Idearium Consultores,
dportoles@idearium.eu
[4] Gobierno de Aragón
{ccano,aperis,jmsubero}@aragon.es

Abstract. During recent years, organizations and governmental bodies are adopting Linked Data as the publication paradigm for publishing statistical data, moving from CSV and PDF files and relational databases to a more interoperable and accessible system for the web. In this work we introduce the data generation system implemented by the statistical office in the Spanish region of Aragón. The system consists on the analysis of the data and its structure, which allows to define an automated process for transforming data from relational databases into Linked Data. We also showcase an application which takes advantage of this approach and allows user to browse and retrieve statistical data.

Keywords: semantics, statistical data, open data, linked data

1 Introduction

In recent years, public administration and governmental bodies have embraced the open data movement, publishing their data into public repositories, usually under different formats and means of access. This brings in universal access to governmental data. An example of such is the Open Data portal of Aragón[5], which currently host more than 2700 open datasets from different organizations belonging to the regional Government.

Among these data, statistical information represents a large and interesting amount of them, exposing data such as census, population density, farming or political representatives. The region of Aragón (Spain), by means

[5] http://opendata.aragon.es/

of its statistical office[6], has been hosting and maintaining the statistical reports of the region since the early 90s. In recent years they have moved from an approach mostly based on data warehousing using relational databases to an approach that combines the existing technology with the publication of Linked data.

Publishing statistical data exposes several challenges, as this type of data is often complex, usually including several dimensions and evolving over time. Several vocabularies have been developed for publishing statistics following the Semantic Web and Linked Data [2] principles, being W3C Data Cube [3] the most relevant one in the area.

In this paper we describe the process of transforming the Aragón statistical data into Linked Data, using Data Cube and SKOS [1], and how this process has been automated using the GitHub[7] collaborative environment as a way to store snapshots of datasets together with their versions, as well as their transformations into RDF. This includes the generation of the data, its publication both as data dumps and at an SPARQL endpoint, producing an adequate documentation, and a public API for developers to consume it.

The rest of the paper is organized as follows. Section 2 introduces the current state of the statistical data portal in Aragón, showcasing how data was consumed by users. Section 3 presents the transformation process and how the resulting data is accessed. Section 4 describes an application for browsing and retrieving the generated Linked Data resources. Finally, Section 5 summarizes our results and identifies future works.

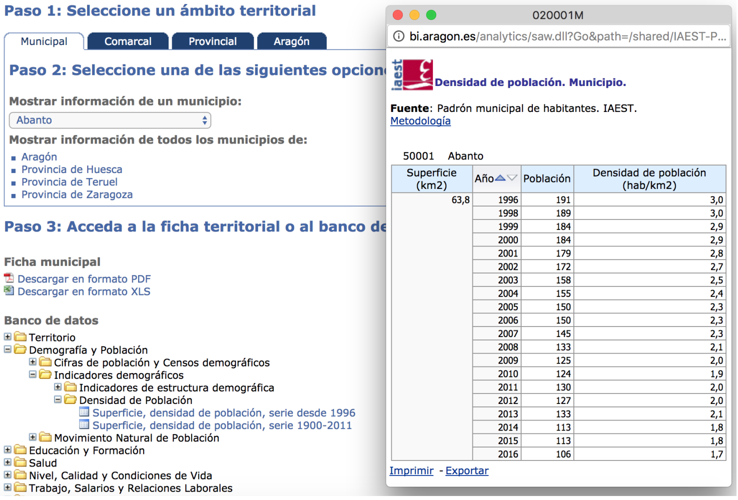
2 Aragón Open Statistical Data

Statistical data collected and produced by IAEst has been available to citizens long before the institution undertook the process of transforming it into Linked Data. Before 2015, on top of a data warehouse infrastructure based on an Oracle Business Intelligence[8] solution, the data portal hosted by IAEst offered browsing capabilities based on a hierarchical taxonomy, allowing users to retrieve tabular data, either as HTML tables, or CSV and PDF files. As depicted in Figure 1 the interface offers three main steps. First, the user selects the administrative division (i.e. Municipality, Region, Province, and Aragon as a whole), then defines the concrete value for each level (i.e. the corresponding municipality or region), and finally, browses the folder structure to reach the target statistical report. Data is then shown as an HTML table.

