Essential Variables workflows for resource efficiency and environmental management

Deliverable 1.4
Semantic services

Creator	A. Folino, A. Caruso, G. Aracri (University of Calabria, Italy)
Creation date	May. 1. 2018
Due date	August. 31. 2019
Last revision date	September. 30. 2019
Status	Final
Type	Other
Description	Terminological resources and ontological model for semantic interoperability services within the Knowledge Platform
Right	Public
Language	English
Citation	Folino A., Caruso A., Aracri G., Semantic services. GEOEssential Deliverable 1.4
Grant agreement	ERA-PLANET No 689443

Executive Summary

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. Earth Observation (EO) data has increased considerably over the last decades, however, access to this data remains difficult for end-users in most domains. As a multitude of heterogeneous data will be made available through the GEOEssential Knowledge Base infrastructure, it is essential to ensure high standards of discoverability, accessibility, and interoperability. The design of this infrastructure involves the alignment and integration of a set of semantic resources defining the specific domain. The latter is important in order to ensure harmonised access to the vast volume of data produced, turning it into usable information and knowledge, and to guarantee semantic interoperability within the infrastructure. This involves the mapping of existing aligned thematic vocabularies (i.e. glossaries, taxonomies, thesauri and ontologies), along with the integration of further domain-specific terminology obtained through a corpus-based approach. The integration of the abovementioned vocabularies in a knowledge base infrastructure will therefore improve the ability of end-users to explore and exploit EO data. Some of the vocabularies employed are the following: GEMET Thesaurus, INSPIRE Feature Concept Dictionary and Glossary, AGROVOC Thesaurus, EARTh Thesaurus.

The GEOEssential Knowledge Base will include an ontological conceptual model which will aide in the transition from Data to Knowledge. On a more abstract level, the ontology schema defines the major concepts of the specific domain (e.g. Essential Variables, Policy Goals, Indicators, Targets) and the relationships between them. The conceptual taxonomy, organizing classes at different hierarchical levels, will be obtained by including concepts and terms defined in the common terminology. The Knowledge Base will be based on OWL/RDF technologies.

Table of contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 From Data to Knowledge

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. Different techniques and methodologies aim to define resources (i.e. models, taxonomies, thesauri[1], ontologies[2]) to ensure data quality and harmonization and to interpret the meaning of data, turning it into usable information and knowledge. Figure 1 illustrates the data-information-knowledge-wisdom (DIKW) pyramid and how it applies to our context. As to EO resources and SDGs, it is possible to recognize the following artefacts: (a) Data: Earth observations and measurements; (b) Information products: EVs and Indicators; (c) Knowledge products: indexes; (d) Wisdom actions: SDGs.

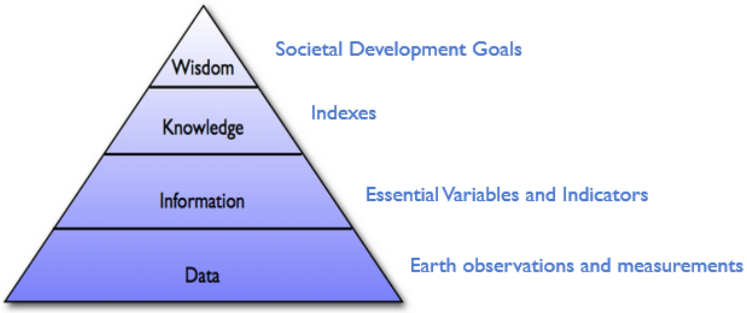
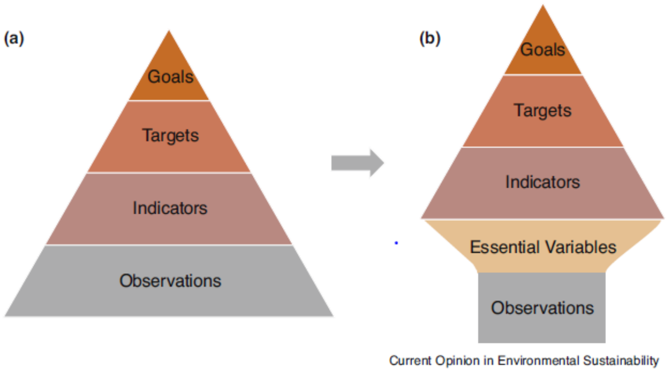

Wisdom
Knowledge
Information
Data
Societal Development Goals
Indexes
Essential Variables and Indicators
Earth observations and measurements

Figure 1. DIKW pyramid

[1] "Controlled and structured vocabulary in which concepts are represented by terms organized so that relationships between concepts are made explicit and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms", ISO 25964-2:2013 Information and documentation - *Thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies,* p. 12.

[2] "explicit formal specifications of the terms in the domain and relations among them" (Gruber 1993).



Current Opinion in Environmental Sustainability

Figure 2. Essential Variables pyramid (Reyers et al. 2017, p. 98)

This is in line with (Reyers et al., 2017, p. 98) who introduce Essential Variables “as a layer between primary observations and indicators”, thus moving from an ever-broadening pyramid (a) to a more streamlined form (b) (Figure 2). Furthermore - and this is important for the construction of the ontological model explained below - image (b) of the figure expresses the concept that “a single EV capturing a key process or structure can potentially contribute to multiple indicators, while similarly 2 or more EVs can direct and use the same primary observations, thus potentially enabling a reduction in the numbers of observations needed to deliver those indicators”.

Information is an added-value product generated by understanding data and working out relations among them and with physical and/or social phenomena. Understanding information and working out valuable patterns generates knowledge, in turn. Models, processing algorithms and workflows as well as lexicon resources play a crucial role for doing that.

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 the domain of Earth Observations (Bodenreider et al., 2002).

In order to address the organization and homogenization of the huge volume of information that Earth Observation research is producing nowadays, scientists 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). Sharing and gaining consensus by the EO community on the categorisations and disambiguation of terms is one of the steps for enabling interoperability among data sets and services that are provided by a heterogeneous set of thematic domains. A further step is that of matching the abovementioned resources so as to facilitate the harmonization of the vocabularies that independent data providers may have adopted for the annotation of resources.

The overall aim of Task 1.4 therefore, is to align (ISO 25964-2:2013)[3] and integrate a set of existing semantic resources and ad hoc vocabularies into an ontological conceptual model,

[3] The alignment or mapping process between semantic resources is described in the ISO 25964:2-2013. It is defined as follows: “process of establishing relationships between the concepts (3.17) of one vocabulary and those of another” (p. 7).

which defines the specific domain and which will facilitate information and knowledge generation from EO data and guarantee semantic interoperability within the GEOEssential Knowledge Base platform, ensuring harmonised access to the vast volume of data produced, turning it into usable information and knowledge.

1.2 EO Platforms, Information Retrieval and Semantic Interoperability

Earth Observation data has increased considerably over the last decades, however, access to this data remains difficult for end-users in most domains. Due to the fact that the nature of this information is heterogeneous and has different levels of granularity, researchers are facing many challenges in analysing all these data. As a multitude of heterogeneous data will be made available through the GEOEssential Knowledge Base, it is essential to ensure high standards of discoverability, accessibility, and interoperability.

Among the activities that have been carried out in Task 1.4 are those that involve the semantic coverage analysis of the thematic vocabularies (i.e. glossaries, taxonomies, thesauri and ontologies), the possibility to integrate further domain specific terminology and the integration of both in the knowledge base infrastructure, thus improving the ability of end-users to explore and exploit EO data.

As suggested above, the purpose of relating independent vocabularies, in particular thesauri, to each other is manifold[4]: on the one hand, relations allow the user to navigate across distinct domains enabling expressive browsing functionalities; on the other hand, thesauri can more efficiently be used for describing and discovering data and services among different disciplines. In fact, relating terms from distinct thesauri creates richer structural information that can be used for query expansion either with terms of a single thesaurus or multiple thesauri (Craglia et al., 2011).

This technique allows for narrowing or broadening the number of results returned by a query by referring to, respectively, more specific and more general terms in the hierarchy or by referring to equivalent terms (synonyms or quasi-synonyms, and exact, inexact or partial equivalence) either in a thesaurus or among multiple thesauri.

The typical EO framework is based on a software platform which supports a variety of geographic data set types as well as tools for data management, analysis and visualisation. Usually, this framework does not provide any mechanism to tackle semantic heterogeneity issues, which arise when the content of information exchange requests is not clearly defined (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. Suppose a user wants access to all data regarding ‘mercury’, and uses the chemical element symbol ‘Hg’ as the search term in the EO platform, the system returns a list of data sources relevant to mercury, but only if the data was indexed using the chemical element symbol. All data for which the descriptive keyword ‘mercury’ had been used would not be retrieved.

[4] For a discussion on the alignment of semantic resources see (Isaac et al., 2009) and (Morshed et al., 2011).

The alignment of distinct vocabularies therefore, allows for the discovery of resources without requiring the metadata used for indexing/describing their contents to be the same and without asking data provider to use compulsory set of predefined keywords.

2 STATE OF THE ART

2.1 EO Vocabularies

The integration of thematic thesauri improves performance and adds coverage of the main thematic areas in any given domain. There are two aspects of data interoperability in the management of semantics-aware data structures: syntactic interoperability and semantic interoperability. Both aspects are required for an integrated exploitation of heterogeneous data. To improve syntactic interoperability, many efforts have already been made, such as standardization of data formats and development of XML-based data encoding rules, i.e. an ISO (International Organization for Standardization) standard and an OGC (Open Geospatial Consortium) standard (Nagai et al., 2012).

Improvement of semantic interoperability requires better consistency among different ontologies, terminologies, taxonomies, and so forth. Several institutions have introduced efforts to propose a standard ontology and/or terminology/taxonomy related to the Environment domain. The SWEET (Semantic Web for Earth and Environment Terminology) ontologies, developed by NASA, constitute an example of such terminologies. FAO (the United Nations' Food and Agriculture Organisation) has been making a similar attempt by creating AGROVOC, a multilingual, structured, and controlled vocabulary designed to encompass all subject fields in agriculture, forestry, fisheries, food, and related domains. GEMET Thesaurus, EARTh Thesaurus and the INSPIRE Feature Concept Dictionary and Glossary also constitute examples of structured thematic vocabularies in the EO domain. The Environmental Thesaurus Server (EnvThs)[5] is an example of initiatives aimed at guaranteeing data sharing and exchange at an international level. EnvThs provides access to controlled vocabularies, taxonomies and ontologies widely used and recognized in the geoscience/environmental informatics community. The Simple Knowledge Organization System (SKOS), which will be explained in Section 3.3, is used for the representation of the controlled vocabularies accessed through EnvThs, while TemaTres[6], an open-source, web-based thesaurus management package is employed and extended for working with them.

SWEET Ontology[7] is an example of a highly modular ontology suite which includes 6,924 concepts (Classes, Object Property, Data Property and Individuals) in 225 separate ontologies[8] covering Earth system science. A modular ontology is defined as a set of ontology modules, where these modules can be integrated through various proposed formalisms (Ensan et al., 2010). Indeed, SWEET is a mid-level ontology and consists of nine top-level concepts that can

[5] http://edscvs.ccny.cuny.edu/

[6] http://www.vocabularyserver.com/

[7] https://sweet.jpl.nasa.gov/

[8] Numbers are accurate as of January 2018.

be used as a foundation for domain-specific ontologies that extend these top-level SWEET components. SWEET has its own domain-specific ontologies, which extend the mid-level ontologies. The former can provide users interested in developing a finer-grained ontological framework for a particular domain with a solid set of concepts to get started.

The **AGROVOC** thesaurus is a multilingual controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment, etc[9]. It has evolved over the years and currently AGROVOC is an SKOS-XL concept scheme and a Linked Open Data (LOD) Dataset edited by VocBench, composed of over 35,000 concepts available in up to 29 languages, containing up to 40,000+ terms in each language. AGROVOC is aligned with 18 other multilingual knowledge organization systems, some are general in scope while others are specific to various domains, e.g. GEMET for environment. These linked resources are mostly available as RDF/SKOS resources.

Besides being widely used in specialized libraries as well as digital libraries and repositories to index content, it is also used as a specialized tagging resource for knowledge and content organization by FAO and other third-party stakeholders (Caracciolo et al., 2010).

GEMET GEneral Multilingual Environmental Thesaurus[10], the reference vocabulary of the European Environment Agency (EEA) and its Network (Eionet) has been developed as an indexing, retrieval and control tool for the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European` Environment Agency (EEA). The basic idea for the development of GEMET was to use the best of the presently available excellent multilingual thesauri, in order to save time, energy and funds. GEMET was conceived as a "general" thesaurus, aimed to define a common general language, a core of general terminology for the environment. Specific thesauri and descriptor systems (e.g. on Nature Conservation, Wastes, Energy, etc.) have been excluded from the first step of development of the thesaurus and have been considered only for their structure and upper level terminology. The merging has been performed both on a conceptual and formal basis. Coinciding concepts in the different thesauri have been identified and scored. Like in other multilingual thesauri, a neutral alphanumerical notation allows the identification of a concept independently of the user's language. The links with the original thesauri are ensured by the respective identifiers or code notations. The resulting 6,562 terms have been arranged in a classification scheme made of 3 super-groups containing a total of 30 groups plus 5 accessory groups of terms, instrumental to the thesaurus use[11]. Each descriptor has been arranged in a hierarchical structure headed by a Top Term. Furthermore, to allow a thematic retrieval of semantically related terms but scattered in different groups, a set of 40 themes[12] have been agreed upon with the EEA and each descriptor has been assigned to as many themes as necessary. There are currently more

[9] http://www.fao.org/agrovoc

[10] https://www.eionet.europa.eu/gemet/en/about/

[11] E.g. Super-group 1: Natural Environment, Anthropic environment - Groups: Environment (natural environment, anthropic environment), Space, Atmosphere (air, climate); Super-group 2: Human activities and products, Effects on the environment - Groups: Wastes, Pollutants, Pollution; Super-group 3: Social aspects, Environmental policy measures - Groups: Legislation, Norms, Conventions, Environmental Policy; Accessory Groups: General Terms, Functional Terms.

[12] E.g. air, climate, energy, environmental policy, pollution, space, water.

than 4,000 definitions available, which provide a useful glossary function. The themes, being complementary to the groups, confer a matrix structure to the thesaurus.

EARTh, Environmental Application Reference Thesaurus, is one of the largest general purpose and structured environment terminological resources available on the Linked Open Data cloud (Albertoni et al., 2014). Its terminological content is derived from various multilingual and monolingual sources of controlled environmental terminology plus other thesauri and documents concerning specific sectors such as inland waters, pollution and climate change, environmental safety and disasters management. EARTh has refined and extended the above mentioned GEMET thesaurus, which is considered the de facto standard with regards to general-purpose thesauri for the environment in Europe. It aims at providing a bridge for the integration of other terminological resources dealing with the environment. It already includes more than 12,000 links to popular LOD datasets as AGROVOC, EUROVOC, DBPEDIA and UMTHES.
EARTh is currently maintained in the context of LusTRE[13], a framework developed within the EU project eENVplus that aims at combining existing thesauri to support the management of environmental resources. LusTRE considers the heterogeneity in scopes and levels of abstraction of environmental thesauri as an asset when managing environmental data, it exploits linked data best practices SKOS and RDF in order to provide a multi-thesauri solution for INSPIRE data themes related to the environment.

The **INSPIRE** Feature Concept Dictionary[14] (IFCD) acts as a common feature concept dictionary for all INSPIRE data specifications. The common feature concept dictionary contains terms and definitions required for specifying thematic spatial object types and it is main role is in particular to support the harmonisation effort and to identify conflicts between the specifications of the spatial object types in the different themes. The INSPIRE Glossary[15] contains general terms and definitions that specify the common terminology used in the INSPIRE Directive and in the INSPIRE Implementing Rules documents. The glossary supports the use of a consistent language in different documents when referring to the terms.

