Copernicus Access Platform Intermediate Layers Small Scale Demonstrator

D2.5 Semantic Search v1

Document Identification			
Status	Final	**Due Date**	31/10/2018
Version	1.0	**Submission Date**	30/10/2018

Related WP	WP2	**Document Reference**	D2.5
Related Deliverable(s)	N/A	**Dissemination Level (*)**	PU
Lead Participant	IRIT-CNRS	**Lead Author**	Cassia Trojahn
Contributors		**Reviewers**	Michelle Aubrun (TAS-FR)
			N/A

Keywords:
Knowledge representation, ontologies, semantic data integration, semantic search, image metadata

This document is issued within the frame and for the purpose of the CANDELA project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 776193. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.

Document Information

List of Contributors	
Name	Partner
Nathalie Aussenac-Gilles	IRIT-CNRS
Catherine Comparot	IRIT-UT2J
Cassia Trojahn	IRIT-UT2J
Ba-Huy Tran	IRIT-CNRS

Document History			
Version	Date	Change editors	Changes
0.1	03/09/2018	Nathalie Aussenac-Gilles (IRIT-CNRS)	Changed the structure of the deliverable
0.2	02/10/2018	Nathalie Aussenac-Gilles (IRIT-CNRS)	Extended with chapters 2, 3 and 4
0.3	18/10/2018	Michelle Aubrun (TAS-FR)	Review of the document
0.4	22/10/2018	Juan Alonso (ATOS ES)	Quality Assessment
0.5	29/10/2018	Nathalie Aussenac-Gilles (IRIT – CNRS)	Minor reviews according to previous controls
0.6	30/10/2018	Juan Alonso (ATOS ES)	Quality Assessment
1.0	29/10/2018	Jose Lorenzo (ATOS ES)	Coordinator approval for submission

Quality Control		
Role	Who (Partner short name)	Approval Date
Deliverable leader	Cassia Trojahn (IRIT-CNRS)	18/10/2018
Quality manager	Juan Alonso (ATOS ES)	22/10/2018
Project Coordinator	Jose Lorenzo (ATOS ES)	30/10/2018

Table of Contents

List of Figures

List of Acronyms

Abbreviation / acronym	Description
D2.5	Deliverable number 2.5 belonging to WP2
EC	European Commission
EOM	Earth Observation Model
LOD	Linked Open Data
NDVI	Normalized Difference Vegetation Index
OGC	Open Geospatial Consortium
OWL	Web Ontology Language
RDF	Resource Description Framework
SOSA	Sensor, Observation, Sample, and Actuator
W3C	World Wide Web Consortium
WP	Work Package

Executive Summary

This deliverable describes the first version of the semantic search module of the CANDELA platform which is the aim of task 2.3 of work package 2.

Semantic search covers a set of services to retrieve images through a semantic description of their content (i.e. places, type of vegetation or occurrence of a forest fire) and to search for data related to their content (i.e. cities and their population, weather measures, fire evolution over time). It relies on a formal representation of data that can be "located on" (or more generally "linked to") images thanks to their date and location. Hence a preliminary work to the design of semantic search facilities is to identify various relevant data to be used to search for images. The use cases defined in tasks 1.1 and 1.2 will provide semantic search scenario and contribute to identify relevant datasets that will enrich the image description. Once various data sources are identified, because each source has its own format and structure, the next stage is to propose a homogeneous representation of this heterogeneous data. This representation requires to define an appropriate data model, which may be a formal vocabulary or an ontology. Then data has to be associated to one or several semantic classes from this vocabulary, and stored in a repository. The semantic search facility can take advantage of this formal representation and of a reasoning engine to support the search for images according to the data that describes it and linked to it.

For this first version of the semantic search software tool, a modular vocabulary has been designed. It can be used as basis for semantically integrating heterogeneous data, including Earth Observation image metadata, data extracted from image processing, open data, and linked open data. This vocabulary aims at reducing the heterogeneities in data representation and granularity, at providing a homogeneous access to these data and at improving the image search task. This first version of the modular vocabulary does not take into account any specific requirement from the other Work Packages, in particular those involving the definitions of use cases (1.1 and 1.2). This will be addressed in the second version of the search module. In this deliverable, after describing several vocabularies that are reused to build the CANDELA data integration vocabulary, the integration process is detailed. It relies on the identification of spatial and temporal relations. The output of the integration process is a knowledge base that can be accessed and searched through an entry point interface. For this first version, this interface is limited to a SPARQL endpoint.

1 Introduction

The present deliverable describes the work done so far by IRIT CNRS in task 2.3 of work package 2. This task consists in developing the first version of the semantic search module (semantic search V1), the main result of task 2.3. This task required to design an ontology for integrating image metadata and various types of open data. The first version of the ontology is a modular vocabulary that provides high level semantic features (classes and properties) to describe the data to be searched.

1.1 Purpose of the document

The advent of the Copernicus program and its wealth of open data (Earth observation images and their metadata) open many economic perspectives thanks to emerging applications in various fields. These applications can benefit from both metadata images (such as cloud cover of an image), results of an automatic analysis of the content of images (for example, calculating the vegetation index, or other indexes of interest) and geo-localized and dated open data (government data, meteorological data, etc.) that can be associates to the images themselves.

One of these applications is semantic search which is the aim of task 2.3 of work package 2. By semantic search we mean services to retrieve images through a semantic description of their content (i.e. places, type of vegetation) and to search for data related to their content (i.e. cities and their population, weather measures). It relies on a formal representation of data that can be "located on" (or more generally "linked to") images thanks to their date and localisation. So, a preliminary work to the design of semantic search facilities is to identify relevant data to be used to search for images, and then to propose a homogeneous representation of this heterogeneous data. We have identified four types of data that can be linked to Earth Observation images: image metadata, data extracted from image processing, open structured data (e.g. data in formats such as CSV, XML, JSON, GeoJSON), and linked open data (in RDF). Our approach could be generalized to private knowledge bases and databases belonging to companies.

The contribution of semantic technologies to facilitate the integration of these different kinds of data has been demonstrated in previous work [1][2], in particular through the use of ontologies as formal representations of domain knowledge. Another advantage of data semantic representation is to allow reasoning on the knowledge base to infer new facts. For instance a `rdfs:subclassOf` relation between two classes A and B (e.g. City and Administrative Unit) leads to infer that any instance of A is also an instance of B. This type of approach relies on one or more ontologies (usually represented using OWL W3C standard language) to represent the data (usually in RDF format) and to link them to each other.

