SMart URBan Solutions for air quality, disasters and city growth

Deliverable D6.4: Semantic resources

Contract Number	Project GA number (ERA-PLANET): 689443	**Acronym of Trans. Project**	SMURBS
Full title of Trans. Project	SMart URBan Solutions for air quality, disasters and city growth		
Trans. Project URL	http://www.smurbs.eu, http://www.era-planet.eu/		
EC Project Officers	Jean Dusart		

Deliverable/Document	**Number**	D6.4	**Name**	Semantic resources		
Work package	**Number**	WP6	**Name**	ERA-PLANET principles and KETs for interoperability		
Date of delivery	**Contractual**		M38	**Actual**	04.01.2021	
Status	Final					
Type	R: Report					
Distribution Level	PU: Public					
Authoring Partner	UNICAL – Università della Calabria					
Prepared by	Assunta Caruso (UNICAL), Antonietta Folino (UNICAL)					
Quality Assurance	Evangelos Gerasopoulos (PC-NOA)					
Contact Person	Dr Evangelos Gerasopoulos		Project Coordinator			
	Metaxa & Vas. Pavlou Str. • 152 36 Penteli, Greece					
	Email	egera@noa.gr	**Phone**	+30-2108109124	**Fax**	+30-2108103236

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 SMURBS Knowledge Base infrastructure, it is essential to ensure high standards of discoverability, accessibility, and interoperability. The design of this infrastructure involves the development of an ontological model, supported by a corpus-based terminological extraction and by a comparison with existing controlled vocabularies. This 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. Some of the vocabularies employed are the following: GEMET Thesaurus, INSPIRE Feature Concept Dictionary and Glossary, AGROVOC Thesaurus, EARTh Thesaurus.

The integration of the abovementioned ontological model in a knowledge base infrastructure will therefore improve the ability of end-users to explore and exploit EO data. 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 Knowledge Base will be based on OWL/RDF technologies.

Project Information

This document is part of a research project funded under the **ERA PLANET - European Union Horizon 2020 Programme.**

Call Identifier: 2nd Joint Transnational Call of ERA-PLANET (SC5-15-2015 - Strengthening the European Research Area in the domain of Earth Observation).

Project GA number: 689443 (ERA-PLANET)

Transnational Project Title: SMURBS - SMart URBan Solutions for air quality, disasters and city growth

Project Beneficiaries:

Nº	Official[1] Nº	Participant Legal Name	Country	Logos
1	26	NATIONAL OBSERVATORY OF ATHENS (NOA) - **Coordinator**	GREECE	
2	3	IDRYMA IATROVIOLOGIKON EREUNON AKADEMIAS ATHINON (AoA)	GREECE	
3	4	ARISTOTELIO PANEPISTIMIO THESSALONIKIS (AUTH)	GREECE	
4	1	CONSIGLIO NAZIONALE DELLE RICERCHE (CNR)	ITALY	
5	7	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (CNRS)	FRANCE	
6	8	CENTRO DE INVESTIGACION ECOLOGICA Y APLICACIONES FORESTALES (CREAF)	SPAIN	
7	18	HELMHOLTZ-ZENTRUM GEESTHACHT ZENTRUM FÜR MATERIAL- UND KÜSTENFORSCHUNG GMBH (HZG)	GERMANY	
8	20	ISTITUTO SUPERIORE PER LA PROTEZIONE E LA RICERCA AMBIENTALE (ISPRA)	ITALY	
9	21	IVL SVENSKA MILJOEINSTITUTET AB (IVL)	SWEDEN	
10	22	INSTITUT JOZEF STEFAN (JSI)	SLOVENIA	
11	24	MASARYKOVA UNIVERZITA (MU)	CZECH REPUBLIC	
12	25	NATIONAL CENTER FOR SCIENTIFIC RESEARCH "DEMOKRITOS" (NCSR)	GREECE	
13	27	PAUL SCHERRER INSTITUT (PSI)	SWITZERLAND	

[1] *As reported in the Grant Agreement AMD-689443-40, Annex 1 - Description of the action (part A).*

14	35	ROMANIAN SPACE AGENCY (ROSA)	ROMANIA	
15	29	SPACE RESEARCH INSTITUTE OF THE NATIONAL ACADEMY OF SCIENCES OF UKRAINE AND THE NATIONAL SPACE AGENCY OF UKRAINE (SRI)	UKRAINE	
16	36	STOCKHOLMS UNIVERSITET (SU)	SWEDEN	
17	30	LEIBNIZ INSTITUT FUER TROPOSPHAERENFORSCHUNG e.V. (TROPOS)	GERMANY	
18	37	HELSINGIN YLIOPISTO (UHEL)	FINLAND	
19	32	UNIVERSITA DELLA CALABRIA (UNICAL)	ITALY	

Table of Contents

Acronyms and Abbreviations 6

List of Figures 8

List of Tables 9

1. Introduction 1

2. Environment and smart cities vocabularies 4

3. Corpus construction 9

4. Comparison between the corpus term list and existing terminologies 12

5. Ontological model development 14

5.1 Use cases 19

5.2 Mapping towards other vocabularies 22

6. Semantic services in SMURBS Knowledge Platform 27

References 28

Acronyms and Abbreviations

Acronym	Description
CHEBI	Chemical Entities of Biological Interest
CORDIS	Community Research and Development Information Service
DIKW	Data-Information-Knowledge-Wisdom
DISIT Lab	Distributed Systems and Internet Technology Lab
EARTh	Environmental Applications Reference Thesaurus
EEA	European Environment Agency
EIONET	European Environment Information and Observation Network
EnvO	Environment Ontology
EnvThs	Environmental Thesaurus Server
EO	Earth Observation
ETC/CDS	European Topic Centre on Catalogue of Data Sources
FAO	Food and Agriculture Organization
GCI	Global City Indicator Foundational Ontology
GEMET	General European Multilingual Environmental Thesaurus
ICT	Information and Communication Technologies
IFCD	INSPIRE Feature Concept Dictionary
INSPIRE	INfrastructure for SPatial InfoRmation in Europe
IoT	Internet of Things
KB	Knowledge Base
KOS	Knowledge Organization System
KOSs	Knowledge Organization Systems
KP	Knowledge Platform

LOD	Linked Open Data
LusTRE	Linked Thesaurus Framework for the Environment
NASA	National Aeronautics and Space Administration
OBI	Ontology for Biomedical Investigations
OGC	Open Geospatial Consortium
OWL	Ontology Web Language
PCO	Population and Community Ontology
RDF	Resource Description Framework
SCO	Smart City Ontology
SDGIO	Sustainable Development Goals Interface Ontology
SKOS	Simple Knowledge Organization System
SKOS-XL	Simple Knowledge Organization System eXtension for Labels
SWEET	Semantic Web for Earth and Environment Terminology
UMTHES	Umweltthesaurus
UNEP	United Nations Environment Program
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

List of Figures

Figure 1. DIKW pyramid (Bucăţa et al., 2019) 2

Figure 2. The Revised Knowledge Pyramid with KM, Big Data and IoT (Jennex 2017) ... 2

