A semantic representation of EO data for image retrieval based on natural language queries

Marco Polignano*[a], Marco de Gemmis[a], Vasilis Kopsacheilis[b], Michail Vaitis[b], Jenny Malig [c], Dominik Grether [c], Ilias Ioannou[d], Anastasia Sarelli [d], Vito De Pasquale[e], Sergio Samarelli[e], Pol Kolokoussis[f], Kleanthis Karamvasis[f], Milto Miltiadou[g], Christiana Papoutsa[g], Olivier Regniers[h], Virginie Lafon [h], Konstantinos Topouzelis[i] and Bogdan Despotov[j]

[a]Dept. of Computer Science, University of Bari Aldo Moro, via E. Orabona 4, 70125, Bari, Italy; [b]Dept of Geography University of Aegean, Lofos Panepistimioy, Mitilini, Geece; [c]TWT GmbH Science and Innovation, Ernsthaldenstraße 17, 70565 Stuttgart, Germany; [d]Planetek Hellas, Leoforos Kifisias 44, Marousi 151 25, Greece; [e]Planetek Italia, Via Massaua 12, I-70132 Bari, Italy; [f]Laboratory of Remote Sensing, National Technical University of Athens, Greece; [g]Civil Engineering and Geomatics, Cyprus University of Technology, Limassol, Cyprus; [h]i-Sea, Bordeaux Technowest 25 Rue Marcel Issartier, 33700 Mérignac, France; [i]Dept of Marine Sciences University of Aegean, Lofos Panepistimioy, Mitilini, Geece; [j]Cloud Sigma, Floor 3, office 308. Sofia 1124, Bulgaria

ABSTRACT

SEO-DWARF (Semantic Earth Observation Data Web Alert and Retrieval Framework) is a project funded by the European Union Horizon 2020 research and innovation programme. The main objective of the project is to realize the content-based search of Earth Observation (EO) images on an application specific basis. The satellite images, which come from EO satellites such as Sentinels 1, 2 and 3, as well as ENVISAT, are distributed with few correlated meta-data which do not describe the phenomena and the objects included in the image. Innovative approaches to process remote sensing images can extract relevant information which semantically describes the land type, the region area border, objects and events such as oil spill. This information can be modeled as structured information through ontologies to be processed by algorithms to perform information retrieval and filtering. The proposed system is aware of the semantic elements which are relevant for final user and will be able to answer natural language queries such as "Show me the images of the Mediterranean Sea which include an algal bloom". The possibility to retrieve a specific set of land images starting from a query expressed by a final user can quickly increase the interoperability and the diffusion of applications able to efficiently use EO data. In this work, we present a brief overview of the most successful application of this formalization strategy focusing on the tools and approaches for creating a robust and efficient domain geo-ontology. Furthermore, we describe the approach adopted to define the specific ontology used in the SEO-DWARF project, including the strategy adopted for implementing and populating it.

Keywords: Ontology, Semantic Web, Question Answering, Earth Observation data

1. INTRODUCTION

Coastal zones are the matter of several studies that have the purpose of individuating tools and methods for their monitoring and management. In the research area, they have been proposed a series of techniques to control and forecast the development of natural phenomena, such as the processing of satellite images, the use of mathematical simulation models, sea truth data and satellite tags. The integration of these techniques within GIS (Geographic information system) allows producing sea temperature maps, sea color maps, maps of spatial distribution of the sea species and thematic maps of the seasonal cycle of microorganisms (based on temperature, sunlight, currents and presence of polluting species). In particular, the Copernicus Marine Service improves ship routing services, offshore operations or search and rescue operations, thus contributing to marine safety. The provision of data on currents, winds and sea ice contribute to the protection and the

sustainable management of living marine resources in particular for aquaculture, fishery research or regional fishery organizations. The growing, interest in smart approaches for retrieving that information has motivated the developing of a strategy for approaching the retrieval process of satellite images with different spatiotemporal and spectral characteristics semantically. The critical challenge is to connect the quantitative information of the EO images (Earth Observation Satellite images) with the qualitative (high-level user queries) and be able to mine these connections in big archives. An essential question arises how to retrieve EO images based on user semantically aware questions. Content-based EO image retrieval techniques have been introduced for bridging the gap between low-level image features and high-level queries. The primary constraint of the existing approaches is the generalization of the problem. The formulated ontologies are not focused on the constraints of EO images. The main objective of SEODWARF is to realize the content-based search of EO images on an application specific basis. The marine application domain and data from Sentinels 1,2,3, ENVISAT will be used. Queries such as "Calculate the rate of increasing chlorophyll in the NATURA area" will be answered by the SEO-DWARF, helping users to retrieve the appropriate EO images for their specific needs or alert them when a specific phenomenon occurs. In this work, we focus on the strategies and tool used for the definition of a geo-ontology for the marine domain able to efficiently describe EO images to allow a semantic retrieval of information.