Domain-specific thematic vocabularies are also emerging in order to accommodate the specific terminology that particular thematic sub-domains may use (e.g. EUROGEOSS Drought Vocabulary, the GEOSS AIP-3 Semantics and Ontology Scenario Water Ontology, the Australian CSIRO Spatial Information Service Stack Vocabulary, InterWATER Thesaurus).
For data interoperability, ontological information – including terminology, taxonomy, thesauri, etc. – must be collected, managed, referred to and compared.

As far as existing domain ontologies are concerned, the Environment Ontology (**ENVO**)[16] is worth mentioning. It is a community-led, open project which seeks to provide an ontology for specifying a wide range of environments relevant to multiple life science disciplines and, through an open participation model, to accommodate the terminological requirements of all those needing to annotate data using ontology classes. ENVO is comprised of classes

[13] http://linkeddata.ge.imati.cnr.it/StartPage.jsp
[14] http://inspire.ec.europa.eu/featureconcept
[15] http://inspire.ec.europa.eu/glossary
[16] http://www.environmentontology.org

(terms) referring to key environment-types that may be used to facilitate the retrieval and integration of a broad range of biological data. In constructing ENVO, the developers recognized the many existing resources which address, among other entities, environment-types and were motivated by the value of unifying such resources in a foundational, or building block, ontology developed within a federated framework and exclusively concerned with the specification of environment types, independent of any particular application. Classes describing natural environments currently dominate ENVO's content as the ontology is geared towards use in the biological domain. Nevertheless, ENVO is suitable for the annotation of any record that has an environmental component.

A significant semantic resource is represented by the Sustainable Development Goals Interface Ontology (**SDGIO**), developed by UNEP (United Nations Environment Program) in collaboration with experts in the domain of knowledge representation[17]. Its importance derives from the closeness of its aims and domain of interest in respect to the ontology we are developing within the GEOEssential project: the objective of SDGIO is to logically represent and define entities relevant for the SDGs so that their meaning could be unambiguously understood and interpreted by the community of experts. Some concept definitions are not universally accepted or are different from one context to another and this can compromise the quality of data and the correct measurement of progress towards the corresponding targets. To this end, concepts included in the ontology have been mapped to the corresponding terminology in resources such as the UN System Data Catalogue and the SDG Innovation Platform. The SDGIO "aims to provide a semantic bridge between 1) the Sustainable Development Goals, their targets, and indicators and 2) the large array of entities they refer to"[18].

The SDGIO currently includes 514 classes, 144 object properties, 27 annotation properties and 702 instances. Several classes are imported from other existing ontologies (e.g. ENVO, CHEBI, OBI, PCO) and are mapped to the concepts contained in the above mentioned GEMET in order to provide a more comprehensive and precise representation of the domain and to guarantee a major interoperability.

3 METHODOLOGY

3.1 Corpus-based terminology extraction

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 such domain (Rospocher et al., 2012). Several studies see acquiring the terminology

[17] http://aims.fao.org/activity/blog/sustainable-development-goals-interface-ontology-sdgio-support-united-nations

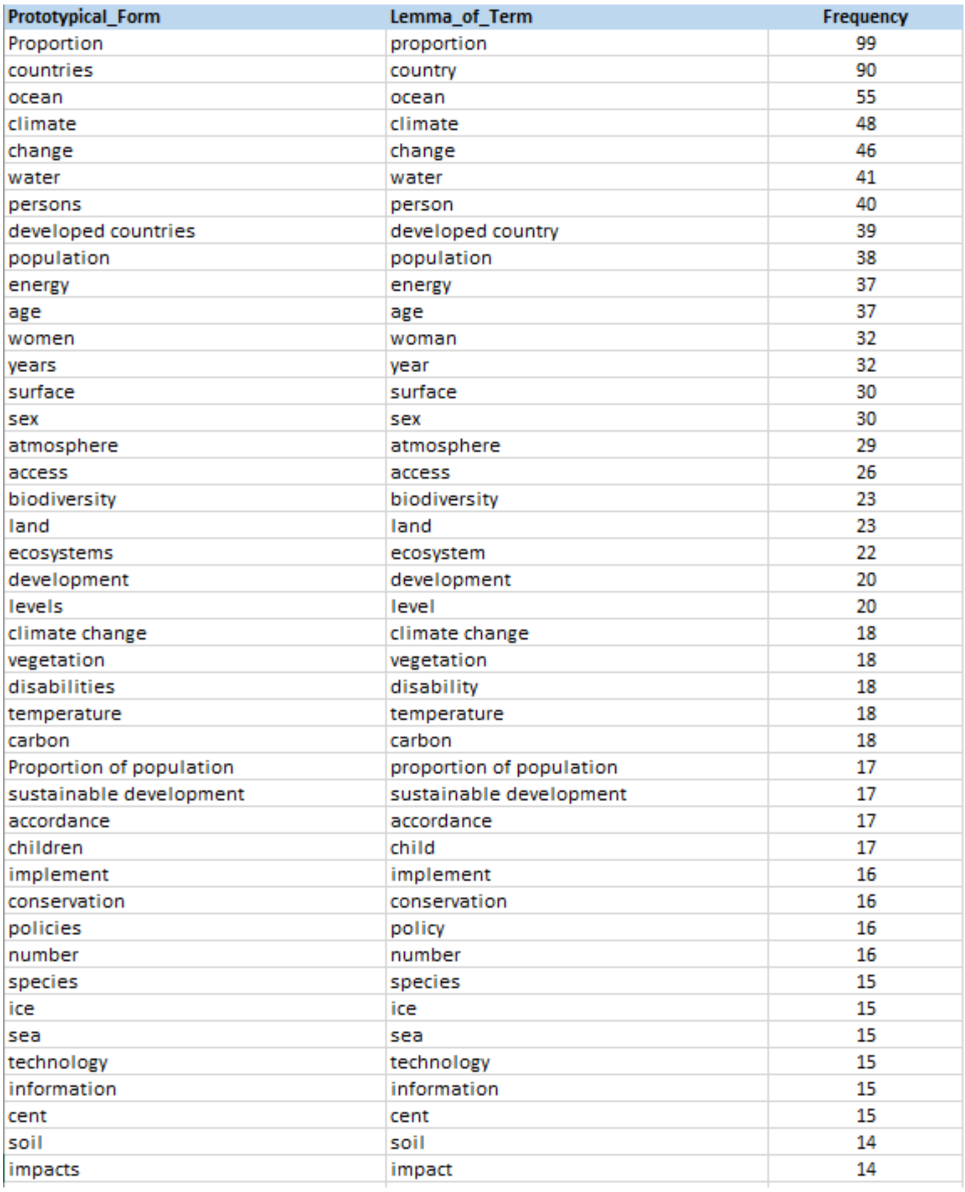
[18] http://www.ontobee.org/ontology/SDGIO

in the specific domain of interest as a useful starting point for the creation of a domain ontology (e.g. Liddle et al., 2003; Navigli et al., 2004; Lee et al., 2005; Wong et al., 2007). Methodologies similar to the one used in the present Task are presented by (Brewster et al., 2004), in which the authors illustrate 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.

When creating a corpus, be it general-purpose or domain-specific, the documents collected should come from authoritative sources. For our purposes the complete corpus will consist of documents describing the specific knowledge fields of Essential Variables and Sustainable Development Goals, such as domain-specific journals and European laws[19].

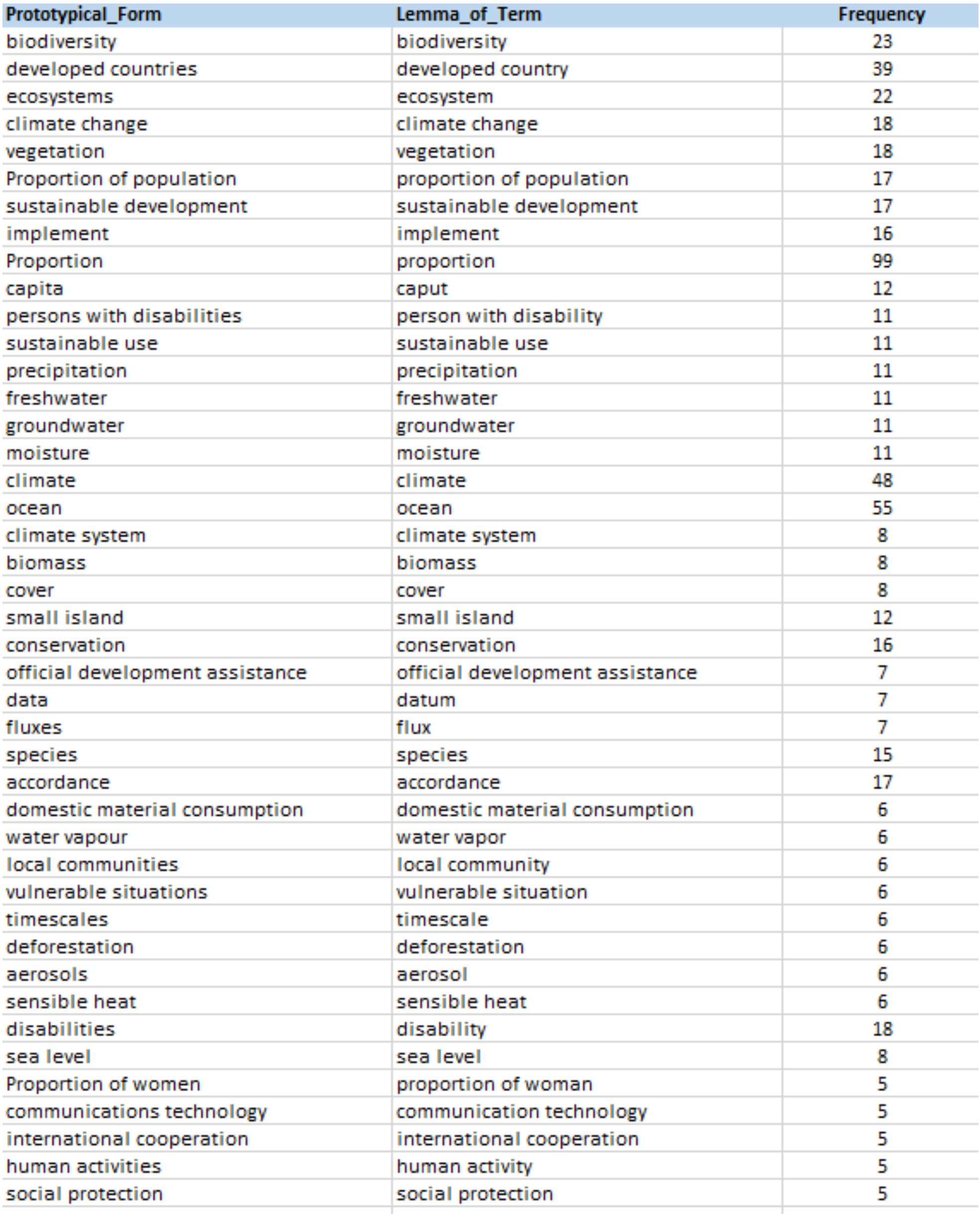
Therefore, the first step will consist in terminological extraction from the current domain-specific corpus. The final aim of the terminology extraction will be to carry out a corpus-based terminological evaluation of the existing vocabularies/thesauri in order to assess whether the latter adequately cover the terminology used in the text corpus which, when completed, will be representative of the domain of interest. A preliminary 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. Indeed, in addition to the set of candidate terms extracted from the documents, the software provides information about lexical and semantic relationships that affect the linguistic units, thus defining a kind of domain ontology made up of clusters of terms organized in a conceptual-semantic network (Caruso et al., 2016). The preliminary terminology extraction was conducted on an initial text corpus including accepted terms and definitions related to Essential Variables and Policies (e.g. SDGs, AICHI, etc.) both with and without the use of a reference corpus. Table 1 below is an extract of the term list obtained without the comparison of a reference corpus and sorted by frequency, while Table 2 illustrates the terms extracted by using a reference corpus and sorted by keyness. The comparison of the specialized term list with a reference corpus - represented by a general language corpus - allows us to compare the frequency of each term in the two lists. If a term is more frequent in the former than in the latter, it is likely that it is representative of the specific knowledge domain. Indeed, in Table 2 the list of terms is rearranged in such a way as to bring to the top of the list terms which are more representative of the domain to which the analysed texts belong.

[19] European laws and domain-specific journals are identified according to some qualitative and quantitative criteria: the domains considered are those related to the specific project WPs, i.e. Biodiversity and Ecosystem, The Food-Energy-Water Nexus, Extractive industry & light monitoring. The time-frame spans from 2016 to 2018.



Prototypical_Form	Lemma_of_Term	Frequency
Proportion	proportion	99
countries	country	90
ocean	ocean	55
climate	climate	48
change	change	46
water	water	41
persons	person	40
developed countries	developed country	39
population	population	38
energy	energy	37
age	age	37
women	woman	32
years	year	32
surface	surface	30
sex	sex	30
atmosphere	atmosphere	29
access	access	26
biodiversity	biodiversity	23
land	land	23
ecosystems	ecosystem	22
development	development	20
levels	level	20
climate change	climate change	18
vegetation	vegetation	18
disabilities	disability	18
temperature	temperature	18
carbon	carbon	18
Proportion of population	proportion of population	17
sustainable development	sustainable development	17
accordance	accordance	17
children	child	17
implement	implement	16
conservation	conservation	16
policies	policy	16
number	number	16
species	species	15
ice	ice	15
sea	sea	15
technology	technology	15
information	information	15
cent	cent	15
soil	soil	14
impacts	impact	14

Table 1. Term list sorted by frequency



Prototypical_Form	Lemma_of_Term	Frequency
biodiversity	biodiversity	23
developed countries	developed country	39
ecosystems	ecosystem	22
climate change	climate change	18
vegetation	vegetation	18
Proportion of population	proportion of population	17
sustainable development	sustainable development	17
implement	implement	16
Proportion	proportion	99
capita	caput	12
persons with disabilities	person with disability	11
sustainable use	sustainable use	11
precipitation	precipitation	11
freshwater	freshwater	11
groundwater	groundwater	11
moisture	moisture	11
climate	climate	48
ocean	ocean	55
climate system	climate system	8
biomass	biomass	8
cover	cover	8
small island	small island	12
conservation	conservation	16
official development assistance	official development assistance	7
data	datum	7
fluxes	flux	7
species	species	15
accordance	accordance	17
domestic material consumption	domestic material consumption	6
water vapour	water vapor	6
local communities	local community	6
vulnerable situations	vulnerable situation	6
timescales	timescale	6
deforestation	deforestation	6
aerosols	aerosol	6
sensible heat	sensible heat	6
disabilities	disability	18
sea level	sea level	8
Proportion of women	proportion of woman	5
communications technology	communication technology	5
international cooperation	international cooperation	5
human activities	human activity	5
social protection	social protection	5

Table 2. Term list sorted by keyness

Having used this function, the term ‘aerosols’ for instance, undoubtedly representative of the domain, has gone up from position 128 in the frequency list to 35 in the keyness list.

Understanding whether a given thesaurus adequately covers the domain of interest is a common and important issue when evaluating a terminological resource. For instance, one may want to understand if a publicly available thesaurus is relevant and adequate for the domain to be modelled, in order to consider its possible adoption.

After having matched and evaluated the abovementioned resources’ semantic/terminological coverage, which will be detailed below in 3.2., terminology unique to the domain corpus, i.e. not present in any of the examined vocabularies, will be considered for inclusion in the ontology to be incorporated in the GEOEssential Knowledge Base platform.