Figure 3. T2K Term configuration 10

Figure 4. Term lists with and without reference corpus 11

Figure 5. Various types of KOSs (Zeng 2008, p. 161) 14

Figure 6. Ontology taxonomy 17

Figure 7. OntoGraph 18

Figure 8. Relations between Indicator 11.1.1 and other ones 20

Figure 9. Relations between Goals and Targets 20

Figure 10. Description of “Informal settlement” concept 21

Figure 11. Description of “Slum” concept 21

Figure 12. Relation between pollutants and health effects 21

Figure 13. “Urban population” concept 22

Figure 14. SKOS Annotations 24

Figure 15. Ontology-GEMET matches 24

Figure 16. GEMET Mappings 25

Figure 17. EARTh Mappings 26

List of Tables

Table 1. Results of term comparison 13

1. Introduction

A Smart city is "*a place where traditional networks and services are made more efficient with the use of digital and telecommunication technologies for the benefit of its inhabitants and business*"[2]. City smartness entails the improvement of efficiency of an urban area by means of digital solutions able to integrate government services with citizen's welfare[3]. The dynamism of both urban environment and urban population impacts on the growth of a city and on its innovative development. Making cities smart is a priority around the world and requires the implementation of best practices able to endorse sustainable mobility and transport, efficient energy use, the social function of consumption, etc. In order to achieve these outcomes and to limit the difficulties originating from uncontrolled urban development, it is necessary to ensure that our cities grow in a sustainable way. ICT innovation can contribute to this goal by providing a valid support in both managing the myriad of available data shared by different networks, software and devices, and adding value to them. "Smart city" is a broad concept (Abid et al., 2016) which includes social and cultural issues in addition to the environmental ones. To fill this gap and harmonize these three levels of sustainability some efforts have been made and several web technologies, based on the principles of sharing information, have been developed (Abid et al., 2016). Understanding the potential value of data and information is important, especially when they effectively support decision-making processes. 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[4], ontologies[5]) to ensure data quality and harmonization and to interpret the meaning of data, turning it into usable information and knowledge.

The data-information-knowledge-wisdom (DIKW) pyramid in Figure 1 illustrates the steps towards the move from data to knowledge and the actions needed to reach this purpose.

[2]<https://ec.europa.eu/info/eu-regional-and-urban-development/topics/cities-and-urban-development/city-initiatives/smart-cities_en>

[3] < https://internetofthingsagenda.techtarget.com/definition/smart-city>

[4] "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.

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

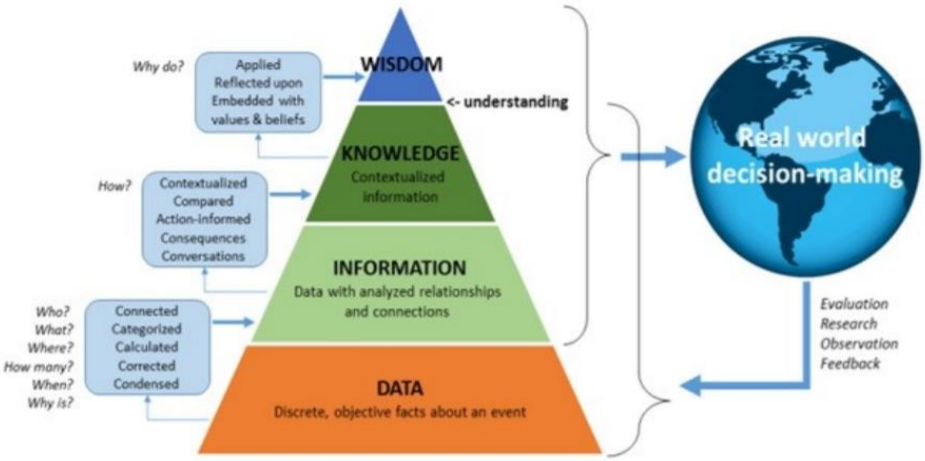

Figure 1. DIKW pyramid (Bucăţa et al., 2019)

The DIKW pyramid has been used as a paradigm to explain how to manage data in order to extract knowledge and to help the prediction of an event. To make sense of a large amount of data distributed in heterogeneous datasets and to generate value from them, it is also important to consider the role that some technologies such as Big Data and the Internet of Things (IoT) could play in data integration. In this perspective, (Jennex 2017) presents an interesting revisiting of the knowledge pyramid which includes the treatment of large amounts of datasets by using innovative technologies to capture, manage and process them.

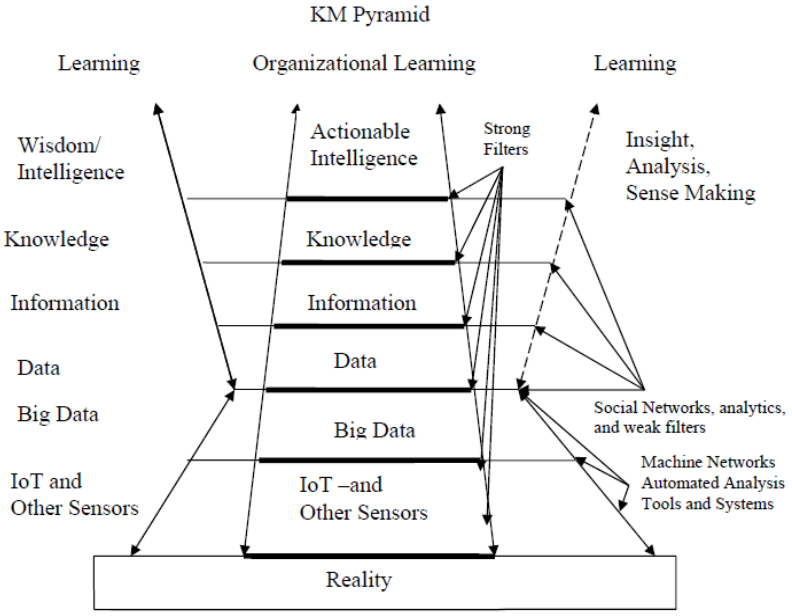

Figure 2. The Revised Knowledge Pyramid with KM, Big Data and IoT (Jennex 2017)

This point of view underlines that information is an added-value product generated by understanding data (Big Data and IoT and sensors) 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 in doing so and in supporting semantic interoperability. 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 this domain. In order to address the organization and homogenization of the huge volume of information, scientists also 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 a community on the categorisations of concepts 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 above-mentioned 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 6.4, therefore, is the development of semantic services specifically represented by an ontology resulting from the specialization of the ontological conceptual model developed to support semantic interoperability in the whole ERA-PLANET Programme. In order to integrate concepts about the smart cities domain within the general model, a set of existing semantic resources and ad hoc vocabularies need to be considered. Moreover, to facilitate the dynamic data integration and to limit difficulties in managing heterogeneous data coming from several sources, a set of semantic relationships will be explained according to the SKOS (Simple Knowledge Organization System)[6] model. This ontology will be integrated into the ERA-PLANET Knowledge Platform (KP) with functionalities specifically tailored to the SMURBS requirements, as described in *Deliverable 6.1 Design of the SMURBS data and service infrastructure* prepared by IIA-CNR. As reported in this document, the KP solution "stems from the need of lowering barriers for both end-users – to easily access the outcomes of models for knowledge generation – and model developers – to easily publish and share their models. Therefore, it aims not only to data sharing but more generally to provide support to multidisciplinary communities-of-practice for providing knowledge for evidence-based policy and informed decision-making". The fulfilment of these objectives is not possible without the integration of semantic services aiming at a formalized and shared representation of the knowledge domain.