Two problems then arise: defining a suitable ontology and managing the large volume of data to be handled by the transformation and integration processes. In the continuity of the work on data access and integration via ontologies [3][4], we have designed a semantic vocabulary to represent the data of various sources with the aim of accessing them homogeneously. This vocabulary is a preliminary stage towards an ontology, i.e. a vocabulary with domain oriented inference rules. It is modular and it reuses several Linked Open Data (LOD) vocabularies recognized as standards to facilitate data reuse. Because data linked to images are geo-localized and may evolve over time, and because we consider metadata of satellite images as sensor data, we reuse GeoSPARQL [5] for the

spatial dimension of data, OWL-Time[1] for the temporal dimension of data, and SOSA[2] for sensor data. To reduce the cost of image indexing, the Earth's surface can be gridded in tiles. ESA provides a grid for Sentinel 2 Single-Title (S2ST) images, in which each tile is a square of 100x100 km. Hence, we use these tiles to link data having a geolocation to Sentinel-2 images.

The modular vocabulary that we have developed so far can be used as a basis for semantically integrating heterogeneous data. The core main modules of this modular vocabulary have been developed in the context of the FUI SparkinData project (http://melodi.irit.fr/sparkindata/). During the first six months of CANDELA we have performed the following subtasks:

- We reworked these modules
- We added the *ndvi* module that aims at representing the vegetation index associated to each tile.
- We also worked on homogenizing the way NDVI and land cover indexes can be represented. In particular we added properties to date them and to store their original sources (Global Land Cover, Corine Land Cover, for instance).
- We published the *ndvi* dataset and made it accessible through a SPARQL endpoint

1.2 Relation to other project work

This deliverable describes the work done so far in task 2.3 by IRIT CNRS within work package 2.

Within the CANDELA project, work package 2 aims at making accessible a large set of Earth Observation data captured with various types of sensors, at providing services to analyse and mine this data at various levels and for different purposes, and at providing semantic search capabilities. Work package 2 will illustrate how to define a pipeline to meet domain specific needs using Earth Observation images and data thanks to several use cases in two domains: agriculture and forestry. Tasks 1.1 and 1.2 are dedicated to the definition and implementation of use cases. They will output their requirements, the images, the open data as well as the services and software tools that will be useful to meet the requirements.

Task 2.3 (described in this deliverable) aims at designing semantic search facilities on the datasets selected by tasks 1.1 and 1.2 in keeping with these requirements. Some of these datasets are built using machine learning techniques in task 2.2.

During the first six months of the project, the information and datasets required by each use-case were not yet available. So, we decided to represent vegetation indexes (NDVI values and land Cover) as semantic data. The semantic representation of NDVI, Land Cover and image metadata enables to search for images of a given place, taken during a specific period, and where the NDVI value or the Land Cover reveals a given type of vegetation using the semantic search engine from a SPARQL endpoint. This dataset also allows temporal analyses of the NDVI index over time on the same area.

1.3 Structure of the document

This document is structured in 4 major chapters:

- **Chapter 2** defines what we mean by semantic search.
- **Chapter 3** describes the vocabularies we reuse and extend to describe image metadata and contextual data (weather information, grid, land cover and NDVI).

[1] https://www.w3.org/TR/owl-time/ (10/2017)

[2] https://www.w3.org/2015/spatial/wiki/SOSA_Ontology

- **Chapter 4** presents the transformation and integration processes that rely on the identification of topological (spatio-temporal) relations, and the dedicated architecture for selecting, integrating, and storing the data into a triple-store knowledge base.
- **Chapter 5** presents the entry point interface for searching on the resulting knowledge base. For this first version, this interface is limited to a SPARQL endpoint. We illustrate our search engine with a use case showing how the semantic representation of these data and their interrogation make it possible to study the evolution of vegetation indexes, and thus the evolution of the vegetation on the ground.

1.4 Glossary adopted in this document

Term	Definition
Concept or class	Representation of a group of entities sharing the same properties
DataSet	Set of data, either in a data-base, a file or a triple store. Can be semantic or not
Entity	individual object or a value
(SPARQL) Endpoint	Web service that provides an interface to write SPARQL queries to search a knowledge base stored in a triple store
Instance	Representation of an entity. Instance of a class: the entity belongs to the class
Knowledge Base	Formal representation of a domain knowledge. In the semantic web, consists in an ontology (the schema), instances and rules.
Knowledge Graph	One of the ways to represent a knowledge base as a graph where nodes are classes and edges are properties.
Linked Open Data	Data represented in semantic format –RDF- and made available on the web
Module	Small vocabulary or ontology that has a semantic consistency, used to represent a sub-domain
Ontology	Formal vocabulary enriched with domain specific rules and respecting good design principles
Open Data	Any database or file made freely available on the web (under licensing conditions)
Property	Labelled link between two resources (values, classes or entities)
Semantic Search	Search engine over semantic graphs
Triple Store	Repository where knowledge graphs can be stored + services for managing the repository
Vocabulary	Set of classes and properties required to represent a domain according specific needs or requirements

2 The semantic search task and data

2.1 Semantic search task

By semantic search we mean services to retrieve images through a semantic description of their content (i.e. places, type of vegetation), their location and date. A request specifies constraints on the data and their values. Results will be not only links towards images but also some data related to their content (i.e. cities and their population, weather measures).

Semantic search relies on a formal representation of data that can be “located on” (or more generally “linked to”) images thanks to their date and localisation. So, a preliminary work to the design of semantic search facilities is to identify relevant data to be used to search for images, and then to propose a homogeneous representation of this heterogeneous data.

2.2 Process to design the semantic search module

The first stage is to **select or build datasets** that are relevant for each use case. In the future, datasets will be selected in keeping with the requirements and scenarios of the use-cases proposed by tasks 1.1 and 1.2. We identified four types of data according to the way they can be obtained: Earth Observation image metadata, data extracted from image processing, open data, and linked open data (open data already available in semantic format). All these datasets are heterogeneous by their content, their structure and their format.

The homogeneity of the representation can be obtained by using the same format for all datasets (RDF in the standard format proposed by the W3C for knowledge graphs). The homogeneity of the representation 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. A vocabulary defines classes (also called concepts) of entities, and their relations (called properties). An ontology includes a vocabulary and domain specific inference rules.

So the second subtask to be carried out is to **design an appropriate ontology** that will provide the right concepts and properties to represent the data and their relations. Once this ontology is available, the third stage is to build the semantic representation of a dataset, that we will call a knowledge base or knowledge repository. This task requires to assign one of the classes of the vocabulary to each data, and to store the relations between the data thanks to properties. Two types of properties are of major importance in the case of Earth Observation data and images: the geometry or location of the data, and their date of capture.

