Transforming Meteorological Data into Linked Data

Ghislain Atemezing [a,*], Oscar Corcho [b], Daniel Garijo [b], José Mora [b], María Poveda-Villalón [b], Pablo Rozas [c], Daniel Vila-Suero [b] and Boris Villazón-Terrazas [b]

[a] *EURECOM, Multimedia Department, Campus SophiaTech, France*
E-mail: atemezin@eurecom.fr
[b] *Ontology Engineering Group, DIA, Facultad de Informática, Universidad Politécnica de Madrid*
E-mail: {ocorcho,dgarijo,jmora,mpoveda,dvila,bvillazon}@fi.upm.es
[c] *Agencia Estatal de Meteorología*
E-mail: prozasl@aemet.es

Abstract. We describe the AEMET meteorological dataset, which makes available some data sources from the *Agencia Estatal de Meteorología* (AEMET, Spanish Meteorological Office) as Linked Data. The data selected for publication are generated every ten minutes by approximately 250 automatic weather stations deployed across Spain and made available as CSV files in the AEMET FTP server. These files are retrieved from the server, processed with Python scripts, transformed to RDF according to an ontology network (which reuses the W3C SSN Ontology), published in a triple store and visualized using Map4RDF.

Keywords: meteorology, ontology, Linked Data, Sensor Networks

1. Introduction

Governments and their associated agencies worldwide are making some of their data sources available under open data licenses, so as to ensure consumption by the general public and other public and private organisations. In this context, AEMET[1], the Spanish Public Weather Service, announced on November 2010 a major change in its data policy, offering a gradual, free and public access to its data in electronic format. As a first step AEMET made publicly available in its website meteorological data registered by its weather stations, radars, lightning detectors and ozone soundings. These data are currently offered with a free and open license[2] as spreadsheets in the AEMET FTP server.

Our work aims at facilitating the use of these data by processing them and offering them as free and open Linked Data (LD)[3]. Following our method for LD generation [13], which we have successfully used in other domains, we start by processing these data and generating RDF according to a meterology ontology network that extends the W3C Semantic Sensor Network Ontology (SSN). The aim of this ontology network is to represent knowledge related to measurements made by weather stations. These measurements represent the state of the atmosphere (humidity, pressure, wind, etc.) in a particular place and time, and is conducted through the sensors equipped at each weather station. Finally we publish these data according to LD principles and visualize it with Map4RDF [5].

The structure of this paper corresponds to the steps followed to create the dataset. Section 2 describes the main features of the data selected to be converted as LD. Section 3 presents the design decisions for the URIs, while Section 4 explains the development of our Ontology Network for Weather Data. Section 5 describes the RDF generation process. Next, Section 6 presents the exploitation of the metereological LD. Finally, we present the conclusions and future lines of work.

[*] Alphabetical order
[1] `http://www.aemet.es/es/portada`
[2] `ftp://ftpdatos.aemet.es/NOTA_LEGAL.txt`
[3] `http://thedatahub.org/dataset/aemet`

2. AEMET Data Sources

Among all of the data made available in the FTP server from AEMET, we have focused on surface meteorological observing stations, and more precisely in measurements taken in ten minute interval times[4]. The data are recorded in several points in the Spanish geography and can be represented as a set of variables (e.g., temperature or pressure) in a map. These stations belong to the Regional Basic Synoptic Network (RBSN) in Spain operated by AEMET, in accordance with the World Meteorological Organization (WMO)[5]. The stations are globally identified by a code and have to accomplish strict quality controls. Their data are used to feed meteorological models and to draw up climate studies. AEMET has around 250 automatic weather stations in this network, registering pressure, temperature, humidity, precipitation and wind data every 10 minutes. Data from the different stations are publicly available online in comma-separated-value (CSV) files, compressed with gzip, updated every hour and kept for seven days. This means that every hour six new files are added, corresponding to periods of ten minutes, and every day a new folder is created to store the files for that day.