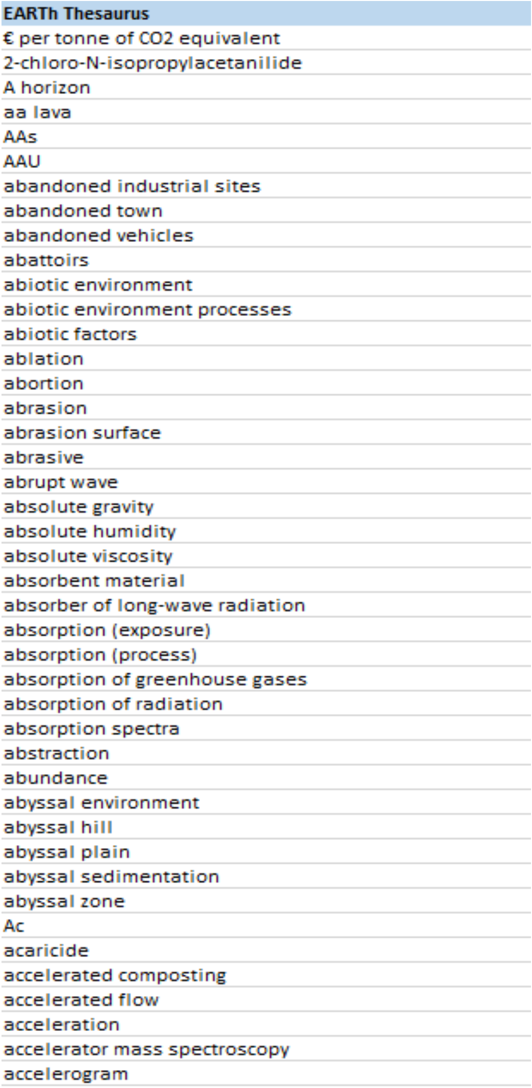
3.2 Evaluation of EO vocabularies semantic coverage

Alongside the corpus construction, relevant existing terminological resources related to the Environment were collected.

The reference thesauri employed to date have been the following: General Multilingual Environmental Thesaurus (**GEMET**); **EARTh** Thesaurus; **AGROVOC** Thesaurus; **INSPIRE** Feature Concept Dictionary and Glossary.

The abovementioned 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.

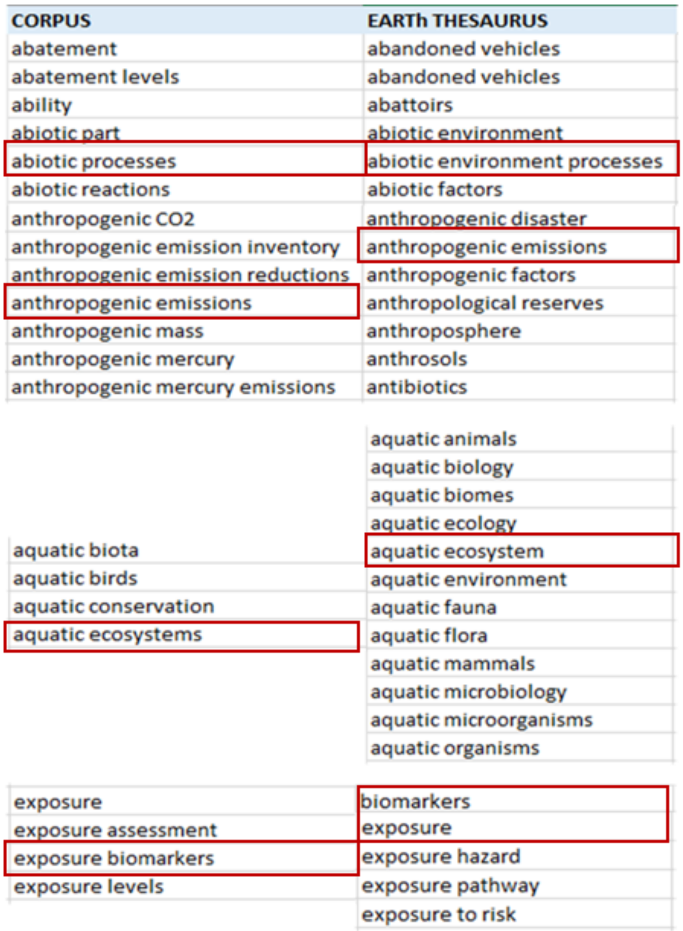
Around 14,000 terms have been extracted from the EARTh thesaurus (Table 3). Meta-terms (Accessory terms - Attributes - Dimensions - Dynamic aspects - Entities), along with Macro areas (Activities - Artificial entities - Biological entities - Complex systems - Composition - Conditions - Data - Equipment and technological systems - General terms - Immaterial entities - Living entities - Material Entities - Measures - Natural entities - Non-living entities - Processes - Properties - Social entities) have not been taken into consideration.



EARTh Thesaurus
€ per tonne of CO2 equivalent
2-chloro-N-isopropylacetanilide
A horizon
aa lava
AAs
AAU
abandoned industrial sites
abandoned town
abandoned vehicles
abattoirs
abiotic environment
abiotic environment processes
abiotic factors
ablation
abortion
abrasion
abrasion surface
abrasive
abrupt wave
absolute gravity
absolute humidity
absolute viscosity
absorbent material
absorber of long-wave radiation
absorption (exposure)
absorption (process)
absorption of greenhouse gases
absorption of radiation
absorption spectra
abstraction
abundance
abyssal environment
abyssal hill
abyssal plain
abyssal sedimentation
abyssal zone
Ac
acaricide
accelerated composting
accelerated flow
acceleration
accelerator mass spectroscopy
accelerogram

Table 3. Extract of EARTh term list

An initial comparison was carried out between the EARTh term list and our specialized term list in order to understand if the concepts relevant to our specific purposes are present in the existing resource.

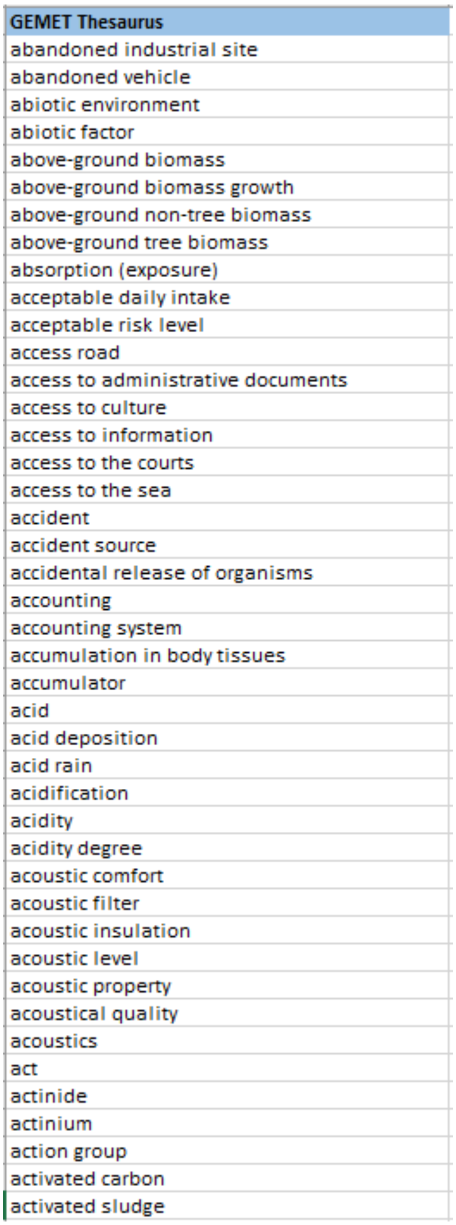


CORPUS	EARTh THESAURUS
abatement	abandoned vehicles
abatement levels	abandoned vehicles
ability	abattoirs
abiotic part	abiotic environment
abiotic processes	abiotic environment processes
abiotic reactions	abiotic factors
anthropogenic CO2	anthropogenic disaster
anthropogenic emission inventory	anthropogenic emissions
anthropogenic emission reductions	anthropogenic factors
anthropogenic emissions	anthropological reserves
anthropogenic mass	anthroposphere
anthropogenic mercury	anthrosols
anthropogenic mercury emissions	antibiotics
	aquatic animals
	aquatic biology
	aquatic biomes
	aquatic ecology
aquatic biota	aquatic ecosystem
aquatic birds	aquatic environment
aquatic conservation	aquatic fauna
aquatic ecosystems	aquatic flora
	aquatic mammals
	aquatic microbiology
	aquatic microorganisms
	aquatic organisms
exposure	biomarkers
exposure assessment	exposure
exposure biomarkers	exposure hazard
exposure levels	exposure pathway
	exposure to risk

Table 4. EARTh term list comparison

As illustrated in Table 4 above, different degrees of equivalence – which will be detailed in Section 3.3 – exist between concepts in both lists.

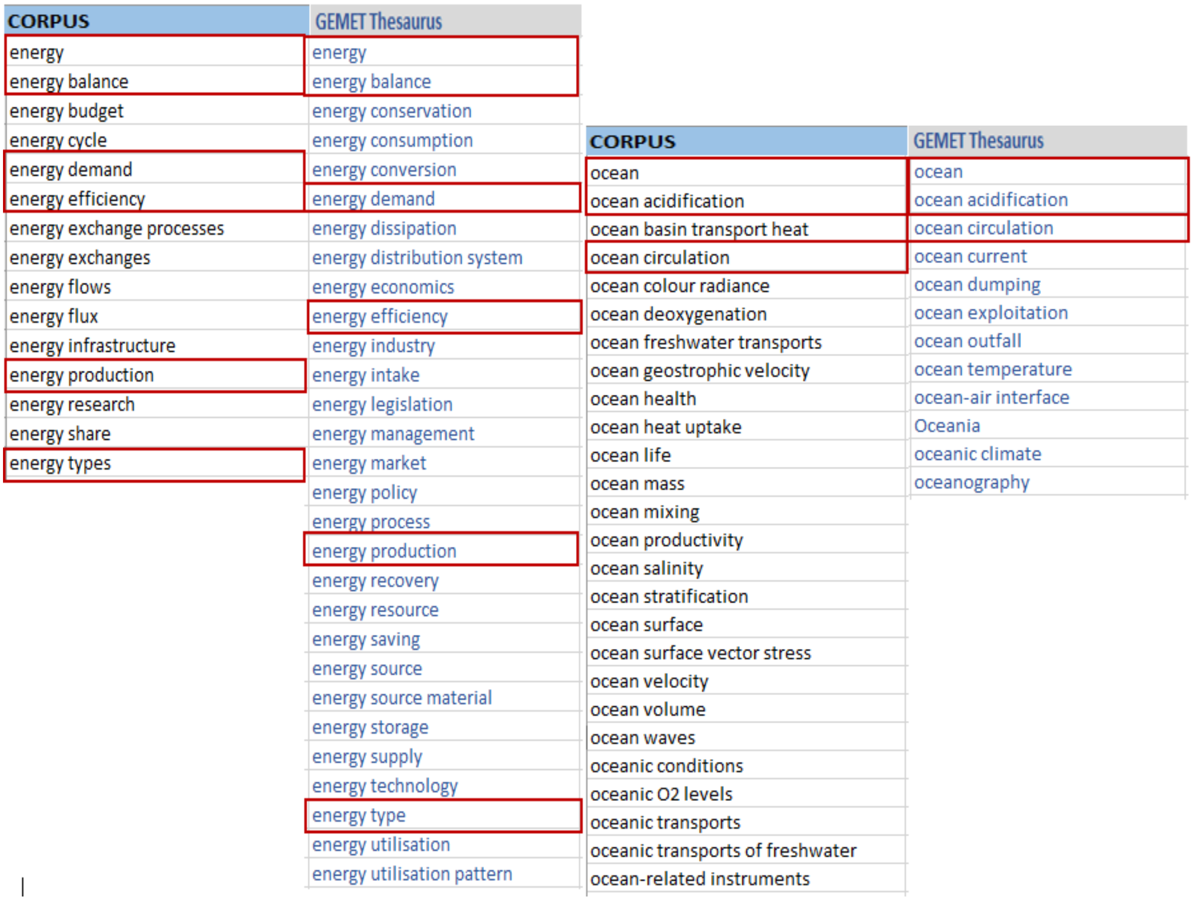
Over 5,550 terms have been extracted from the GEMET Thesaurus (Table 5).



GEMET Thesaurus
abandoned industrial site
abandoned vehicle
abiotic environment
abiotic factor
above-ground biomass
above-ground biomass growth
above-ground non-tree biomass
above-ground tree biomass
absorption (exposure)
acceptable daily intake
acceptable risk level
access road
access to administrative documents
access to culture
access to information
access to the courts
access to the sea
accident
accident source
accidental release of organisms
accounting
accounting system
accumulation in body tissues
accumulator
acid
acid deposition
acid rain
acidification
acidity
acidity degree
acoustic comfort
acoustic filter
acoustic insulation
acoustic level
acoustic property
acoustical quality
acoustics
act
actinide
actinium
action group
activated carbon
activated sludge

Table 5. Extract of GEMET term list

Table 6 shows once again how concepts can be mapped from one resource to another.

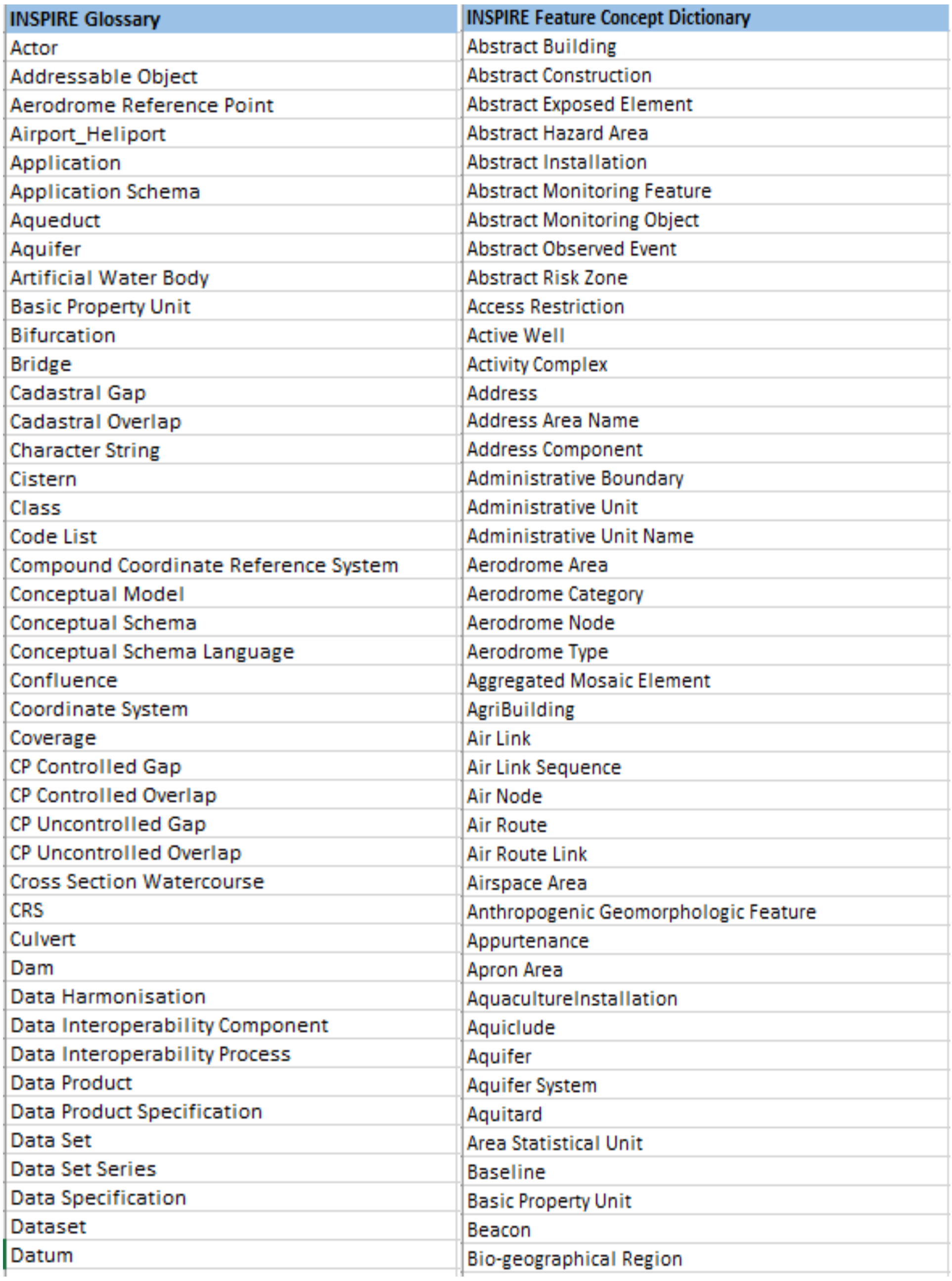


CORPUS	GEMET Thesaurus
energy	energy
energy balance	energy balance
energy budget	energy conservation
energy cycle	energy consumption
energy demand	energy conversion
energy efficiency	energy demand
energy exchange processes	energy dissipation
energy exchanges	energy distribution system
energy flows	energy economics
energy flux	energy efficiency
energy infrastructure	energy industry
energy production	energy intake
energy research	energy legislation
energy share	energy management
energy types	energy market
	energy policy
	energy process
	energy production
	energy recovery
	energy resource
	energy saving
	energy source
	energy source material
	energy storage
	energy supply
	energy technology
	energy type
	energy utilisation
	energy utilisation pattern