2. RELATED WORK

2.1 Basic concepts of Information Retrieval

In the last 20 years, the Information Retrieval (IR) task become very common for the new digital users. The massive diffusion of search engines such as Google, Bing or Yahoo! made the research of relevant documents (web pages) in the web straightforward and available for everyone. The user can perform queries in a language understood by the retrieval system that can "retrieve information which might be useful or relevant to the user" [1]. Information Retrieval Systems often deal with natural language text which is not structured (such as a regular expression used in a data retrieval task), consequently the retrieved object might be inaccurate or unexpected. The main problem to face is not how to extract relevant information but knowing what is relevant to the user. The relevance is the main aspect to consider when a retrieval system is designed because an accurate IR system is also a successful one. Nowadays the IR is a mature research area and includes modeling, object classification and categorization, filtering, data visualization, natural language processing and more. IR Systems are constantly gaining popularity. The user interacts with the system only through the "User Interface" module, which is strictly connected to other modules that work together in order to identify the user needs and filter the domain elements. The item provided as sources of the retrieval system, have to be stored inside a knowledge base. A frequent problem with this operation is the significant quantity of elements to manage. Different approaches can be adopted, but the most used is the creation of indexes that will be used to summarize the items and then retrieve it accurately. The indexes on semi-static items can be updated with regular intervals, for example daily, and do not require mandatory real-time updates. After the construction of the indexes, the more straightforward strategy for a retrieval task is the search of the query words in the index matrix and the retrieval of all the items that are listed as occurrences. When the correct items are retrieved, it is important to rank them with a utility function. Also in this area of application, we can identify many different strategies based on various aspects, such as the user's past choices, the collected user preferences, the serendipity, the cost if present and more. One easy and often used strategy is based on a similarity function. This category of functions, which include as an example the Cosine Similarity, the Euclidean Similarity, the Manhattan Similarity and more, can define numerically how much the user query is similar to each item in the collection. The final ranking will follow these numerical values ordering the result from the most similar to the less similar. Following the presented schema, the user can make a query in natural language and obtain real-time documents relevant to her needs. The most challenging task to perform, when a retrieval system has to work with multimedia (included images), is their conversion from visual data into semantically textual content. This process is called semantic annotation, and it can use ontologies as a standard for formalizing the semantics of the annotations.

2.2 Semantic annotation of images

The formalization of information for a semantic retrieval approach requires a module able to map semantic concepts and image metadata. Some of the most common in literature automatic strategies of images annotation are described by the survey proposed by Rajam [2]. The first approach presented uses techniques of Neural Networks, Decision Trees and Support Vector Machines (SVMs). They are based on an artificial intelligence strategy called "supervised learning" that can automatically identify relevant features starting from a set (training data) of annotated low-level features that describe the image. These data are manually annotated to be sure that the algorithm can learn the correct abstract definition of each.

The supervised learning strategy needs a significant amount of training data to obtain a reasonable level of precision for the annotation prediction of new data. When this happens, the algorithm can generalize concepts and apply them on never seen data.
All the strategies of images annotation need a shared and self-reliant vocabulary for describing elements of the domain with the correct concept. A popular approach is based on the use of Ontologies as structured vocabularies of terms. This solution can represent the specific domain following the shared common sense of each element. An ontology can be defined as a structured model that is able to describe semantically the concept that we want to consider. The knowledge used in them must be shared and largely adopted in order to make it reusable by others. As described by Neches in 1991 [3], "The cost of this duplication of effort has been high and will become prohibitive as we attempt to build larger and larger systems. To overcome this barrier and advance the state of the art, we must find ways of preserving existing knowledge bases and of sharing, reusing, and building on them." Very often heterogeneous representations of the same object are present, and every time it is required in a new project, a new representation had to be constructed. Shared ontology can provide a basis for packing knowledge in a unique representation and shared terminology. In the same work, Neches [3] proposes ontologies as the "Architectures of the future". Now after more than twenty years, this architecture became an essential instrument for knowledge representation and sharing. Nowadays, ontologies are used in many applications and very often, they are presented as an important component of the Semantic Web. The knowledge sharing on the web can support processes of retrieval and exploration more accurately, because it can provide to the user information about the real semantics of the parts of a website, removing all the problems of ambiguity. It is common to face problems of ambiguity because the analysis of the webpage is conducted considering only the single terms that compose the document. Ontologies, then, are useful to the retrieval task because on them it possible to use a "logic language" that allows reasoning and interrogation tasks.

