An Ontological Model for Semantic Interoperability Within an Earth Observation Knowledge Base

Abstract:
This paper presents an ontological model which will be included in an EO Knowledge Base (KB) covering four thematic strands. As a multitude of heterogeneous data will be made available through the KB, it is essential to ensure high standards of discoverability, accessibility, and interoperability. The overall aim therefore, is to align and integrate a set of existing semantic resources and ad hoc vocabularies into a single ontological conceptual model, which defines the specific domains and which will facilitate information and knowledge generation from EO data. Thus, guaranteeing semantic interoperability within the KB platform, ensuring harmonised access to and retrieval of the vast volume of data produced, turning it into usable information and knowledge.

1.0 Aim and scope of the study
Integrating complex data, dynamic in nature, from heterogeneous resources, and without broadly applied standards constitutes a real challenge for users trying to make sense of the increasing amount of information made publicly available in any domain (Bodenreider et al. 2002). Earth Observation (EO) data, in particular, has increased considerably over the last decades and end-users are facing many challenges in accessing and analysing this data. In order to address the organisation and homogenisation of the huge volume of information that Earth Observation research is producing nowadays, scientists, and not only, need support from lexical and semantic tools, such as terminologies, vocabularies, nomenclatures, code and synonym sets, lexicons, thesauri, ontologies, taxonomies and classifications (De la Iglesia et al. 2013). Information Science typically defines information in terms of data, knowledge in terms of information, and wisdom in terms of knowledge (Rowley 2007). Generating information and knowledge from data is about understanding and connecting. Some initiatives, such as Copernicus, Group on Earth Observation (GEO), INSPIRE and others 2, focusing on air quality monitoring, atmospheric conditions and pollution emissions, have generated a large volume of data. These data have been collected by using different devices simultaneously and dissimilar modalities, thus access to them remains difficult.

On the basis of the above actions, the ERA-PLANET Programme 3 with the enrolment of the most active experts in the EO field aims to develop, among other objectives, a Knowledge Platform to ensure harmonized access to the vast amount of data produced.

1 Although the authors have cooperated in the research work and in writing the paper, they have individually devoted specific attention to the following sections: Aracri: 3.3, 3.4 and 4.0; Caruso: 1.0 and 3.2; Folino: 2.0 and 3.1.
The purpose is to gather distributed data coming from in-situ sensors and satellite-based remote sensing technologies, by means of practices and infrastructures able to organize, interpret and summarize them. To meet this challenge, the Programme is made up of 4 strands\(^4\), each one corresponding to a project\(^5\). All the projects try to provide more reliable information to policy makers concerning the status of the Earth in order to discuss and identify common strategies able to limit climate change, which is dangerous both for environmental and human health, by promoting sustainable development. Such data and information need to be made available through an interoperable system for data sharing and management able to ensure data quality and to interpret the meaning of data, turning them into usable information and knowledge.

Existing EO frameworks support a variety of geographic data set types as well as tools for data management, analysis and visualisation, but often they do not provide any mechanism to tackle semantic heterogeneity issues (Fugazza et al. 2010). Even when EO platforms do include a number of vocabularies, if they are not aligned, the retrieval of all the information regarding a certain topic will not be guaranteed. Relating terms from distinct vocabularies creates richer structural information that can be used for improving search and query expansion (Craglia 2011). To fulfill this requirement an ontology is under development and a set of semantic resources are being integrated and aligned through a corpus-based approach. The overarching goal is to organize and to give access to integrate complex data, dynamic in nature, from heterogeneous resources, and without broadly applied standards, in order to improve the ability of end-users to explore and exploit EO data. The fulfillment of these ambitious goals, as well as the necessity to integrate our semantic resource into a comprehensive platform oriented us towards the choice of an ontology since it provides a conceptual framework which is more structured, adaptable and reusable.