CORPUS	GEMET Thesaurus
ocean	ocean
ocean acidification	ocean acidification
ocean basin transport heat	ocean circulation
ocean circulation	ocean current
ocean colour radiance	ocean dumping
ocean deoxygenation	ocean exploitation
ocean freshwater transports	ocean outfall
ocean geostrophic velocity	ocean temperature
ocean health	ocean-air interface
ocean heat uptake	Oceania
ocean life	oceanic climate
ocean mass	oceanography
ocean mixing	
ocean productivity	
ocean salinity	
ocean stratification	
ocean surface	
ocean surface vector stress	
ocean velocity	
ocean volume	
ocean waves	
oceanic conditions	
oceanic O2 levels	
oceanic transports	
oceanic transports of freshwater	
ocean-related instruments	

Table 6. GEMET term list comparison

As previously mentioned, INSPIRE includes both a *Glossary* and a *Feature Concept Dictionary*. Almost 200 terms have been extracted from the former and 360 from the latter (Table 7).

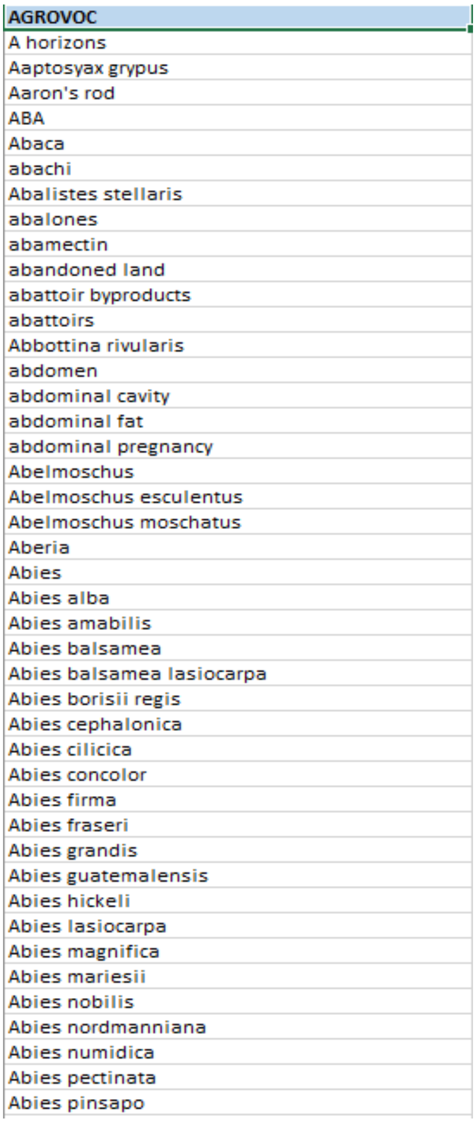


INSPIRE Glossary	INSPIRE Feature Concept Dictionary
Actor	Abstract Building
Addressable Object	Abstract Construction
Aerodrome Reference Point	Abstract Exposed Element
Airport_Heliport	Abstract Hazard Area
Application	Abstract Installation
Application Schema	Abstract Monitoring Feature
Aqueduct	Abstract Monitoring Object
Aquifer	Abstract Observed Event
Artificial Water Body	Abstract Risk Zone
Basic Property Unit	Access Restriction
Bifurcation	Active Well
Bridge	Activity Complex
Cadastral Gap	Address
Cadastral Overlap	Address Area Name
Character String	Address Component
Cistern	Administrative Boundary
Class	Administrative Unit
Code List	Administrative Unit Name
Compound Coordinate Reference System	Aerodrome Area
Conceptual Model	Aerodrome Category
Conceptual Schema	Aerodrome Node
Conceptual Schema Language	Aerodrome Type
Confluence	Aggregated Mosaic Element
Coordinate System	AgriBuilding
Coverage	Air Link
CP Controlled Gap	Air Link Sequence
CP Controlled Overlap	Air Node
CP Uncontrolled Gap	Air Route
CP Uncontrolled Overlap	Air Route Link
Cross Section Watercourse	Airspace Area
CRS	Anthropogenic Geomorphologic Feature
Culvert	Appurtenance
Dam	Apron Area
Data Harmonisation	AquacultureInstallation
Data Interoperability Component	Aquiclude
Data Interoperability Process	Aquifer
Data Product	Aquifer System
Data Product Specification	Aquitard
Data Set	Area Statistical Unit
Data Set Series	Baseline
Data Specification	Basic Property Unit
Dataset	Beacon
Datum	Bio-geographical Region

Table 7. Extract of INSPIRE term list

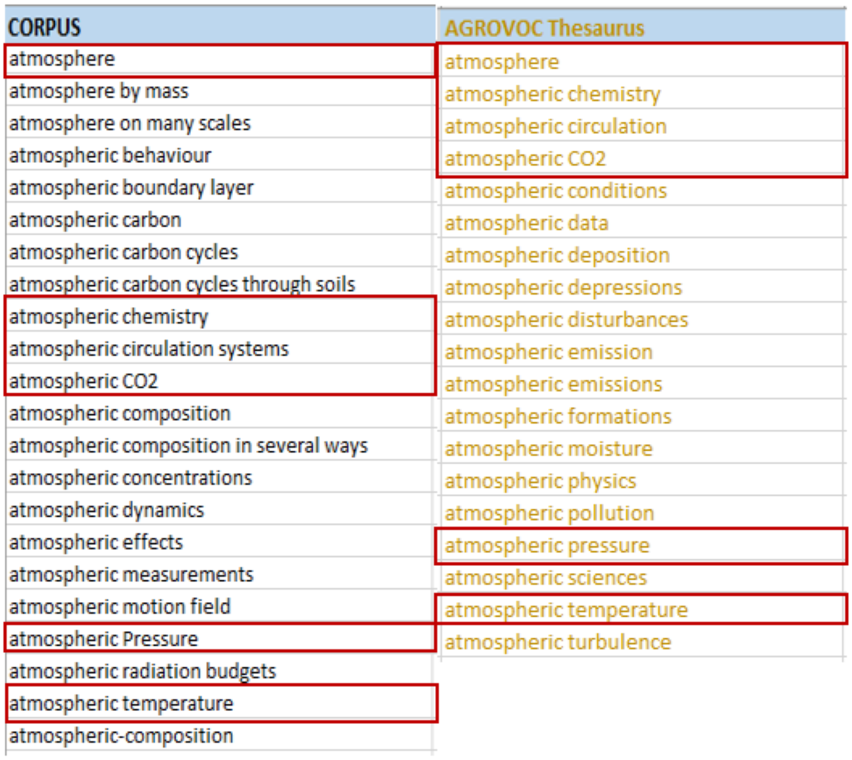
An initial comparison between the specific term list and *INSPIRE Glossary* and *Feature Concept Dictionary* terminology results in very few mappings. However, INSPIRE terminology does present potentially interesting terms to be included in the ontological model. A subsequent step will be the analysis of *INSPIRE Data Specification* and their *Technical Guidelines*.

The AGROVOC thesaurus includes approximately 45,500 terms (Table 8).



AGROVOC
A horizons
Aaptosyax grypus
Aaron's rod
ABA
Abaca
abachi
Abalistes stellaris
abalones
abamectin
abandoned land
abattoir byproducts
abattoirs
Abbottina rivularis
abdomen
abdominal cavity
abdominal fat
abdominal pregnancy
Abelmoschus
Abelmoschus esculentus
Abelmoschus moschatus
Aberia
Abies
Abies alba
Abies amabilis
Abies balsamea
Abies balsamea lasiocarpa
Abies borisii regis
Abies cephalonica
Abies cilicica
Abies concolor
Abies firma
Abies fraseri
Abies grandis
Abies guatemalensis
Abies hickeli
Abies lasiocarpa
Abies magnifica
Abies mariesii
Abies nobilis
Abies nordmanniana
Abies numidica
Abies pectinata
Abies pinsapo

Table 8. Extract of AGROVOC term list



CORPUS	AGROVOC Thesaurus
atmosphere	atmosphere
atmosphere by mass	atmospheric chemistry
atmosphere on many scales	atmospheric circulation
atmospheric behaviour	atmospheric CO2
atmospheric boundary layer	atmospheric conditions
atmospheric carbon	atmospheric data
atmospheric carbon cycles	atmospheric deposition
atmospheric carbon cycles through soils	atmospheric depressions
atmospheric chemistry	atmospheric disturbances
atmospheric circulation systems	atmospheric emission
atmospheric CO2	atmospheric emissions
atmospheric composition	atmospheric formations
atmospheric composition in several ways	atmospheric moisture
atmospheric concentrations	atmospheric physics
atmospheric dynamics	atmospheric pollution
atmospheric effects	atmospheric pressure
atmospheric measurements	atmospheric sciences
atmospheric motion field	atmospheric temperature
atmospheric Pressure	atmospheric turbulence
atmospheric radiation budgets	
atmospheric temperature	
atmospheric-composition	

Table 9. AGROVOC term list comparison

Table 9 illustrated possible alignments between our domain-specific term list and AGROVOC. However, what appears even more interesting is a list of missing terminology in AGROVOC, and not only, which is representative of the domain and will be integrated in the final common terminology.

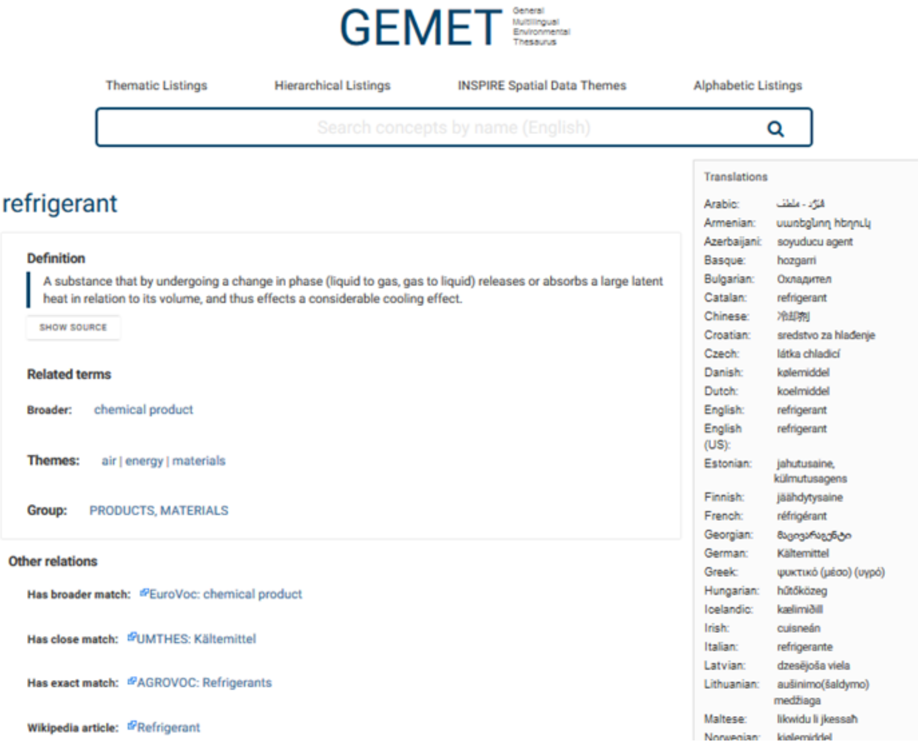
3.3 Thematic vocabularies alignment and integration

Considering their partial semantic overlapping, some vocabularies are already mapped to each other in order to allow federated access to information (Table 10).

abyssal environment	EARTH
abyssal environment	AGROVOC
abyssal zone	EARTH
abyssal zone	AGROVOC
acid deposition	EARTH
acid deposition	GEMET
acid deposition	AGROVOC
aquifer	GEMET
Aquifer	INSPIRE

Table 10. Existing vocabularies term list matching

For instance, as can be seen in Figure 3 below, GEMET is aligned with other existing vocabularies such as the AGROVOC, EUROVOC and UMTHES thesauri. The figure includes examples of different matching types: ‘refrigerant’ has an exact match with AGROVOC:Refrigerants, since the two concepts have a fully equivalent meaning. Furthermore, it has a close match with UMTHES:Kältemittel, in other words, their meaning is partially equivalent, and a broader match with EUROVOC:chemical product, which specifies that ‘refrigerant’ is a more specific concept.


GEMET
General Multilingual Environmental Thesaurus
Thematic Listings
Hierarchical Listings
INSPIRE Spatial Data Themes
Alphabetic Listings
Search concepts by name (English)
SHOW SOURCE

refrigerant

Definition

A substance that by undergoing a change in phase (liquid to gas, gas to liquid) releases or absorbs a large latent heat in relation to its volume, and thus effects a considerable cooling effect.

Related terms

Broader: chemical product

Themes: air | energy | materials

Group: PRODUCTS, MATERIALS

Other relations

Has broader match: EuroVoc: chemical product

Has close match: UMTHES: Kältemittel

Has exact match: AGROVOC: Refrigerants

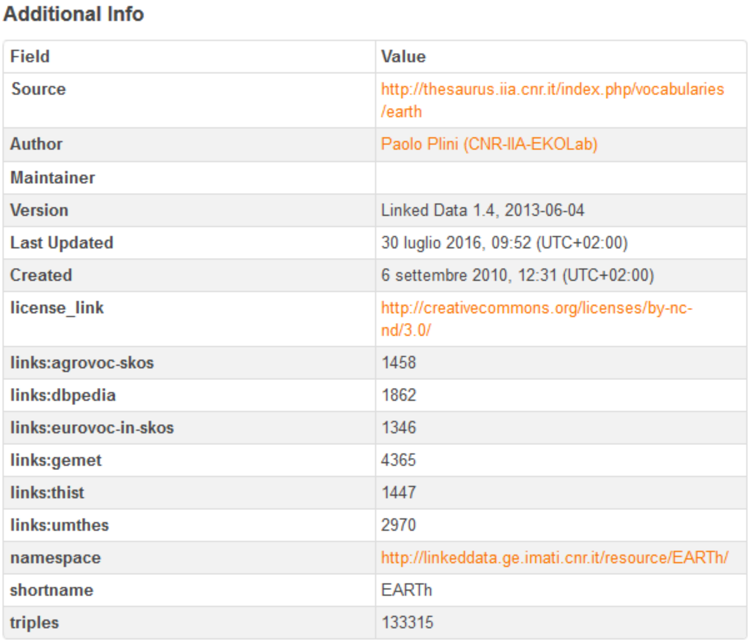
Wikipedia article: Refrigerant

Translations

Arabic:	مبرّد - ملطف
Armenian:	սառեցնող հեղուկ
Azerbaijani:	soyuducu agent
Basque:	hozgarri
Bulgarian:	Охладител
Catalan:	refrigerant
Chinese:	冷却剂
Croatian:	sredstvo za hlađenje
Czech:	látka chladicí
Danish:	kølemiddel
Dutch:	koelmiddel
English:	refrigerant
English (US):	refrigerant
Estonian:	jahutusaine, külmutusagens
Finnish:	jäähdytysaine
French:	réfrigérant
Georgian:	მაცივარაგენტი
German:	Kältemittel
Greek:	ψυκτικό (μέσο) (υγρό)
Hungarian:	hűtőközeg
Icelandic:	kælimiðill
Irish:	cuisneán
Italian:	refrigerante
Latvian:	dzesējoša viela
Lithuanian:	aušinimo(šaldymo) medžiaga
Maltese:	likwidu li jkessaħ
Norwegian:	kjølemiddel

Figure 3. Thesauri matchings

In the same way, the EARTh Thesauri is aligned with other vocabularies as illustrated in Figure 4 below.