2.3 An overview about ontologies

In literature, different definitions of "ontology" were proposed. One of the first was provided by Neches [3]: "An ontology defines the basic terms and relations comprising the vocabulary of a topic area". This definition describes what is needed to define an ontology. In particular, the author proposes "terms of a vocabulary" that are the concepts that describe the represented world and "relations" among them that describe the correlations. Each object of the ontology is called "concept," and it represents an abstraction of a real word element. A "class" is a high-level concept that can be specialized into one or more specific concepts. This specialization relation is called "is-a relation" and allows to define the hierarchical structure of the ontology. The properties of a class are acquired by all the child concepts that can add more properties to the specific concept. Relations that describe the componential task are known as "a-part-of relations". The ontological model will be used to define "instances" of the world, that are objects with real values for each property described for the concept of membership. Finally, we can define "axioms" that are rules of the domain, which help the applications to work with formal reasoning and induction.
The ontologies can be classified [4] by the level of abstraction used

- **Top-level ontology**: it is an ontology that describes general and abstract concepts not correlated to a specific problem or domain.
- **Domain ontology**: it is an ontology that describes the vocabulary of a specific domain; if possible all the concept will specify a more general one defined in a top-level ontology. Domain ontologies can be subdivided into:
    - **Application Ontology**: it is an ontology that describes the applications in all their parts.
    - **Task Ontology**: it is an ontology that describes a specific task of the domain. It has many relations "a-part-of" with the application ontology concepts.

The top-level ontologies are too abstract to be used directly in an application, but at the same time, they are simple to extend and specify in many domains. On the contrary, a specific domain ontology is very difficult to reuse because it is strictly connected to a particular application or specific task. Different top-level ontologies are available for specializations in more specific ontologies. As an example, Cyc is a significant knowledge base of common sense concepts developed by la Microelectronics and Computer Technology Corporation (MCC). This ontology is made of almost 5 million assertions and more than 500,000 concepts [5]. A famous methodology for the design of new ontologies is the Methontology, proposed in 1997 by FIPA (Foundation for Intelligent Physical Agents) [6]. This framework proposes developing and maintenance strategies for the definition of ontologies. The ontology development strategy starts by defining the needed activities for planning the main tasks and how much time, and resources must be allocated. After that, it is essential to identify the 'end users' of the domain ontology and model the requirements of each type of user. If the ontology engineer

is not a domain expert, she can first construct a toy ontology to start familiarizing with the specific concepts. Once the ontology engineer acquires enough knowledge, she can begin the conceptualization of the specific domain that includes problems and solutions faced in the user scenarios. When the ontology is correctly defined, it is possible to formalize it. For this task, ontological languages are used, such as frame-based or first-order representation languages. The formalized model must be expanded and integrated with existing sources to support the reuse of common knowledge. In this activity, the knowledge is represented in a conceptual model that describes the problem and its solution using the domain vocabulary. A similar strategy of acting can be used for reusing already existing ontologies to specialize them in domain-related ontologies. The main issues concerning the management of duplicates and the resolutions of conflictual information shared among concepts. These limits make challenging to work with large ontologies. Many tools for supporting the resolution of these issues are available in the literature. Some of them are implemented in a development environment for ontologies such as Protégé [7] or Chimaera [8]. Others are provided in a stand-alone mode such as FCA-Merge [9], or Prompt [10]. For our purposes, we are going to work with Protégé.

3. DESIGN OF THE SEO-DWARF ONTOLOGY

The retrieval module of EO images is the core part of the SEO-DWARF project. It is the gateway between the user needs and the services offered by the system. The purpose of this module is to collect the user needs to be expressed through a query, extract the relevant information and provide them in a format useful for the user. The input of the module are questions expressed in natural language and sometimes formalized in keywords obtained by the interaction of the user with a guided form (combo boxes, data pick-up, sliders, etc.). Pivotal words of the user requests are used to identify the concepts in the Ontology relevant for answering to the query. The pivots are used to navigate the ontological graph and to determine the specific classes of objects to extract from the knowledge base. When the classes are identified, the pivots are translated into functions able to retrieve all the elements that match the user requests. This step requires a complete formalization of all the elements observable in EO images in a computational semantic model. The model has to be easy to expand, to define and to navigate. These properties are all available in a classical ontological representation. Moreover, for supporting the reuse of shared knowledge, the design of the seo-dwarf ontology has been faced starting from already available top-level ontologies which already defines marine domain phenomena, measure metrics, and geographical concepts.

