An Approach for Integrating Earth Observation, Change Detection and Contextual Data for Semantic Search

Ba-Huy Tran, Nathalie Aussenac-Gilles, Catherine Comparot, Cassia Trojahn dos Santos

AN APPROACH FOR INTEGRATING EARTH OBSERVATION, CHANGE DETECTION AND CONTEXTUAL DATA FOR SEMANTIC SEARCH

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

IRIT - CNRS and University of Toulouse
118, route de Narbonne, 31062 Toulouse Cedex 9 - France
firstname.lastname@irit.fr

ABSTRACT

This paper presents an integration process of open data and Earth Observation (EO) data for supporting EO semantic search. This process relies on an ontology that describes spatial and temporal dimensions of data. The resulting dataset provides rich contextual information about EO and makes possible the search of EO data according to this contextual information through a semantic search interface. The approach is illustrated on the integration of different datasets: change detection, administrative unit, land register and land cover.

***Index Terms*—** EO, change detection, data integration

1. INTRODUCTION

The series of satellites launched by the European Copernicus program, generating free EO data, opens up many economic perspectives, contributing to the emergence of new applications in various fields. In particular, the CANDELA[1] project aims at creating a platform that provides building blocks and services allowing users to quickly use, manipulate, explore, and process Copernicus data together with large sets of open data. One of the motivations to build such a platform is that searching for images just with their original meta-data, mainly using the sensor type, the capture date and location is not sufficient to find relevant images for a specific purpose. Contextual information, coming from different heterogeneous data sources, may be useful as well as a "Semantic search" module. By semantic search we mean services to retrieve images through a semantic description of their content (i.e. kind of vegetation, change detection results), their location and date, or any semantic feature coming from open data (weather measures, places, etc.).

Semantic search relies on a formal representation of data that can be linked to images thanks to their validity period and localization. So, a preliminary work to the design of semantic search facilities for a specific use-case is to identify relevant datasets to be used as contextual data to describe images, and then to propose a homogeneous representation of this heterogeneous data. We propose an ontology-based approach for integrating heterogeneous data, including EO data, such as Sentinel image metadata, change detection results on Sentinel images, and contextual open data, such as administrative unit data, weather measurements or report data, land register or land cover. Contextual information is linked to EO data based on spatial and temporal relations. To illustrate our approach, we explain how the above datasets were semantically integrated. The resulting semantic data is stored and published as a semantic database. It is exploitable through a SPARQL endpoint and a semantic search interface.

We present first the ontology-based data integration approach, in particular the integration model and the system architecture, through a use-case. Then we give some details about the semantic search interface. We finally summarize our achievements and highlight future work.

2. SEMANTIC INTEGRATION

Our approach for data integration is illustrated with a set of data sources. We selected three datasets that provide information about territorial units or cadastre parcels and the land cover over areas of interest. This data is not directly accessible on EO images and comes as contextual information that support a better interpretation of image analysis. The change detection results on Sentinel images forms the fourth dataset. These four datasets are heterogeneous by their content, their structure and their format. Data integration is needed to use these sources altogether as information documenting EO images. Our hypothesis is that semantic technologies provide a good support to data integration. At the heart of semantic data integration is an ontology that acts as a mediator for re-conciliating the conceptual and terminological variations between different data sources [1]. We defined an ontology that serves as basis for data representation and integration together with a process that converts data into instances of this ontology.

The semantic data integration process required two main

This work is partially funded by the CANDELA project under convention number 776193 of the H2020 EC Research and Innovation Program.