The name of each file follows the naming convention "`yyyymmddhhmm_datos.csv`". This allows processing directly a set of specific files without parsing (or even downloading) all of them in order to check whether they are relevant for a specific time range or not. For example, the file named "`201102121900_datos.csv`" contains data from the year 2011, the month of February, on the day 12 at 19:00h. The files are structured in rows and columns: each row corresponds to a single station, while each column has the ID of the observation type and the value recorded in that period of time.

3. URI design for meteorological data

Every resource in our dataset is identified using a URI[6] (Uniform Resource Identifier). URIs have been designed with simplicity, stability and manageability principles in mind, following common guidelines for their effective use[7],[8],[9]. Similarly, naming conventions for ontology concepts and their labels [10,7,2] have a tantamount importance when linking not only data but also its explicit model, as an ontology.

This section presents the design decisions and conventions used in the project: (1) base URIs structure and conventions; (2) URIs and naming conventions of the AEMET Ontology (TBox), described in section 4; and (3) URIs of the generated instances (ABox).

Base URI structure. The base URI, common to all elements in the knowledge base, is `http://aemet.linkeddata.es/`. TBox and ABox have been separated into two URI schemes, respectively: `http://aemet.linkeddata.es/ontology/` and `http://aemet.linkeddata.es/resource/`. In case of ABox resources, the class name is appended as part of the URI (e.g. `http://aemet.linkeddata.es/resource/Point/`).

TBox URIs. Names assigned to TBox elements follow camel case style, using English names. Class and property labels have been added in two different languages (Spanish and English). These natural language descriptions are useful in ontology mapping, information extraction and ontology verbalization.

ABox URIs. Ensuring the "uniqueness" of the URIs assigned to distinct real-world objects is important to mitigate issues such as *co-reference* [9] and the *instance unification problem* [4,1]. A series of decisions have been made for the URIs within the generated dataset in order to preserve the integrity and semantics of the data. These decisions followed the recommendations and design patterns proposed by the semantic web community where applicable; in the remaining cases the team has analysed and discussed the best options and identified certain patterns, as it will be discussed later in this section.

The first decision was to make ABox URIs use *Patterned URIs*[10]. This solution mitigates *co-reference* problems between two distinct individuals of different type and same local ID by adding the class name to the base URI. The second decision was to use *Natural keys*[11] whenever the source group of resources already had a unique identifier that could be used (e.g., the synoptic weather stations and their unique identifier: *INDSINOP*). Finally, some other ad-hoc URI structures have been created in order to uniquely identify the resources. Most of them could be referred to as

[4] `http://bit.ly/NGYvhE`

[5] `http://www.wmo.int/`

[6] `http://tools.ietf.org/html/rfc3986`

[7] `http://www.w3.org/TR/cooluris/`

[8] `http://www.w3.org/Provider/Style/URI`

[9] `http://www.w3.org/TR/chips/`

[10] `http://bit.ly/NKiUIB`

[11] `http://bit.ly/bltOSi`

Resource class	Local ID pattern	Example
Station	INDSINOP	http://aemet.linkeddata.es/resource/Station/08202
Point	INDSINOP	http://aemet.linkeddata.es/resource/Point/08202
Observation	at_<timestamp>_of_<INDSINOP>_on_<observedProperty>	http://.../Observation/at_1306446000000_of_08202_on_PREC
Interval	tenMinutes_since_<timestamp>	http://.../Interval/tenMinutes_since_1306446000000
Instant	timestamp	http://aemet.linkeddata.es/resource/Instant/1306446000000

Table 1

ABox URI patterns

a "composite" pattern, where the local ID is formed by several interconnected pieces of information related to the resource being identified (e.g., "Observation" in Table 1, which presents the different local IDs assigned to each type of individual in the ABox).

4. An Ontology Network for Weather Data

The development of the AEMET ontology network has been performed following an iterative approach based on the reuse of existing knowledge resources, both ontological (including ontologies and Ontology Design Patterns[12]) and non-ontological, as proposed by the NeOn methodology [11].