[6] https://www.w3.org/TR/skos-primer/

2. Environment and smart cities vocabularies

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 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 made 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)[7] 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.

SWEET Ontology[8] 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[9] 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 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, a solid set of concepts to start with.

The **AGROVOC** thesaurus is a multilingual controlled vocabulary covering all areas of interest of the FAO of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment, etc[10]. It has evolved over the years and currently AGROVOC is an SKOS-XL (Simple Knowledge Organization System eXtension for Labels) concept scheme and a Linked Open Data (LOD) Dataset edited by VocBench, composed of over 35,000

[7] http://edscvs.ccny.cuny.edu/
[8] https://sweet.jpl.nasa.gov/
[9] Numbers are accurate as of January 2018
[10] http://www.fao.org/agrovoc

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[11], 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 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 alphanumeric 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[12]. 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[13] have been agreed upon with the EEA and each descriptor has been assigned to as many themes as necessary. There are currently more 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

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

[12] 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.

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

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 such as AGROVOC, EUROVOC, DBPEDIA and UMTHES. EARTh is currently maintained in the context of LusTRE (Linked Thesaurus Framework for the Environment)[14] 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[15] (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 in particular, its main role is to support the harmonisation effort and to identify conflicts between the specifications of the spatial object types in the different themes. The INSPIRE Glossary[16] 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).

As far as existing domain ontologies are concerned, the Environment Ontology (**ENVO**)[17] 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 consists of classes (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

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

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[18]. Its importance derives from the similarity of its aims and domain of interest with those of the ontology we are developing within the SMURBS 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"[19]. 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 greater interoperability.

Considering the specific environment sub-themes dealt with in the SMURBS project, alongside these vocabularies which concern general issues related to environment, our search for existing terminologies has also focused on smart cities, natural disasters and urban growth. To our knowledge, not many vocabularies have been defined in these domains. Most of them, especially in the form of ontologies, concern the field of smart cities, which is analysed and represented from different perspectives with the aim to support and improve functionalities of the digital applications and solutions, widely used in energy and transport monitoring, to inform citizens about traffic, park services, emission level of CO_2 in the air, etc. In this perspective the Smart City Ontology (SCO) (Komninos et al., 2015) represents a valid solution to be considered in the specialization of our general conceptual model. SCO is a general ontology that tries to map overall aspects related to smart cities and guarantees interoperability by formalizing the multidimensionality of the domain and the heterogeneous systems from which information is captured. Regarding the structure, the first release consists of ten super-classes about space structuring, urban functions and city characteristics, 708 entities, 422 classes, 62 object properties, 190 data properties and 27 individuals. It is interesting to note that this ontology has been applied to several smart cities applications in order

[18] http://aims.fao.org/activity/blog/sustainable-development-goals-interface-ontology-sdgio-support-united-nations
[19] http://www.ontobee.org/ontology/SDGIO

to verify which classes and properties have been used to generate knowledge and which functionalities derive from their combination.

A comprehensive representation is also provided by the "km4city, the DISIT Knowledge Model for City and Mobility"[20], developed by the DISIT Lab (Distributed Systems and Internet Technology Lab) of the University of Florence and oriented to the "the description of smart cities, leveraging interconnection, storage and interrogation of data from many different sources, such as various portals of the Tuscan region (MIIC, Muoversi in Toscana, Osservatorio dei Trasporti), Open Data and Linked Data, provided by individual municipalities (mainly Florence)". The main and most general classes represent *Administration* (*Municipality, Province, Region*, etc.); *Street Guide* (*Road, StreetNumber, RoadLink, Junction*, etc.); *Points of Interest* (i.e. services and activities useful to the citizen); *Local Public Transport* (*Ride, Route, BusStop*, etc.); *Sensors* (*SensorSite, CarParkSensor, Weather_sensor*, etc.); *Temporal* (concepts related to time); *Metadata* (triples providing the context of each dataset).

Another interesting ontology is the "gci- Global City Indicator Foundation Ontology" (Fox 2013), defined at the University of Toronto and oriented "to work with cities globally in identifying a common set of indicators and establishing standardized definitions and methodologies that can be consistently applied globally". It is based on a set of foundational ontologies, i.e. high-level domain-independent models, which provide many of its basic concepts and it integrates ontologies including geonames, measurement theory, statistics, time, provenance, validity and trust. More than 150 cities have been involved in the ontology development process.

Other examples, taken from the linked open vocabularies (LOV) repository[21] are represented by "shw Smart Home Weather", "SAREF: the Smart Appliances REFerence ontology", "ha- Home Activity" mainly concerning the sub-theme of smart houses and "smg- CERISE CIM Profile for Smart Grids", "OPM: Ontology for Property Management", "SEAS ontology", related to the issue of smart energy.

[20] http://wlode.disit.org/WLODE/extract?url=http://www.disit.org/km4city/schema

3. Corpus construction

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 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), while (Brewster et al., 2004) illustrate a method for evaluating an ontology by comparing it with a domain-specific corpus and (Cui 2010) 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 collected documents are represented by deliverables and scientific papers related to the projects listed in *Deliverable 2.2 Smart city's state of the art*, which consists in an inventory of available smart city solutions, best practices, success stories and ongoing projects directly proposed by project partners. From a methodological point of view, this guarantees the authoritativeness and the adherence and representativeness of the corpus in respect to the project themes. In particular, most of the project specific documents have been downloaded from the CORDIS (Community Research and Development Information Service) website, that is the "European Commission's primary source of results from the projects funded by the EU's framework programmes for research and innovation"[22]. This allowed us to retrieve more complete and official documents for several of the proposed projects. In the absence of available information in this repository, we referred to the project websites. The constructed corpus contains 799 pdf files. They have been converted into txt format in order to be automatically processed. Indeed, the following step has consisted in terminological extraction from the domain-specific corpus. The main aim of the terminology extraction has been 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 should be representative of the domain of interest. The 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. 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

[22] https://cordis.europa.eu/about

organized in a conceptual-semantic network (Caruso et al., 2016). The obtained term list consists of 148,920 candidate terms. It has been obtained by applying the *Term configuration* represented in the following Figure. Apart from the grammatical structure of the compound terms, we have set quantitative parameters, the most significant of which is the frequency threshold, equal to 2. It indicates the minimum value of the occurrence frequency of each candidate term in the document from which it has been extracted and in the whole corpus. The chosen value depends on the corpus dimension and on the possibility to extract potential interesting terms even if less frequent.