The third stage is **to link this semantic data to images** whatever the data source. This relation is based on the notion of ‘interest’ to select the relevant data, on a spatial dimension to link the data according to their location, and the date or temporal feature to link data according to their date Given a set of data that satisfy a set of criteria of interest, thanks to the spatial and temporal links, one can retrieve the images that show the Earth at the time and location when and where the data has been collected or computed. It is important to note that temporal and spatial relations are not necessarily stored in the knowledge base. They can be evaluated at the time of querying the knowledge base.

The fourth stage will be to **design a search interface adapted to each use-case**. The interface will take into account the kind of data that is part of a query. It should provide a relevant way to express the user's interest, the searched area and the time period.

2.3 Semantic search V1 datasets

For this first version of the semantic search software tool, we demonstrate the feasibility of the approach on a simplified case, independently of the actual needs that will emerge from the use-case definitions in tasks 1.1 and 1.2. We have selected 4 datasets (the Sentinel-2 image metadata, Sentinel-2 grid, NDVI evaluated by image analysis and Land Cover collected from open datasets). Then we have designed a modular vocabulary to build a semantic representation of these data that should be relevant for the two use-cases. Thanks to this vocabulary, we have made available a semantic representation of the data in a triple-store. We also propose a SPARQL endpoint to query the triple-store.

2.3.1 Image metadata

We collect some of the metadata from the metadata records of the Sentinel-2 Single Tile images[3] from the PEPS platform made available by CNES (https://peps.cnes.fr/). The revisit time for Sentinel-2 is five days. Every night, we run a process that calls the RESTO API, a data service managed by CNES [6]. RESTO returns files of metadata records in a GeoJSON format. Among the metadata that we collect are the names of the image raster files, the cloud cover, the capture time, the corresponding tile. The RESTO query that retrieves these meta-data records makes it possible to specify some parameters to get a set of images, i.e. a maximum value for the cloud cover, an interval of time, an area of interest, etc.

The following URI calls the RESTO API to get all the metadata records from the S2ST collections that of images located in France, taken between 19/09/2017 23:00 and 25/09/2017 00:00:

```
https://peps.cnes.fr/resto/api/collections/S2ST/search.json?q=France\
&startDate=17-09-19T23:00:00\&completionDate=2017-09-25T00:00:00 .
```

2.3.2 Sentinel-2 grid

The S2ST grid was obtained from ESA as KML (link to the KML file). In total the dataset comprises 56984 tiles. However, in our research we focus only on France although the procedure can be extended to all the tiles if necessary. The tiles that cover France is covered by a total of 132.

2.3.3 NDVI

The NDVI vegetation index (Normalized Difference Vegetation Index) is calculated using S2ST images at L1C level. We chose the images with a cloud cover less than 3% so as not to distort the results. This index is obtained by a calculation on infrared (PIR) and red (R) sensors, which can detect the chlorophyll level. The result of this operation gives a matrix of values between -1 and 1 characterizing the NDVI of each pixel of the image. The values between -1 and 0 represent the elements composed of water, values between 0 and 0.25 represent the compound earth elements. We are interested here only in the values concerning vegetation, that is, values between 0.25 and 1. In order to

[3] https://sentinel.esa.int/web/sentinel/missions/ (07/2016)

associate an index with each pixel of an image, we have classified the indices into three categories: LowVegetationIndex (NDVI between 0.25 and 0.5), MidVegetationIndex (NDVI between 0.5 and 0.75), and HighVegetationIndex (NDVI between 0.75 and 1), corresponding each approximately to grass, culture, and forest. To get a representation of these categories as a percentage of the image the number of pixels in each category must be calculated in relation to the total number of pixels in the image.

2.3.4 The land cover

Global Land Cover SHARE (GLC-SHARE) dataset was created by FAO in 2014. It is provided in raster format as a TIFF file. The pixel resolution is 30 arc-seconds, approximately 1 sqkm. In the resulting raster dataset, 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. Using all the pixel of an image, a percentage of each GLC-SHARE Land Cover class can be associated to each image; percentages are calculated using a Django module from the geometry of the tile. The module creates a temporary file with the fragment of the GLC-SHARE that overlaps the geometry of the tile and creates a frequency table for the raster pixel values of the image linked to this tile. The date of this NDVI value is the one of the image. As a consequence, various NDVI values with different dates are assigned to each single tile.

3 Modular vocabulary for the semantic integration of Earth Observation data

Our vocabulary for integrating data with image metadata is based on the fact that each data is localized and dated. For each of them we know at least a point defined by its latitude/longitude, or a geolocated zone. Such a point or zone is called a "geometry" in Figure 1. Thanks to this geometry it is possible to link data to an image or to a part of an image when the geometries of the image and the data intersect. In Figure 1 for instance, meteorological measures (humidity, temperature, pressure) and administrative units (cities, regions, etc.) are geolocated and cover a well-defined surface. By comparing data and image geometries, it is possible to know which part of an image is concerned by these data. Images are dated too. As long as other data (e.g. meteorological measures) are also dated, data and images can be linked by temporal relations (e.g. to link an image to the available weather information captured one week after the image was created).

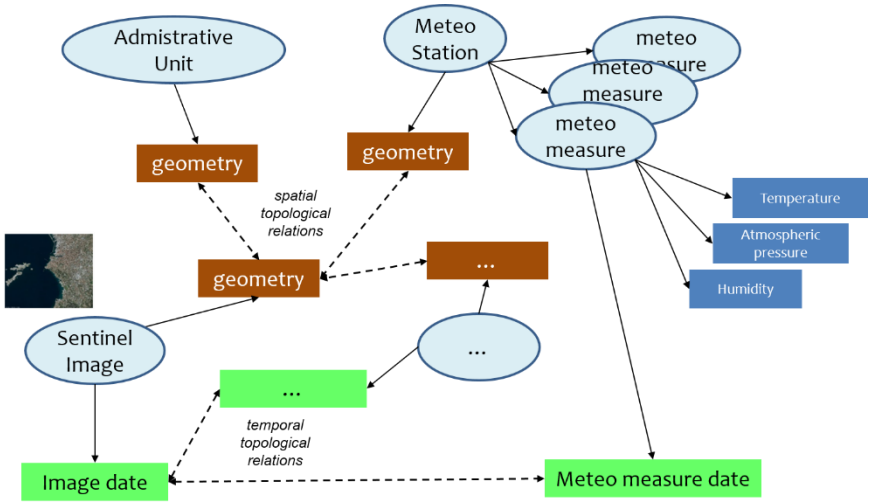

Figure 1: Integration of data using their spatial and temporal properties

To make the integration vocabulary compliant with existing standards, we have reused various existing vocabularies. This chapter first describes the reused vocabularies and then presents the vocabulary that we have designed to support this approach (shown in Figure 2). To allow the integration of diverse data types, this model is composed of a generic part (*time*, *sosa* and *geo* frames of Figure 2), with classes and properties reused from existing vocabularies (GeoSPARQL, OWL-Time, and SOSA), and a specific part dedicated to data to be integrated (such as the *mfo* and *eom* frames in Figure 2). The specific part of the model contains at least a vocabulary to describe image metadata (the *eom* frame for "Earth Observation Metadata") and as many vocabularies as data types to be integrated (only the *mfo* vocabulary which represents weather data, *mfo* standing for "Météo France Observation", is mentioned here).