The AEMET ontology network[13] follows a modular structure [11] consisting of a central ontology that links together a set of ontologies that describe different sub domains involved in the modelling of meteorological measurements.

Figure 1 presents the main classes and properties of this ontology network. This model, which can be considered as the AEMET core, contains four modular ontologies: Sensor, Time, Location, and Measurement.

The current version of the AEMET ontology network has been implemented in OWL DL and contains 83 classes, 102 object properties, 80 datatype properties, and has a SROIQ(D) expressiveness. Next, we present a brief description of each of the sub domains, as well as the knowledge resources that were reused in each ontology.

Measurement ontology models the knowledge related to meteorological observations. Main concepts in this ontology are: ssn:Observation, ssn:FeatureOfInterest and ssn:Property, reused from the W3C SSN ontology [8]. As part of the customization of SSN to our use case, the concept ssn:Property has been extended and populated according to the properties that AEMET gathers in its observations. For this purpose, we have used the tool NOR$_2$O [12] to transform the non-ontological resource "Describe_VAR.csv" file provided by AEMET to an ontological resource.

Location ontology models knowledge about locations, such as administrative limits and coordinates. The *wgs84_pos* vocabulary [14] has been reused with the aim of supporting the representation of geometric positioning by means of the concept "Point" and defining a mapping between the relationships aemet:locatedIn and wgs84_pos:location. The choice of "wgs84_pos" is motivated because it is widely accepted by the community[15] to represent simple geo-coordinates. In addition, we reused part of the GeoBuddies ontology network[16] to represent knowledge about the Spanish administrative division.

Time ontology models knowledge about time such as temporal units, temporal entities, instants, intervals, etc. This ontology reuses the OWL Time ontology[17].

Sensor ontology models the network of sensors and weather stations. It extends SSN, which supports the description of the physical and processing structure of sensors. Here, a sensor is anything that can estimate or calculate the value of a phenomenon, so as a device or a computational process or a combination of both. The representation of a sensor in the ontology links together what it measures (the domain phenomena), the physical sensor (the device) and its functions and processing (the models). During the alignment process between the SSN ontology and the AEMET ontology network the concept "ssn:Sensor" has been extended by means of a hierarchy of types of sensors used by AEMET.