\section*{2.0 Literature review}

In this section we provide a brief literature review regarding methods and projects oriented to align and integrate sets of existing semantic resources and ad hoc vocabularies, regardless of the domain within which they have been applied. As for the alignment, the term used for referring to the establishment of a variable degree of correspondence between concepts that belong to different controlled vocabularies is ontology mapping\(^6\). The increasing number of ontology matching methods and tools and the necessity of reaching a consensus on their evaluation determined the advent of the Ontology Alignment Evaluation Initiative (OAEI)\(^7\).

\begin{flushleft}
\footnotesize\(^7\) http://oaei.ontologymatching.org/
\end{flushleft}
In general, in order to be aligned and subsequently published as linked open data on the Semantic Web, several existing controlled vocabularies have been previously converted in SKOS format. Some of them have been mapped without the systematic application of automatic approaches or the direct involvement of domain experts. The indexing languages RAMEAU, Library of Congress Subject Headings (LCSH) and Subject Headings Authority File (German: Schlagwortnormdatei SWD) (Landry 2009), as well as the Dewey Decimal Classification (DDC), the Library of Congress Classification (LCC) and the Medical Subject Headings (MeSH) (Vizine-Goetz et al. 2004), the Thesaurus for Economics (STW) and the Thesaurus for the Social Sciences (TSS) (Mayr and Petras 2008) and the Thésaurus du Tourisme et des Loisirs (Caruso and Folino 2015) have been mainly aligned without exploiting automatic approaches. Some other research works propose the semi-automatic detection of exact matches between concepts, and several matching systems are based on the computation of string similarity measures, rather than on semantic criteria. However, in almost all studies, the evaluation phase is performed manually. As stated in Morshed et al. (2011), some tools, such as S-match, use external resources (i.e. WordNet) as a background for recognising semantic relationships, but this approach seems less suitable for domain-specific terminologies. The approach here explained concerns the alignment between AGROVOC and other six KOSs more or less related to the field of agriculture. For each pair of concepts some string similarity measures are computed and the average value is taken into consideration for the subsequent manual validation. In order to perform this phase, experts have considered the status of the term (preferred or not), the hierarchy, the equivalent labels in other languages and the notes associated to concepts. The use of hierarchy as a disambiguation technique has also been used in presence of one-to-many alignments in Tordai et al. (2009) and, for each concept to be mapped, it takes into consideration both broader and narrower alignments. The authors have adopted a combination of techniques for establishing mappings between concepts coming from thesauri belonging to the cultural heritage domain: syntactic exact match techniques, a linguistic analysis and a technique deploying the ontology structure.

The need for a manual evaluation depends on several problems generating incorrect matches when using automatic approaches. Some of them are listed in Kempf et al. (2014): terms share the same lexical value but their broader and narrower terms or their scope notes are different; terms in different domains seems to be similar but their meaning is different; the matching between a synonym and a preferred term generates an incorrect equivalence.

To sum up, we refer to one of the most complete classifications of matching techniques (Euzenat and Shvaiko 2013), shown in Figure 1. Furthermore, starting from the assumption that a domain of interest can be represented through a corpus of text documents, it can be assumed that the knowledge domain that should be encoded into an ontology is represented through a domain corpus, and that the evaluation should output some measures that express the coverage and the adequacy of the ontology with respect to said domain (Rospocher et al. 2012). The integration of ad hoc vocabularies should therefore begin with the acquisition of domain-specific terminology (ex. Liddle et al. (2003); Navigli and Velardi (2004);  

---

8 https://www.w3.org/TR/skos-reference/
Methodologies similar to the one used in the present paper are brought forth by Brewster et al. (2004), in which the authors present a method for evaluating an ontology by comparing it with a domain-specific corpus, and by Cui (2010), who compares the coverage, semantic consistency, and agreement of four thematic ontologies by checking them against a corpus of domain literature.

Figure 1. Classification of matching techniques

3.0 Method

3.1 Corpus-based terminology extraction

New domains with specific conventions and new terminology are continuously appearing. The development of terminology lists is a previous step in KOS building that allows the use of automatic or semi-automatic techniques that could facilitate the labour (Peñas et al. 2001). Terminology extraction deals with the identification of terms which are frequently used to refer to the concepts in a specific domain, and therefore most likely representative of it. The main aim of term extraction here was to carry out a corpus-based terminological evaluation of the existing semantic resources in the EO domain in order to assess whether they adequately cover the terminology used in the domain corpus. Terminology extraction has been undertaken using the T2K² (text-to-knowledge) tool (Dell’Orletta et al. 2014), specifically conceived to identify and extract simple and compound terms from unstructured texts. The main assumption on which T2K², along with most terminology extraction software, is based, is that the relevant concepts of a text are conveyed by the terms that will occur most frequently. The tool performs a linguistic analysis of the texts, the result of which consists of a terminological vocabulary accompanied by semantic and conceptual information about the terms themselves, which add to the value of the output. The list of candidate terms has been sorted by frequency and subsequently manually revised. Understanding whether a given KOS adequately covers the domain of interest is a common and important issue when evaluating a semantic resource. After having matched and evaluated the existing resources’ semantic/terminological coverage, which will be detailed in the following section, the terms appropriate to the sub-domains and objectives involved were used to produce/integrate the ontology. Both terms common to the corpus and to the existing vocabularies, as well as terminology unique to the domain corpus, i.e. not present in any of the examined vocabularies, was used to integrate the ontology to be incorporated in the EO Knowledge Base platform.
3.2 Evaluation of EO vocabularies semantic coverage