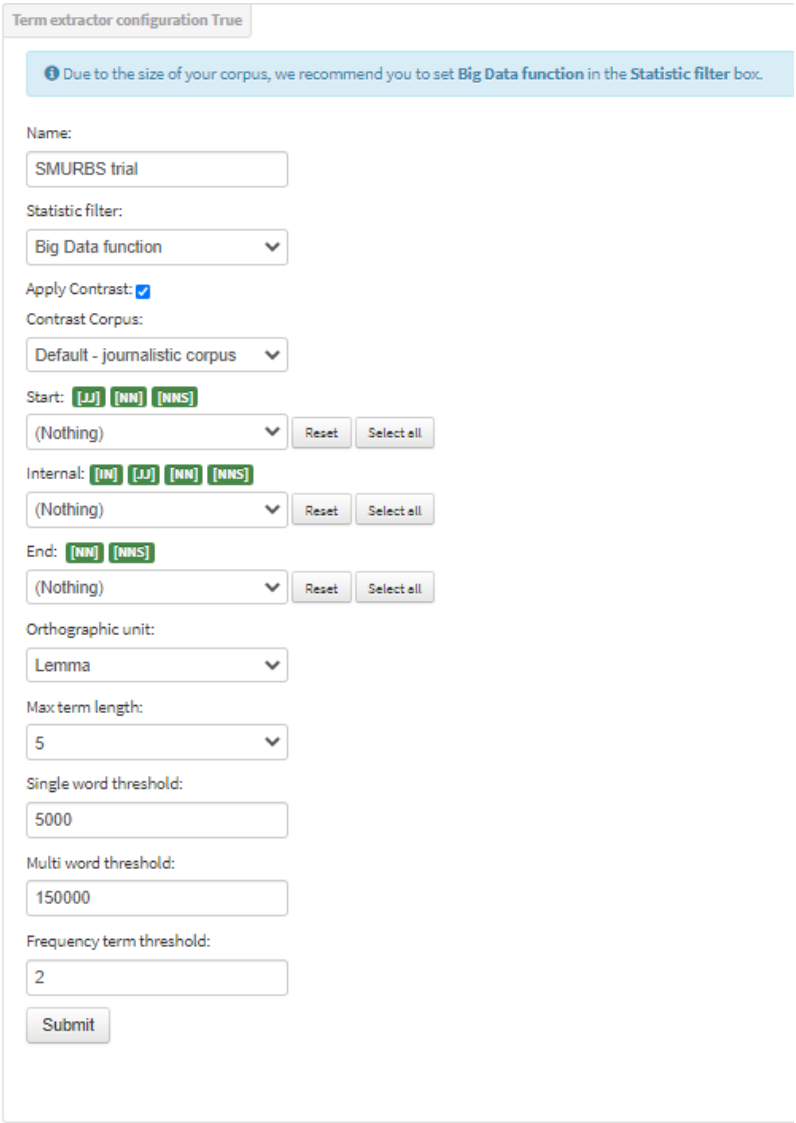

Figure 3. T2K Term configuration

The terminology extraction was conducted both with and without the use of a reference corpus. In Figure 4, the first table shows an extract of the term list obtained without the comparison of a reference corpus and sorted by frequency, while the second table 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 the second table 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.

	Prototypical_Form	Lemma_of_Term	Frequency
1	data	datum	30297
2	project	project	22590
3	city	city	16462
4	measures	measure	12836
5	energy	energy	12360
6	results	result	12269
7	air	air	11667
8	area	area	11097
9	information	information	9778
10	users	user	9686
11	model	model	9510
12	time	time	8367
13	order	order	8231
14	air quality	air quality	5165
15	years	year	7965
16	buildings	building	7858
17	case	case	7462
18	use	use	7462
19	implementation	implementation	7228
20	number	number	7151
21	study	study	7092
22	platform	platform	7061
23	citizens	citizen	7025
24	sensors	sensor	6870
25	services	service	6609
26	solutions	solution	6396

	Prototypical_Form	Lemma_of_Term	Frequency
1	data	datum	30297
2	air quality	air quality	5165
3	ecosystem	ecosystem	4538
4	stakeholders	stakeholder	4449
5	dataset	dataset	3022
6	vegetation	vegetation	3010
7	workshop	workshop	2937
8	doi	doi	2860
9	implementation	implementation	7228
10	climate change	climate change	2705
11	mobility	mobility	2694
12	grant agreement	grant agreement	2559
13	replication	replication	2291
14	simulations	simulation	2211
15	metadata	metadata	1990
16	biodiversity	biodiversity	1963
17	species	species	4989
18	particles	particle	1872
19	aerosol	aerosol	1795
20	pixels	pixel	1712
21	social media	social medium	1708
22	engagement	engagement	1668
23	parameters	parameter	1622
24	website	website	1562
25	variables	variable	1558
26	ecosystem services	ecosystem service	1536
27			

Figure 4. Term lists with and without reference corpus

Having used this function, the term “climate change” for instance, undoubtedly representative of the domain, has gone up from position 59 in the frequency list to 11 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, 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 SMURBS Knowledge Base platform.

4. Comparison between the corpus term list and existing terminologies

As already mentioned in previous sections, before and alongside the corpus construction, relevant existing terminological resources related to environment and to smart cities were taken into consideration. These resources, because of their inner conceptual structure and functional characteristics, have been used to achieve two different objectives: on the one hand, thesauri - where concepts reach a deeper level of granularity - have been collected and compared with the term list obtained by the above mentioned terminological extraction in order to evaluate their semantic coverage in relation to the corpus and vice-versa; on the other hand, ontologies - characterized by a more abstract conceptual organization, have been analysed with a view towards the development of our ontological model. In this sense, the terminological extraction and comparison are functional to the ontology construction and enrichment.

Concerning the first group of resources, 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. They 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. The term list extracted from our corpus (148,920 terms) has been compared with the term lists obtained from each one of the abovementioned vocabularies. Hereafter we present the quantitative results obtained from this comparison. They include only exact matches, i.e. matches based on the exact correspondence between the lexical labels coming from the compared lists. This represents a subset of matches which should be integrated by correspondences between terms having differences in lexical aspects (e.g. climate change and climatic change) or between synonyms. The presence of polysemy is rarer, but it should also be considered, in particular when the meaning of terms is too generic. Thus, a manual evaluation of the automatically obtained results is required in order to obtain a more realistic estimation of the semantic overlap between vocabularies.

The obtained results are shown in Table 1:

Vocabulary	N. of vocabulary terms	N. of vocabulary terms found in our list	N. of vocabulary terms found in our list/N. of vocabulary terms (%)	N. of vocabulary terms found in our list/N. of terms from our list (%)
EARTh Thesaurus	13,969	3,031	21.7%	2.04%
GEMET	5,527	1,998	36%	1.34%
AGROVOC	45,501	1,664	3.66%	1.18%
INSPIRE	559	155	27.73%	0.10%

Table 1. Results of term comparison

Concerning the EARTh thesaurus, 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. In regards to INSPIRE, as previously mentioned, it includes both a *Glossary* and a *Feature Concept Dictionary*. The comparison has involved both of them, encompassing 200 and 360 terms respectively.

