Implementation of a Linked Open Data Solution for the Statistics Agency of Cantabria's Metadata and Data Bank

Alejandro Villar Fernández
Spain
contacto@alejandro-villar.es

Ana Santurtún Zarrabeitia
University of Cantabria,
Spain
ana.santurtun@unican.es

Abstract

Statistics is a fundamental piece inside the Open Government philosophy, being a basic tool for citizens to know and make informed decisions about the society in which they participate. Due to the great number of organizations and agencies that collect, process and publish statistical data all over the world, several standards and methodologies for information exchange have been created in recent years in order to improve interoperability between data producers and consumers, of which SDMX is one of the most renowned examples. Despite having been developed independently of this, the global Semantic Web effort (backed mainly by the W3C-driven Linked Open Data initiatives) presents itself as an extremely useful tool for publishing both completely contextualized metadata and data, therefore making them easily understandable and processable by third parties.

This report details the changes made to the IT systems of the Statistical Agency of Cantabria (Instituto Cántabro de Estadística, ICANE) with the purpose of implementing a Linked Open Data solution for its website and statistical data bank, making all data and metadata published by this Agency available not only to humans, but to automatized consumers, too. Multiple standards, recommendations and vocabularies were used for this task, ranging from Dublin Core metadata RDFa tagging, through the creation of several SKOS concept schemes, to providing statistical data using the RDF Data Cube vocabulary.

Keywords: Linked Open Data; statistics; ICANE; Spain; Cantabria; SKOS; RDF; RDFa; Semantic Web.

1. Background

Legally created in 1998, but not officially operative until 2004, the Statistics Agency of Cantabria (Instituto Cántabro de Estadística, ICANE) is the official public organization in Cantabria (Northern Spain) in charge of the production and dissemination of statistical data regarding every aspect of Cantabria's economy and society.

Despite its young age, ICANE's face to the public has been in constant evolution, from its inception as a static web site populated with Microsoft Excel files, to its current state of a machine-navigable, metadata-annotated portal serving on-the-fly generated data in 7 different formats.

1.1 Initial Situation and Motivation

ICANE's web portal and data bank underwent a major visual and functional overhaul in 2011, replacing a legacy OpenCMS[1] installation with Liferay Enterprise Portal[2] 5.2, and an in-house-developed web application for data selection and display with a Pentaho Mondrian[3] 3.1 +

[1] http://www.opencms.org/
[2] http://www.liferay.com/es/
[3] http://mondrian.pentaho.com/

https://doi.org/10.23106/dcmi.952136202

JPivot[4]1.8 combination, which also involved converting its previous data warehouse to an OLAP[5]-compatible setup.

On the second half of 2012, ICANE decided to implement a solution for improving its data and metadata availability and to publish and link its resources on the Semantic Web. At the time, all internal, structural metadata about data series (OLAP cubes, dimensions, measures, etc.) was stored in XML Schema files and deployed onto the data bank application servers, and all descriptive metadata (including hierarchical organization into folders and sections) resided in a relational database, which was used by the web portal and by the data bank application to present a hierarchically navigable view to the end user. This setup relied on the following domain model classes to represent navigational and descriptive metadata:

- Area: subject area for a group of series.
- Type: type of series (time series, historical data or municipal data).
- Group: series group (subject area subdivision), belonging to an area and a type.
- Nodes: hierarchical layout of folders and time series. Includes all descriptive metadata for every series and folder.

Most series metadata (modification dates, periodicity, data sources, etc.) had not been yet normalized. In addition, no URL path hierarchy was used when accessing the different sections of the web portal or data bank application (with the latter relying on user-provided HTTP GET parameters for time series selection and descriptive metadata retrieval).

2. Metadata Reorganization, Normalization and Centralization

A decision was made to reorganize and normalize all database-stored metadata. The first step was dropping the old area/group/type schema in favor of a section/subsection/category one so that:

- Four main sections (economy, population, society, territory and environment) would exist.
- Each section would contain several, more specific subsections, which would not be dependent on type or category.
- Inside each subsection the end user would find up to three different series/folder trees, one for each category.