[6] http://www.aragon.es/iaest

[7] http://github.com/

[8] https://www.oracle.com/solutions/business-analytics/business-intelligence/



Paso 1: Seleccione un ámbito territorial

Paso 2: Seleccione una de las siguientes opcion

Mostrar información de un municipio:

Abanto

Mostrar información de todos los municipios de:

- Aragón
- Provincia de Huesca
- Provincia de Teruel
- Provincia de Zaragoza

Paso 3: Acceda a la ficha territorial o al banco de

Ficha municipal

Descargar en formato PDF
Descargar en formato XLS

Banco de datos

- Territorio
- Demografía y Población
  - Cifras de población y Censos demográficos
  - Indicadores demográficos
    - Indicadores de estructura demográfica
    - Densidad de Población
      - Superficie, densidad de población, serie desde 1996
      - Superficie, densidad de población, serie 1900-2011
  - Movimiento Natural de Población
- Educación y Formación
- Salud
- Nivel, Calidad y Condiciones de Vida
- Trabajo, Salarios y Relaciones Laborales

020001M

bi.aragon.es/analytics/saw.dll?Go&path=/shared/IAEST-P...

Densidad de población. Municipio.

Fuente: Padrón municipal de habitantes. IAEST.
Metodología

50001 Abanto

Superficie (km2)	Año	Población	Densidad de población (hab/km2)
63,8	1996	191	3,0
	1998	189	3,0
	1999	184	2,9
	2000	184	2,9
	2001	179	2,8
	2002	172	2,7
	2003	158	2,5
	2004	155	2,4
	2005	150	2,3
	2006	150	2,3
	2007	145	2,3
	2008	133	2,1
	2009	125	2,0
	2010	124	1,9
	2011	130	2,0
	2012	127	2,0
	2013	133	2,1
	2014	113	1,8
	2015	113	1,8
	2016	106	1,7

Imprimir - Exportar

Fig. 1: Previous version of the GUI, which is still maintained.

During 2015 and 2016, the Aragón Statistical Office (IAEst)[9] undertook the process of transforming the statistical datasets from the Aragón region, stored in relational databases and CSV files, into a more standard and accessible format, both for the final users (citizens mainly) as well as for city developers and civic hackers.

The main goal of this process is to make data accessible following Linked Data principles. In the following section we introduce how the original data sources were analyzed and transformed to meet those principles, and how the process has been automated in order to manage not only current datasets but also the ones that will be added in the future. The transformation and data generation processes are built on top of GitHub, allowing to track the process itself and the collaborative management of the data.

3 Data transformation and publication

The need of migrating the statistical data hosted by IAEst to a more integrated format was identified several years ago. The vast majority of this data suffered from a dispersion problem, associated to the large amount of databases containing information at different administrative levels (region, municipalities, etc.), located usually into different servers.

Rather than supporting just a data distribution solution, the final goal of the process is to obtain a dataset that includes structured and normalized data, on top of which to built rich user experiences and allow its re-usability,

[9] http://www.aragon.es/iaest/EstadisticaLocal

both for developers and citizens. This adds to the data warehouse system, based on relational databases, which is still available. The new portal aims to provide a dataset in which the data is published following the Linked Data principles.

The process is divided into two main phases, in which the data is analyzed to derive the corresponding transformation processes, including the generation of mappings and error detection, and the system is deployed to support a continuous publication cycle.

Data extraction and characterization phase The first phase of the transformation process aims to identify the data sources to be published, and characterize them, defining the statistical dimensions to be part of the final cubes and the rules for the transformation of the data. This characterization will guide the data transformation process during the second phase. All the scripts of the process are available online[10], each of them numbered according to their invocation order.

The process itself starts by identifying the statistical reports from the IAEst site that need to be transformed to Linked Data. This is done by means of a set of web services that are hosted by the institution[11] and powered by their data warehouse solution (Oracle BI). The first step in this stage is the retrieval of the codes used to refer to those reports (e.g., 01-010001A, which refers to the report on housing according to their types), which amount to approximately one thousand.

Making use of these codes, data dumps in the form of CSVs were obtained using the download services provided by IAEst, so as to get the initial version of the reports to be transformed. This was required in order to start the bulk characterisation of the whole set of datasets. These data dumps needed to be preprocessed further, by transforming them into UTF-8 encoding and verifying that there were no further errors on such retrieved files.

The bulk set of data dumps was analysed to identify the dimensions and measurements used in the datasets. For that purpose, we looked at both the column header names and their associated data. Initially, the set of automatically generated dimensions was of approximately 700 dimensions for the whole dataset, which was further refined later together with experts from IAEst, reducing it to the current number. At this stage, several errors were identified. For example, the same dimension was published with different column headers, such as "género" and "sexo" to refer to the gender dimension; typos were also found in column header names; there were several columns with no associated values; etc. Other typical errors that were identified were related to the values that the columns had for existing dimensions.

[10] https://github.com/aragonopendata/local-data-aragopedia/tree/master/src

[11] http://bi.aragon.es

For each of these dimensions, the corresponding SKOS concept schemes were generated, and are published together with the rest of data, as well as available as an RDF data dump[12].

