Integrated Spatio-temporal Data for Earth Observation: A RDF dataset of Territorial Units with their Land Cover

Ba-Huy Tran, Nathalie Aussenac-Gillles, Catherine Comparot, Cassia Trojahn

IRIT, Université de Toulouse et CNRS, France
prenom.nom@irit.fr

Abstract. In this paper, we propose an ontology-based approach that relies on data spatial and temporal dimensions to semantically integrate open datasets together with Earth Observation data. The resulting dataset provides rich contextual information about EO and makes it possible to search EO images according to this contextual information. We illustrate the approach on three French datasets: administrative units, land registries and their land cover dataset. The resulting dataset is a semantic database and it is exposed through a semantic search interface or can be accessed via a SPARQL endpoint.

1 Introduction

The European Copernicus program developed a series of satellites called Sentinel to collect Earth observation (EO) data. The vast majority of the program data is open to the public and made available on the web through Data Integration and Analysis Systems (DIAS). The data accessible opens up many economic prospects thanks to new applications in various fields. Among the European projects started to exploit such data are BETTER [1], openEO [2], Perceptive Sentinel [3] or EOPEN [4] and CANDELA [5]. CANDELA aims at creating a platform that provides building blocks and services and that allows users to quickly use, manipulate, explore, and process Copernicus data together with large sets of open data.

In the context of the CANDELA project, a "Semantic search" module is developed to enable the search for enriched EO data through different heterogeneous data sources. These sources are selected in keeping with the requirements and scenarios of use-cases. In this paper, we propose an approach that relies on data spatial and temporal dimensions to semantically integrate these datasets together with Earth Observation data. The resulting dataset provides rich contextual information about EO and makes it possible to search EO images according to this contextual information.

We distinguish three categories of sources of interest:

— Sentinel image metadata: these metadata are available together with the Sentinel EO images in the DIAS.

[1] https://www.ec-better.eu/
[2] https://openeo.org/
[3] http://www.perceptivesentinel.eu/
[4] https://eopen-project.eu/
[5] http://www.candela-h2020.eu/

— Sentinel-image-related data: this data is extracted from Sentinel images through image processing. It may be provided as a dataset (for instance, change detection results provided by our partners) or computed by formulas (for example, the vegetation index).

— Contextual datasets: these datasets contain open data or linked open data. A number of open datasets can provide contextual information for EO studies on France, for instance French administrative unit data, weather measure data, weather bulletin data, land register and its land cover.

Data in all these sources is characterized by its spatial and temporal dimensions. But the sources are heterogeneous by their content, their structure and their format. The format may be databases, JSON, CSV or XML structured files. The representation can be made more homogeneous by using the same format for all datasets. This homogeneity is purely syntactic and it doesn't guarantee the quality of the integration. Homogeneity is also necessary at a semantic level. Then it requires to define and use a single and unifying vocabulary or better an ontology.

Another challenge is that several preliminary processes must be performed on the dataset itself or on the data. For instance, it may be required to sample or check the data, to average some values, to turn numeric data into qualitative values (or the reverse), or to select only a part of the properties because not all properties are relevant. Other kinds of processes may be required to merge similar data from various sources, or to aggregate information available at the pixel level and compute it for larger aeras, like the geometry of parcels or administrative units. Then a spatial processing is needed. For instance, we implemented a spatial process to compute at parcel level a vegetation index or a change value (from change detection dataset) which are originally available at the pixel level .

In this paper, we describe a part of our work that aims to propose an ontology-based integration process for datasets in relation with EO, based on spatial and temporal features. To illustrate our approach, we explain how we semantically integrated three French datasets: administrative units, land registries and land covers. The resulting dataset is stored and published as a semantic database and it is exposed through a semantic search interface. It can be accessed via a SPARQL endpoint. Cadastral parcels at a given period and the evolution of their dominant land covers can thus be retrieved using the village to which they belong or a given area thanks to their geometry.

The rest of this paper is organized as follows. First, the data sources are described. Next, we present the ontology-based data integration process, in particular the integration model and the system architecture, through a use-case. Finally, the conclusion summarizes the achieved progress and our future work.