Moreover, the following efforts were made, with regards to metadata normalization and preparation for the Linked Open Data implementation:

- *URI tags* (friendly URL components) were defined for all user-referenceable classes. A concrete URL hierarchy based on these tags was laid out to make it easier for users to access and understand the web portal and data bank application layouts. More information on the use of URI tags can be found in section 3.2.
- Folder and series metadata were normalized whenever possible. This included series periodicities (which were linked to their DC CDF Vocabulary, 2007, equivalent) and initial and final periods, data sources, reference areas, etc.
- A link repository was created to store all entities' semantic relationships towards external resources.

Finally, ICANE's IT department developed a Web Service to provide centralized access to this newly-created metadata database through a REST API, and avoid directly querying the database server containing it, thus protecting the clients from the specific implementation details. This Web Service was to be used by all internal applications, if possible.

[4] http://jpivot.sourceforge.net/

[5] Online Analytical Processing. An introduction to OLAP can be found in Pedersen, T.B (2001).

https://doi.org/10.23106/dcmi.952136202

3. Proposed Solution

Alejandro Villar's proposal consisted of 5 different, but interdependent, phases:

1. Definition of a RDF model for all publicly accessible entities, centered around the DCMI model.
2. Implementation of a Linked Open Data strategy to serve entity metadata.
3. Development of an RDF export filter for the data bank application.
4. Implementation of a SPARQL endpoint to access metadata.
5. Semantic linking of entities to external resources, in order to provide context to automatized data consumers.

3.1. RDF Model

In addition to property mapping, the definition of the RDF model had to take into consideration the hierarchical nature of the time series and folders tree. Therefore, a SKOS ontology was conceived, comprising several Concept Schemes (one for each section and for each subsection), which would in turn contain a hierarchical array of folders (represented by SKOS Concepts). Time series and folders would be linked using the DCMI subject property. Given that SKOS does not provide an explicit mechanism to define a hierarchy between Concept Schemes, its inScheme property (whose domain is effectively any RDF resource) was used to this effect.

A property summary of this model, using CURIE syntax, can be seen in Table 1, while table 2 contains a list of all prefixes used throughout this project report.

In addition, a special ICANE vocabulary[6] was created to enhance entity classification and to provide additional means of entity interlinking.

TABLE 1: Property summary of the RDF model

Entity	Property	Description
Section A section has many subsections.	rdf:type	Class of this resource. Values: icane:Section and skos:ConceptScheme.
	skos:prefLabel and rdfs:label	Resource label.
Subsection A subsection belongs to a single section and has many folders and time series.	rdf:type	Class of this resource. Values: icane:Subsection and skos:ConceptScheme.
	skos:prefLabel and rdfs:label	Resource label.
	icane:section and skos:inScheme	Section that this subsection belongs to.
Category A category has many folders and time series.	rdf:type	Class of this resource. Value: icane:Category.
	rdfs:label	Resource label.
	icane:acronym	Acronym for this category.
Folder A folder is a branch node in the subject matter tree. It has many child folders and time series, and it belongs to one subsection (and therefore to one section) and to one category.	rdf:type	Class of this resource. Value: skos:Concept.
	skos:prefLabel and rdfs:label	Resource label.
	icane:section	Section that this folder belongs to.
	icane:subsection	Subsection that this folder belongs to.
	skos:inScheme	Concept Schemes that this folder belongs to (Its section and subsection).
	icane:category	Category that this folder belongs to.
	skos:broader / skos:narrower	Used for creating a hierarchical structure of folders.