Finally, measure properties were also proposed taking into account those columns where the number of distinct values was much larger than average. These properties were also checked together with IAEst experts.

A summary of all the errors that were identified in this first phase is available in GitHub[13]. This includes empty columns (i.e. columns with no data), data belonging to wrong municipalities and districts, etc. This analysis allowed to develop proper fixing mechanisms in the following stage, automating the process.

Automating phase The goal of this phase is to produce and automatize processes based on the results obtained in the first stage, as well as to handle potential errors during the process. This process is meant to support a continuous data production cycle, updating the resultant RDF and Linked Data as new statistical reports are produced or existing ones are modified or removed.

The process, which is run every night, starts by retrieving the information about the statistical reports from the IAEst database, following the same process used in the previous phase. That is, it retrieves the list of available report codes, checks whether the reports that correspond to those codes have been already transformed into data cubes or not, and whether they contain new data.

Hash signatures are generated for each data cube during each iteration, allowing to track their data evolution and identify when the cubes must be regenerated or new dimensions should be added in their configuration files for transformation. In each iteration, the system compares the new hash codes with the previous ones, and when codes do not match, the cube is marked for being processed correspondingly.

As introduced before, the system uses GitHub as a management tool for handling data evolution. A GitHub issue is generated on each iteration, listing the cubes that have must be created or modified[14]. These issues can be resolved automatically by the system, generating new data cubes that can be stored. Each cube will define a new graph, as well as new graphs for the updated Data Structure Definition (DSD), properties and SKOS information. The generated data cubes are also committed and pushed to the GitHub repository, including a report of the errors found during the process.

[12] https://github.com/aragonopendata/local-data-aragopedia/tree/master/data/dump/DatosTTL/codelists

[13] https://github.com/aragonopendata/local-data-aragopedia/blob/master/data/dump/errorReport.txt

[14] https://github.com/aragonopendata/local-data-aragopedia/issues/93

3.1 Data access

As stated before, the data generated as a result of the transformation process is stored in GitHub, along with the error reports, being automatically linked to the GitHub issues that were involved on the process. Besides this, data is published to be accessed using a SPARQL endpoint[15] as well as by means of an API[16]. These mechanisms are the basis for the development of the platform introduced on the following section.

The SPARQL endpoint is hosted on a Virtuoso[17] 7.20 server, in which the generated data is loaded automatically at the end of the transformation process. This way the updated information is continuously fed into the server, being available for querying. Besides this, the resources included in the dataset are accessible by their URI, as the system supports content negotiation for agents accessing the data.

The statistical information generated during the process is also available by means of a programmatic API, developed using ELDA[18]. This API supports the retrieval of information, both hierarchical and non hierarchical information, obtaining the data as JSON objects. The current version of the API supports retrieving data cubes based on different criteria, such as those containing given dimensions, belonging to a certain municipality, or having an specific average value of a measurement. It also allows retrieving the information describing a dataset, a cube, or the list of labels associated to the data.

4 An application on top of our Statistical Linked Data

As a result of the data transformation, a new version of the portal was developed, improving the previous version (which is still available). Under the name of Aragopedia[19], this new version offers access to the same data, but provides new browsing features in a more integrated and interoperable fashion. Following the same line of design as implemented in the previous interface, the new one was structured under three main steps, namely *where*, *what* and *when* (*dónde*, *qué* and *cuándo* in Spanish), as shown in Figure 2.

On the *where* step the user selects the administrative division from which to obtain data, as in the previous case. On the second step, the system allows to select the statistical dimensions to be explored, including search functionalities based on the descriptions associated to data. This improves the way data is queried compared to the previous version, as the user is able to explore dynamically the available information. On the final step, time ranges

[15] http://opendata.aragon.es/sparql/

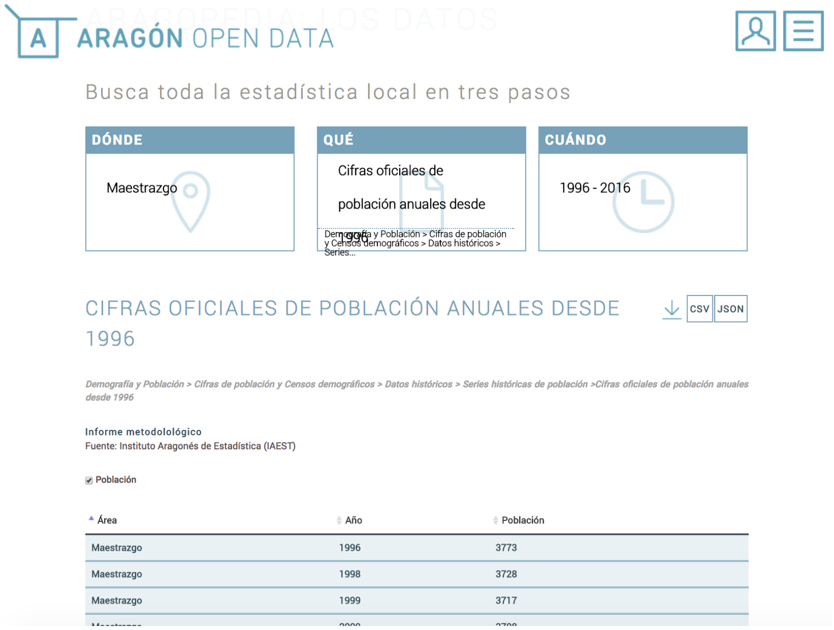
[16] http://opendata.aragon.es/datos/api