Additional Info

Field	Value
Source	http://thesaurus.iia.cnr.it/index.php/vocabularies/earth
Author	Paolo Plini (CNR-IIA-EKOLab)
Maintainer	
Version	Linked Data 1.4, 2013-06-04
Last Updated	30 luglio 2016, 09:52 (UTC+02:00)
Created	6 settembre 2010, 12:31 (UTC+02:00)
license_link	http://creativecommons.org/licenses/by-nc-nd/3.0/
links:agrovoc-skos	1458
links:dbpedia	1862
links:eurovoc-in-skos	1346
links:gemet	4365
links:thist	1447
links:umthes	2970
namespace	http://linkeddata.ge.imati.cnr.it/resource/EARTh/
shortname	EARTh
triples	133315

Figure 4. EARTh matchings

Thanks to thesauri alignment, if, for example, a term in the EARTh thesaurus is linked with a term in the GEMET thesaurus, all documents indexed by the same term in the document repositories related to EARTh and GEMET are also potentially linked.

The development of new vocabularies or the integration of new concepts and terms in the existing ones will depend on how well the GEOEssential domain of interest is covered - from a semantic perspective - by the available terminological resources. Further correspondences will be established starting from the terms identified in the corpus, and terms not present in the existing terminologies but considered relevant for the topics covered in the project will be integrated with the objective of defining a common and shared terminology which will be formalized in SKOS language.

SKOS offers a model for semi-formally representing the structure of different kinds of KOSs (thesauri, classification schemes, taxonomies, and so on). It is based on Resource Description Framework (RDF)[20] and it allows to publish vocabularies on the World Wide Web, to link concepts with other data on the Web and to integrate them with other concept schemes. “In basic SKOS, conceptual resources (concepts) can be identified with URIs, labelled with lexical strings in one or more natural languages, documented with various types of note, semantically related to each other in informal hierarchies and association networks and aggregated into concept schemes”[21].

The elements which can be modelled in SKOS language are the following: concepts; labels (Preferred Lexical Labels, Alternative Lexical Labels, Hidden Lexical Labels); semantic relationships (Broader/Narrower and Associative); documentary notes and concept schemes. Moreover, it allows to establish interlinks between two or more concept schemes by connecting concepts coming from each one of them. This possibility of creating a network

[20] https://www.w3.org/RDF/

[21] https://www.w3.org/TR/skos-primer/#secmapping

between existing and/or new resources is fundamental for our purposes, in particular for guaranteeing information retrieval processes based on several KOSs semantically related.

The mapping process consists in stating that two concepts coming from different vocabularies have a comparable meaning and in specifying the level of comparability. Different properties can be used to explicit mappings: skos:exactMatch, skos:closeMatch, skos:broadMatch, skos:narrowMatch and skos:relatedMatch. Semantic relationships that can be defined between concepts belonging to different vocabularies are the same of those that can be defined between terms and concepts within a vocabulary: skos:exactMatch/ skos:closeMatch for equivalence mappings; skos:broadMatch/skos:narrowMatch for hierarchical relationships; skos:relatedMatch for associative relationships.

Table 11 illustrates some examples of matching identified between concepts coming from the preliminary corpus we collected and concepts coming from EARTh:

- *anthropogenic emissions*: the meaning is exactly equivalent in the two term lists. This kind of mapping is transitive and symmetric;
- *abiotic processes* and *abiotic environment processes*: the two concepts are partially equivalent but they can be used interchangeably;
- *acidification* and *soil acidification*: the latter is more specific than the former;
- *exposure biomarkers* is a complex concept given by the combination of *exposure* and *biomarkers*. It is absent in EARTh so it is mapped to the union of the two single terms. In this specific example the mapping property comes from the iso-thes schema, a SKOS extension based on the ISO norm 25964-2:2013[22];
- *abiotic reactions* does not have any equivalent concepts in EARTh. If not found in the other explored terminologies, it will be added in the final common terminology.

Corpus	Mapping type	EARTh
abiotic processes	skos:closeMatch	abiotic environment processes
abiotic reactions	/	*abiotic reactions (to be added)*
acidification	skos:narrowerMatch	soil acidification
anthropogenic emissions	skos:exactMatch	anthropogenic emissions
exposure biomarkers	iso-thes:CompoundEquivalence	exposure + biomarkers

Table 11. Alignment and integration

3.4 Design of the ontological conceptual model

Ontologies, as shown in Figure 5, belong to the category of the so-called Knowledge Organization Systems (KOSs), items that “have been designed to support the organization of knowledge and information in order to make their management and retrieval easier”[23]. Because of their high level of structuring and formalism in representing knowledge, ontologies are both human and machine-readable and they are therefore considered as the

[22] SKOS does not have native elements for representing compound equivalences, groups and arrays of concepts and for specifying the kind of semantic relations (e.g. hierarchical relationship can be generic, partitive and instantiative). ISO-THES reuses SKOS and SKOS-XL and defines properties and classes able to represent thesauri according to the data model provided by the ISO 25964 standard.

[23] http://www.isko.org/cyclo/kos

main component of the Semantic Web and of several other application contexts (e.g. e-commerce, problem solving, data integration, etc.) that require a common sharing and understanding of information, the reuse of the modelled knowledge and the advanced capability of reasoning and making assumptions by interpreting the information explicitly formalized in the ontology itself.

The main concepts of the domain knowledge are represented through *classes*, which can be further subdivided into *sub-classes* at different hierarchical levels[24]. The resulting taxonomy is therefore based on the establishment of *is-a* and *kind-of* relationships. Other types of relationships between classes are explicitly expressed by means of binary *object properties*, which are defined between a source class (*domain*) and a target one (*range*) and can be organized in a taxonomy, while attributes are associated to classes by *data properties*. In order for an ontology to become a Knowledge Base, a set of *individuals* should be added: they represent specific instances of classes and subclasses and they inherit all the properties established at the class level. Inheritance is also valid when creating subclasses: they inherit properties defined for their superclasses, therefore it is not necessary to create properties for each one of them.

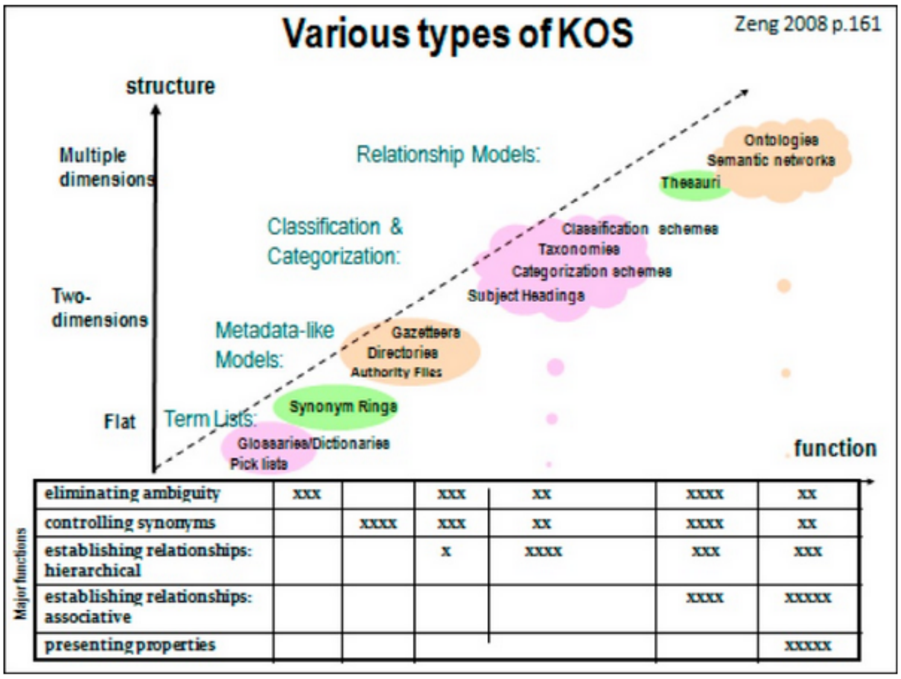
It is also possible to explicitly model further information about classes, properties and individuals: two (or more) classes should be declared as disjoint when the individuals of a class cannot belong at the same time to another class (if not explicitly stated, the same individuals can be associated to both classes); object properties can be inverse, transitive, functional, symmetric, reflexive, and so on. Moreover, some restrictions can be defined about classes, in particular about the individuals that can belong to a specific class. Existential and universal restrictions are quantitative: the former "defines a class as the set of all individuals that are connected via a particular property to another individual which is an instance of a certain class" and is represented by the symbol "∃", the latter "is used to describe a class of individuals for which all related individuals must be instances of a given class"[25] and is represented by "∀".

The standard formal language used to express ontologies is the Ontology Web Language (OWL), developed by the World Wide Web Consortium (W3C)[26]. It is mainly based on RDF and has foundations in Description Logics, which allows programs called reasoners to check if the ontology is consistent.

[24] For a description of ontology structure and construction see (Noy et al., 2001; Capuano 2005).

[25] https://www.w3.org/TR/owl2-primer/

[26] https://www.w3.org/OWL/



eliminating ambiguity	xxx		xxx	xx	xxxx	xx
controlling synonyms		xxxx	xxx	xx	xxxx	xx
establishing relationships: hierarchical			x	xxxx	xxx	xxx
establishing relationships: associative					xxxx	xxxxx
presenting properties						xxxxx

Figure 5. Various types of KOSs (Zeng 2008, p. 161)

The ongoing definition of the ontological conceptual model refers to the description provided by *Deliverable 1.1. Knowledge Services Architecture*[27] and is based on an Entity-Relationship database schema and on the *GEOEssential Reference Table* describing almost the same concepts and prepared by the University of Geneva (*Deliverable 6.1. Description of Food-Water-Energy EVs*). The latter in order to harmonize work related to the modelling of domain concepts. The existing ontologies and semantic resources, described in the previous sections, are also being taken into consideration. The model will be refined based on the *Deliverable 2.2. EVs list*.

In the ontology the more abstract concepts related to our specific domain are *classes* while the more specific ones have been introduced as *subclasses*. The outcome is a taxonomy structured on different hierarchical levels, where each narrower concept is completely included in the broader concept and provides deeper information regarding the level immediately above. The main general classes and subclasses defined in the ontology are the following:

- *Algorithm*;
- *Anatomical entity;*
- *Area;*
- *Dataset;*
- *Ecosystem;*
- *Essential Variable* (physical parameter which is necessary to describe the Earth system status)[28];
- *Essential Variable Category;*

[27] Mazzetti P., Santoro M., Nativi S., *Knowledge services architecture*. GEOEssential Deliverable 1.1.

[28] Concept definitions have been taken from the *Deliverable 1.1 Knowledge services architecture*.

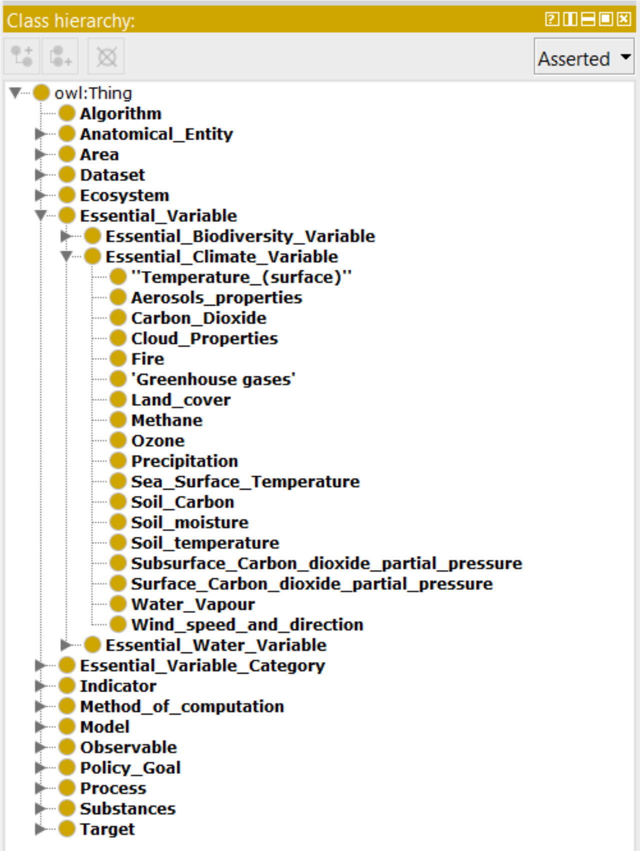
- *Indicator* (a derived parameter summarizing the status of the system under consideration. (Note that we use the term Indicator in a broad sense to include both an Indicator in strict sense – a physical parameter indicating the status of a system for decision-making purposes – and an Index – a figure summarizing multiple parameters to represent the status of a system for decision-making purposes);
- *Method of computation*;
- *Model*;
- *Observable* (a physical parameter which can be directly observed with proper instruments);
- *Policy Goal* (the desired outcome or what is to be achieved by implementing a policy);
- *Process;*
- *Substance;*
- *Target.*

From the analysis of the classes it can be deduced that the model is not exclusively GEOEssential centred: it considers concepts concerning other subdomains or concepts (e.g. Substance, Anatomical entity, etc.) which are specific of two other projects included in the ERA-PLANET program, SMURBS and iGOSP.

The choice to develop a single general model and to specialize it by integrating a conceptualization of information regarding each project has been made in agreement with the other partners involved in the development of the KB and is oriented towards guaranteeing interoperability between one project and the others.

This will allow to avoid the use of different models in the KP and to facilitate the retrieval of data and information from different sources.

Figure 6 represents the taxonomy under the class Essential Variable, illustrating in particular the Essential Climate Variables.



```
owl:Thing
    Algorithm
    Anatomical_Entity
    Area
    Dataset
    Ecosystem
    Essential_Variable
        Essential_Biodiversity_Variable
        Essential_Climate_Variable
            "Temperature_(surface)"
            Aerosols_properties
            Carbon_Dioxide
            Cloud_Properties
            Fire
            'Greenhouse gases'
            Land_cover
            Methane
            Ozone
            Precipitation
            Sea_Surface_Temperature
            Soil_Carbon
            Soil_moisture
            Soil_temperature
            Subsurface_Carbon_dioxide_partial_pressure
            Surface_Carbon_dioxide_partial_pressure
            Water_Vapour
            Wind_speed_and_direction
        Essential_Water_Variable
    Essential_Variable_Category
    Indicator
    Method_of_computation
    Model
    Observable
    Policy_Goal
    Process
    Substances
    Target
```

Figure 6. Extract of ontology taxonomy

Contrary to what was initially assumed, classes are not populated by *Individuals* (instances of classes) but they are enriched by subclasses which specialize the super classes. In this way each subclass is at the same time member of the superclass and root of another subclass. The result is an ontology consisting of a set of logically connected classes and subclasses and a list of *Properties* concerning the type of relationship between two or more classes.