It is important to note that percentages shown in the last column are low because of the large amount of candidate terms extracted from our corpus. It can be defined as a rough list, considering that it has not undergone any terminological analysis oriented to delete common language terms or terms not representative of the specific domain. Once analysed, this datum will be more representative of the semantic overlapping and coverage of the extracted terminology. In general, both the term extraction and the comparison allowed us to gather a rich terminology and to identify semantic correspondences with existing vocabularies which can be a significant support in the construction of the ontological model described in the following section. The analysis of those candidate terms which have not been found in the existing terminologies will allow us to identify concepts specifically related to the main themes of the project with the aim of disposing of a common and shared terminology. Among these "new" concepts we can mention: *smart climate shells*; *sustainable urban mobility*; *illegal migration surveillance*; *natural disaster risk mitigation*; *smart building management systems; smart mobility; smart lighting;* and so on.

5. Ontological model development

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 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.

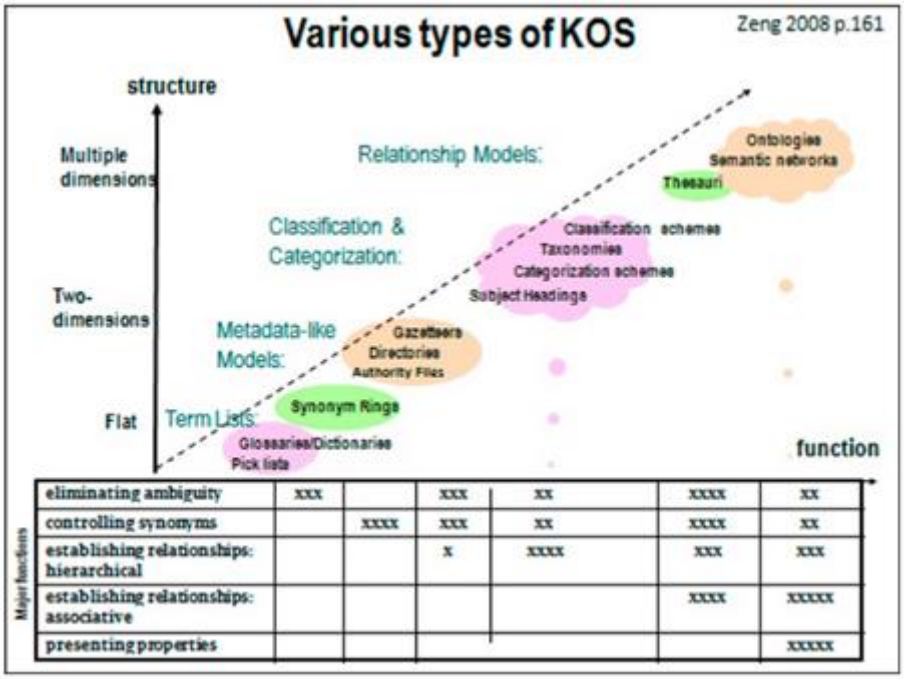

Figure 5. Various types of KOSs (Zeng 2008, p. 161)

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

[23] http://www.isko.org/cyclo/kos
[24] For a description of ontology structure and construction see (Noy et al., 2001; Capuano 2005)

when creating subclasses: they inherit properties defined for their super-classes, 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" and is represented by "∀"[25].

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 Resource Description Framework (RDF)[27] and has foundations in Description Logics, which allows programs called reasoners to check if the ontology is consistent.

The ontological model developed within the context of the SMURBS project is interconnected with the Knowledge Platform (KP) described in *Deliverable 6.1: Design of the SMURBS data and service infrastructure*. The KP infrastructure is conceived to be transversal to the four ERAPLANET projects, with functionalities and contents specifically tailored to the main themes of each one of them. Similarly, the ontological model is characterized by a more general level, which encompasses the main concepts concerning ERA-PLANET domains or subdomains, and by more specific levels, focused on SMURBS. 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, more specifically with CNR-IIA Florence, 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. In particular, the development of an ontological model within the context of the SMURBS project, is motivated by the need of addressing the semantic mismatch due to the large number of datasets; the issue of non-formalized knowledge, which should be provided by the domain experts; the semantic interoperability, particularly needed when experts themselves do not agree on some concept definition or on some term use. Furthermore, it will be implemented in the Virtual Laboratory under development by CNR-IIA, in order to guide users in annotating models and workflows and to discover data and information.

[25] https://www.w3.org/TR/owl2-primer/
[26] https://www.w3.org/OWL/
[27] https://www.w3.org/RDF/

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. Indeed, we have decided not to populate classes by *Individuals* (instances of classes) but to enrich them 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. The main general classes and subclasses defined in the ontology are the following:

- Agent;
- Algorithm;
- Anatomical Entity;
- Area;
- Dataset;
- Ecosystem;
- Essential Variable;
- Essential Variable Category;
- Indicator: a derived parameter summarizing the status of the system under consideration (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;
- Process;
- Sensor;
- Substance;
- Target.

Figure 6 represents the taxonomy under some of the more representative classes, such as Essential Urban Variable, which includes those proposed by the project partners within SMURBS research activities, Anatomical Entity, related to the effect of pollution on human health and Urban Area, related to the concept of smart city underlying the whole project:

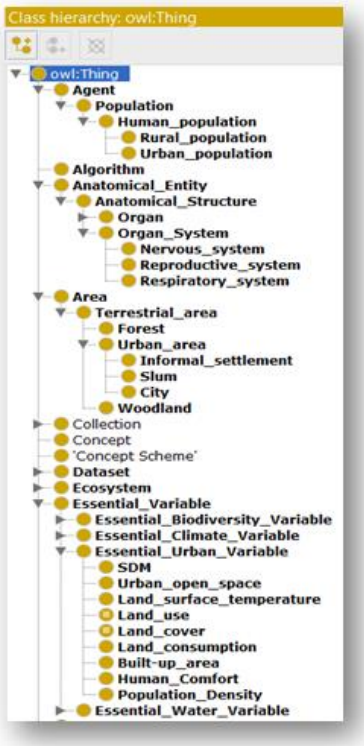

Figure 6. Ontology taxonomy

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);
- causes (domain: Pollutant; range: Pollution);
- computes (domain: Method_of_Computation; range: Indicator);
- concerns (domain: Process; range: Area);
- 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);
- measures (domain: Indicator and Sensor; range: Target and Observable);
- uses (domain: Model and Indicator; range: Dataset, Essential_Variable, Observable and Indicator);
- hasIndex (domain: Land_degradation; range: Observable);
- hasPart (domain: Organ System; range: Organ);
- hasTarget (domain: Policy; range: Target);
- isCausedBy (inverse of causes);
- isTargetOf (inverse of hasTarget);
- isMeasuredBy (inverse of measures);

- isPartOf (inverse of hasPart);
- 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);
- isSettledIn (domain: Agent; range: Area);
- isPopulatedBy (inverse of isSettledBy);
- isUsedBy (inverse of uses).

The OWL ontology can be interactively navigated through the OntoGraph plug-in provided by Protégé (Figure 7)[28], the tool used for its development. 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 expresses a Statement in the form of *subject - predicate - object* expressions. For example: "*Air Pollution hasIndex Particulate Matter" and vice versa "Particulate Matter isIndexOf Air Pollution*". 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.