[12] http://ontologydesignpatterns.org

[13] http://aemet.linkeddata.es/models.html

[14] http://www.w3.org/2003/01/geo/wgs84_pos

[15] http://bit.ly/R3vfrj

[16] http://bit.ly/T2MM1P

[17] http://www.w3.org/TR/owl-time/

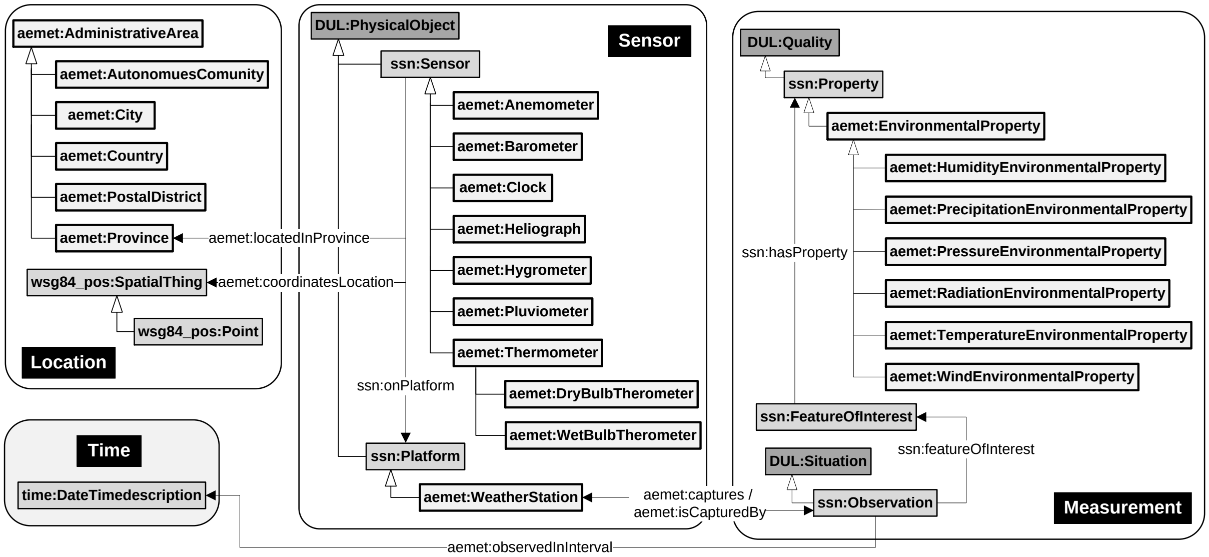

Fig. 1. AEMET ontology network detailed model

5. RDF Generation

RDF has been generated with ad-hoc Python scripts[18] that are executed in two processing steps, integrating with ease the generation of RDF and tasks such as crawling the FTP server where the CSV files are located.

The first step generates the information about the automatic stations, which is static, and thus needs to be executed only once. Having a script is also more maintainable than manually generating the RDF in case there are changes, which is usual in an evolutionary prototype context.

The second step generates the information about the observations on a regular basis, keeping the data updated. The observations are obtained by crawling the FTP server from AEMET. When new files are added or old files are modified, these are downloaded and processed. The process is very similar to the one described in the first step and it can be reused when using plain text files to generate RDF.

Both steps use templates that model the RDF to be generated, expressed in the N3 notation. These templates have been generated manually in this agile approach. For each step two tasks are performed: first the files are parsed and the relevant information is extracted from them; then, we instantiate the templates and write the generated RDF to the corresponding files, using the information extracted in the previous task.

The generated RDF is stored in Virtuoso[19], which integrates with Pubby[20] to publish the results, making them available for humans and computers. The number of triples generated per day is $24\frac{hours}{day} * 6\frac{entries}{hour} * (31\frac{triples}{observation} * 9\frac{observation}{station} * 261\frac{stations}{entry} + 19\frac{triples}{entry}) = 10488672\frac{triples}{day}$. Following the same update process as AEMET does, only the data from the last seven days are kept on the online system.

Finally, we have used the Silk framework[21] to link the location of the weather stations to other datasets in LD. In total, 153 links[22] were stablished with GeoLinkedData [6], and 82 with DBpedia[23].

6. Exploitation

The AEMET dataset exposes the metereological sensor data in RDF allowing for queries that otherwise would require a lot of time by just looking at the original files. This section aims to show how the different types of observations can be retrieved and filtered by submitting SPARQL queries to the endpoint[24]. To

[18] http://bit.ly/Qo0NX5

[19] http://virtuoso.openlinksw.com/

[20] http://www4.wiwiss.fu-berlin.de/pubby/

[21] http://bit.ly/OGn2S

[22] http://aemet.linkeddata.es/links/

[23] http://dbpedia.org/

[24] http://aemet.linkeddata.es/sparql

make the text more readable, we introduce the next prefix declarations:

```
ssn: <http://purl.oclc.org/NET/ssnx/ssn#>
aemet: <http://aemet.linkeddata.es/ontology/>
w3ctime: <http://www.w3.org/2006/time#>
geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
```

6.1. Retrieving the contents from the observations

Thanks to the modeling proposed in section 4, each observation can be accessed by any of its attributes. For example, given the location (latitude and longitude) of a weather station and a date and an hour, we could retrieve the observations for that exact time with the following query:

```
SELECT distinct ?station ?obs ?est ?prop ?value ?q
WHERE {
    ?station geo:location ?position ;
       aemet:stationName ?est .
   ?position geo:lat <latitude> .
   ?position geo:long <longitude> .
   ?obs ssn:observedBy ?station ;
       ssn:observedProperty ?prop ;
       aemet:valueOfObservedData ?value ;
       aemet:observedDataQuality ?q ;
       aemet:observedInInterval ?inter .
   ?inter w3ctime:hasBeginning ?instant .
   ?instant w3ctime:inDateTime ?dateTime .
   ?dateTime w3ctime:inXSDDateTime <myDateTime>
. }
```