Table continued

[6] http://www.icane.es/opendata/vocab

https://doi.org/10.23106/dcmi.952136202

Entity	Property	Description
Time series A time series belongs to one folder, to one subsection (and therefore to one section) and to one category. It has a single source and a reference area.	rdf:type	Class of this resource. Values: icane:TimeSeries and qb:DataSet.
	rdfs:label and dcterms:title	Resource label.
	dcterms:subject	Subject of this resource, which will be its parent folder in the hierarchy.
	icane:section	Section that this series belongs to.
	icane:subsection	Subsection that this series belongs to.
	icane:category	Category that this series belongs to.
	dcterms:modified	Date of last data update.
	dcterms:accrualPeriodicity	Update periodicity for this series. Linked to the DC CDF Vocabulary.
	dcterms:spatial	Reference area relative to this series.
	dcterms:temporal	A classless resource with icane:initialPeriod and icane:finalPeriod properties to represent the period for which data is available.
	dcterms:source	Data source for this series.
	rdfs:comment	Comments or footnotes about this series or its data.
	void:dataDump	URI for a dump of this series' data.
Reference area A reference area is a location or geographical zone that a time series refers to. It has many time series.	rdf:type	Class of this resource. Values: icane:ReferenceArea and dcterms:Location.
	rdfs:label	Resource label.
Source A source is a survey or organization from which data is retrieved for publishing. It has many time series.	rdfs:label and dcterms:title	Resource label.
	foaf:page	Main website for the source survey or organization.

TABLE 2: CURIE prefixes used

Prefix	Reference	Description
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF vocabulary
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema vocabulary
dcterms	http://purl.org/dc/terms/	DCMI metadata terms
foaf	http://xmlns.com/foaf/0.1/	FOAF vocabulary
skos	http://www.w3.org/2004/02/skos/core#	SKOS vocabulary
void	http://rdfs.org/ns/void#	VoID vocabulary
qb	http://purl.org/linked-data/cube#	RDF Data Cube vocabulary
icane	http://www.icane.es/opendata/vocab#	ICANE's own RDF vocabulary

Dates are generally published as XML Schema date types, except for initial and final periods, which are presented as text literals. This discrepancy is explained by the existence of several time series whose initial and/or final periods cannot be represented by simple dates (such as academic courses, which are identified by their duration period ["2008-2009"], i.e., a single initial or final date cannot be pinpointed).

All entites also have an extra property, icane:metadataApiUri, that links them to their ICANE Metadata API representation.

Apart from the aforementioned issue regarding initial and final periods, the conceived metadata model aligned with the selected, well-known vocabularies without the need to introduce any changes. All model properties and values designated for third party consumption were easily mapped to the RDF model.

Additionally, a VoID document was generated to aid in the task of navigating the site, and to provide supplementary metadata about the whole publication. This document resides in

https://doi.org/10.23106/dcmi.952136202

http://www.icane.es/opendata/void, but can also be accessed via http://www.icane.es/.well-known/void.

3.2. Serving Linked Open Data

With the intention of minimizing the impact on aspect and functionality that the addition of the newly-created RDF metadata would suppose, an XHTML+RDFa solution was proposed. The following strategy was outlined:

- URIs would have to include a fragment identifier in order to correctly identify individual resources inside the XHTML+RDFa documents. A list of such URI patterns for the entities introduced in Table 1 is shown in Table 3.
- Metadata properties already available to users would be annotated using RDFa.
- New metadata properties would be included using RDFa, but only for RDFa processors and not for human (HTML) consumption, with the goal of preventing user interface clutter, while at the same time preserving these attributes for automatic processing.
- For some entities (Category, Reference area and Source), having an HTML view could be unnecessary or even counter-productive, since human consumers can gather all their information from just their name or title. However, they would still be assigned a dereferenceable URI with an RDF+XML description in order to provide context to RDF consumers.
- To maintain URI uniqueness, all requests directed to http://icane.es would use an HTTP 301 (moved permanently) redirect code to http://www.icane.es, making the latter the canonical HTTP host.
- Even though XHTML+RDFa version 1.1 would be used, backward-compatible markup would be added to improve interoperability with version 1.0 clients.

TABLE 3 Entity URI patterns. Variables in {braces} represent URI tags

Entity	URI Pattern
Section	http://www.icane.es/{section}#section
Subsection	http://www.icane.es/{section}/{subsection}#section
Category (RDF/XML only)	http://www.icane.es/opendata/categories#{category}
Folder	http://www.icane.es/{section}/{subsection}#{category}-{folder}
Time Series	http://www.icane.es/data/{category}/{section}/{subsection}/{time-series}#timeseries
Reference Area (RDF/XML only)	http://www.icane.es/opendata/reference-areas#{reference-area}
Source (numeric ID is used)	http://www.icane.es/data/{category}/{section}/{subsection}/{time-series}#source_{id}

3.3. Exporting RDF Data

An export filter for RDF/XML data was created for the data bank application, following the RDF Data Cube (2012) recommendation. The RDF Data Cube vocabulary supersedes the Statistical Core Vocabulary (SCOVO), created by Hausenblas et al. (2009), and provides an approach that aligns transparently with the existing OLAP data model.