As already mentioned classes and subclasses have been related to each other through *ObjectProperties* defined according to the type of relationship it is useful to explicit. Illustrated below are the *ObjectProperties* that have been defined:

- affects (domain: Substance; range Anatomical_Entity);
- belongsTo (domain: Observable and Essential Variable; range: Essential Variable Category);
- computes (domain: Method_of_Computation; range: Indicator);
- examines (domain: Indicator_Generation_Model and EV_Generation_Model; range: Observable and Essential_Variable);
- generates (domain: EV_Generation_Model and Indicator_Generation_Model; range: Essential_Variable and Indicator);
- hasIndex (domain: Land_degradation; range: Observable);
- hasTarget (domain: Policy; range: Target);
- isTargetOf (inverse of hasTarget);

- isMeasuredBy (inverse of measures);
- isAffectedBy (inverse of affects);
- isComputedBy (inverse of computes);
- isExaminedBy (inverse of examines);
- isGeneretedBy (inverse of generates);
- isIndexOf (domain: Substances; range: Process);
- isRelatedTo (domain: Indicator; range: Indicator);
- isUsedBy (inverse of uses);
- measures (domain: Indicator and Sensor; range: Target and Observable)
- uses (domain: Model and Indicator; range: Dataset, Essential_Variable, Observable and Indicator).

The OWL ontology can be interactively navigated through the OntoGraph plug-in provided by Protégé (Figure 7)[29]. The graph approach permits to display the ontology as a set of nodes (classes and subclasses) and direction lines (Properties) which can express direct or inverse relationships. Each pair of Class and Properties express a Statement in the form of subject - predicate - object expressions. For example: "*Indicator 15.3.1 measures Target 15.1" and vice versa "Target 15.1 isMeasuredBy Indicator 15.3.1*". In this way a single statement can be made explicit and interlinked to other statements in order to create a rich and interconnected structure which unambiguously represents the conceptualization of the domain with regard to the specific tasks the ontology should accomplish.

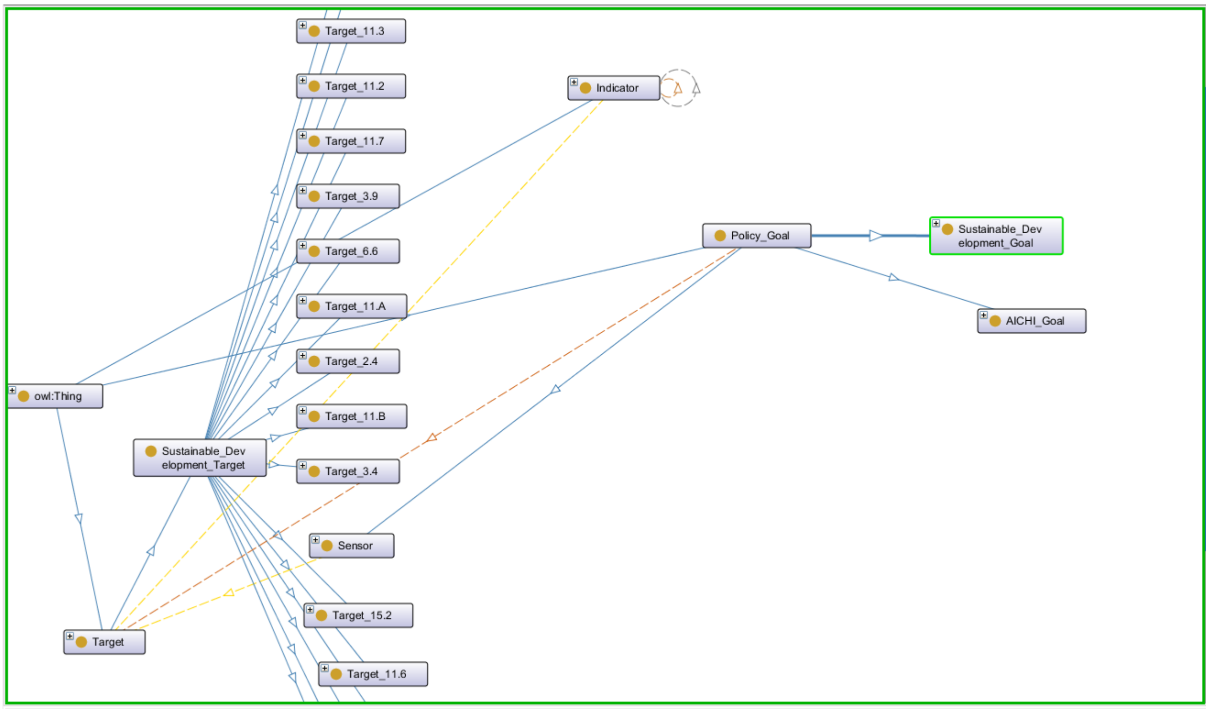


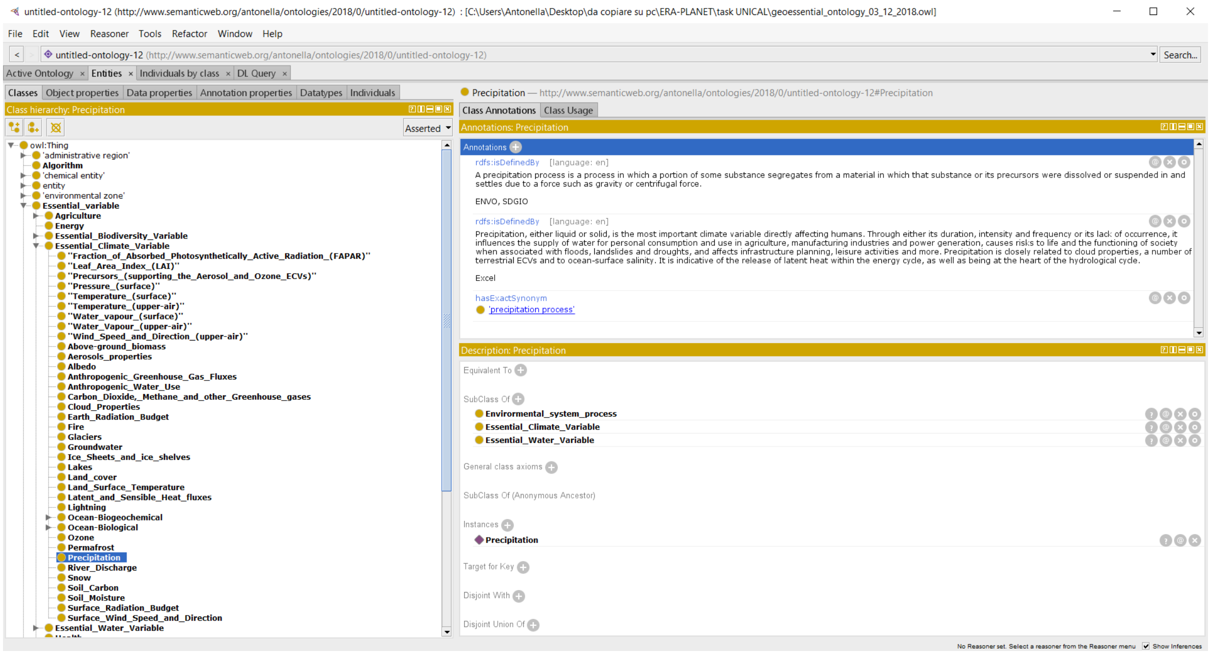
Figure 7. OntoGraph

The ongoing construction of the ontology schema and taxonomy is being enriched by the inclusion - as classes, subclasses or individuals - of concepts represented by the terms

[29] https://protege.stanford.edu/

obtained in the previous phases: both those taken from existing terminologies and those taken from the corpus-based terminology extraction.
For instance, the following concepts “acid precipitation”, “artificial precipitation”, “atmospheric precipitation”, “chemical precipitation”, “cyclonic precipitation”, extracted from the analysed existing terminologies, could all be related to the main concept “Precipitation”, already included in the ontology taxonomy.

Moreover, we are currently evaluating the potential reuse or integration of part of the SDG Interface Ontology for our specific aims. Figure 8 illustrates an example of an ontology import: the left-hand side shows the taxonomy that integrates both the abovementioned classes and those coming from SDGIO; the right-hand side illustrates the mapping (hasExactSynonym) between our concept “Precipitation” and the SDGIO concept “Precipitation process”. Nevertheless, compared to our ontology, SDGIO has a more general structure and a broader conceptual coverage, thus its adoption should be evaluated in respect to the specific aims of GEOEssential.


untitled-ontology-12 (http://www.semanticweb.org/antonella/ontologies/2018/0/untitled-ontology-12) : [C:\Users\Antonella\Desktop\da copiare su pc\ERA-PLANET\task UNICAL\geoessential_ontology_03_12_2018.owl]
File
Edit
View
Reasoner
Tools
Refactor
Window
Help
untitled-ontology-12 (http://www.semanticweb.org/antonella/ontologies/2018/0/untitled-ontology-12)
Search...
Active Ontology
Entities
Individuals by class
DL Query
Classes
Object properties
Data properties
Annotation properties
Datatypes
Individuals
Class hierarchy: Precipitation
Asserted
Precipitation — http://www.semanticweb.org/antonella/ontologies/2018/0/untitled-ontology-12#Precipitation
Class Annotations
Class Usage
Annotations: Precipitation
Annotations
Description: Precipitation
Equivalent To
SubClass Of
General class axioms
SubClass Of (Anonymous Ancestor)
Instances
Target for Key
Disjoint With
Disjoint Union Of
No Reasoner set. Select a reasoner from the Reasoner menu
☑ Show Inferences

```
owl:Thing
  'administrative region'
  Algorithm
  'chemical entity'
  entity
  'environmental zone'
  Essential_variable
    Agriculture
    Energy
    Essential_Biodiversity_Variable
    Essential_Climate_Variable
      "Fraction_of_Absorbed_Photosynthetically_Active_Radiation_(FAPAR)"
      "Leaf_Area_Index_(LAI)"
      "Precursors_(supporting_the_Aerosol_and_Ozone_ECVs)"
      "Pressure_(surface)"
      "Temperature_(surface)"
      "Temperature_(upper-air)"
      "Water_vapour_(surface)"
      "Water_Vapour_(upper-air)"
      "Wind_Speed_and_Direction_(upper-air)"
      Above-ground_biomass
      Aerosols_properties
      Albedo
      Anthropogenic_Greenhouse_Gas_Fluxes
      Anthropogenic_Water_Use
      Carbon_Dioxide,_Methane_and_other_Greenhouse_gases
      Cloud_Properties
      Earth_Radiation_Budget
      Fire
      Glaciers
      Groundwater
      Ice_Sheets_and_ice_shelves
      Lakes
      Land_cover
      Land_Surface_Temperature
      Latent_and_Sensible_Heat_fluxes
      Lightning
      Ocean-Biogeochemical
      Ocean-Biological
      Ozone
      Permafrost
      Precipitation
      River_Discharge
      Snow
      Soil_Carbon
      Soil_Moisture
      Surface_Radiation_Budget
      Surface_Wind_Speed_and_Direction
    Essential_Water_Variable
```

rdfs:isDefinedBy [language: en]
A precipitation process is a process in which a portion of some substance segregates from a material in which that substance or its precursors were dissolved or suspended in and settles due to a force such as gravity or centrifugal force.

ENVO, SDGIO

rdfs:isDefinedBy [language: en]
Precipitation, either liquid or solid, is the most important climate variable directly affecting humans. Through either its duration, intensity and frequency or its lack of occurrence, it influences the supply of water for personal consumption and use in agriculture, manufacturing industries and power generation, causes risks to life and the functioning of society when associated with floods, landslides and droughts, and affects infrastructure planning, leisure activities and more. Precipitation is closely related to cloud properties, a number of terrestrial ECVs and to ocean-surface salinity. It is indicative of the release of latent heat within the energy cycle, as well as being at the heart of the hydrological cycle.

Excel

hasExactSynonym
'precipitation process'

SubClass Of
Enviromental_system_process
Essential_Climate_Variable
Essential_Water_Variable

Instances
Precipitation

Figure 8. Example of SDGIO Import

3.5 Use case: Indicator 15.3.1

In order to test the consistency of the ontology as compared to the GEOEssential objectives and to verify if the specific characteristics of its domain of interest are well represented, it has been necessary to identify a representative case study intended as an investigation of a real phenomenon that often occurs in the domain. In accordance with CNR-IIA and the University of Geneva, Indicator 15.3.1 “Proportion of land that is degraded over total land area” has been chosen, so the ontology has been tailored to the modelling of the information related to it.

This indicator represents a case study also for other experimentations within the project, therefore we have been able to collect a great deal of information about it and in the coming future it will be possible to test the ontological model in the VLab running workflows representative of the Land degradation domain. In order to specialize the general structure of the ontology so that it could represent this specific subject, we analysed different authoritative documents, mainly taken from the United Nations website[30].

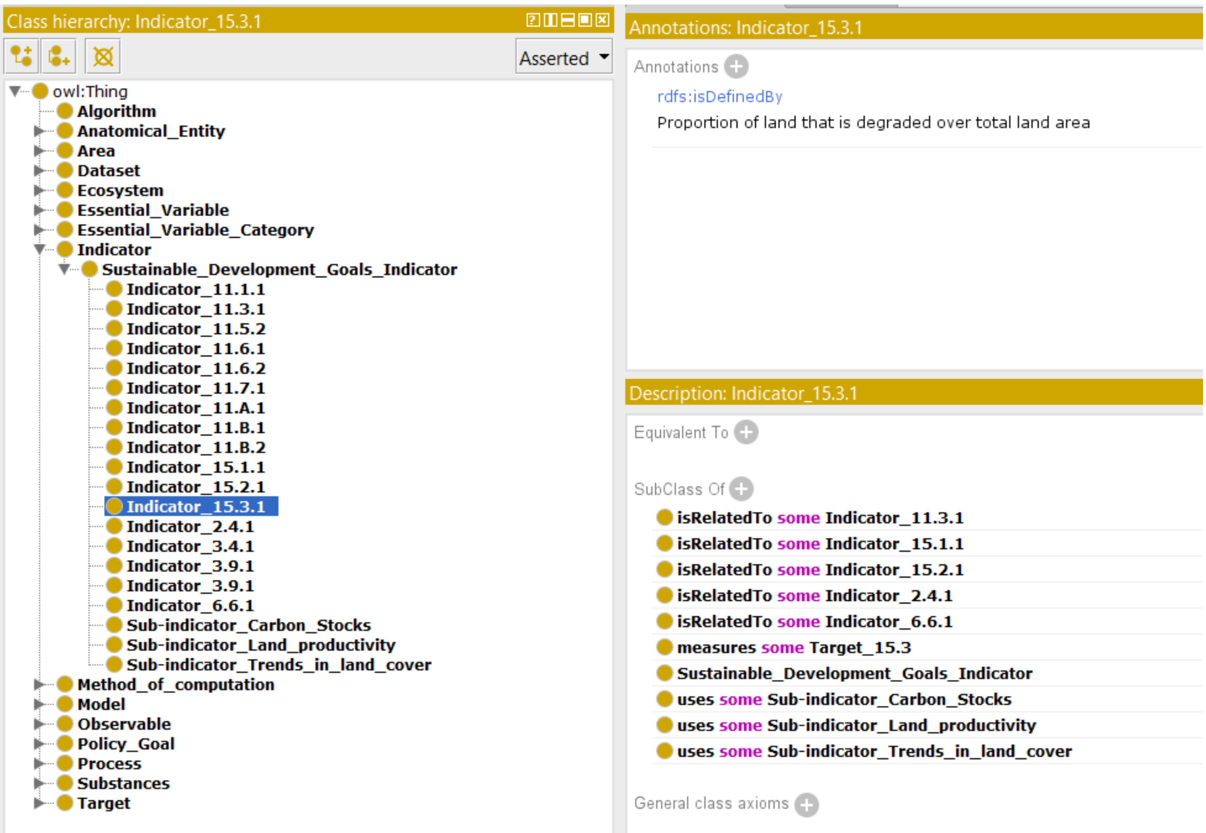
Knowledge about Indicator 15.3.1 has been modelled in the ontology thanks to the support of the abovementioned partners, who provide us with specific notions, strictly related to the domain knowledge. Nevertheless, the correctness and the completeness of the model are not fully guaranteed at the moment, as further validation is being carried out by CNR-IIA. The involvement of experts, both from a technical and from a domain point of view, is mandatory for the development of such a system.