2 Data sources

As introduced below, we aim at integrating three datasets from different sources. In particular, data describing French administrative units can be obtained from open datasets:

— OpenStreetMap based dataset[6]: the dataset is available as a shape files. It is updated yearly based on open cartographic data of OpenStreetMap.

— GeoZones dataset[7]: this dataset comes from a certified public service. The purpose of

[6] https://www.data.gouv.fr/en/datasets/decoupage-administratif-communal-francais-issu-d-openstreetmap/

[7] https://www.data.gouv.fr/en/datasets/geozones/

this dataset is to provide a common geospatial and administrative repository based on open data. It is available in JSON format.

We selected GeoZones because it provides more details about administrative units compared to the first ones. Beside the basic information, for instance the id, code, name and geometry, the dataset provides additional information such as the area, population and especially links to other open datasets (for example, Geonames, INSEE, Wikipedia or Wikidata).

Along with administrative units, land register data is also available from the French government data website [8] in GeoJSON format or shapefiles. The dataset indicates the identification and the localization of parcels from land register.

Land cover data is also considered in order to provide more contextual information for Earth Observation studies. Various land cover datasets are available as open data, each of them havinf its own way to evaluate the land cover and its own set of land cover classes:

- Global Land Cover SHARE dataset [9]: Created by FAO in 2012, it is provided in raster format as a GeoTIFF file. The value of each pixel is an integer that represents the identifier of the most prevalent land cover class for the area that is covered by the pixel.
- Cesbio and cover dataset [10] is updated yearly. It is only available for France, in raster format as a GeoTIFF file.
- Corine Land cover dataset [11] is rather the most standard one as it uses the Corine Land cover vocabulary [12]. The two most recent versions of the dataset were published in 2012 and 2018.

Currently the Cesbio dataset and the corresponding land cover classification are integrated to our system. The Corine dataset will be considered in further developments.

3 Semantic integration

The three datasets described above are heterogeneous by their content, their structure and their format. In order to integrate them and eventually to perform data pre-processing, we rely on a semantic approach for data integration. At the heart of semantic data integration is the ontology that acts as a mediator for re-conciliating the heterogeneities between different data sources (Wache et al., 2001; Cruz and Xiao, 2005). Our approach relies on defining an ontology that serves as basis for the integration together with a process that converts the different formats and data to instances of this ontology.

The semantic data integration process can be divided into two main stages: semantic representation and data integration. The first stage aims to build a modular ontology with specific parts adapted from each source schemas, while at the second stage, we integrate the data sources based on this ontology and a set of transformation rules.

[8] https://cadastre.data.gouv.fr/datasets/cadastre-etalab

[9] http://www.fao.org/geospatial/resources/detail/en/c/1036591/

[10] http://osr-cesbio.ups-tlse.fr/ oso/

[11] https://www.data.gouv.fr/en/datasets/corine-land-cover-occupation-des-sols-en-france