RDF/XML files would be created on the fly, and as with other ICANE export filters, the user would have the option of either configuring a custom selection or using a default one containing, for most series, all available data.

ICANE's data was already structured according to the OLAP model, which aligns perfectly with the RDF Data Cube one.

- An OLAP Cube represents an RDF Data Cube DataSet.
- For every OLAP Dimension, a Concept Scheme, a Concept class and an RDF Data Cube DimensionProperty are created. The Concept Scheme will be populated with resources of

https://doi.org/10.23106/dcmi.952136202

the aforementioned class, each one corresponding to a Member of the Dimension. The Concept class will also be used as the RDF Schema range of the Property.

- For every OLAP Measure, an RDF Data Cube MeasureProperty is created.
- For every OLAP data cell, an RDF Data Cube Observation is created. The observation will specify value members for all dimensions, and a value for each measure, using the properties created in the previous steps.

The icane:ObservableValue class was defined as a 'tagging class' for all observable values, serving as the Measure Properties range, with two subclasses: icane:NullValue (for null, empty or otherwise missing values) and icane:ConfidentialValue (for values subject to statistical confidentiality).

3.4. SPARQL Endpoint

It was deemed at design time that only series metadata but not data could be served via a SPARQL endpoint. This limitation arose from the fact that, even if a SPARQL engine could be configured to query all available data with an acceptable performance, all structural metadata was contained in Mondrian XML files, which would hinder the goal of serving reasonably up-to-date information.

Initially, a D2RQ Platform[7]based solution was considered. While tests seemed satisfactory at first, this path was eventually abandoned due to the following complications:

- The similarity of many of the URI patterns meant most of the possible optimizations could not be applied.
- Complex SPARQL queries resulted in high SQL query redundancy.
- Any database structure modification would imply a change to the mapping model, and a redeployment (or at least a restart) of the D2RQ application.

For the aforementioned reasons, a custom solution was developed, in the form of a web application that, using the ICANE Metadata API client, generates an RDF Jena[8] model on the fly according to the definition in section 3.2, and serves it through an Apache Tomcat[9] hosted Joseki[10] server. The application checks the ICANE Metadata API Web Service for modifications periodically, and re-generates the model if necessary.

In addition, a SPARQL form Liferay portlet was developed, to provide a user-friendly way to query the endpoint and control output formatting.

3.5. Connecting to the Semantic Web

The final phase of the implementation consisted of populating the link repository, selecting external resources similar to the existing ICANE entities, and connecting them through the pertinent RDF properties. This task was performed mostly manually, and the use of tools for automatic or semi-automatic link creation were discarded due to the heterogeneity of the datasets involved, the specificity of the domain at hand, and the fact that all texts and labels were written in Spanish, while most of the linking targets were English-only collections.

Table 4 summarizes the different properties used and the number of links generated for every entity type. Since foaf:page is defined for data sources by default, and categories are a special ICANE classification with no external counterpart, no additional links were created for these entities. No links were created either for time series, because of their abundance and domain specificity, and due to the fact that they were already connected to folders via dcterms:subject.

[7] http://d2rq.org/
[8] http://jena.apache.org/
[9] http://tomcat.apache.org/
[10] http://joseki.sourceforge.net/

https://doi.org/10.23106/dcmi.952136202

TABLE 4: Properties and number of links generated for each entity type

Entity	Entity Count	Property	Number of Links
Section	4	dcterms:subject	18
		rdfs:seeAlso	1
Subsection	27	dcterms:subject	141
		rdf:seeAlso	43
Category	3	*None*	*None*
Folder	703	skos:closeMatch	161
		rdfs:seeAlso	199
Time Series	2707	*None*	*None*
Reference Area	6	owl:sameAs	10
		rdf:seeAlso	15
Source	2694	*None*	*None*

The following external datasets where used as link targets:

- GeoNames (4 links).
- DBpedia (45 links) and Spanish DBpedia (47 links).
- National Statistics Institute (Instituto Nacional de Estadística, INE) (251 non-RDF links only).
- Eustat (22 non-RDF links only).
- Lista de Encabezamientos de Materia (LEM) para las Bibliotecas Públicas (168 links).
- Library of Congress Subject Headings (151 links).