[1] http://www.candela-h2020.eu/

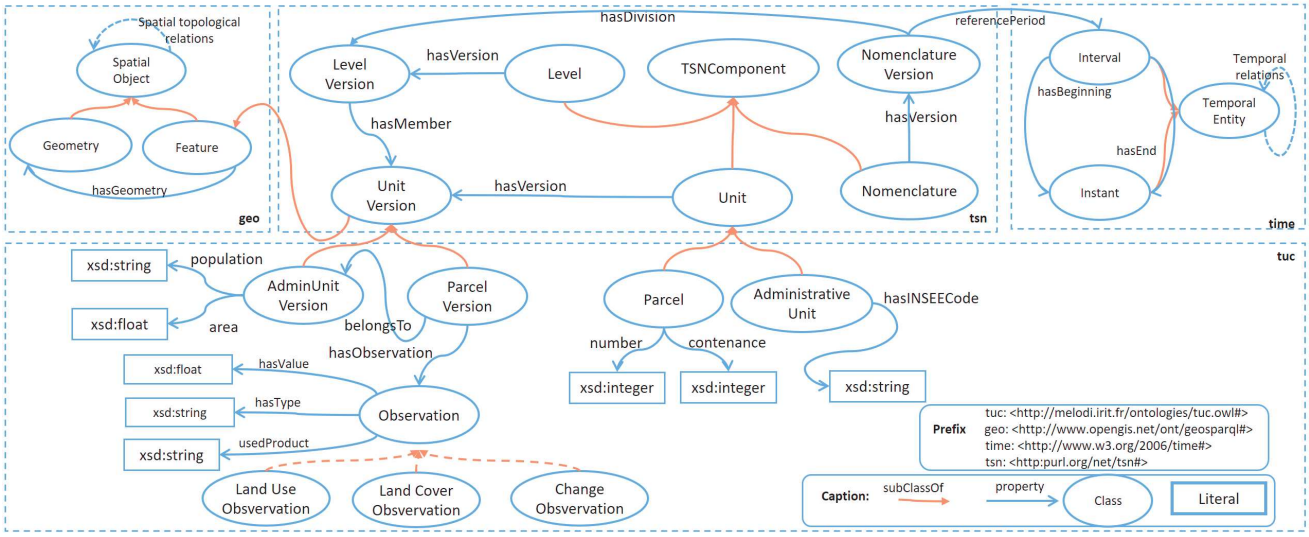

Fig. 1. Modular ontology for integrating heterogeneous datasets.

tasks: i) semantic representation where we built a modular ontology representing the data sources and ii) data integration where we defined a set of transformation rules to be used for data conversion together with the ontology.

2.1. Semantic representation

We built a modular ontology composed of a generic part and a specific part. The generic modules structure the concepts and properties required to represent any data with spatial and temporal dimensions in relation with territorial entities. The specific part is dedicated to data to be integrated. Fig.1 represents the corresponding modular ontology where the specific part fits our use case. The generic part of the ontology reuses three existing vocabularies:

• **TSN Ontology**: the Territorial Statistical Nomenclature Ontology [2] describes any territorial statistical nomenclature. The ontology adopts the notion of *perdurant* from ontologies for fluents [3] to describe the TSN elements that vary in time; however, the authors rather use the term *version* while other ontologies for fluents use *timeslice*. We reused some of its concepts (top middle box in Fig.1) to represent administrative units and their various states throughout time. Indeed, since the presented datasets are updated regularly, a version of the dataset forms a version of the nomenclature.

• **OWL-Time ontology** [4]: recommended by the W3C, the OWL-Time ontology is dedicated to concepts and temporal relationships as defined in the theory of Allen. We reused three main concepts: `time:Instance`, `time:Interval` and `time:TemporalEntity` (top-right box in Fig.1) to date nomenclature versions with the `tsn:referencePeriod` property.

• **GeoSPARQL ontology** [5]: From GeoSPARQL ontology, an OGC standard, we reuse the `geo:SpatialObject` concept composed of two subclasses, `geo:Feature` and `geo:Geometry` (top-left box in Fig.1). The first one represents an entity of the real world and the later represents all geometric forms defined on a spatial coordinate reference system. An entity is associated to its geometries by the `geo:hasGeometry` property. Unit versions are spatial entities.

The specific part of the ontology is presented in the bottom box in Fig.1 as a single module with `tuc` as name space prefix. Two classes, `tuc:AdministrativeUnit` and `tuc:Parcel` extend the `tsn:Unit` class from TSN to represent administrative units and cadastral parcels respectively. In order to take into account different states of these entities over time, we specialized the `tsn:Version` class with two sub-classes: `tuc:AdminUnitVersion` and `tuc:ParcelVersion`. A *Parcel Version* can be associated to an `tuc:Observation` made on the parcel during the period, or analysis results, such as land cover or changes computed from raster files.

2.2. Data integration

To study land use evolution or the identification of agricultural productions from EO images, we use contextual data from four sources: i) The GeoZones dataset[2] provides the French administrative units in JSON format. This dataset comes from a certified French public service that provides a common geospatial and administrative repository for France based on open data. For each unit, beside basic information

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

(id, code, name and geometry), we can get its area, its population and links to other open datasets (like Geonames, INSEE, Wikipedia or Wikidata). ii) Land register data is also available from the French government data website[3] in GeoJSON format or shapefiles. The dataset indicates the identification and the localization of parcels from land register. iii) Various land cover datasets are available as open data, each of them having its own way to evaluate the land cover from image rasters and its own set of land cover classes. Currently, we use the CESBIO source[4], which is updated yearly and available for France, in raster format as a GeoTIFF file. Other datasets such as Corine land cover will be considered later on. iv) Change detection results, available in raster format, from deep learning algorithms developed by Candela partners [6].