3.1 Reused vocabularies

Several spatial and temporal extensions to RDF have been proposed and implemented. Our vocabulary relies on two of these vocabularies: GeoSPARQL and OWL-Time. In our previous work [7],

image metadata records and meteorological observations were represented with two other vocabularies, DCAT and SSN, respectively. Instead, we have now adopted SOSA as a core ontology that can be reused for different types of data, as detailed below. We also reuse other vocabularies that we do not detail here, such as PROV-O[4] to add provenance information (agent or activity) to entities (e.g. ESA Tiles).

GeoSPARQL, an OGC standard, defines an ontology for the representation of features, spatial relations and functions [5]. While alternative vocabularies exist, such as GeoRDF which allows for representing simple data like latitude, longitude, and altitude as properties of points (using WGS84 as reference datum) and GeoOWL, which allows for expressing spatial objects (lines, rectangles, polygons), we opted for GeoSPARQL because it offers good reasoning capabilities to compare geometries and then establish relations (such as *contains, touches, overlaps*, etc.) between them. The frame *geo* of Figure 2 presents the main classes of GeoSPARQL. The `geo:Feature` class represents any entity having a spatial component. This spatial component is described as a "geometry" (point, polygon, etc.), instance of the `geo:Geometry` class, and related to its feature via the property `geo:hasGeometry`.

OWL-Time[5] is a W3C standard ontology for representing the temporal component of data. The *time* frame of Figure 2 shows its main classes. The `time:TemporalEntity` class represents any entity having a temporal component, i.e. a start date (`time:hasBeginning` property) and an end date (`time:hasEnd` property), and thus a duration (`time:hasDuration` property). Temporal entities can be linked with binary relations (such as *meets, overlaps, during*) coming from the Allen's interval algebra. They are used for spatio-temporal reasoning. OWL-Time provides a generic property called `time:hasTime` which may be used to associate a temporal entity to anything (such as a calculated ndvi).

Several alternative exists to OWL-Time. One of these is worth mentioning, because it combines time and space representation: the stRDF (for Space Time RDF) model accompanied by the stSPARQL query language [8]. StRDF allows representing and querying geospatial data that change over time. Unfortunately, stRDF is not supported by open source triplestores.

SOSA (Sensor, Observation, Sample, and Actuator) is a light-weight but self-contained core ontology representing elementary classes and properties of the ontology SSN[6] (Semantic Sensor Network). SOSA describes sensors, their observations and their procedures. It has been largely adopted in a range of applications, and more recently, satellite imagery. In the SOSA vocabulary (frame *sosa* in Figure 2), an observation (`sosa:Observation`) is considered as a sensor activity providing an estimation of a property value using a given procedure. It allows to describe the related features of interest and the observed properties as well. SOSA reuses OWL-Time to date observations (`sosa:phenomenomTime`). We have hence adopted SOSA for describing image metadata and meteorological observations as respectively, *Earth observations* (*eom* module) and *meteorological observations* (*mfo* module). We choose to specialize SOSA in order to better type the instances of these concepts, although the trend in domains largely adopting SOSA, such as IoT, is to avoid this kind of construction and to directly use SOSA as main vocabulary. It allows for better dealing with the different kinds of data described with this core ontology (e.g., image metadata and meteorological observations, for instance).

[4] https://www.w3.org/TR/prov-o/ (10/2018)
[5] https://www.w3.org/TR/owl-time/ (10/2017)
[6] https://www.w3.org/TR/vocab-ssn/

3.2 Data integration model

We can link GeoSPARQL to any domain ontology by specializing the class `geo:Feature` with a class of the considered ontology (e.g. `eom:Footprint`).

For each type of observation measured by sensors a new vocabulary has to be defined, with a specific namespace (used to prefix the vocabulary terms), dedicated to this source of data. This vocabulary contains at least a class which specializes both `geo:Feature` and `sosa:FeatureOfInterest` classes. Hence each observation defined according to this principle is localized by its "feature of interest" which corresponds to the observed area on Earth and which is characterized by a geometry. The frames *eom* and *sosa* of Figure 2 each describe a vocabulary developed to represent the satellite image metadata (*eom*) and the weather station measures (*mfo*).

To reduce the cost of image indexing, we propose a vocabulary to represent grids made of tiles, such as the grid provided by ESA for Sentinel 2 Single-Title (S2ST) images. Tiles are also localized entities; we represent them as instances of the `grid:Tile` (sub-frame *grid* in Figure 2) that specializes the `geo:Feature` class. As mentioned earlier, they could be dated thanks to the `time:hasTime` property if needed (if a new grid was defined by ESA).