When linking to non-RDF resources, only rdfs:seeAlso and foaf:page properties were used. The rest of the properties are all guaranteed to have an RDF resource URI as their object.

4. Conclusions

This project report has shown how a Linked Open Data solution can be designed, developed and deployed in a reasonable amount of time (less than 6 months), and can be "attached" to an existing data publication without any aspect or functionality impact. End users may continue to explore the site as if the changes had never taken place, and at the same time other types of consumers have been provided with options to discover and make use of the published data; for example, data can be aggregated by automatic web-crawling tools, using the numerous existing links to other datasets for context.

Additional improvements also open the door to the possibility for other developers to integrate their projects with our own, creating new ways to combine and exploit data. The SPARQL endpoint, for example, allows user to query the full dataset in order to retrieve statistics pertaining to a specific knowledge domain or territory, using resources from well-known vocabularies (such as DBpedia or GeoNames) as starting points.

4.1. Future Improvements

The following is a non-exhaustive list of improvements that could be performed on the project described in this report:

- Publishing RDF metadata directly from the ICANE Metadata API Web Service.
- Adding RDFa markup for describing other less-data-bank-oriented published resources.
- Centralizing time series structural metadata in a manner similar to descriptive metadata.
- Creating a SPARQL data repository.
- Internationalization for text literals.

https://doi.org/10.23106/dcmi.952136202

- Increasing the number and quality of links (e.g., for time series), using automatic or semi-automatic entity analyzers such as Google Refine[11], Free Your Metadata Named Entity Extraction[12] or the Silk Framework[13].

Acknowledgements

The achievements presented in this report would have never materialized without the efforts and dedication of ICANE's IT department, with special mention of Miguel Expósito and Alberto Lezcano.

References

CURIE. (2009). A syntax for expressing Compact URIs. W3C Candidate Recommendation January 16, 2009. Retrieved March 20, 2013, from http://www.w3.org/TR/2009/CR-curie-20090116/.

DC CDF Vocabulary. (2007). Dublin Core Collection Description Frequency Vocabulary, March 9, 2007, version. Retrieved March 20, 2013, from http://dublincore.org/groups/collections/frequency/2007-03-09/.

DCMI. (1998). Dublin Core Metadata Element Set, version 1.0: Reference description. Retrieved March 20, 2013, from http://www.dublincore.org/documents/1998/09/dces/.

FOAF. (2010). FOAF Vocabulary Specification 0.98. Retrieved March 20, 2013, from http://xmlns.com/foaf/spec/20100809.html.

Hausenblas, Michael, Wolfgang Halb, Yves Raimond, Lee Feigenbaum and Danny Ayers. (2009). SCOVO: Using Statistics on the Web of Data. Lecture Notes in Computer Science, Volume 5554, 708-722.

Pedersen, Torben Bach, and Christian S. Jensen (2001). Multidimensional database technology. Computer, December 2006, Volume 34, Issue 12, 40-46.

RDF. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation February 10, 2004. Retrieved March 20, 2013, from http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

RDF Data Cube Vocabulary. (2012). W3C Working Draft April 05, 2012 Retrieved March 20, 2013 from http://www.w3.org/TR/2012/WD-vocab-data-cube-20120405/.

RDF Schema. (2004). RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation February 10, 2004. Retrieved March 20, 2013, from http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

SKOS. (2009). SKOS Simple Knowledge Organization System Reference, W3C Recommendation August 18, 2009. Retrieved March 20, 2013 from http://www.w3.org/TR/2009/REC-skos-reference-20090818/.

[11] http://code.google.com/p/google-refine/

[12] http://freeyourmetadata.org/named-entity-extraction/

[13] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

https://doi.org/10.23106/dcmi.952136202