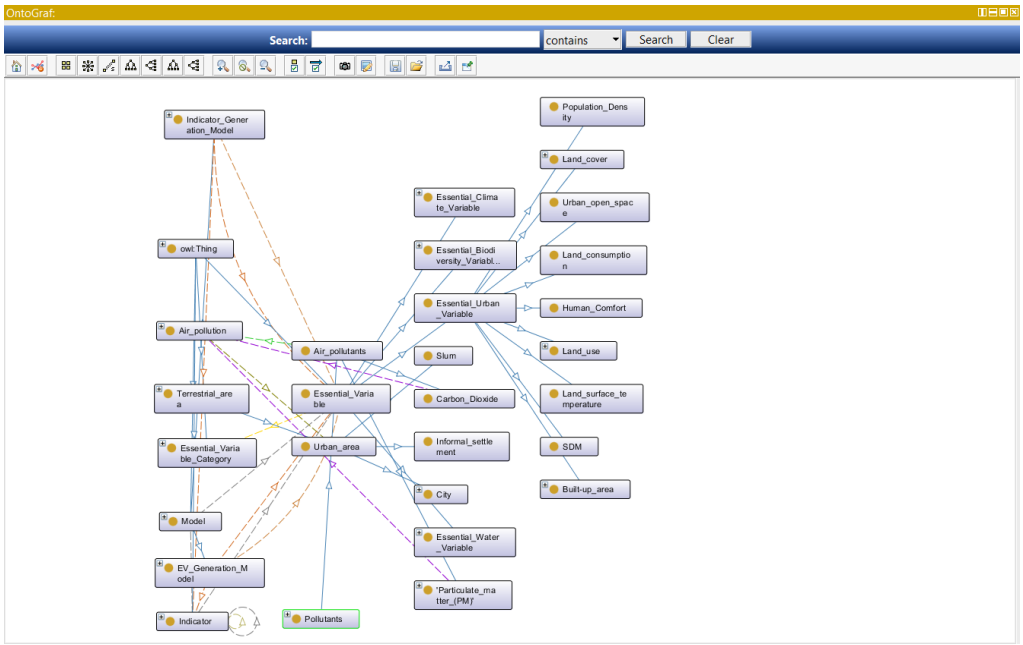

Figure 7. OntoGraph

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

5.1 Use cases

In order to test the consistency of the ontology compared to the SMURBS objectives and to verify if the specific characteristics of its domain of interest are well represented, it has been necessary to identify some representative case studies intended as an investigation of real phenomenon that often occur in the domain. The indicators which have been chosen in accordance with CNR-IIA and NOA are 11.1.1, 11.6.2 and 11.3.1. Their conceptual modelling and, consequently, the specialization of the general structure of the ontology, has been based on the analysis of the official and authoritative documentation provided by the United Nations and containing specific metadata describing each indicator and of some of the project deliverables submitted by the project partners and concerning the specific themes the indicators deal with. Knowledge about the identified indicators has been modelled in the ontology thanks to the support of the abovementioned partners, who provided 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 and will be provided by domain experts. Indeed, their involvement, both from a technical and from a domain point of view, is mandatory for the development of such a system. Furthermore, the chosen indicators represent a case study also for other experimentations within the project, therefore we have been able to collect sufficient information about them and in the coming future it will be possible to test the ontological model in the VLab by running workflows.

Apart from the indicator analysis, the conceptual model is being specialized and enriched also by the representation of concepts related to the Essential Variables which can be classified as “Urban”. They are being identified and proposed by some of the project partners within their activities in the different WPs of the project itself. Thus, even in this case, the analysed documents are represented by deliverables, in which these variables are properly described. As shown in the above Figure 7, some of these Essential Urban Variables are Built-up area; Human Comfort; Land Consumption; Land cover; Land surface temperature; Land use; Population density; Spatial Distribution of regular Migrant population (SDM); Urban open space.

Indicator 11.1.1 is defined as the “Proportion of urban population living in slums, informal settlements or inadequate housing”. As anticipated, most concepts have been identified through the analysis of the indicator related metadata described in the document provided by the UN[29].

The following images, extracted from the Protégé tool, show the information currently available in the model regarding Indicator 11.1.1:

- it has been linked to various other indicators, some of which refer to other SDGs (e.g. Indicator 11.1.1 *isRelatedTo* Indicator 6.6.1 “Proportion of population using safely managed drinking water services”) (Figure 8);

[29] https://unstats.un.org/sdgs/metadata/files/Metadata-11-01-01.pdf

- the corresponding Target has been specified (Indicator 11.1.1 measures Target 11.1);
- the relationship between Indicator, Target and Goal has been formalized (Figure 9);
- information about the urban areas considered by the indicator have been included and defined (Figure 10 and Figure 11);