3.1 Analysis of geographical top-level ontologies

Ontologies for EO images are rare because, in this domain, many research projects that produce a semantic representation of geospatial phenomena do not share the generated schema through open channels.
One of the most famous projects that involve Earth and Environmental aspects is SWEET (The Semantic Web for Earth and Environmental Terminology) [11]. It is promoted by NASA to improve the use of Earth science data in semantic applications. For this project, 200 separated ontologies were created, and more than 6000 concepts subdivided into nine categories that cover aspects of space, time, Earth realms, physical quantities, etc. and integrative science knowledge concepts such as phenomena, events, etc.). They can be used, extended and adapted to the specific domain. The starting point of this ontology development was the collection of keywords in the NASA Global Change Master Directory that contains about 1000 controlled terms structured as a taxonomy. Moreover, other 20.000 terms, often synonymous with the previous, where extracted by free-text. The level of granularity used is high, and this group of ontologies can be seen as a group of top-level ontologies. For example, the term “air temperature” was not defined at a specific concept but only as a composition of “air” and “temperature” term. SWEET Ontologies are written in OWL 2 [12] and can be easily edited in Protégé after the download from the official project site [11].
European Environment Agency (EEA) is the driving force of a consortium of organizations that provide CORINE Land Cover methodology, technology, and data [13]. Land cover and land use in Europe is derived from satellite imagery, then classified, and provided for download (as shapefiles) to the public. The classification is used to characterize areas, e.g., as Green urban areas, code 141. The classification scheme dates back to 1994, was refined in 2000 and is applied to all data in a homogeneous way using a sound methodology.
On top of the EEA maintained classification an ontology is modeled [13]. This ontology is developed to cover the CORINE nomenclature. The ontology is defined in three levels and describes concepts of natural and artificial elements that can be visualized in a geographical image. Analyzing the ontology macroscopically, we can identify these five classes: Artificial areas, Agricultural areas, Forest and semi-natural areas, wetlands, water bodies. Marine waters are also described such as Oceanic and continental shelf waters, bays and narrow channels including sea lochs or loughs, fiords or fjords, rye straits

and estuaries. The ontology is written in OWL 2 [12], and it is available online for free download, use, and extension.
Koubarakis [14] proposed another ontology for Earth Observation Images called DLR Ontology, and it was developed to annotate TerraSAR-X images for the European project TELEIOS [15]. This ontology is different from the previous because it was used to describe EO images and also it presents concepts about the image acquisition metadata. In particular, the following macro sections are described:

- **Image metadata**: this section includes predicates that describe image properties. A small number of metadata are included such as time and area of acquisition, sensor, image mode, incidence angle.
- **Elements of annotation**: this section includes classes about patches, images, vectors used to describe an EO Image after the knowledge discovery step.
- **Concepts about land cover**: this section includes object that are visible in a EO image such as agriculture areas, bare grounds, forests, transport areas, urban areas, water bodies.

The ontology is not very specific but covers only some macro-concepts that can be further specialized and extended for specific domain applications.

Considering the proposed overview, it is possible to observe that SWEET [11] ontology covers many concepts but not in a particular application domain. Consequently, if it is adopted for the seo-dwarf ontology, it needed to be specialized and adapted to the specific application. Nevertheless, the specialization cost is mitigated by the low complexity for future extensions supported by the frequent available updates. CORINE [13] contains many useful concepts for EO topic, but it is less detailed than SWEET. However, the absence of many concepts can make laborious the extension process for future new applications. DLR [14] has specific concepts for the application domain. It covers water and land concepts, and the three levels structure make possible the extension or the specification of concepts. The strong limitation is the difficulty in accessing to that ontology because it is private and no future updates are confirmed.

As consequences of these considerations, it has been decided to adopt Nasa SWEET Ontology as top-level domain ontology and to extend it when needed with new concepts. SWEET is continuously updated and extended by NASA, collecting in this way a large community of supporters. The considerable quantity of information available allows covering a domain that is not only restricted to the Marine Areas but also to the Landscapes, with natural or artificial areas. The dimension makes not easy the handling of it because the density of interconnection does not allow selecting only a branch of the entire Ontology to formalize only the closed domain of application. On the contrary, it makes easy the extension of the project for all the different kinds of areas because more of the information needed is fully covered.

3.2 Design of the ontology

The specialization of concepts has been approached through a top-down decomposition strategy with a conceptualization approach which follows the one proposed in the Methontology. We start from a general concept, and consequently, we specialize them where needed. For the scenario of application considered, we started from the general concept of image and with the main properties which can describe it as meta-data considering the INSPIRE metadata directive [16]:

- ID of image.
- **Temporal timestamp of acquisition.**
- Satellite of provenience.
- Textual description of the image (Abstract).
- Spatial Resolution.
- Coordinates of the Area of Interest (Bounding Box: North Bound Latitude, East Bound Longitude, South Bound Latitude, West Bound Longitude).
- **Names and coordinates** (Polygon Coordinates) of all the **relevant geographical places** in the area like: Seas, Islands, Coast/Beach, Rivers, Lakes, Mountains, Hills, Level grounds, Forests, Cities, Regions, and Nations. They are described by: a proper name, a unique id obtained from GeoNames, coordinates of the polygon that describe the area, the type of the area obtained from the elements described in GEMET Thesaurus (Table 1.).
- **Phenomena present in the relevant area** and their numerical properties' values and coordinates.