[12] http://dd.eionet.europa.eu/vocabulary/landcover/clc

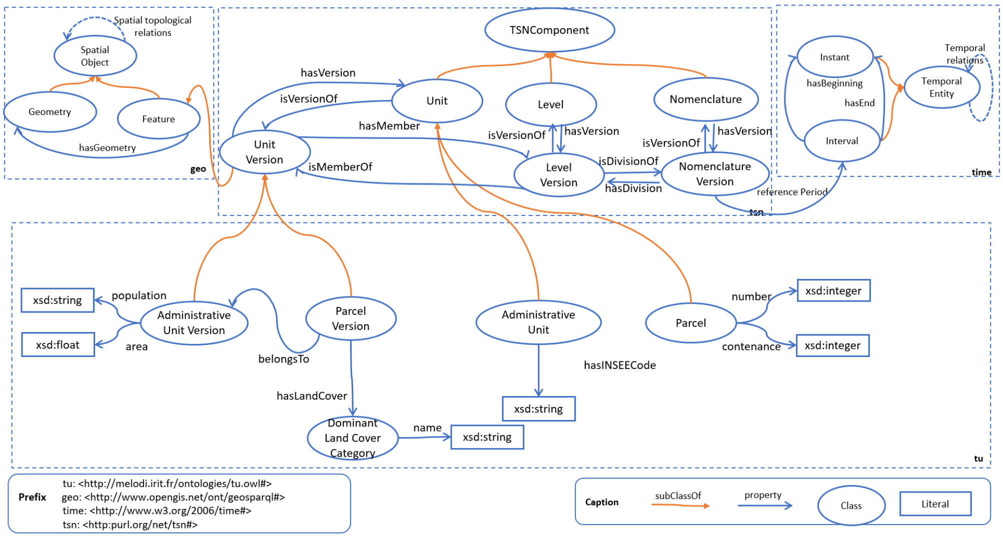


FIG. 1 – *A modular ontology to represent French administrative units, land registry parcel, their dominant land cover, and their history.*

3.1 Semantic representation

The modular ontology required for data integration was developed by reusing existing vocabularies, which forms a generic part, and by defining specific classes and properties dedicated to the data to be integrated, which form the specific part of the ontology. cadastral and CESBIO LandCover data are updated yearly. With the TSN Ontology it is possible to manage different versions of each cadastral parcel The TSN Ontology (Territorial Statistical Nomenclature Ontology (Bernard et al., 2018)) describing any territorial statistical nomenclature is reused to represent French administrative units, cadastral parcels and their land cover; and their history.

The TSN Ontology adopts the perdurantist approach of ontologies for fluents (Welty and Fikes, 2006) to describe the TSN elements that vary in time; however, the authors rather use the term *Version* while other ontologies for fluents use *TimeSlice*. As territorial units are geolocalized and dated entities, the ontology reuses in turn the OWL-Time ontology (Hobbs and Pan, 2004) and the GeoSPARQL ontology (Battle and Kolas, 2012).

The OWL-Time ontology, dedicated to concepts and temporal relationships as defined in the theory of Allen, is used first to describe the temporal content of Web pages and the temporal properties of web services. The ontology is recommended by the W3C for modeling temporal concepts due to its vocabulary for expressing topological relations between instants and intervals.

The GeoSPARQL ontology, an OGC standard, introduced the geo:SpatialObject class composed of two primary subclasses, *geo:Feature* and *geo:Geometry*. The first one represents an entity of the real world while the later represents all geometric forms defined on a spatial coordinate reference system. An entity is associated to its geometries by the *hasGeometry* relation.

Our modular ontology introduces two classes, *Administrative Unit* and *Parcel* that extend the TSN *Unit* to represent administrative units and cadastral parcels respectively. To take into account different timeslices of these entities through time we specialized the *TSN Unit Version class* with the *Administrative Unit Version* and *Parcel Version* classes. A dominant land cover is associated to each timeslice of the parcels.

3.2 Data integration

The data integration stage can be accomplished in two ways based on the constructed ontologies and their relations established in the previous stage: either a mediator is built for virtual systems (on-demand mapping) or data materialization is accomplished for materialized systems.

— On-demand mapping: Data remain located in their source, as a consequent, semantic queries must be rewritten into SQL ones at the query evaluation step. The approach is well suited in the context of very large datasets that would hardly support centralization due to resource limitations.

— Data materialization: Like in warehouse approaches, data sources are transformed into RDF graphs that are thereafter loaded into a triple store and accessed through a SPARQL [13] query engine. The whole process is often referred to as Extract-Transform-Load (ETL). The major advantage of the approach is to facilitate further processing, analysis or reasoning on the materialized RDF data. Third-party reasoning tools can be used to infer complex entailments since materialized data are made available at once. Furthermore, complex queries can be answered without compromising the run-time performances because the reasoning task has been performed at an earlier stage.

We believe that the materialization approach is more appropriated for integrating spatio-temporal data in general and for our project in particularly, because:

— It isn't easy to accomplish an on-demand mapping since along with other datasets, the three presented sources are available in JSON format, GeoTIFF image, shapefile or even remote compressed files.