Alongside the corpus construction, relevant existing terminological resources related to the Environment have been collected to be used as references. To date, thesauri such as the General Multilingual Environmental Thesaurus (GEMET)\(^9\), the EARTH Thesaurus\(^10\), the AGROVOC Thesaurus\(^11\), along with the INSPIRE Feature Concept Dictionary and Glossary\(^12\), have been taken into consideration. These terminologies have been downloaded in an easily computable format and in the form of flat-lists in order to compare all their terms to those extracted in the previous phase.

Over 5,550 terms have been extracted from the GEMET Thesaurus, while around 14,000 terms from the EARTH thesaurus. Meta-terms along with Macro areas have not been taken into consideration. The AGROVOC thesaurus includes approximately 45,500 terms. Almost 200 terms have been extracted from the INSPIRE glossary and 360 from the Feature Concept Dictionary. Considering their partial semantic overlapping, some vocabularies are already mapped to each other in order to allow federated access to information. An initial comparison has been carried out between the four term lists and those extracted from our sub-corpora in order to identify exact matches and understand if the concepts relevant to our specific purposes are present in the existing resources.

The comparison has been carried out through the use of WordSmith Tools\(^13\), a software tool that allows the comparison of the wordlists generated from the existing resources and the corpus. For instance, 450 candidate exact matches were found when comparing GEMET with the sub-corpus terminology, 26 exact match candidates in INSPIRE, 886 exact matches between AGROVOC and the specialized term list, and 763 candidate exact matches were automatically identified between the terminology present in EARTH and the terms extracted from one of the sub-corpora. A further manual analysis however, results in the identification of close matches, broader and narrower matches.


\(^{10}\) https://vocabularyserver.com/cnr/ml/earth/en/.


\(^{13}\) https://lexically.net/wordsmith/.
3.3 Definition of an ontological conceptual model

In this section we present the methodological choices and the steps involved in the development of the ontological model covering the thematic strands, which are strictly related but not necessarily overlapping. In a nutshell, an ontology is a complex formal vocabulary\(^\text{14}\) which requires background knowledge of the domain in order to explicit an efficient conceptual framework. Within the ontology the main issue is how concepts are logically implicated and which kind of information/knowledge we would like to get out from concepts and relationships, even if there are no \textit{a-priori} links between them. EO data and information are both extremely heterogeneous and dynamic in terms of formats, meanings, languages, etc., and captured by using different technologies, applications and systems\(^\text{15}\). For the ERA-PLANET program, the major challenge regards providing decision support tools aimed at guaranteeing access to relevant information, whenever needed and in a comprehensible format. Thus, data and information can be reused and managed for various purposes, depending on the challenge.

The ontology to be included in the ERA-PLANET Knowledge Platform has been developed through the use of Protégé, an open tool used to design and integrate ontologies. The aim is to support decision making in accordance with the Sustainable Development Goals (SDGs) and to monitor activities in evaluating measures to support environmental policies. To provide a complete and detailed representation of the concepts faced in the projects a top down approach has been adopted by implementing a high level conceptualization due to the inclusive nature of the ontology. Hence, at the top level there are the main classes which are key concepts across the domains and the access points to explore and browse the entire conceptual structure without evident boundaries among the sub-domains. Classes are, in turn, organized in further sub-classes

\(^14\) https://www.w3.org/standards/semanticweb/ontology.

\(^15\) Big Data in Earth Observation.

at different hierarchical levels\textsuperscript{16} and are at the same time member of the superclass and root of another subclass. The result is a taxonomy in which concepts are hierarchically organized by means of the specification of generic relationships in a universally acceptable manner. Even if concepts included in the ontology are related to the subdomains, there are those that are strictly domain-based (ex. Ecosystem - Terrestrial_ecosystem - Anthropogenic_terrarestrial_ecosystem - Cropland), thus they present a very high level of specialization, while others, whose meaning is less domain-specific, present few hierarchical levels (ex. Dataset - Copernicus scenes). Therefore, at the moment, the same granularity cannot be guaranteed in all classes. Classes are not populated by individuals, but are related to each other by means of direct and inverse properties (ObjectProperties) defined according to the type of relationship that is useful to explicit. The match between classes and properties generates a statement in the form of subject-predicate-object expressions (ex. “Indicator 15.3.1 measures Target 15.1” and vice versa “Target 15.1 isMeasuredBy Indicator 15.3.1”).