Concepts are linked with the SWEET Ontology when possible and added to it with their properties when they are not defined in the original version. The conceptualization step computed is about the identification of concepts relevant for the domain to map them with the SWEET ontology in order to produce the "SEO-DWARF Environment Ontology". The Environment Ontology will describe all the concept of interest with a relation of father-child, linking them, where possible, with external sources. The ontology the namespace "**seo**" is used for referring to **seodwarf.eu/ontology/v1.0/**

The external prefixes and references of the sources involved in the process are shown in Table 1.

Table 1. Namespaces used for the ontology formalization

Namespace	Name of Reference	URL
rdf	Framework	http://www.w3.org/1999/02/22-rdf-syntax-ns
rdfs	RDF Schema	http://www.w3.org/2000/01/rdf-schema
xsd	XML Schema	https://www.w3.org/2001/XMLSchema
dbo	DBPedia Ontology	http://dbpedia.org/resource/
geo	GEMET Thesaurus	http://www.eionet.europa.eu/gemet
dct	Dublin core DCMI	http://dublincore.org/documents/2012/06/14/dcmi-terms
swe	Nasa Sweet Ontology	https://sweet.jpl.nasa.gov/sweet2.3
seo	SEO-DWARF Ontology	http://seodwarf.eu/ontology/v1.0
strdf	stRDF	http://strdf.di.uoa.gr/ontology
gn	Geonames	http://www.geonames.org/ontology

During this design task, we have identified and formalized the following concepts:

- **image**: the concept refers to a satellite image, used and archived in the project. It is the source used for responding to the user queries. Searching in the SWEET Ontology, we can identify the concept of "image" under the "Representation" class. The concept of "image" is extended generating the class of "Satellite Image" and the class "seo:image".
- **time**: the concept identifies a specific time instant, it is measured in seconds as default. It is defined in the SWEET Ontology (swe:Time), but it is expanded with a property that explicitly assigns a season name to the timestamp.
- **geographical_coordinate**: the concept refers to the coordinate concept used to describe a point or a polygon in a geographical map. It has a reference to the DBpedia "Geographic_coordinate_system" concept for creating a linking point with the Linked Open Data Cloud. The specific concept will be added as child of the "swe:Coordinate System" concept.
- **image_content_region**: the concept refers to a specific region of the image that it is relevant to describe it. The region has a name, a set of coordinates, a type, a geoname_id if it is universally identified and a list of institutional areas of reference (nations, regions, province, cities). It refers to the DBpedia concept "Location_(geography)" for creating a linking point with the Linked Open Data Cloud. It is collocated in the SWEET Ontology as child of the "swe:Region" concept.
- **region_type**: the concept refers to a specific description able to geographically characterize the region of interest. The concept is found in the SWEET Ontology as "swe:type", but it is extended as following for creating a specific class to represent GEMET Thesaurus types.
- **phenomenon**: the general concept "phenomenon" is specialized into the "seo:ocean_phenomenon" and where possible, concepts and relations from other ontologies are reused, while new concepts and relations are added if required. It refers to the DBpedia concept "Phenomenon" for creating a linking point with the Linked Open Data Cloud. It is collocated in the SWEET Ontology as child of "swe:OceanPhenomena".