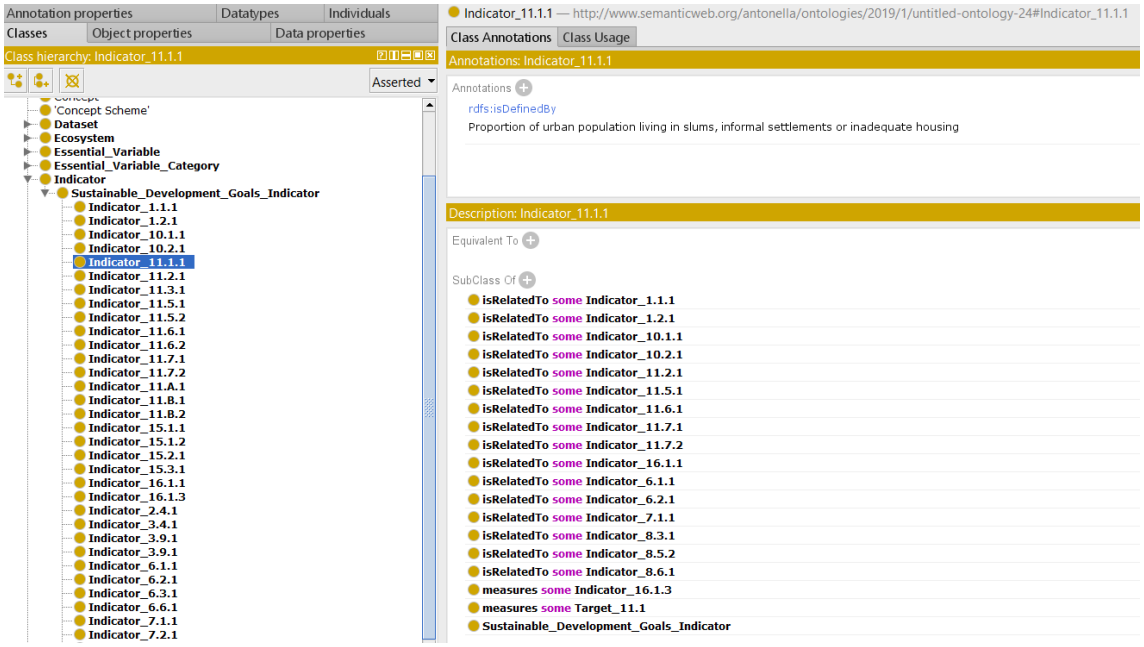

Figure 8. Relations between Indicator 11.1.1 and other ones

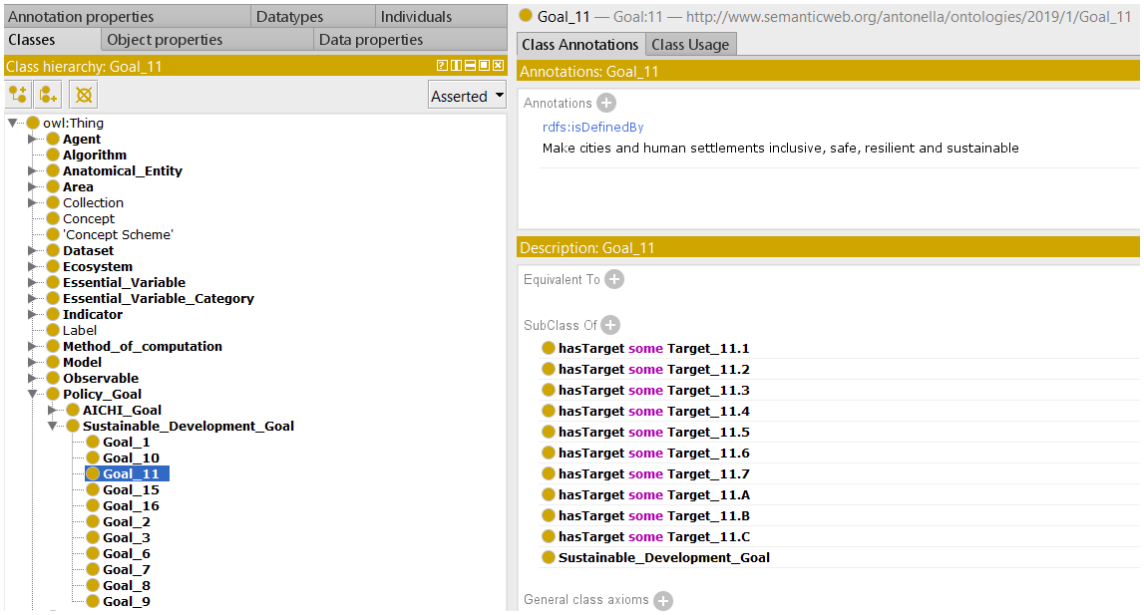

Figure 9. Relations between Goals and Targets

Figure 10. Description of "Informal settlement" concept

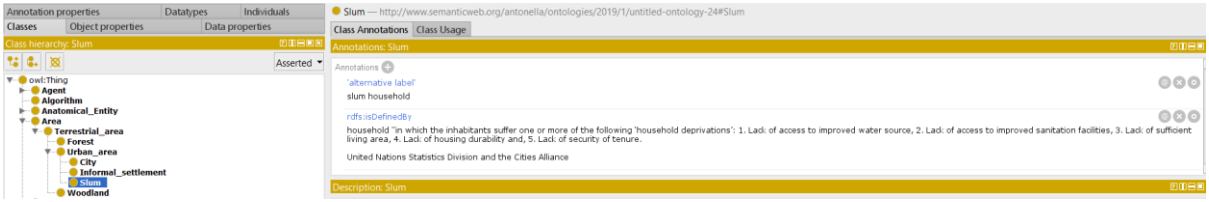

Figure 11. Description of "Slum" concept

Indicator 11.6.2 is defined as the "Annual mean levels of fine particulate matter (e.g. PM2.5 and PM10) in cities (population weighted)"[30]. As for the previous indicator, it has been linked to various other indicators, the corresponding Target has been specified (Indicator 11.6.2 measures Target 11.6), the relationship between Indicator, Target and Goal has been formalized. Furthermore, information about pollutants and health effects have been structured, as well as datasets used, as shown in Figure 12:

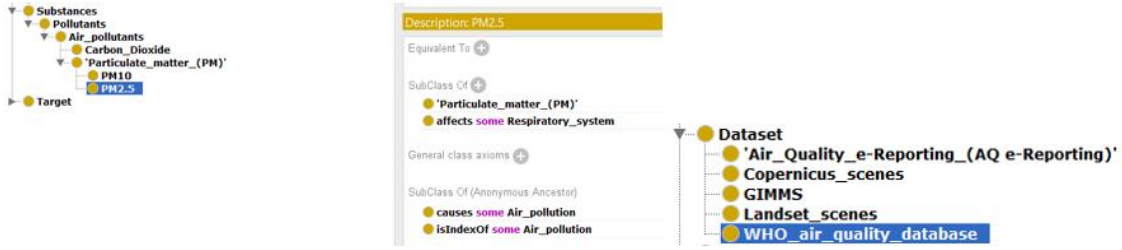

Figure 12. Relation between pollutants and health effects

Indicator 11.3.1 is defined as the "Ratio of land consumption rate to population growth rate"[31]. Also for this indicator information about the links to other indicators and to the corresponding Target and Goal has been structured. More specific information

[30] https://unstats.un.org/sdgs/metadata/files/Metadata-11-06-02.pdf
[31] https://unstats.un.org/sdgs/metadata/files/Metadata-11-03-01.pdf

concerning the population and the areas where people live, as shown in Figure 13, have been included:

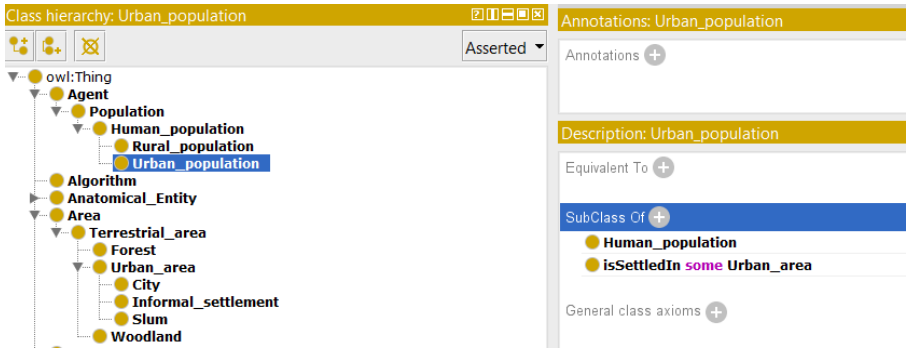

Figure 13. "Urban population" concept

In the following few months the ontological model will be improved, finalized and validated with the help of domain experts.

5.2 Mapping towards other vocabularies

In order to guarantee the semantic interoperability with other databases or platforms used by the EO community of experts, in particular with GEOSS, the activity of ontology development is accompanied by the establishment of mappings, i.e. semantic correspondences, between concepts included in the ontology and concepts coming from the existing vocabularies illustrated in previous sections[32]. The main reference vocabulary is represented by GEMET, considered both its wide use by experts and the already established mappings towards some of the other resources taken into consideration. However, this choice does not prevent from considering other vocabularies as well. Mappings are being included directly within the ontology, through the import of the SKOS model into the Protégé tool.