— The stable performances of the approach in querying and reasoning tasks. In fact, the performance of the on-demand mapping system is heavily dependent on the ontology, the schema of the database, and the mappings.

— Moreover, a geospatial triplestore can also be used as a warehouse to store semantic data so that data enrichment.

— Finally, federated queries that perform spatial joins spanning different geospatial endpoints are not yet supported in any federated system (Brüggemann et al., 2016).

3.3 System architecture

The system is developed through docker technology to be in compliance with other partners tools and deployed on the project platform. There are two dockers as described in Figure 2.

[13] http://www.w3.org/TR/sparql11-query/

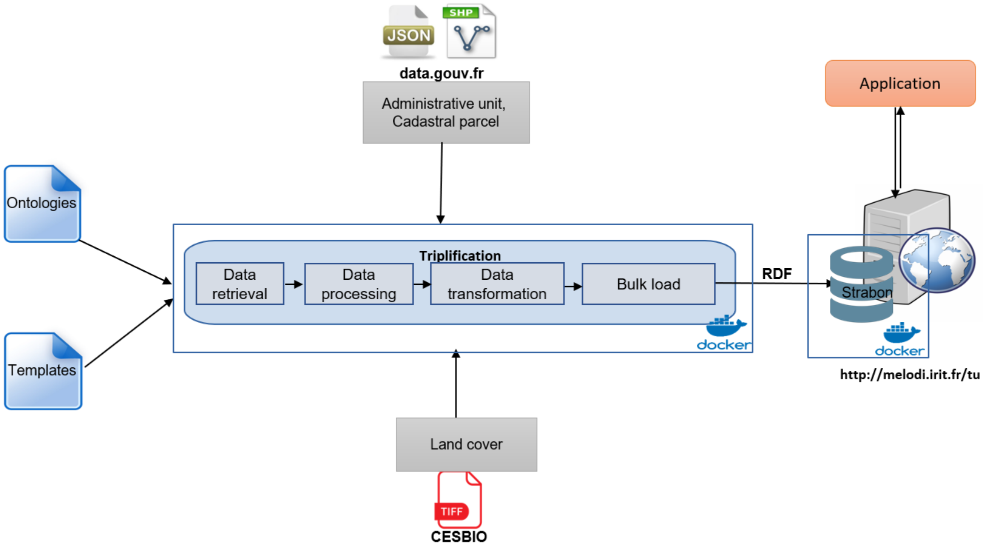


FIG. 2 – *The system architecture based on docker technology.*

The triplification docker is dedicated to the semantic data integration while the other is used for hosting the knowledge-base.

3.3.1 Triplification

The docker contains Python scripts for data retrieval, data processing, data transformation and triplestore bulk load. The three later operations correspond to ETL approaches.

1. Data retrieval: The module is used to download the remote datasets of interest with spatial and temporal filters.For example, the dataset containing information of a village can be retrieved based on its insee code and the publication year.
2. Data processing: This step is used for several purpose depending on the dataset. For example, some of the feature properties are chosen because not all properties are relevant; aggregation is done on pixels to produce new properties; or spatial mask is made on raster image to eliminate undesired area. Here the dominant land cover of each parcel is computed by three steps:
   (a) Apply the cadastral parcel (its geometry at a given date) as a mask on the CESBIO land cover raster image file of the same year.
   (b) Determinate the most probable land cover for each pixel inside the mask .
   (c) Determinate the most dominant land cover of the parcel based on the corresponding number of pixels.
3. Data transformation: This step aims to transform the processed data into semantic one. Templates that define the mapping between the source schema and the ontologies are

refered to perform the process. The templates are usually hand written based on the developed ontologies and the data in each data source. They make explicit the mappings between the ontologies and the data source schemas. In other words, the template is an explicit writing of how entities of some category in the schema of the data source will be represented with the ontology classes and properties.

A data translation tool, such as D2RQ [14], Ultrawrap [15], Morph [16] Ontop [17], or GeoTriples [18], that makes use of such mappings can be used. However, we have chosen to evolve the mapping template and processing mechanism of our recent work (Arenas et al., 2016) because it contains functions helping to perform more sophisticated operations that are not possible in alternative approaches.