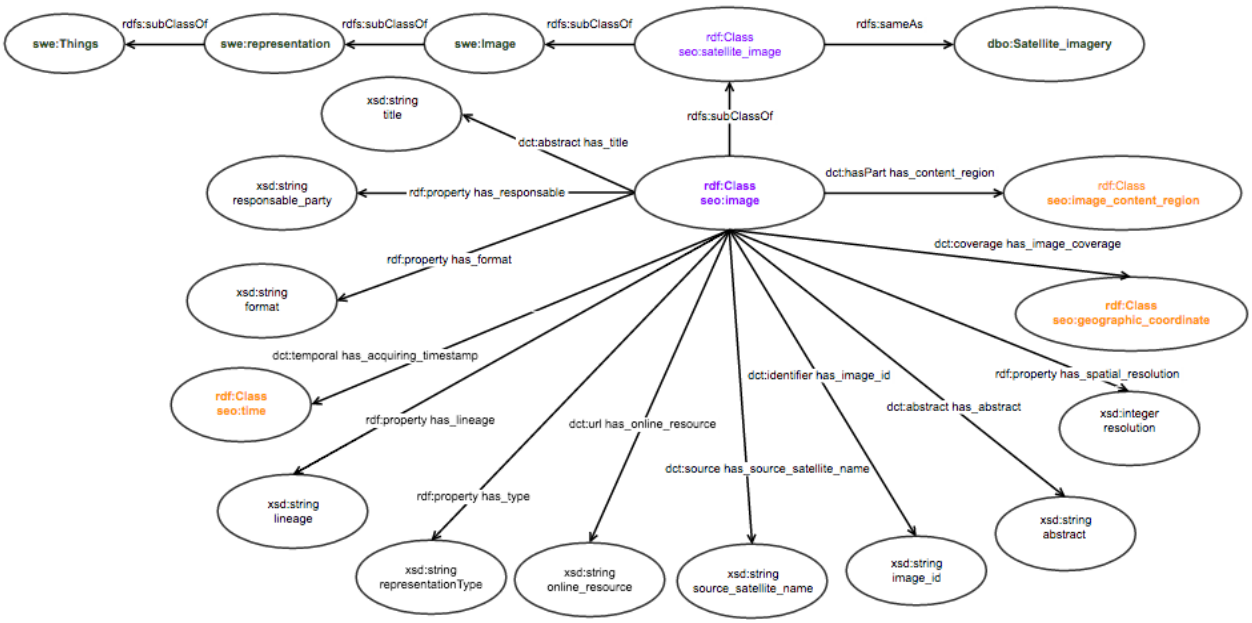

Figure 1. The big picture of seo-dwarf ontology. The "rdf:class" in orange have been better detailed in each specific portions of the ontology.

Each phenomenon of the marine domain, considered in the project, is then defined as a "rdfs:subClassOf" the "seo:phenomenon" concept acquiring all its shared properties. We considered as valid a conceptualization of the following marine phenomena:

- **hot_spot***:* it is an area with higher water temperature than the temperatures recorded usually. Sweet has a class called HotSpot, but this is describing areas of volcanism. Thus, the concept seo:hot_spot is added.
- **algal_bloom**: it refers to the rapid increase or accumulation in the population of algae in a water system. The concept is placed as subclass of "seo:ocean_phenomenon" but it is linked with the SWEET Ontology concept "algal bloom" and with the DBpedia concept "Algal_Bloom". "Harmful Algal Bloom" is a specific type of Algal Bloom correlated with a toxic species of algae.
- **upwelling**: it refers to areas where water with Sea Surface Temperature (SST) lower than surrounding one is recorded. The concept is placed as subclass of "seo:ocean_phenomenon" but it is linked with the SWEET Ontology concept "Upwelling" and with the DBpedia concept "Upwelling".
- **oil_spill:** it refers to areas where liquid petroleum is released into the environment, especially marine areas. The concept is placed as subclass of "seo:ocean_phenomenon" but it is linked with the SWEET Ontology concept "oil spill" and with the DBpedia concept "Oil_spill".
- **turbidity**: it refers to the cloudiness or haziness of a fluid caused by large numbers of individual particles. The concept is placed as subclass of "seo:ocean_phenomenon" but it is linked with the SWEET Ontology concept "Turbidity" and with the DBpedia concept "Turbidity".
- **wind***:* the concept refers to a sea surface wind that is a crucial parameter for operational use and for studies in coastal regions. The concept is placed as subclass of "seo:ocean_phenomenon" and it is linked with the SWEET Ontology concept "Ocean Wind" and with the DBpedia concept "Wind".
- **front**: it refers to a boundary separating two water masses moving in two different directions. The concept is placed as subclass of "seo:ocean_phenomenon" and it is linked the DBpedia concept "Front_(oceanography)".
- **trophic_status_index**: it refers to the increase of nutrient concentrations followed by corollary increases in subsequent trophic levels, such as eutrophication. The concept is placed as subclass of "seo:ocean_phenomenon" and it is linked the DBpedia concept "Trophic_state_index".
- **wave**: it refers to the waves present on the sea surfaces. The concept is placed as subclass of "seo:ocean_phenomenon" and it is linked the DBpedia concept "Wind_wave".
- **shallow_water**: it refers to phenomenon about the disruption of the natural orbital motion of the wind. The

concept is placed as subclass of “seo:ocean_phenomenon” and it is linked the DBpedia concept “Waves_and_shallow_water”.

The concepts defined in this task have been enriched with properties able to describes numerically or with semantical valid labels its detectable features. As an example, an EO image of the Mediterranean sea with an algal bloom inside will be described using the concept “image” with an “image_content_region” described using “geographical_coordinate” and an “algal_bloom” “seo:phenomenon”. The “algal_bloom” concept is then described by properties such as “has_algae” which describes the name of the algae, “has_chlorophyll_concentration” which describes the value of chlorophyll concentration and “has_temerature” which describes the water temperature. An instance of the ontology is then a satellite image, described using the concepts defined below (Figure 1.) and with concepts properties instantiated with values which are referred to the image considered.