The following images, extracted from the Protégé tool, show the information currently available in the model about Indicator 15.3.1:

- it has been linked to different indicators, some of which belong to other SDGs (e.g. Indicator 15.3.1 isRelatedTo Indicator 6.6.1 “Change in the extent of water-related ecosystems over time”);
- the corresponding Target has been specified (Indicator 15.3.1 measures Target 15.3);
- the related sub-indicators have been listed[31];

[30] https://sustainabledevelopment.un.org/sdg15.

[31] “SDG indicator 15.3.1 is a binary - degraded/not degraded - quantification based on the analysis of available data for three sub-indicators to be validated and reported by national authorities. The sub-indicators (Trends in Land Cover, Land Productivity and Carbon Stocks) were adopted by the UNCCD’s governing body in 2013 as part of its monitoring and evaluation approach”, Metadata 15-03-01, https://sustainabledevelopment.un.org/sdg15.

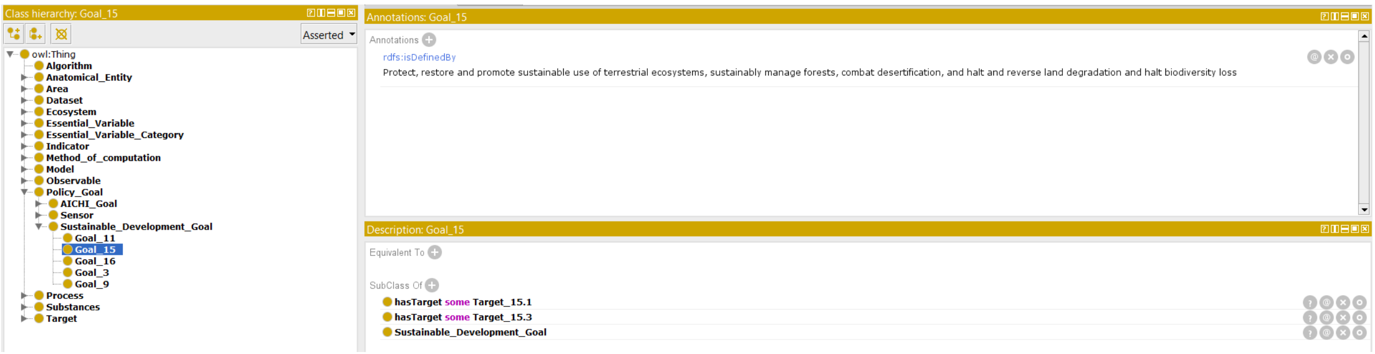


```
Class hierarchy: Indicator_15.3.1
Asserted
owl:Thing
    Algorithm
    Anatomical_Entity
    Area
    Dataset
    Ecosystem
    Essential_Variable
    Essential_Variable_Category
    Indicator
        Sustainable_Development_Goals_Indicator
            Indicator_11.1.1
            Indicator_11.3.1
            Indicator_11.5.2
            Indicator_11.6.1
            Indicator_11.6.2
            Indicator_11.7.1
            Indicator_11.A.1
            Indicator_11.B.1
            Indicator_11.B.2
            Indicator_15.1.1
            Indicator_15.2.1
            Indicator_15.3.1
            Indicator_2.4.1
            Indicator_3.4.1
            Indicator_3.9.1
            Indicator_3.9.1
            Indicator_6.6.1
            Sub-indicator_Carbon_Stocks
            Sub-indicator_Land_productivity
            Sub-indicator_Trends_in_land_cover
    Method_of_computation
    Model
    Observable
    Policy_Goal
    Process
    Substances
    Target
```

```
Annotations: Indicator_15.3.1
Annotations
    rdfs:isDefinedBy
    Proportion of land that is degraded over total land area

Description: Indicator_15.3.1
Equivalent To

SubClass Of
    isRelatedTo some Indicator_11.3.1
    isRelatedTo some Indicator_15.1.1
    isRelatedTo some Indicator_15.2.1
    isRelatedTo some Indicator_2.4.1
    isRelatedTo some Indicator_6.6.1
    measures some Target_15.3
    Sustainable_Development_Goals_Indicator
    uses some Sub-indicator_Carbon_Stocks
    uses some Sub-indicator_Land_productivity
    uses some Sub-indicator_Trends_in_land_cover

General class axioms
```

Figure 9. Indicator 15.3.2 relationships

- the relationship between Indicator, Target and Goal has been formalized;

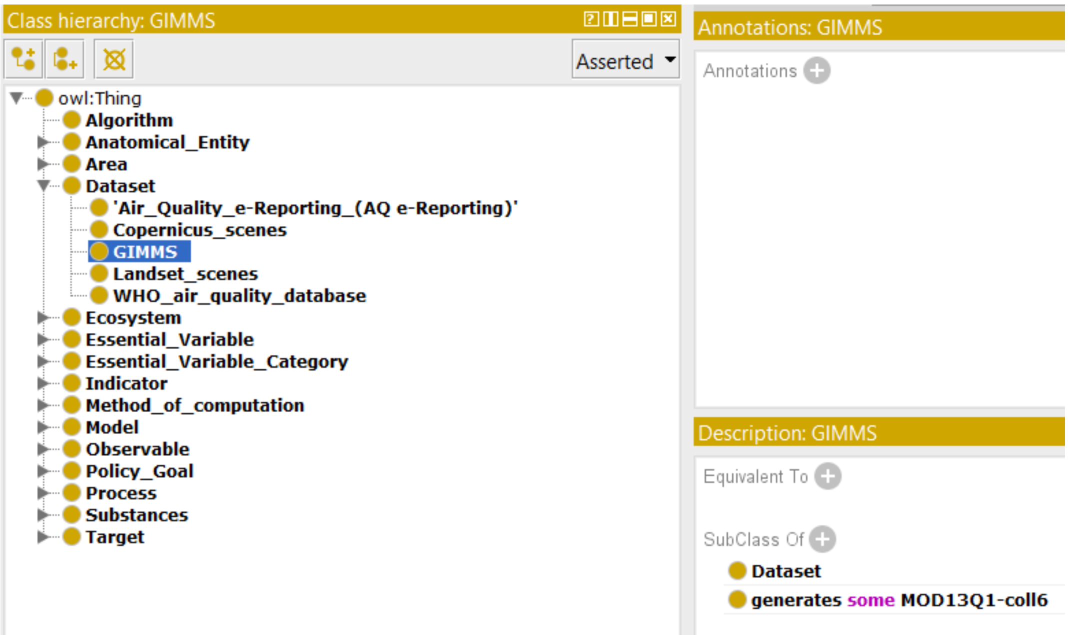


```
Class hierarchy: Goal_15
Asserted
owl:Thing
    Algorithm
    Anatomical_Entity
    Area
    Dataset
    Ecosystem
    Essential_Variable
    Essential_Variable_Category
    Indicator
    Method_of_computation
    Model
    Observable
    Policy_Goal
        AICHI_Goal
        Sensor
        Sustainable_Development_Goal
            Goal_11
            Goal_15
            Goal_16
            Goal_3
            Goal_9
    Process
    Substances
    Target
```

```
Annotations: Goal_15
Annotations
    rdfs:isDefinedBy
    Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss

Description: Goal_15
Equivalent To

SubClass Of
    hasTarget some Target_15.1
    hasTarget some Target_15.3
    Sustainable_Development_Goal
```

Figure 10. SDG

- the datasets providing the data useful for the computation of the Indicator have been listed and linked to the model they are able to generate (e.g. *GIMMS generates MOD13Q1*, which is an EV generation model);



Class hierarchy: GIMMS

Asserted

- owl:Thing
  - Algorithm
  - Anatomical_Entity
  - Area
  - Dataset
    - 'Air_Quality_e-Reporting_(AQ e-Reporting)'
    - Copernicus_scenes
    - GIMMS
    - Landset_scenes
    - WHO_air_quality_database
  - Ecosystem
  - Essential_Variable
  - Essential_Variable_Category
  - Indicator
  - Method_of_computation
  - Model
  - Observable
  - Policy_Goal
  - Process
  - Substances
  - Target

Annotations: GIMMS

Annotations

Description: GIMMS

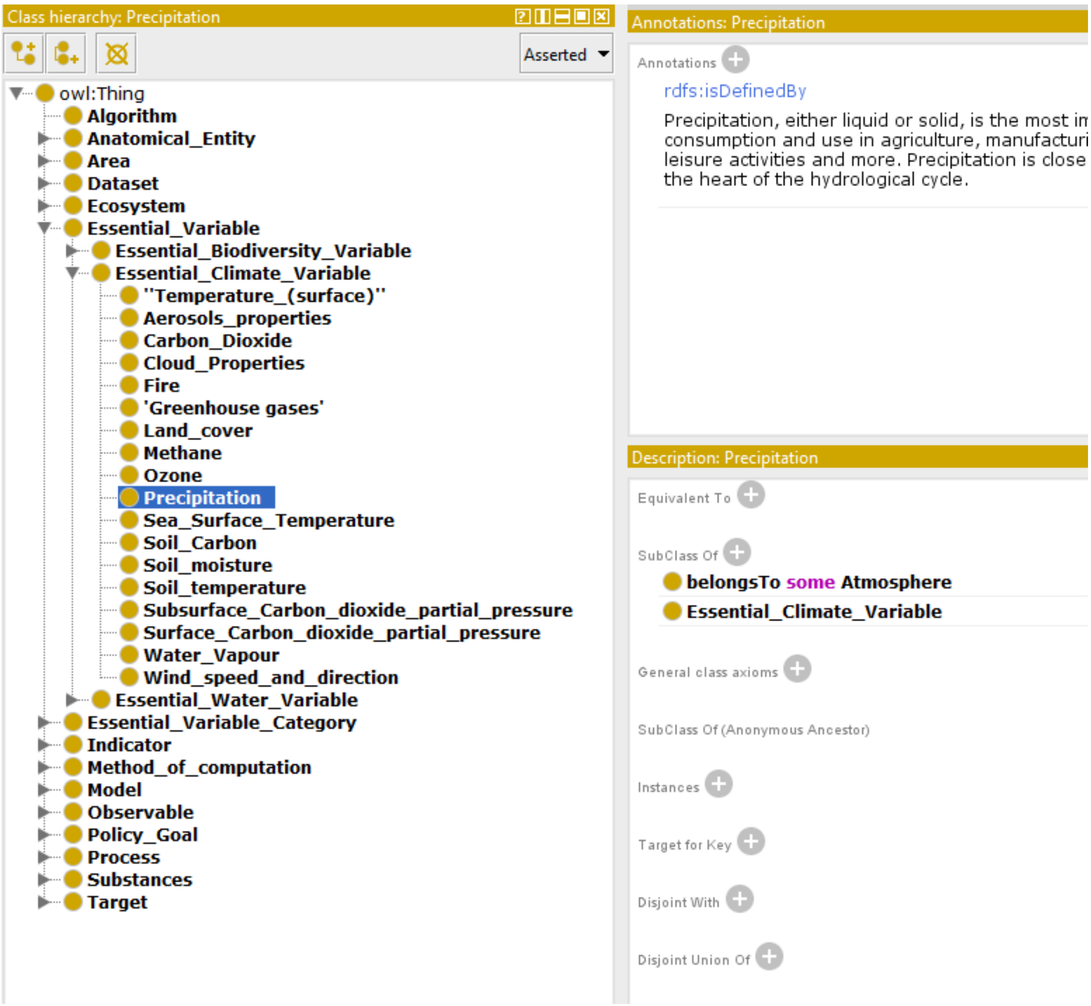
Equivalent To

SubClass Of

- Dataset
- generates some MOD13Q1-coll6

Figure 11. Dataset

- Essential Variables potentially related to the Indicator have been identified and organized according to the corresponding Category (e.g. Precipitation is an Essential Climate Variable and belongs to the Atmosphere category);



Class hierarchy: Precipitation

Asserted

```
▼ owl:Thing
    Algorithm
  ▶ Anatomical_Entity
  ▶ Area
  ▶ Dataset
  ▶ Ecosystem
  ▼ Essential_Variable
    ▶ Essential_Biodiversity_Variable
    ▼ Essential_Climate_Variable
        "Temperature_(surface)"
        Aerosols_properties
        Carbon_Dioxide
        Cloud_Properties
        Fire
        'Greenhouse gases'
        Land_cover
        Methane
        Ozone
        Precipitation
        Sea_Surface_Temperature
        Soil_Carbon
        Soil_moisture
        Soil_temperature
        Subsurface_Carbon_dioxide_partial_pressure
        Surface_Carbon_dioxide_partial_pressure
        Water_Vapour
        Wind_speed_and_direction
    ▶ Essential_Water_Variable
  ▶ Essential_Variable_Category
  ▶ Indicator
  ▶ Method_of_computation
  ▶ Model
  ▶ Observable
  ▶ Policy_Goal
  ▶ Process
  ▶ Substances
  ▶ Target
```

Annotations: Precipitation

Annotations

rdfs:isDefinedBy

Precipitation, either liquid or solid, is the most im
consumption and use in agriculture, manufacturi
leisure activities and more. Precipitation is close
the heart of the hydrological cycle.

Description: Precipitation

Equivalent To

SubClass Of

belongsTo some Atmosphere

Essential_Climate_Variable

General class axioms

SubClass Of (Anonymous Ancestor)

Instances

Target for Key

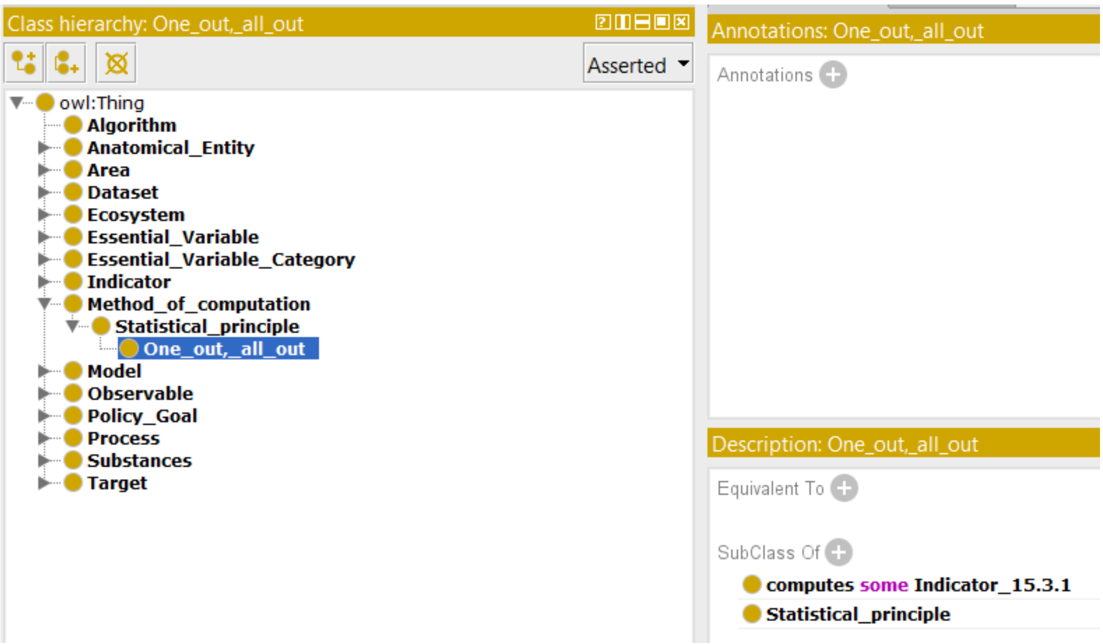
Disjoint With

Disjoint Union Of

Figure 12. Essential Variable

– the Indicator has been related to the method of computation generally used to calculate it[32];

[32] "The method of computation for this indicator follows the 'One Out, All Out' statistical principle and is based on the baseline assessment and evaluation of change in the sub-indicators to determine the extent of land that is degraded over total land area", Metadata 15-03-01, https://sustainabledevelopment.un.org/sdg15.