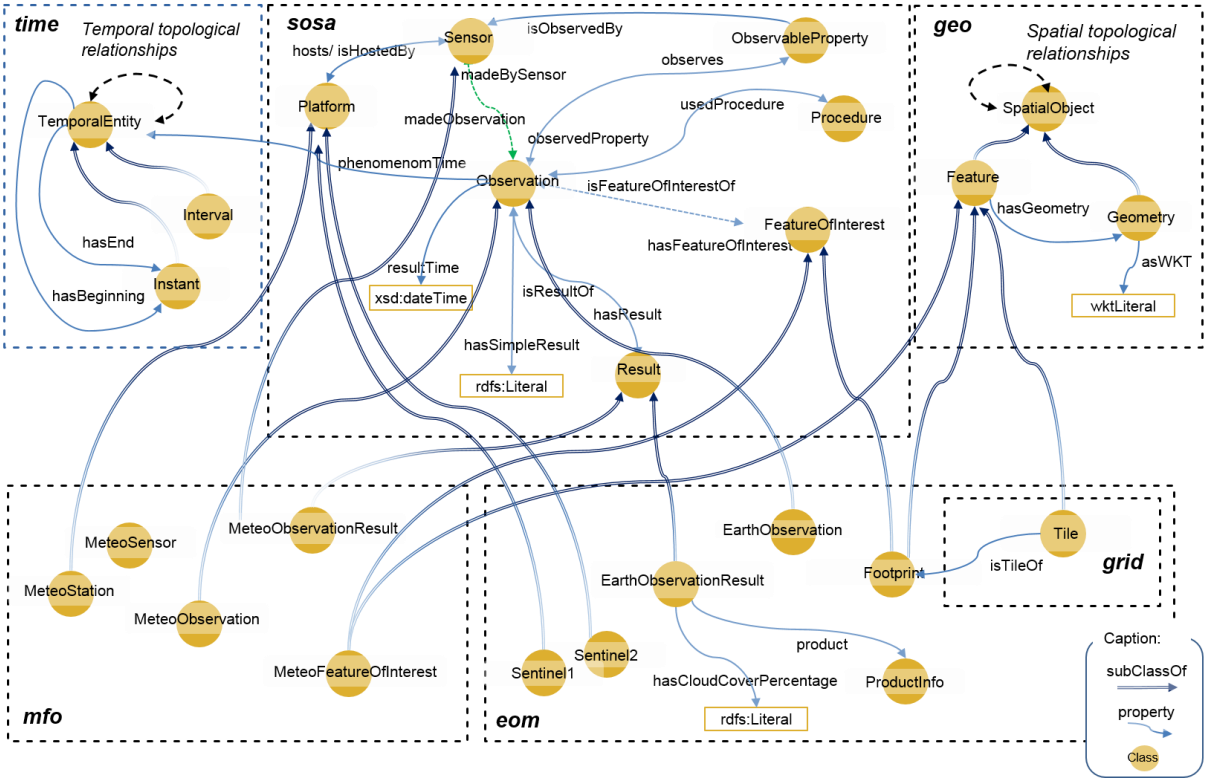

Figure 2: The integration model relying on the OWL-Time, GeoSPARQL and SOSA vocabularies, which are specialized in modules dedicated to different knowledge sources (image metadata and weather information)

Any index associated to an image can be represented in a similar way: a new vocabulary module needs to be added. We illustrate the process with two such properties: the land cover, which is available as an open data, and the vegetation index (NDVI), that we evaluated for each image using the NDVI value of each image pixel. These features are represented thanks to the *lci and ndvi* frames in Figure 3. As said earlier, we associate these data to tiles, that are represented using the `grid:Tile` concept from the *grid* module.

We have represented three categories in the *ndvi* module as subclasses of `ndvi:NDVI` (Figure 3). Vegetation indices calculated on images are instances of one of these classes, and are dated from the date of the image using the `time:hasTime` property. They are geo-localized using the geometry of the tile associated with the image.

For the land cover representation, we define the lci module and the `lci:LandCover` class which must be specialized for each considered Land Cover (Cropland, Forest, Baresoil, ArtificialSurface, …). The percentage of each GLC-SHARE Land Cover class is associated to a title via the `lci:hasLandCoverInfo` property using the `lci:LandCoverInfo` class.

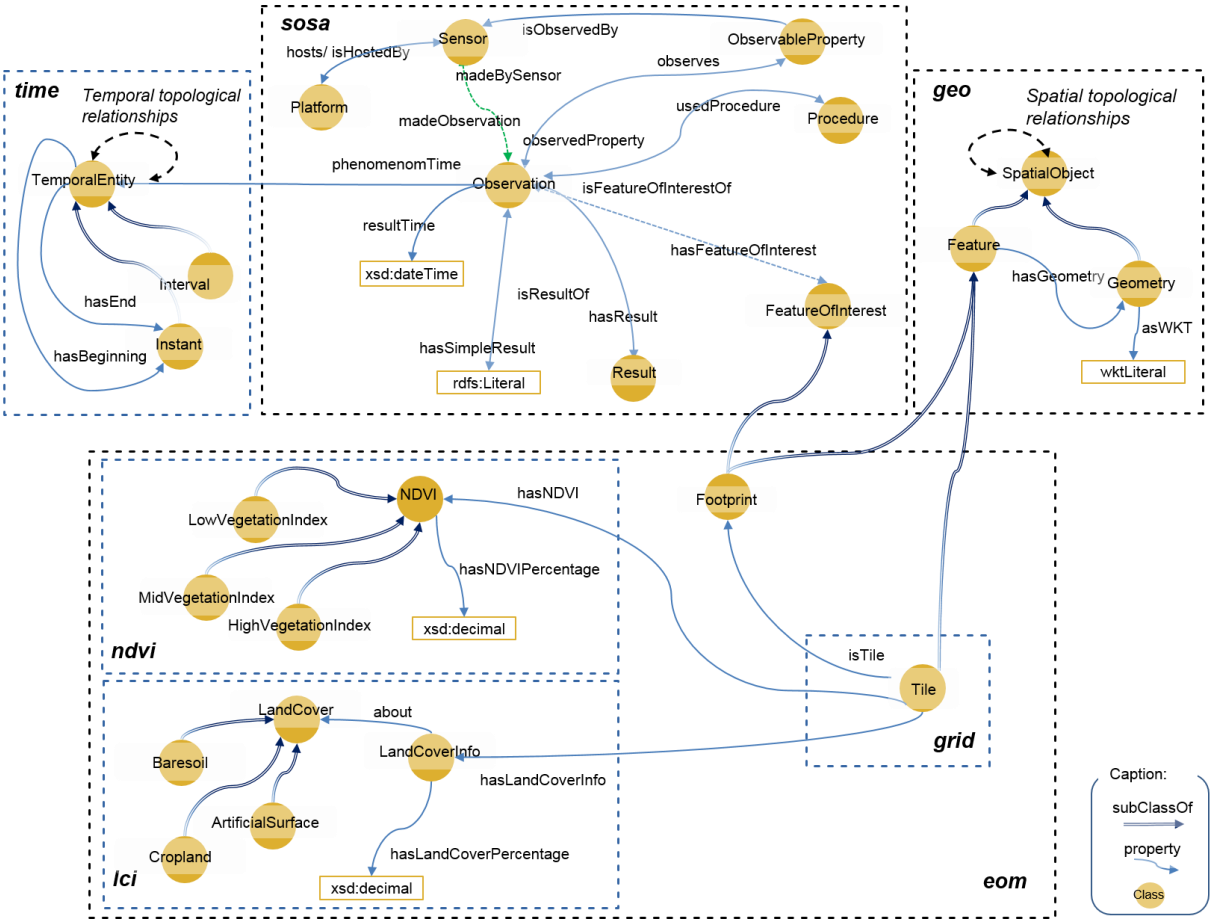

Figure 3: The integration model extended to represent tiles and their land cover and ndvi