The data integration process is based on data materialization, where data from each of the sources is transformed into RDF graphs using the shared vocabulary of the modular ontology for class and property description. All the RDF graphs are later loaded into a single triplestore. The major advantage of the approach is to facilitate further processing, analysis or reasoning on the materialized RDF data. The data integration process is carried out for each data set in 4 stages:

1. **Data retrieval**: the remote contextual datasets of interest are downloaded using predefined URL containing spatial and temporal criteria. For example, the dataset containing information of a village can be retrieved based on the village INSEE code and the publication year. Other datasets take the form of raster files providing indices or parameters like the land cover classes or change values.
2. **Data pre-processing**: this stage aims at selecting some of the data in the datasets and at computing relevant parameter values, depending on the dataset and the user's goal. For example, only some relevant properties may be chosen; value aggregation may be done on raster pixels to produce new properties; or spatial masks may be mapped on raster files to eliminate undesired areas. In our use cases, the dominant land cover class or change level of each parcel is computed in three steps: (a) the parcel geometry (at a given period) is used as a mask and mapped on the raster files (CESBIO land cover or change detection result) of the same date; (b) determinate the land cover class or the change level for each pixel inside the mask; (c) determinate the most dominant land cover class or dominant change level of the parcel based on the corresponding number of pixels of each land cover class.
3. **Data transformation**: The processed data is next converted into a semantic format. To support this task, we defined templates that map the source schema with the ontology classes and properties. These templates are hand written based on the integration ontology and the data in each data source. We chose to evolve the mapping template and processing mechanism of our previous work [7] because it contains functions helping to perform sophisticated operations that are not possible in alternative approaches. The output of this step is a set of RDF files. An extract of the templates and the RDF files is available here[5].
4. **Data bulk load**: RDF files are uploaded in the triplestore. The RDF data is managed by Strabon[6] [8], a geospatial triplestore. Strabon extends the Sesame triplestore with the capacity of storing spatial RDF data in the PostgreSQL DBMS enhanced with PostGIS. Strabon has a good overall performance thanks to optimization techniques that allow spatial operations to take advantage of PostGIS functionality instead of relying on external libraries [9]. It also provides an endpoint to access the content of the triplestore with GeoSPARQL queries.

3. SEMANTIC SEARCH

The endpoint of the triplestore is accessible on a server[7]. Currently, data of all French administrative units is available for 2016, 2017 and 2019, but parcel data is only available for the department 33 due to the limited system resources. Fig.2 represents the tool GUI when querying land register data: April 2017 is the temporal filter (the "When"); an area located in the department 33 is the spatial filter (the "Where"); and a middle change level (the "What") is requested. A SPARQL query is formulated and sent to the endpoint. Filters are made based on spatio-temporal quantitative information of the parcels (the geometry of parcel versions and the validity period of the nomenclature version). The resulting parcels are next displayed on the map. When a parcel is selected, all related information is displayed on the right. In this example, the parcel number (33227000AK0477) and the land cover class ("vignes" - vineyard) appear on top of the tables; the parcel details, the corresponding village, and the change detection result are described by tables. This information can be used or next combined with the other sources for further analysis, for example, discovering changes occurred on vineyards after some frosts or investigating the land cover and administrative information of damaged parcels during a certain period.

4. CONCLUSION

This paper has presented an ontology-based approach for integrating various spatial and temporal data sources. The approach has been used to integrate contextual open datasets together with the results of change detection algorithms on EO images. As future work, we consider to apply the approach on other available datasets. For example, weather forecast measures or vegetation index can be attached as observations to parcels or administrative units; Sentinel image metadata can be used to provide context to these observations.