The query also retrieves the name of the weather station measuring the observation, the measured property, the value of the measure and the quality of the value. <latitude> and <longitude> would correspond to the latitude and longitude of the station (e.g. “43.3672222222”, “-8.41944444444”), and <myDateTime> would be the date at which we are interested to recover observations, in xsd:DateTime format (e.g. “2011-05-27 01:40:00”).

Additional filters can be added, e.g. if the observations to retrieve from a weather station must be during a time interval, the final part of the query can be replaced with:

```
?dateTime w3ctime:inXSDDateTime ?dt .
FILTER (?dt >= xsd:dateTime("timeIntervalStart")).
FILTER (?dt <= xsd:dateTime("timeIntervalEnd")).
```

If the objective is to recover a single property of the weather station, the next SPARQL fragment can be added to the query:

```
?obs ssn:observedProperty <desiredProperty> .
```

Where <*desiredProperty*> would correspond to the different properties that describe the sensor measurements in the model. For example, we could choose the velocity of the wind in m/s (e.g., http://aemet.linkeddata.es/resource/WindAmbientProperty/VV10m). Other possibilities are filtering by the quality of the observation (the numeric "q" value in the query), proximity to a location (by comparing latitudes and longitudes), etc.

6.2. An application for representing and retrieving meteorological data

By extending the queries introduced in the previous section, we have built an application[25] for showing the contents of the AEMET dataset. The application consists on a viewer based on Map4RDF [5], a tool to visualize and access to LD resources.

Each station is represented with a marker in a map. When a user clicks on one of the markers, a new pop up window appears showing the last stored measure and the value of each of its observations (wind, temperature, pressure, etc.). The value of the last measure is retrieved dynamically from the endpoint, enabling the application to show up an updated value at the specific point where it was measured. Additionally, the windows show a menu to display the graphics of the values for an observation by the specific weather station.

Another reported use of our dataset is the evaluation of meteorological models and streams of sensor data. As an example, the work described in [3] builds a self-contained evaluation system which links raw sensor measurement to high-level semantics using our dataset.

Finally, other potential uses of the dataset show off when we combine it with other datasets in LD. For example, it can be used with the GeoLinkedData dataset exploiting its geospatial information (e.g., measure the wind speed in stations close to airports to analyze its relation to flight delay, check the temperature close to certain rivers, villages or provinces for biology studies, etc.).

7. Conclusions

In this paper we have presented the AEMET LD dataset, which exposes public meteorological data from AEMET resources as LD. We have reused ontologies like the W3C Time ontology and the SSN ontology (widely used in other systems); we have shown how to retrieve the data, create and update and expose the triples automatically in our endpoint and show the

[25] `http://aemet.linkeddata.es/browser.html`

results in a generic visualizer that considers the geoposition of the data.

Other systems can consume this data and offer visualizations, aggregations, reason over it, use it for evaluations, etc. Once the data has been extracted and is offered as LD, applications that use LD can instantly benefit from it. These data can be linked with data from sectors like agriculture or tourism to find correlations and study its impact. Moreover, publishing the dataset as LD allows modularity and reusability. For example, less time would be required for the development of applications based on this data (in particular, it would save time understanding its format and parsing without errors, given the fact that it is standarized and there are many tools to process it) as we have already shown with the visualizer. Finally, by complying with well established standards, interoperability is enhanced (e.g., within other sensor networks and applications using these ontologies), and the linking in other domains such as medical, geographical, biological, etc.