4. IMPLEMENTATION STRATEGY

The task of conceptualization, performed in the previous chapter, produces the general design of the ontology which needs to be translated using a descriptive language in order to be usable as schema for storing domain products (EO images) into a computable representation (knowledge-base). RDF [17] is a very common language adopted for ontologies development. It is also promoted by W3C (World Wide Web Consortium) as an instrument for the Semantic Web. This framework is based on three simple rules:

- All the described resources must have an identification link called ‘IRI’ (International Resource Identifier a subclass of URI)
- To describe a resource is necessary to use the less expressive language available
- It is possible to describe everything

Each resource defined using RDF is a statement <subject> <predicate> <object> called ‘triple’. The predicate describes the relation between the subject and the object, which are both resources. The predicate is also commonly called ‘property’. This representation is independent of the resources, and the formalized relations can be easily visualized as a graph. The serialization of the predicates is commonly an XML file stored in a database or in a File System (RDF Store/ Endpoint).
To translate the design of the ontology in RDF language, in this project has been used Protégé [7], an open source tool provided by the University of Stanford that allows developing ontologies and intelligent systems. It supports all the W3C standards previously presented and supports plug-in to extend its functionality. The big community that supports the project provides frequent updates. For these reasons, it is suitable for every type of use, from the development of small ontologies to the implementation of significant commercial intelligent systems. The coding of the ontology designed is guided by the user interface which allows, using few clicks, to code the classes, the properties and the relations among them. The final object produced by this step is a structured document which includes the translation of the ontology design into RDF triple.

The RDF document produced by this task is used as the schema of an RDF triple store which allows memorizing observed elements of the domain annotated with the ontological concepts. Famous software which can supply this need is Jena, Sesame, and Virtuoso [18]. In particular, we decided to use the last one. Virtuoso is a hybrid storage server architecture that allows covering the following areas: traditional relational RDBMS, RDF stores, Content Managers of web resources, and web application servers. It supports the interaction with data using different strategies. It provides libraries to use data from the code written in Java, .Net, C++ and more. Moreover, it supports different strategies of interaction such as SPARQL, SOAP, and JDBC. Finally, it also provides strategies to interact with other DBMS and make a federation of data. Virtuoso is a commercial tool.

In order to get the EO images available in a Spatial Metadata Catalogue (CSW) and to store them in RDF triple store using the ontological schema defined, the module CSWToRDF has been developed. The purpose of the CSWToRDF utility is the synchronization of the Spatial Metadata Catalog and the RDF Store operating as middleware between them. In seo-dwarf context, the Spatial Metadata Catalog is responsible for storing, documenting and disseminating products resulted by the Remote Sensing processing able to detect observable phenomena in the image. Each product is described based on an extended INSPIRE metadata template and is kept in the catalog as a metadata record. Spatial Metadata Catalog was built on GeoNetwork which provides an implementation of a CSWserver. CSW is an “open geospatial consortium” standard for exposing metadata records through web services. The CSW Server accepts CSW requests (e.g. getRecords,

getRecordById etc) and returns XML responses. On the other end, the RDF Store is responsible for keeping information about SEO-DWARF products using the RDF data model, adding thus semantic searching capabilities to seo-dwarf end-user applications. The RDF triples structure is in accordance with the seo-dwarf ontology. The CSWToRDF utility keeps (in an automated way) the RDF Store updated with the Spatial Metadata Catalogue.

To accomplish this task, it performs the following operations:

1. Harvesting of the Spatial Metadata Catalog by issuing CSW requests to get newly created or modified metadata records as XML Documents
2. Conversion of the retrieved XML Documents to RDF documents. The conversion is based on a manual defined mapping among INSPIRE metadata and seo-dwarf ontologies concepts
3. Storing the obtained RDF documents to the RDF Store

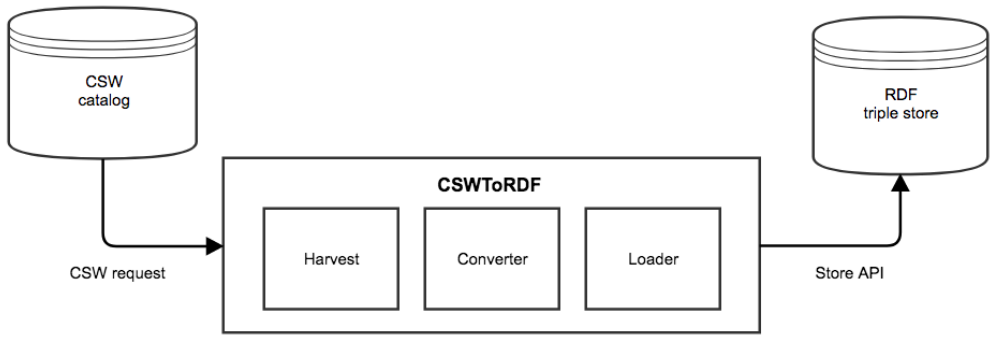

Figure 2 Middleware module for the population of the RDF triple store

The products of the Spatial Metadata Catalogue stored in an RDF format are then consultable using a query language such as SPARQL [19] which allows the retrieval module to correctly extract products which are relevant for the question asked by the final user of the system navigating the ontological structure proposed.