![Ontological model](image)

**Figure 4. Ontological model**

### 3.4 Ontology and vocabulary alignment

The ontology development benefits from the terminology extraction and comparison phases described above. Some concepts have been added to expand the semantic granularity, and in order to better contextualize each concept, some additional information, such as definitions coming from other vocabularies or links to multiple sources, has been included. In defining mappings a semi-automatic approach has been adopted and a unidirectional mapping has been implemented from our ontology to the above mentioned controlled vocabularies. In particular, our main reference is represented by GEMET, because of its consolidated use within the scientific community and the alignments it already has with other vocabularies.

More specifically, automatic procedures are used to discover and establish exact and sometimes close matches across concepts coming from different vocabularies, while human mediation was necessary to validate the output of automatic procedures and to

\textsuperscript{16} For a description of ontology structure and construction see (Noy and McGuinness 2001; Capuano 2005).
identify both hierarchical and associative mappings when it is not possible to provide a valid equivalence match. This is done to ensure that all concepts included in the ontology have an external reference in at least one vocabulary involved in ontology mapping. With regard to equivalence, using an hybrid method allows to control and manage inexact matches (ex. Homographs, Synonyms), because, as within the same vocabulary, also in cross-mapping, the equivalence relationship can be exact, inexact and partial (ISO 25964-2:2013). The different degrees of equivalence mapping have been expressed by means of the SKOS data model that has been imported in Protégé and its properties skos:exactMatch and skos:closeMatch have been used as AnnotationProperties. For the sake of completeness, some skos:broadMatch, skos:narrowMatch and skos:relatedMatch have been included in our ontology, as can be seen in Figure 5 below.

![Figure 5. Examples of mapping](image)

To overcome differences in vocabulary structures, ISO 25964-2:2013 provides some recommendations concerning the suitable model or combination of models to be used. Indeed, SKOS and the ISO 25964 data model are aligned\(^\text{17}\) so that it is possible to create a mapping compliant to both. As regards our mapping process, we have opted for a combination of the hub model, because our ontology is the core and the external vocabularies act as satellites, and the selective mapping, because links are established only in one direction and solely for concepts used in the ontology.

The vocabularies involved in the mapping partially overlap with the ERA-PLANET thematic strands, as illustrated in Figure 6 below.

![Figure 6. Mapping percentages](image)

The preliminary results, exclusively regarding exact and close matches, therefore show the necessity to extend this mapping involving other vocabularies.

### 4.0 Conclusion

Ontology mapping is a challenge in some strategic domains when a large volume of data need to be managed. The ongoing initiatives and actions highlight the importance of interoperability, discoverability, accessibility and reusability of data. These issues are central to our project because all data must be made available to policy decision-makers, who have the hard task of designing the strategies to be used to support sustainable development. In this sense, the ontology and the matching with external vocabularies,

\(^{17}\) [https://www.niso.org/schemas/iso25964](https://www.niso.org/schemas/iso25964).
represent a valid semantic support because they allow to expand and specialize an information request. The ontology evolves continuously, therefore a periodic update is necessary in order to revise mappings and add new links to further vocabularies. To improve semantic interoperability in the KP it is expected to define links to some further existing vocabularies (e.g., AGROVOC) and ontologies (e.g., Sustainable Development Goals Interface Ontology (SDGIO), Chemical Entities of Biological Interest (ChEBI)) which cover some other areas of interest such as health, agriculture and chemistry, that are strictly related to the environmental domain.

Acknowledgments
This activity was funded by the European Commission in the framework of program “The European network for observing our changing planet (ERA-PLANET)”, Grant Agreement: 689443. The authors would like to thank colleagues from IIA-CNR Florence for their collaboration in defining the ontological model.

References
Fugazza, Cristiano, Sören Dupke, and Lorenzino Vaccari. 2010. “Matching SKOS Thesauri for Spatial Data Infrastructures.” In Metadata and Semantic Research: Research Conference on