There are two known shortcomings of our dataset. The first one is the maintenance and discovery of high quality links to external datasets, which is a daunting task due to the large size and diverstity of such datasets. In order to tackle this issue, we plan to re-explore the existing datasets in the cloud and try to find the commonalities with those that expose additional information of sensors and locations.

The second shortcoming is addressing the scalability of the data, due to the growth of the triples per day. We currently address scalability issues by keeping only the latest week in the dataset (following AEMET policy), and storing the rest of the files as compressed unprocessed data. As future work, we plan to produce summaries of older data in a separate endpoint, so as to be able to store and query a wider range of days efficiently.

As other lines of future work, we plan to publish additional types of measurements recently exposed by AEMET (e.g., solar radiation, weather forecasts for the current day, etc.), which could be very useful when combined with biology datasets (how species migrate) or for users planning trips.

Acknowledgments This work has been supported by the Spanish project myBigData (TIN2010-17060). We thank A. de León, I. Santana, M. A. García, R. Alcazar, and S. Hurtado.

References

[1] N. Aswani, K. Bontcheva, and H. Cunningham. Mining information for instance unification. In *The Semantic Web - ISWC 2006*, volume 4273 of *Lecture Notes in Computer Science*, pages 329–342. Springer Berlin Heidelberg, 2006.

[2] D. Booth. URIs and the Myth of Resource Identity. In *Workshop on Identity, Meaning and the Web at International World Wide Web Conference. (Edinburgh, Scotland) ACM*, 2006.

[3] J.-P. Calbimonte, Z. Yan, H. Jeung, O. Corcho, and K. Aberer. Deriving Semantic Sensor Metadata from Raw Measurements. In *Proceedings of the 5th International Workshop on Semantic Sensor Networks SSN2012*, 2012.

[4] J. de Bruijn and A. Polleres. Towards an ontology mapping specification language for the semantic web. In *DERI Technical Report 2004-06-30*, 2004.

[5] A. de León, F. Wisniewki, B. Villazón-Terrazas, and O. Corcho. Map4rdf - Faceted Browser for Geospatial Datasets. In *Proceedings of the First Workshop on USING OPEN DATA*. W3C, June 2012.

[6] A. De León, V. Saquicela, L. M. Vilches, B. Villazón-Terrazas, and F. Priyatna. Geographical linked data : a spanish use case. In *I-SEMANTICS 6th International Conference on Semantic Systems*, September 2010.

[7] G. Fliedl, C. Kop, and J. Vöhringer. From owl class and property labels to human understandable natural language. In *Natural Language Processing and Information Systems*, volume 4592 of *LNCS*, pages 156–167. Springer Berlin Heidelberg, 2007.

[8] R. García-Castro, O. Corcho, and C. Hill. A Core Ontology Model for Semantic Sensor Web Infrastructures. *International Journal on Semantic Web and Information Systems*, 2012.

[9] A. Jaffri, H. Glaser, and I. Millard. Uri disambiguation in the context of linked data. In *Linked Data on the Web*, Beijing, China, 2008.

[10] D. Schober, B. Smith, S. E. Lewis, W. Kusnierczyk, J. Lomax, C. Mungall, C. F. Taylor, P. Rocca-Serra, and S.-A. Sansone. Survey-based naming conventions for use in OBO Foundry ontology development. *BMC Bioinformatics*, 10(125), 2009.

[11] M.-C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi. *Ontology Engineering in a Networked World*. Springer, Berlin, 2012.

[12] B. Villazón-Terrazas, A. Gómez-Pérez, and J. P. Calbimonte. NOR$_2$O: a Library for Transforming Non-Ontological Resources to Ontologies. In *ESWC*, volume 5554 of *LNCS*. Springer, 2010.

[13] B. Villazón-Terrazas, L. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez. Methodological Guidelines for Publishing Government Linked Data Linking Government Data. In *Linking Government Data*, chapter 2, pages 27–49. Springer New York, New York, NY, 2011.