SKOS (Simple Knowledge Organization Systems) 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 RDF 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"[33]. 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

[32] Mapping: "process of establishing relationships between the concepts of one vocabulary and those of another", or "relationship between a concept in one vocabulary and one or more concepts in another", ISO 25964-2:2013, p. 7.
[33] https://www.w3.org/TR/skos-primer/#secmapping

schemes by connecting concepts coming from each one of them. This possibility of creating a network between existing and/or new resources is fundamental for our purposes, in particular for guaranteeing information retrieval processes based on several semantically related KOSs. 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 express mappings[34]:

- skos:closeMatch: indicates that two concepts can be considered similar and interchangeable in both the concept schemes they belong to;
- skos:exactMatch: it is a sub-property of skos:closeMatch because it indicates a higher degree of closeness between concepts having equivalent meaning;
- skos:broadMatch: is used to assert that one concept is broader in meaning than another;
- skos:narrowMatch: is used to assert the inverse, namely when one concept is narrower in meaning (i.e. more specific) than another;
- skos:relatedMatch: asserts an associative relationship between two concepts.

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.

The SKOS model imported in Protégé allows the definition of such mappings through the use of Annotations, as illustrated in Figure 14:

[34] https://www.w3.org/TR/skos-primer/#sechierarchy

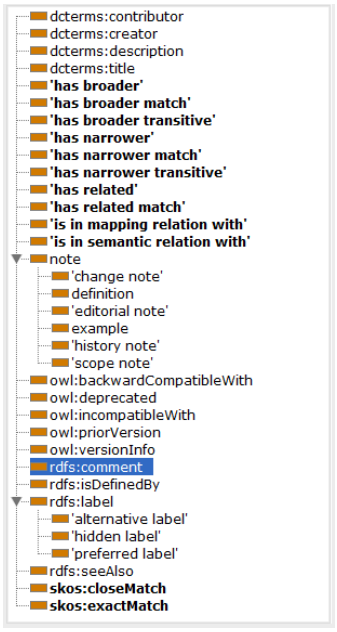

Figure 14. SKOS Annotations

Currently, most of the established mappings are represented by exact matches. Some of them, indeed, are easier to detect because the concepts involved share the same lexical aspects. Figure 15 shows an example of exactMatch between the concept *Air Pollution* and the same concept taken from GEMET and identified by a unique URI:

Figure 15. Ontology-GEMET matches

Other exactMatches with GEMET involve concepts such as *Land use, Built-up area, Population density, Evapotranspiration, Land cover*, *Indoor air pollution, Outdoor air pollution*, etc.

Further effort is required in the recognition of exact matches between lexically different terms and of close, broader, narrower and related mappings. In order to correctly identify these correspondences a dt analysis of the existing terminologies is required in

order to understand if concepts within them are used with the same meaning of that identified in the ontology under development. Examples of other kinds of matching are:

- *Forest* skos:closeMatch with *Forested area* from ENVO ontology;
- *Human comfort* skos:closeMatch with *Human well-being* from GEMET;
- *Cropland* skos:relatedMatch with *Cropland management* from GEMET;
- *Soil temperature* skos:broaderMatch with *Temperature* from GEMET.

Concerning the mapping activity, it is important to underline that some of the existing vocabularies, thanks to the partial semantic overlapping among them, are already mapped to each other in order to allow federated access to information: 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. As can be seen in Figure 16 below, GEMET is aligned with other existing vocabularies such as the AGROVOC, EUROVOC and UMTHES thesauri. The figure includes examples of different matching types: 'urban area' has an exact match with AGROVOC and EUROVOC 'urban areas', since the three concepts have a fully equivalent meaning. Furthermore, it has a close match with UMTHES 'Stadtgebiet', in other words, their meaning is partially equivalent.

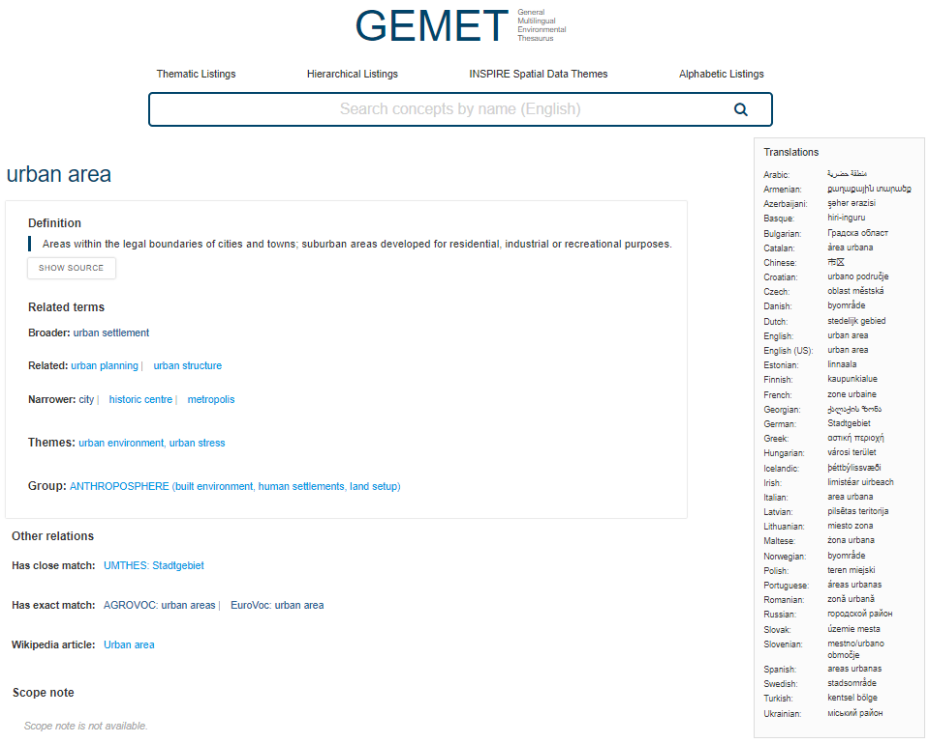

Figure 16. GEMET Mappings

In the same way, the EARTh Thesaurus is aligned with other vocabularies as illustrated by the quantitative information recapitulated in Figure 17:

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 17. EARTh Mappings

6. Semantic services in SMURBS Knowledge Platform

As already stated, the users of the 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 SMURBS 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. Moreover, the organization of the knowledge domain in an ontological model will allow advanced discovery and modelling services for answering complex queries.

References

Abid T., Zarzour H., Laouar M.R., *et al.*, *Towards a smart city ontology*, in "2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)", ISSN: 2161-5330, DOI: 10.1109/AICCSA.2016.7945823.

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.

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.

Bucăţa G., Rizescu M.A., *Improving the quality and efficiency of higher education systems based on the knowledge-management approach*, in "Proceedings of the International Conference on Knowledge-Based Organization", vol. XXV, n. 1, 2019, pp. 199-205, DOI: 10.2478/kbo-2019-0032.

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.

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., *T2K²: 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.

Fox M. S., *A Foundation Ontology for Global City Indicators*, Technical Report, August 2013.

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

Komninos N., Bratsas C., Kakderi C., et al., Smart city ontologies: Improving the effectiveness of smart city applications, in «Journal of Smart Cities», September 2015, DOI: 10.18063/JSC.2015.01.001.

Jennex M.E., *Big Data, the Internet of Things, and the Revised Knowledge Pyramid*, in «ACM SIGMIS Database», vol. XLVIII, n. 4, pp. 69-79, doi: 10.1145/3158421.3158427.

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.

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.

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.

END OF DOCUMENT