Output of this step are semantic data files in N-triples format.

4. Data bulk load: The final step is used to import the semantic data generated to the triplestore.

3.3.2 The triplestore

The second docker is used to deploy the geospatial triplestore that manage the knowledge-base. An endpoint is also hosted. Triplestores are DBMS for data modeled in RDF. Currently, several triplestores support storing and querying spatial data using the GeoSPARQL or stSPARQL query language. Those open-source that manage the best are Parliament [19] (Battle and Kolas, 2012) and Strabon [20] (Kyzirakos et al., 2012). Only the two triplestores explicitly adopt the existing geospatial geometry standard although many triplestores now support spatial queries of different complexity (Scheider et al., 2017). Strabon has been chosen in our project as it has many advantages:

— Strabon extends the Sesame triplestore with the capacity of storing spatial RDF data in the PostgreSQL DBMS enhanced with PostGIS. The triplestore works over the stRDF data model, a spatio-temporal extension of RDF. The triplestore has a good overall performance thanks to particular optimization techniques that allow spatial operations to take advantage of PostGIS functionality instead of relying on external libraries (Patroumpas et al., 2014).

— Strabon also provides a SPARQL endpoint that helps to access the content of the triplestore with both complex stSPARQL and GeoSPARQL queries. The interface also provides an additional possibility to manage the knowledge base, for instance storing and updating functionality with SPARQL Update.

The triplestore endpoint that receives SPARQL queries for knowledge base discovery is accessible on a server [21]. Currently, only the department 33 and its cadastral data is made available due to the lack of system resources.

[14] http://d2rq.org/

[15] https://capsenta.com/ultrawrap/

[16] http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/technologies/315-morph-rdb/

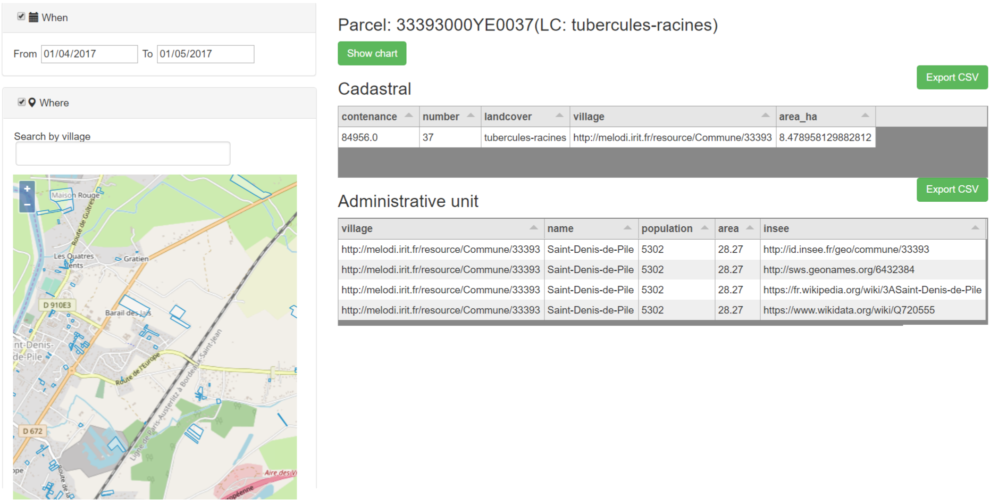
[17] http://ontop.inf.unibz.it/

[18] http://geotriples.di.uoa.gr/

[19] http://parliament.semwebcentral.org/

[20] http://strabon.di.uoa.gr/

[21] http://melodi.irit.fr/tu



☑ When

From 01/04/2017 To 01/05/2017

☑ Where

Parcel: 33393000YE0037(LC: tubercules-racines)

Cadastral

contenance	number	landcover	village	area_ha
84956.0	37	tubercules-racines	http://melodi.irit.fr/resource/Commune/33393	8.478958129882812