5. CONCLUSION

Information retrieval systems have been demonstrated to be, in the last decade, an excellent tool for identifying information relevant to the user quickly and efficiently. The standard application on web-pages is moving toward the use of them in new domains such as the one of Earth Observation Images retrieval. An EO image is a complex object which includes graphical information and semi-structured metadata, but it does not include the description of phenomena and semantic objects observable by a human in the image. The semantic annotation of images overcome this gap. Automatic approaches can annotate images with a structured set of concepts and properties which are defined into an ontology. In this work, it has been described the design strategy used for the project SEO-DWARF (Semantic Earth Observation Data Web Alert and Retrieval Framework) funded by the European Union Horizon 2020 programme. The seo-dwarf ontology has extended the top-level NASA SWEET geographical ontology, to include specific marine phenomena necessary for the project (i.e., oil spills, algal bloom, fronts). The image has been serialized as a set of classes, properties and relations able to describes the content computationally. This schema has been used as the storing structure of an RDF triple store able to memorize Spatial Metadata Catalogue products into RDF triples using the CSWToRDF middleware. The computational object obtained have then been used as a tool for a semantic-aware data exploration by the retrieval module of the SEO-DWARF project.

ACKNOWLEDGEMENTS

This research has received funding from the European Union's Horizon 2020 research and innovation programme under

the Marie Skodowska-Curie grant agreement N. 691071.

REFERENCES

[1] Baeza-Yates, R., & Ribeiro-Neto, B. , [Modern information retrieval] (Vol. 463). New York: ACM press (1999).

[2] Rajam, F., & Valli, S., "A survey on content based image retrieval", Life Science Journal, 10(2), 2475-2487, (2013).

[3] Chang, N. S., & Fu, K. S., "Query-by-pictorial-example", IEEE Transactions on Software Engineering, (6), 519-524, (1980).

[4] Uschold, M., & Gruninger, M., "Ontologies: Principles, methods and applications", The knowledge engineering review, 11(2), 93-136, (1996).

[5] Matuszek, C., Cabral, J., Witbrock, M. J., & DeOliveira, J., "An Introduction to the Syntax and Content of Cyc", In AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering, 44-49, (2006).

[6] Fernández-López, M., Gómez-Pérez, A., & Juristo, N., [Methontology: from ontological art towards ontological engineering], (1997).

[7] Knublauch, H., Fergerson, R. W., Noy, N. F., & Musen, M. A., "The Protégé OWL plugin: An open development environment for semantic web applications", In International Semantic Web Conference, 229-243, Springer, Berlin, Heidelberg ,(2004).

[8] Noy, N. F., & McGuinness, D. L., [Ontology development 101: A guide to creating your first ontology], (2001).

[9] Stumme, G., & Maedche, A., "FCA-Merge: Bottom-up merging of ontologies", In IJCAI Vol. 1, 225-230, (2001).

[10] Noy, N. F., & Musen, M. A., "Algorithm and tool for automated ontology merging and alignment", In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00). Available as SMI technical report SMI-2000-0831, (2000).

[11] Raskin, R. G., & Pan, M. J., "Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)", Computers & geosciences, 31(9), 1119-1125, (2005).

[12] Vassiliadis, V., Wielemaker, J., & Mungall, C., "Processing OWL2 ontologies using Thea: An application of logic programming", In Proceedings of the 6th International Conference on OWL: Experiences and Directions-Volume 529, pp. 89-98, CEUR-WS. org, (2009).

[13] Bossard, M., Feranec, J., & Otahel, J., [CORINE land cover technical guide: Addendum 2000], (2000).

[14] Koubarakis, M., & Kyzirakos, K., "Modeling and querying metadata in the semantic sensor web: The model stRDF and the query language stSPARQL", In Extended Semantic Web Conference, 425-439, Springer, Berlin, Heidelberg, (2010).

[15] Koubarakis, M., Datcu, M., Kontoes, C., Di Giammatteo, U., Manegold, S., & Klien, E., "TELEIOS: a database-powered virtual earth observatory", Proceedings of the VLDB Endowment, 5(12), 2010-2013, (2012).

[16] Bordogna, G., Bucci, F., Carrara, P., Pagani, M., Pepe, M., & Rampini, A., "Extending INSPIRE Metadata to imperfect temporal descriptions", International Journal of Spatial Data Infrastructures Research, 4, (2009).

[17] Klyne, G., & Carroll, J. J., [Resource description framework (RDF): Concepts and abstract syntax], (2006).

[18] Rohloff, K., Dean, M., Emmons, I., Ryder, D., & Sumner, J., "An evaluation of triple-store technologies for large data stores.", In OTM Confederated International Conferences On the Move to Meaningful Internet Systems, 1105-1114, Springer, Berlin, Heidelberg, (2007).

[19] Quilitz, B., & Leser, U., "Querying distributed RDF data sources with SPARQ", In European Semantic Web Conference, pp. 524-538, Springer, Berlin, Heidelberg, (2008).