4 Process of semantic data integration

4.1 Overview and architecture

The architecture of our system is modular (Figure 4). Its different levels allow decoupling stages in the process from raw data to semantic data. It consists of different modules:

- Data selection: the first step of the data integration process is to identify and access the data sources to be collected. A data set is either a file or the result of a query to retrieve data, from a data store. The formats of files currently considered are CSV, RDF, XML, TIFF, Shape files.
- Data conversion: the data gathered are then converted into a JSON pivot representation. To do so, we have reused dedicated scripts or developed customised ones, according to the specific kind of data source. The intermediate JSON files are stored in a MongoBD data base as a security back-up.
- Data alignment: from the data in JSON files, we generate instances of classes of the vocabulary presented in Chapter 2. We have defined a mapping template and a processing mechanism implemented as a Python module in which customised functions use the values in JSON documents as input data or parameters to define RDF triples. Thanks to these functions we can perform more sophisticated operations that are not possible in alternative approaches such as RML[7] (RDF Mapping Language).
- Data integration: the integration process relies on the spatial and temporal relationships between the instances of the model classes. At this point, all the instances in the knowledge base have a spatial representation. Then it is possible to pre-process the spatial relationships and store them as declarative statements in the triple store. It is also possible to evaluate the spatial relationships on the fly, however this generates a high computing cost that we consider unnecessary due to the nature of the data that we consider (they have a fixed positions). Then temporal relationships can be established on demand using SPARQL.

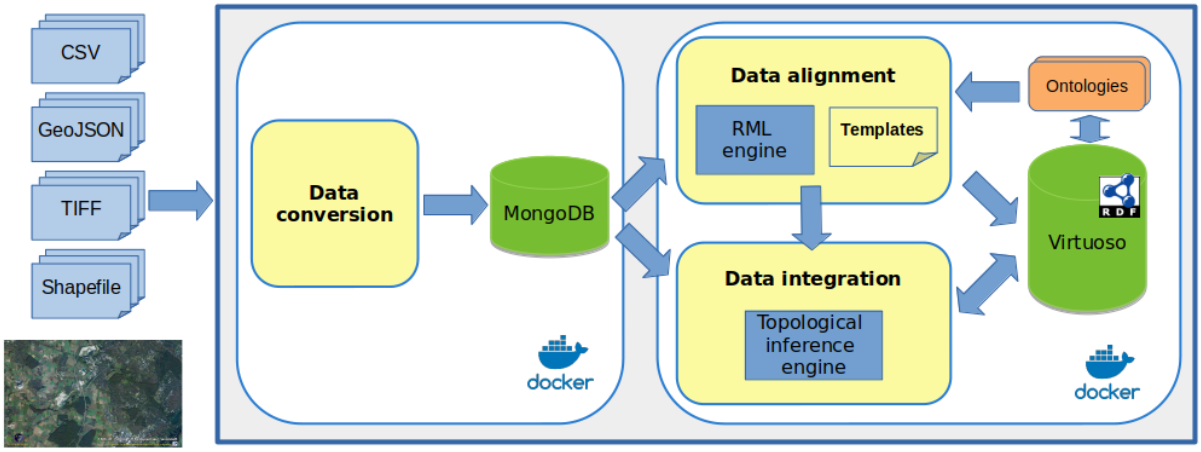

Figure 4: Architecture of the services

[7] http://rml.io/RMLsoftware.html (octobre 2018)

4.2 Data alignment

Each ontological module from Figure 1 to Figure 3 is described in an OWL file. When instantiating each sub-model we create separate RDF datasets. Processing each dataset relies on Python functions that exploit the template and values in JSON files to generate RDF triples.

- **S2ST metadata records**: Image metadata records are downloaded as GeoJSON files. These files are turned into RDF graph files using the *eom* vocabulary. Each new data is defined as an Earth Observation, I.E. an instance of an `eom:EarthObservation` (*eom* frame in Figure 2).
- **NVDI**: For each vegetation index calculated on images, depending on the index value, the process creates an instance of one of the three classes (Low, Mid or HighVegetationIndex). This instance is dated with the date of the image using the `time:hasTime` property. Its localization can be known thanks to the geometry of the tile associated with the image. So a direct property ndvi:hasNDVIPercentage is added between the tile and the ndvi instance.
- **Land cover**: the process is similar for Land Cover data. Land Cover values are linked to tiles. So for each tile, an instance of the `lci:LandCoverInfo` class is created and linked to the title via the `lci:hasLandCoverInfo` property. For each considered each GLC-SHARE Land Cover class (cropland, Forest, Baresoil, ArtificialSurface, …), an instance is created and linked to the instance of `lci:LandCoverInfo and` to the percentage value thanks to the `lci:hasLandCoverPercentage` property.
- **Grid**: as said earlier, each tile is represented as an instance of the `grid:Tile` class. It is linked to all the images that are taken on this area every 5 days via the `grid:isTile` property that links the tile with the `geo:footprint` of the images.

4.3 Data integration

It is then possible to calculate the spatial and temporal relationships to link the data of each of these datasets. By linking entities with a spatial dimension to the ESA tiles, one can connect the related knowledge to the images taken on this part of the Earth. This approach consists in making the Cartesian product of two sets. Hence when the datasets to be combined are large, calculating such relations while querying the dataset can be extremely time-consuming and unacceptable for real-time applications. That is why it's best to pre-calculate them. Some data, i.e. the position of weather stations, of cities and most of administrative places, and even land cover, are valid for a very long period, larger than the one of the application, and can be considered as stable or static. In contrast, some data streams are continuously providing new data at regular time spans. For instance, temperature measures are given every 3 hours by Meteo France weather reports, and tens of new EO images and their metadata are available on the PEPS server every day. We have thus distinguished the case of space-only data from those with a temporal component.

Integration of data with fixed spatial component. For the static data sources, it is reasonable to consider the materialization of the spatial relationships between the corresponding RDF datasets in the triple-store as long as these have a reasonable size. If the volume is too large it is necessary to consider some optimization techniques. Thus, by relying on the spatial indexing provided by the tiling of the S2ST images, we can link the contextual data to images of this type, by calculating the relationships between each of the datasets and the one containing the tiles. For instance, we would calculate the spatial relationships between administrative units and ESA tiles. Similarly, it is possible to link meteorological observations to images by calculating only the spatial topological relationships between weather stations and ESA tiles.

Integration of data with temporal dimension. It is possible to establish temporal relationships between an image metadata record (collected every five days on a given area) and any data collected regularly after a period of interest on the same area. Weather measurements are such data. First we identify and define the period of interest and then we compute on the fly temporal relations using SARQL queries. If no entity can serve as a time reference (i.e., the ESA tiles or weather stations have no date, no temporal feature), we use the time interval defined by a user when searching for images, as a temporal buffer that provides context to select image metadata records. Then we compute on the fly temporal relations using SARQL queries.