Class hierarchy: One_out,_all_out

Asserted

- owl:Thing
  - Algorithm
  - Anatomical_Entity
  - Area
  - Dataset
  - Ecosystem
  - Essential_Variable
  - Essential_Variable_Category
  - Indicator
  - Method_of_computation
    - Statistical_principle
      - One_out,_all_out
  - Model
  - Observable
  - Policy_Goal
  - Process
  - Substances
  - Target

Annotations: One_out,_all_out

Annotations

Description: One_out,_all_out

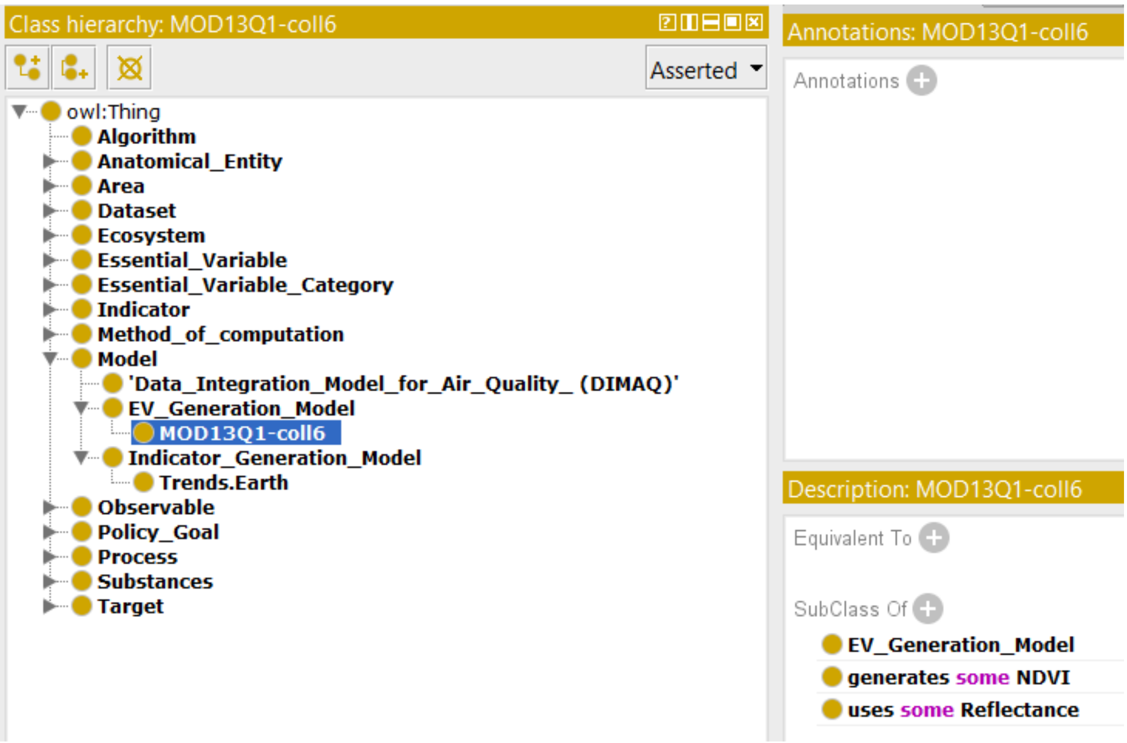
Equivalent To

SubClass Of

- computes some Indicator_15.3.1
- Statistical_principle

Figure 13. Method of computation

– The Model class specifies both EV generation models and Indicator generation models[33];



Class hierarchy: MOD13Q1-coll6

Asserted

- owl:Thing
  - Algorithm
  - Anatomical_Entity
  - Area
  - Dataset
  - Ecosystem
  - Essential_Variable
  - Essential_Variable_Category
  - Indicator
  - Method_of_computation
  - Model
    - 'Data_Integration_Model_for_Air_Quality_ (DIMAQ)'
    - EV_Generation_Model
      - MOD13Q1-coll6
    - Indicator_Generation_Model
      - Trends.Earth
  - Observable
  - Policy_Goal
  - Process
  - Substances
  - Target

Annotations: MOD13Q1-coll6

Annotations

Description: MOD13Q1-coll6

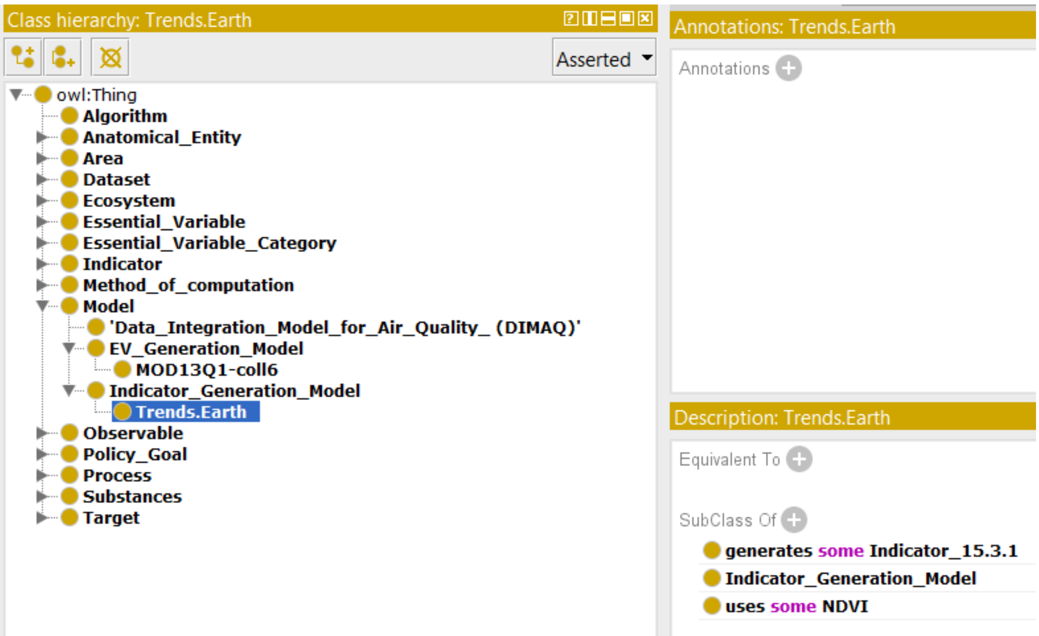
Equivalent To

SubClass Of

- EV_Generation_Model
- generates some NDVI
- uses some Reflectance

Figure 14. EV Generation Model

[33] "The MOD13Q1 product provides two primary vegetation layers. The first is the Normalized Difference Vegetation Index (NDVI) which is referred to as the continuity index to the existing National Oceanic and Atmospheric Administration-Advanced Very High Resolution Radiometer (NOAA-AVHRR) derived NDVI. The second vegetation layer is the Enhanced Vegetation Index (EVI), which has improved sensitivity over high biomass regions", https://lpdaac.usgs.gov/products/mod13q1v006/.



Class hierarchy: Trends.Earth

Asserted

- owl:Thing
  - Algorithm
  - Anatomical_Entity
  - Area
  - Dataset
  - Ecosystem
  - Essential_Variable
  - Essential_Variable_Category
  - Indicator
  - Method_of_computation
  - Model
    - 'Data_Integration_Model_for_Air_Quality_ (DIMAQ)'
    - EV_Generation_Model
      - MOD13Q1-coll6
    - Indicator_Generation_Model
      - Trends.Earth
  - Observable
  - Policy_Goal
  - Process
  - Substances
  - Target

Annotations: Trends.Earth

Annotations

Description: Trends.Earth

Equivalent To

SubClass Of

- generates some Indicator_15.3.1
- Indicator_Generation_Model
- uses some NDVI

Figure 15. Indicator Generation Model

During the following months, other case studies, suggested by the project partners, will be included in the ontological model. This will allow to further test our ontological model or to improve it if some important information is missing or is not correctly modelled. Furthermore, we will include the whole list of Essential Variables, defined in the Excel Sheet provided by the University of Geneva, in the model.

4 KETs and Semantic Services in GEOEssential KP

As stated in *Deliverable 1.1. Knowledge services architecture*, "the open and interoperable access to data and knowledge is assured by an ERA-PLANET Knowledge Platform, fully integrated with GEOSS, with functionalities specifically tailored to the GEOEssential requirements". The users of this Knowledge Platform (KP) will be represented by both decision makers, who should be able to take decisions and to adopt knowledge-based policies, and domain experts or data providers who want to share or search for information. A high degree of interoperability should be guaranteed in the organization of data and knowledge in the KP, hence in keeping with the ERA-PLANET and GEOEssential interoperability principles, the objective of the present task, embedded in the Key Enabling Technologies (KETs), is to ensure technical and semantic interoperability to facilitate data discovery, access and integration. The implementation of these patterns and technologies will ensure a full horizontal interoperability with relevant EO initiatives and programmes (e.g. GEOSS, Copernicus) and especially with the other ERA-PLANET projects.

The main advantages deriving from the implementation of semantic services are related to improving the information retrieval process, both for end users and for policy makers. On the one hand, the organization of the knowledge domain in an ontological model will allow advanced discovery and modelling services for answering complex queries. On the other hand, the provision of a unique terminology to be used for filling in metadata and searching for information, will enable query expansion techniques by exploiting semantic relationships (i.e. equivalence and hierarchical) between terms and concepts.

REFERENCES

Albertoni R., De Martino M., Di Franco S., De Santis V., Plini P., *EARTh: an Environmental Application Reference Thesaurus in the Linked Open Data Cloud*, in «Semantic Web», vol. V, n. 2, 2014, pp. 165-171.

Bodenreider O., Mitchell J.A., McCray A.T., *Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics*, in "Proceedings of the AMIA Symposium", 2002, pp. 61-65.

Brewster C., Alani H., Dasmahapatra A., Wilks, Y., *Data driven ontology evaluation*, in "Proceedings of the International Conference on Language Resources and Evaluation (LREC)", Lisbon, Portugal, 2004.

Capuano N., *Ontologie OWL: Teoria e Pratica*, in «Computer Programming», n. 148 - Luglio/Agosto 2005.

Caracciolo C., Stellato A., Morshed A., Johannsen G., Rajbhandar S., Jaques Y., Keizer J., *The AGROVOC Linked Dataset*, in «Semantic Web», vol. IV, n. 3, 2013, pp. 341-348.

Caruso A., Folino A., *Corpus-based knowledge representation in specialized domains*, in *Corpus-based studies on language varieties*, edited by F. Alonso Almeida, L. Cruz Garcia, V. Gonzalez Ruiz, Peter Lang, 2016, pp. 11-35.

Craglia M., Nativi S., Santoro M., Vaccari L., Fugazza C., *Inter-disciplinary Interoperability for Global Sustainability Research*, in "International Conference on GeoSpatial Sematics", edited by C. Claramunt, S. Levashkin, M. Bertolotto, LNCS, vol. (V)MDCXXXI, Springer-Verlag Berlin Heidelber, 2011, pp. 1-15.

Cui H., *Competency evaluation of plant character ontologies against domain literature*, in «Journal of the American Society for Information Science and Technology», vol. LXI, n. 6, 2010, pp.1144–1165.

De la Iglesia D., Cachau R.E., Garcìa-Remaesal M., Maojo V., *Nanoinformatics knowledge infrastructures: bringing efficient information management to nanomedical research*, in «Computer Science Discovery», vol. VI, n. 1, 2013.

Dell'Orletta F., Venturi G., Cimino A., Montemagni S., *T2K^2: a System for Automatically Extracting and Organizing Knowledge from Texts*, in "Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)", 26-31 May, Reykjavik, Iceland, 2014.

Ensan F., Du W., *A Modular Approach to Scalable Ontology Development*, in *Canadian Semantic Web: Technologies and Applications*, edited by W. Du, F. Ensan, Springer Science+Business Media, 2010.

Fugazza C., Dupke S., Vaccari L., *Matching SKOS Thesauri for Spatial Data Infrastructures*, in "Metadata and Semantic Research", edited by S. Sanchez-Alonso, I.N. Athanasiadis, CCIS, vol. CVIII, 2010, pp. 211-221.

Gruber, T.R., *A Translation Approach to Portable Ontology Specification*, in «Knowledge Acquisition», vol. V, 1993, pp. 199-220.

Isaac A., Wang S., Zinn C., Matthezing H., Van der Meij L., Schlobach S., *Evaluating Thesaurus Alignments for Semantic Interoperability in the Library Domain*, in «IEEE INTELLIGENT SYSTEMS», 2009, pp.76-86.

ISO 25964-2:2013, Information and documentation - *Thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies*.

Lee G., Mariam T., Ahmad K., *Terminology and the construction of ontology*, in «Terminology», vol. XI, n. 1, 2005, pp. 55-81.

Liddle S. W., Hewett K. A., Embley D. W., *An Integrated Ontology Development Environment for Data Extraction*, in "Proceedings of Information Systems Technology and its Applications, International Conference (ISTA)", Kharkiv, Ukraine, 2003.

Morshed A., Caracciolo C., Johannsen G., Keizer J., *Thesaurus alignment for Linked Data publishing*, in "Proceedings of the International Conference on Dublin Core and Metadata Applications", 2011.

Nagai M., Ono M., Shibasaki R., *Earth Observation Data Interoperability Arrangement with Ontology Registry*, in "Information Search, Integration, and Personalization: International Workshop", edited by A. Kawtrakul et al., CCIS, vol. CDXXI, 2012, pp. 128-136.

Navigli R., Velardi P., *Learning Domain Ontologies from Document Warehouses and Dedicated Websites*, in «Computational Linguistics», vol. XXX, 2004, pp. 151–179.

Noy N. F., McGuinness D. L., *Ontology Development 101: A Guide to Creating Your First Ontology*, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

Rospocher M., Tonelli S., Serafini L., Pianta E., *Corpus-based terminological evaluation of ontologies*, in «Applied Ontology», vol. VII, n. 4, 2012, pp. 429-448.

Rowley J., *The wisdom hierarchy: Representations of the DIKW hierarchy*, in «Journal of Information Science», vol. XXXIII, 2007, pp. 163-180.

Reyers B., Stafford Smith M., Erb K-H., et al., *Essential Variables help to focus Sustainable Development Goals monitoring*, in «Current Opinion in Environmental Sustainability», vol. XXVI-XXVII, June 2017, pp. 97-105.

Wong W., Liu W., Bennamoun M., *Determining termhood for learning domain ontologies using domain prevalence and tendency*, in "IProceedings of the sixth Australasian conference on Data mining and analytics - AusDM '07", Darlinghurst, Australia, vol. LXX, Australian Computer Society, Inc., 2007, pp. 47–54.

Zeng, M. L., *Knowledge Organization Systems*, in «Knowledge Organization», vol. XXXV, n. 2-3, 2008, pp. 160-182.