[3] https://cadastre.data.gouv.fr/datasets/cadastre-etalab
[4] http://osr-CESBIO.ups-tlse.fr/ oso/
[5] http://melodi.irit.fr/share/IGARSS2020
[6] http://strabon.di.uoa.gr/
[7] http://melodi.irit.fr/tuc

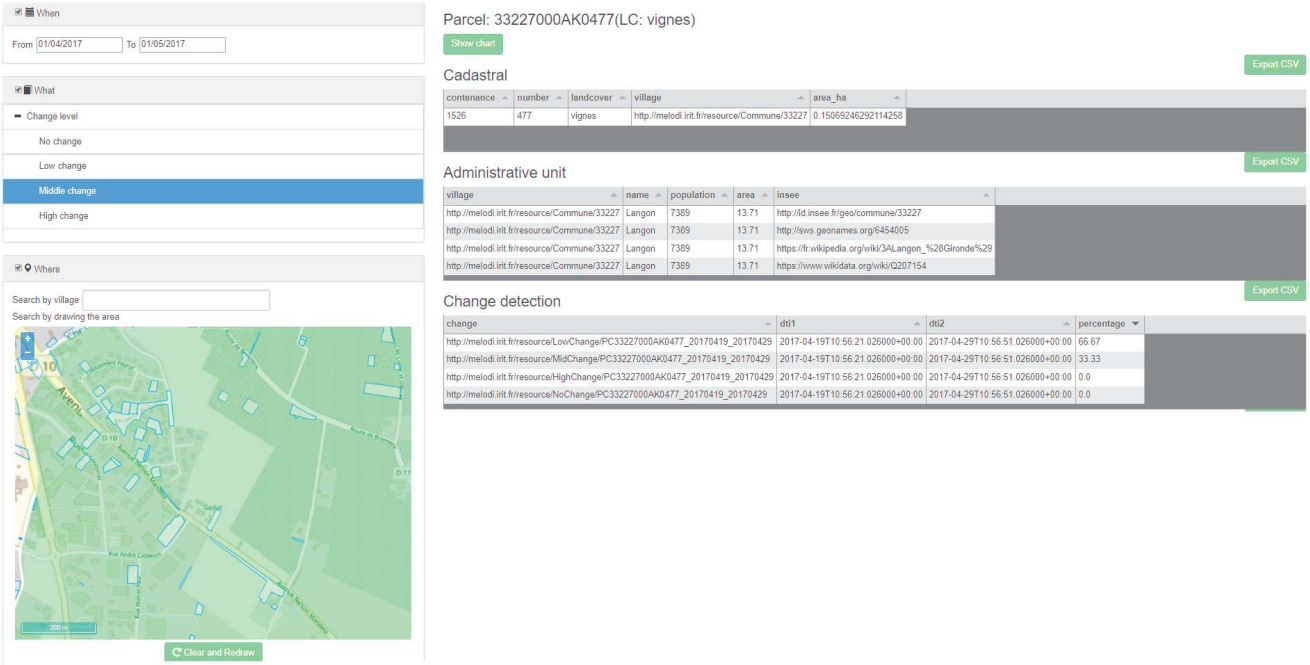

Fig. 2. A use-case in the CANDELA project that makes use of the semantic database.

5. REFERENCES

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

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

[3] C. Welty and R. Fikes, "A reusable ontology for fluents in owl," in *Proc. of the 2006 Int. Conf. on Formal Ontology in Information Systems (FOIS 2006)*, Amsterdam, NL, 2006, pp. 226–236, IOS Press.

[4] J. R. Hobbs and F. Pan, "An ontology of time for the semantic web," *ACM Transactions on Asian Language Information Processing*, vol. 3, pp. 66–85, 2004.

[5] R. Battle and D. Kolas, "Enabling the geospatial semantic web with parliament and geosparql," *Semant. web*, vol. 3, no. 4, pp. 355–370, Oct. 2012.

[6] M. Aubrun, A. Troya-Galvis, M. Albughdadi, R. Hugues, and M. Spigai, "Unsupervised learning of robust representations for change detection on sentinel-2 earth observation images," in *Proc. of the 13th Int. Symp. on Environmental Software Systems*, Wageningen (NL), 2020.

[7] H. Arenas, N. Aussenac-Gilles, C. Comparot, and C. Trojahn, "Semantic Integration of Geospatial Data from Earth Observations," in *20th Int. Conf. on Knowledge Engineering and Knowledge Management (EKAW 2016)*, Bologne (I), 2016, pp. 97–100.

[8] K. Kyzirakos, M. Karpathiotakis, and Koubarakis M., "Strabon: A semantic geospatial dbms," in *The Semantic Web ISWC 2012*, Berlin, 2012, pp. 295–311, Springer.

[9] K. Patroumpas, G. Giannopoulos, and S. Athanasiou, "Towards geospatial semantic data management: Strengths, weaknesses, and challenges ahead," in *Proc. of the 22Nd ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems*, New York, USA, 2014, pp. 301–310, ACM.