5 Prototype

The prototype that we developed in CANDELA as a first version of the search module aims at searching for images and their vegetation index variation over time. For representing such indexes, the *ndvi* module has been added to the integration vocabulary, as presented in Chapter 3. Here is an example of RDF graph using the vocabulary and representing the vegetation index (instance of `ndvi: LowVegetationIndex)` for the tile 31TCJ (g-grid:31TCJ) on April 21, 2018.

```
g-ndvi:ndvi_31TCJ_20180421T105031_low a ndvi:LowVegetationIndex .
g-grid:31TCJ a grid:Tile .
g-grid:31TCJ ndvi:hasNdvi g-ndvi:ndvi_31TCJ_20180421T105031_low .
g-ndvi:ndvi_31TCJ_20180421T105031_low ndvi:hasNdviPercentage "15.11"^^xsd:decimal .
g-ndvi:ndvi_31TCJ_20180421T105031_low time:hasTime ndvi:instant_1524300631 .
ndvi:instant_1524300631 a time:Instant .
ndvi:instant_1524300631 time:inXSDDateTime "2018-04-21T10:50:31.026000"^^xsd:dateTime .
ndvi:instant_1524300631 time:inXSDDateTimeStamp "1524300631"^^xsd:dateTimeStamp .
```

A graph of this type is generated for each index category of an image thanks to a Python script that takes the image's jpeg2000 as input, calculates the corresponding NDVI and populates the vocabulary with instances.

Each RDF graph generated for each data source (ndvi, grid, metadatas and Land Cover) is stored in a triple-store. For instance, the graph containing S2ST image NDVI data is located at http://melodi.irit.fr/lod/ndvi/ and the graph containing S2ST image metadata records is http://melodi.irit.fr/ontologies/eom.owl

The query language dedicated to RDF data is SPARQL. To access to data in these RDF triple stores we have implemented a SPARQL endpoint. A SPARQL endpoint is a REST service; a response to a SELECT query is in a RDF serialized format if the answer is requested as an RDF graph or in XML, JSON or CSV if not. The endpoint for accessing these data is available at http://melodi.irit.fr/sparql.

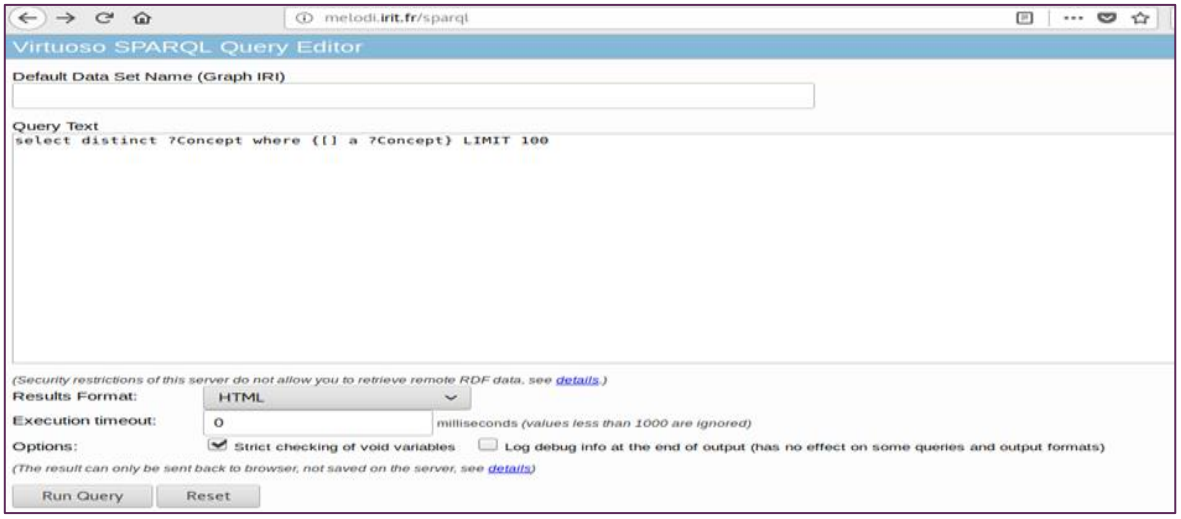

Figure 5: Endpoint for querying the knowledge base

Here below is an example of SPARQL query that asks for data to show the evolution of the NDVI of the 31TCJ tile over a year (filtering on the temporal component of NDVI). The query asks for a table showing, for all the images linked to tile 31TC (which becomes "for all the times when NDVI information has been estimated"), all the NDVI values (percentages) of the 3 types of percentage (Mid, low and High) at a given date (?ndviTimeInstant) during the selected period (between January 2017 and January 2018).

```
prefix ndvi: <http://melodi.irit.fr/ontologies/ndvi.owl#>
prefix time: <http://www.w3.org/2006/time#>
select distinct ?time ?ndviHighPercent ?ndviMidPercent ?ndviLowPercent
FROM <http://melodi.irit.fr/lod/ndvi/>
WHERE{
 ?ndviHigh a ndvi:HighVegetation .
 < http://melodi.irit.fr/lod/grid/tile_31TC> ndvi:hasNdvi ?ndviHigh .
 ?ndviHigh time:hasTime ?ndviTimeInstant .
 ?ndviTimeInstant time:inXSDDateTime ?time.
 ?ndviHigh ndvi:hasNdviPercentage ?ndviHighPercent .
 ?ndviMid a ndvi:MidVegetation .
 ?ndviMid ndvi:hasNdviPercentage ?ndviMidPercent .
 ?ndviMid time:hasTime ?ndviTimeInstant .
 ?ndviLow a ndvi:LowVegetation .
 ?ndviLow time:hasTime ?ndviTimeInstant.
 ?ndviLow ndvi:hasNdviPercentage ?ndviLowPercent.
 FILTER (?time > "2017-01-01T00:00:00.00"^^xsd:dateTime AND ?time < "2018-01-
01T00:00:00.00"^^xsd:dateTime)
```

The answer table to this query is shown in Figure 6.

time	ndviHighPercent	ndviMidPercent	ndviLowPercent
2017-05-26T10:50:31.026	18.22	47.09	16.91
2017-07-05T10:50:31.026	16.11	35.22	23.81
2017-07-15T10:50:31.026	16.42	34.78	24.81
2017-08-14T10:50:31.026	8.19	36.38	31.73
2017-10-08T10:50:09.027	2.66	37.25	23.83
2017-10-13T10:50:31.026	0.45	35	26.2
2017-10-28T10:51:29.027	0.38	31.12	26.38
2017-11-07T10:52:29.027	0.33	29.3	28.77
2017-11-22T10:53:41.026	0.16	14.7	38.44
2017-11-27T10:53:59.027	0.23	10.77	43.09

Figure 6: Results of a SPARQL query about all the NDVI percentages of images linked to tile 31TCJ of the collection

Thanks to these values, one can observe the evolution of the vegetation indexes, and somehow the vegetation itself, according to the seasons. In this example, during spring (between the end of February and the end of May) the value of the category HighVegetationIndex increases,

corresponding to the growth of foliage forests; conversely we note a decrease in this percentage during the winter period. The categories MidVegetationIndex and LowVegetationIndex are also impacted by the season change.