Administrative unit

village	name	population	area	insee
http://melodi.irit.fr/resource/Commune/33393	Saint-Denis-de-Pile	5302	28.27	http://id.insee.fr/geo/commune/33393
http://melodi.irit.fr/resource/Commune/33393	Saint-Denis-de-Pile	5302	28.27	http://sws.geonames.org/6432384
http://melodi.irit.fr/resource/Commune/33393	Saint-Denis-de-Pile	5302	28.27	https://fr.wikipedia.org/wiki/3ASaint-Denis-de-Pile
http://melodi.irit.fr/resource/Commune/33393	Saint-Denis-de-Pile	5302	28.27	https://www.wikidata.org/wiki/Q720555

FIG. 3 – *A use-case in the CANDELA project that makes use of the semantic dataset.*

In the context of the CANDELA project, several use-cases defined in the project can make use of the semantic database. Figure 3 represents a part of our tool that queries land register data with a spatio-temporal filter.

4 Conclusion

We described an approach integrate various spatio-temporal data source including French administrative units and land registry along with its land cover. As future work, we consider to apply the same approach for other available datasets. For example, vegetation index and change detection information can be modeled, processed and transformed so that they can be attached to parcels throughout their life. We also plan to perform inferences on the constructed knowledge-base.

References

Arenas, H., N. Aussenac-Gilles, C. Comparot, and C. Trojahn (2016). Semantic Integration of Geospatial Data from Earth Observations. In *20th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2016)*, Bologne, Italy, pp. pp. 97–100.

Battle, R. and D. Kolas (2012). Enabling the geospatial semantic web with parliament and geosparql. *Semant. web 3*(4), 355–370.

Bernard, C., M. Villanova-Oliver, J. Gensel, and H. Dao (2018). Modeling changes in territorial partitions over time: Ontologies tsn and tsn-change. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, New York, NY, USA, pp. 866–875. ACM.

Brüggemann, S., K. Bereta, G. Xiao, and M. Koubarakis (2016). Ontology-based data access for maritime security. In H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, and C. Lange (Eds.), *The Semantic Web. Latest Advances and New Domains*, Cham, pp. 741–757. Springer International Publishing.

Cruz, I. F. and H. Xiao (2005). The Role of Ontologies in Data Integration. *Journal of Engineering Intelligent Systems 13*, 245–252.

Hobbs, J. R. and F. Pan (2004). An ontology of time for the semantic web. *ACM Transactions on Asian Language Information Processing 3*, 66–85.

Kyzirakos, K., M. Karpathiotakis, and K. M. (2012). Strabon: A semantic geospatial dbms. In *The Semantic Web ISWC 2012*, Berlin, pp. 295–311. Springer.

Patroumpas, K., G. Giannopoulos, and S. Athanasiou (2014). Towards geospatial semantic data management: Strengths, weaknesses, and challenges ahead. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14, New York, NY, USA, pp. 301–310. ACM.

Scheider, S., A. Degbelo, R. Lemmens, C. van Elzakker, P. Zimmerhof, N. Kostic, J. Jones, and G. Banhatti (2017). Exploratory querying of sparql endpoints in space and time. *Semantic web 8*(1), 65–86.

Wache, H., T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner (2001). Ontology-Based Integration of Information - A Survey of Existing Approaches. In *IJCAI-01 Workshop: Ontologies and Information*, pp. 108–117.

Welty, C. and R. Fikes (2006). A reusable ontology for fluents in owl. In *Proceedings of the 2006 Conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006)*, Amsterdam, The Netherlands, The Netherlands, pp. 226–236. IOS Press.

Résumé

Dans cet article, nous proposons une approche basée une ontologie et sur les propriétés spatiales et temporelles de données pour intégrer sémantiquement des jeux de données ouvertes à des images d'observation de la Terre. Le jeu de données résultat fournit des informations contextuelles riches sur les images d'observations de la Terre et rendent possible une recherche de ces images sur la base de ces informations. Nous illustrons cette approche sur trois jeus de données français : les unités administratives, les registres du cadastre et leur couverture terrestre. Ce jeu de données est accessiblevia une interface de recherche sémantique ou via un point d'accès SPARQL.