[17] https://virtuoso.openlinksw.com/

[18] https://www.epimorphics.com/technology/elda/

[19] http://opendata.aragon.es/apps/aragopedia/datos

can be defined, filtering the data to be retrieved. Users can specify a time period of one or more years. This adds to the previous version, in which only the latest version of the data was available.



ARAGÓN OPEN DATA

ARAGOPEDIA: LOS DATOS

Busca toda la estadística local en tres pasos

DÓNDE: Maestrazgo

QUÉ: Cifras oficiales de población anuales desde 1996

Demografía y Población > Cifras de población y Censos demográficos > Datos históricos > Series...

CUÁNDO: 1996 - 2016

CIFRAS OFICIALES DE POBLACIÓN ANUALES DESDE 1996

Demografía y Población > Cifras de población y Censos demográficos > Datos históricos > Series históricas de población >Cifras oficiales de población anuales desde 1996

Informe metodolológico
Fuente: Instituto Aragonés de Estadística (IAEST)

☑ Población

Área	Año	Población
Maestrazgo	1996	3773
Maestrazgo	1998	3728
Maestrazgo	1999	3717

Fig. 2: Current version of the Aragopedia GUI

The results obtained are automatically displayed on the same page and can be downloaded as a CSV file, as in the previous version, but in this case containing the dereferenceable URIs generated on the transformation process. It also support the generation of JSON documents that can be retrieved, containing as well proper URIs. Listing 1.1 shows the JSON content of the file for a query about *Maestrazgo* region (*where*), about its population (*what*) during 1999 (*when*). As shown, every resource is identified with its corresponding URI and typed properly, including labels for names tagged with their language.

Listing 1.1. JSON excerpt for the Maestrazgo region during 1999.

```
{
    "refArea":{
       "type":"uri",
       "value":"http://opendata.aragon.es/recurso/territorio/
           Comarca/Maestrazgo"
    },
   "nameRefArea":{
      "type":"literal",
      "xml:lang":"es",
      "value":"Maestrazgo"
   },
   "refPeriod":{
      "type":"uri",
      "value":"http://reference.data.gov.uk/id/year/1999"
   },
   "nameRefPeriod":{
      "type":"literal",
      "value":"1999"
   },
   "poblacion":{
      "type":"typed-literal",
      "datatype":"http://www.w3.org/2001/XMLSchema#int",
      "value":"3717"
   }
}
```

5 Conclusions and future work

This work follows the current trend in publishing statistical data openly under the Linked Data principles. Many statistical agencies around Europe have undertaken similar processes during recent years, promoting a more structured way of accessing data and allowing to develop more interoperable data interfaces.

The transformation process exposed in this work aligns with such initiatives, transforming statistical data from relational databases into Linked Data. The process itself is fully open and is built around GitHub, which works as a layer for facilitating the collaboration around data distributively.

This data is generated using the W3C Data Cube and SKOS vocabularies, what makes data more interoperable with other statistical datasets, for instance, by the usage of some common SDMX codes. Also, the datasets themselves are documented using DCAT, so as to be published in the open data portal of Aragón, facilitating their discoverability and usability. Largely due to this effort, the region of Aragon, by means of its statistical office, is now one of the leaders in statistical Linked Data, and in Open Data in general in Spain.

As for the graphical interface introduced in this work, early usability results with users, both with and without an statistical background, show that even when the same results could be obtained with the previous version of the application, the new one offers better performance in terms of time and ease of use. These evaluations are being currently performed and we are planning to work on them in the near future.

References

1. Antoine Isaac, E. S. SKOS Simple Knowledge Organization System Primer. `https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/`, 2019. [Online; accessed 11-September-2017].
2. Bizer, C., Heath, T., and Berners-Lee, T. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst. 5*, 3 (2009), 1âĂŞ22.
3. Richard Cyganiak, D. R. The RDF Data Cube Vocabulary. `https://www.w3.org/TR/vocab-data-cube/`, 2014. [Online; accessed 11-September-2017].