A similar query could provide the same information for all the tiles. It would start with the following command lines

SELECT distinct ?time **?tileId** ?ndviHighPercent ?ndviMidPercent ?ndviLowPercent

tileId a tile …

The query returns as a result a large table of which we give an extract in Figure 7.

time	tileId	ndviHighPercent	ndviMidPercent	ndviLowPercent
2017-05-26T10:50:31.026	http://melodi.irit.fr/lod/grid/tile_31TCJ	18.22	47.09	16.91
2017-07-05T10:50:31.026	http://melodi.irit.fr/lod/grid/tile_31TCJ	16.11	35.22	23.81
2017-07-15T10:50:31.026	http://melodi.irit.fr/lod/grid/tile_31TCJ	16.42	34.78	24.81
2017-08-14T10:50:31.026	http://melodi.irit.fr/lod/grid/tile_31TCJ	8.19	36.38	31.73
2017-10-08T10:50:09.027	http://melodi.irit.fr/lod/grid/tile_31TCJ	2.66	37.25	23.83
2017-10-13T10:50:31.026	http://melodi.irit.fr/lod/grid/tile_31TCJ	0.45	35	26.2
2017-10-28T10:51:29.027	http://melodi.irit.fr/lod/grid/tile_31TCJ	0.38	31.12	26.38
2017-11-07T10:52:29.027	http://melodi.irit.fr/lod/grid/tile_31TCJ	0.33	29.3	28.77
2017-11-22T10:53:41.026	http://melodi.irit.fr/lod/grid/tile_31TCJ	0.16	14.7	38.44
2017-11-27T10:53:59.027	http://melodi.irit.fr/lod/grid/tile_31TCJ	0.23	10.77	43.09

Figure 7: Part of the results of a SPARQL query about all the NDVI percentages of all the images of the collection

Most of the URI that identify the images and the tiles are dereferenceable: so a click on a URI in the table (a tileId for example in Figure 7) triggers a SPARQL query which returns a new table with all the triples where the resource is referenced. This principle allows to go from one tile or image to another within the data repository.

6 Conclusions

Task 2.3 of the CANDELA project aims at demonstrating that the integration of EO data from heterogeneous sources with satellite image metadata can be a means to promote the use of these images, and that it can be achieved thanks to semantic web technologies. Publishing some data sets and image metadata as LOD opens new opportunities to use satellite images in a variety of applications by providing an easier access to linked Earth observations. Moreover, for large and dynamic data sets, using SPARQL queries to jointly search various types of data (among which Linked Data) together with EO images enables to create RDF triples on the flow and avoids to convert huge data sets into RDF triples.

This deliverable reports the work carried out by IRIT in the scope of task 2.3 during the first 6 months of the project. The results presented in the document concern version V1 of the semantic search system. The deliverable presents the models, datasets and web service that we have developed. It details the methodology that we followed to develop a semantic repository in relation with Earth Observation images, and search facilities on these data. The main result is a framework to integrate and search data with spatial and temporal features. Several of our contributions improve this process: we designed a vocabulary to represent EO data and image metadata; we proposed an RDF conversion process using resource specific templates and a Python library that overcomes some of the RML limitations; we have also proposed an integration process that exploits the data geometry and date, OWL-time and GeoSPARQL ontologies to link spatial data, and finally SPARQL queries to get dynamic data linked to images according to spatial and temporal features.

We expect CANDELA partners that could be future users of the search function and that are involved in tasks 1.1 and 1.2 to evaluate and provide feed-back about this first version, available at: http://melodi.irit.fr/sparql.

As future work, we plan to consider domain-oriented sources as well as scenarios and requirements from the use-cases proposed by tasks 1.1 and 1.2. We will select relevant data sources for these use-cases, adapt the integration vocabulary to model data from these data sources, and take into account the kind of search expected in these use-cases. We also plan to develop for a friendly, more intuitive search interface, on top of SPARQL, to support searching in the semantic repositories. Once evaluated, the second version will be integrated as a service within the CANDELA platform.

References

[1] F. Reitsma and J. Albrecht. Modeling with the semantic web in the geosciences. IEEE Intelligent Systems, 20(2), pages 86–88, 2005.

[2] D. Sukhobok, H. Sanchez, J. Estrada, and D. Roman. Linked data for common agriculture policy: Enabling semantic querying over sentinel-2 and lidar data. In Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th Int. Semantic Web Conference, Vienna, Austria, Oct. 23-25, 2017.

[3] M. Console, Lenzerini M. Reducing global consistency to local consistency in ontology-based data access - extended abstract. In Informal Proceedings of the 27th International Workshop on Description Logics, p. 496–499. Vienna, Austria, 2014.

[4] M. Lefrançois, A. Zimmermann, N. Bakerally. A SPARQL extension for generating RDF from heterogeneous formats. In Proceedings of the 2017 Extended Semantic Web Conference, Part I, p. 35–50. Portoroz, Slovenia, 2017.

[5] D. Kolas, M. Perry, and J. Herring. Getting started with GeoSPARQL. Technical report, OGC, 2013.

[6] Gasperi J. Semantic Search Within Earth Observation Products Database Based on Automatic Tagging of Image Content [Conference] // Proc. of the Conf. on Big Data from Space. - ESA/ESRIN, Frascati, Italy: EU Publications, 2014. - pp. 4-6

[7] H. Arenas, N. Aussenac-Gilles, C. Comparot, and C. Trojahn. Semantic integration of geospatial data from earth observations. In Knowledge Engineering and Knowledge Management - EKAW 2016 Satellite Events, pages 97–100, 2016.

[8] M. Koubarakis and K. Kyzirakos. Modeling and querying metadata in the semantic sensor web: The model stRDF and the query language stSPARQL. In the Semantic Web: Research and Applications, pages 425–439. Springer Berlin Heidelberg, 2010.