



NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE

DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

POSTGRADUATE STUDIES

“ADVANCED INFORMATION SYSTEMS”

MASTER THESIS

**Semantic Search and Discovery for Earth Observation
Products using Ontology Services**

Maria I. Karpathiotaki

Supervisor: Manolis Koubarakis, Professor

ATHENS

SEPTEMBER 2014

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Σημασιολογική Αναζήτηση Δεδομένων από την Παρατήρηση της Γης
με χρήση Οντολογιών**

Μαρία Ι. Καρπαθιωτάκη

A.M.: M1279

ΕΠΙΒΛΕΠΩΝ :

Μανόλης Κουμπαράκης, Καθηγητής ΕΚΠΑ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Ευστάθιος Χατζευθυμιάδης, Αναπληρωτής Καθηγητής ΕΚΠΑ

**ΑΘΗΝΑ
ΣΕΠΤΕΜΒΡΙΟΣ 2014**

Περίληψη

Η πρόσβαση σε δεδομένα που έχουν προέλθει από την παρατήρηση της Γης παραμένει δύσκολη για τους περισσότερους απλούς χρήστες μέχρι και σήμερα. Οι υπάρχουσες μηχανές αναζήτησης απευθύνονται σε ειδικούς του πεδίου παρατήρησης της Γης, αδυνατώντας να καλύψουν τις ανάγκες επιστημονικών κοινοτήτων από άλλα πεδία, καθώς και απλών χρηστών που δεν είναι εξοικιωμένοι με τα δεδομένα παρατήρησης της Γης. Στα πλαίσια αυτής της διπλωματικής αναπτύχθηκαν σημασιολογικές τεχνολογίες οι οποίες ενσωματώθηκαν σε μια πλατφόρμα αναζήτησης EO-netCDF δεδομένων. Οι τεχνολογίες αυτές με τη χρήση οντολογιών επιτρέπουν την εύκολη αναζήτηση και πρόσβαση σε δεδομένα που έχουν προέλθει από την παρατήρηση της Γης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Σημασιολογικός Ιστός

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ : σημασιολογική αναζήτηση, δεδομένα παρατήρησης της γης, EO-netCDF, οντολογίες, συσχέτιση οντολογιών

Abstract

Access to Earth Observation products remains difficult for end-users in most domains. Although various search engines have been developed, these are targeted for advanced Earth Observation users, and fail to support scientific communities from other domains, as well as casual users not familiar with the concepts of Earth Observation. In the context of this thesis, we developed semantic technologies that were used to semantically enhance a search engine for EO-netCDF product. We present how these technologies utilize ontology services to substantially improve the ability of end-users to explore, understand and exploit the vast amount of Earth Observation data that is available nowadays.

SUBJECT AREA: Semantic Web

KEYWORDS: semantic search, earth observation products, EO-netCDF, ontologies, ontology matching

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή μου και επιβλέποντα αυτής της διπλωματικής, Μανόλη Κουμπάρακη, που με την καθοδήγηση και τη συνεχή υποστήριξή του έκανε δυνατή την εκπόνηση της εργασίας αυτής. Επίσης, θα ήθελα να ευχαριστήσω ιδιαίτερα την συνάδερφό μου Καλλιρρόη Δογάνη, καθώς από την άριστη συνεργασία μας προέκυψε η παρούσα διπλωματική εργασία. Θα πρέπει, ακόμη, να ευχαριστήσω τον συνεργάτη μου Bernard Valentin, από την εταιρεία Space Applications Services, για τις πολύτιμες συμβουλές του και την πολύ καλή συνεργασία που είχαμε. Παράλληλα, θέλω να ευχαριστήσω τον Δρ. Κωστή Κυζηράκο για την βοήθεια και συμβολή του στο κομμάτι της διπλωματικής εργασίας που περιλαμβάνεται στο Κεφαλαίου 8. Τέλος, θα πρέπει να ευχαριστήσω τους συναδέλφους μου, και μέλη της ερευνητικής ομάδας που ανήκω, για την αγαστή συνεργασία που έχουμε καθώς και την συνεχή υποστήριξή τους.

Μαρία Καρπαθιωτάκη

Αθήνα, Σεπτέμβριος 2014

Contents

1	Introduction	23
2	A General Overview of Earth Observation Practices	25
2.1	Basic Concepts	25
2.2	European and Global Initiatives	27
2.2.1	Copernicus	27
2.2.2	GEOSS	29
2.2.3	INSPIRE	31
2.2.4	Heterogeneous Missions Accessibility	32
2.3	Standards Organisations	33
2.3.1	ISO	33
2.3.2	OGC	34
2.3.3	W3C	35
2.4	EO Metadata Profile of Observations and Measurements	35
2.4.1	General Concepts	36
2.4.2	Observation and Measurements	37
2.4.3	EO Metadata Mapping on Observation and Measurements	38
2.5	Semantic Annotations	39
2.6	Summary	42
3	Technical Background	43
3.1	Vocabularies and Ontologies	43
3.1.1	Controlled Vocabularies	43
3.1.2	Taxonomies	44
3.1.3	Thesauri	44

3.1.4	Ontologies	46
3.1.5	SKOS and OWL	46
3.2	Data Formats and Models	48
3.2.1	NetCDF	49
3.2.2	NetCDF Conventions	54
3.2.3	SAFE	57
3.3	Related Technologies	61
3.3.1	OpenSearch in Earth Observation	61
3.3.2	GI-cat	66
3.3.3	The GEOSS Discovery and Access Broker	68
3.3.4	The Spatiotemporal RDF Store Strabon	68
3.4	Summary	69
4	Related Activities	71
4.1	RARE	71
4.2	SMADD	73
4.3	OTE/OTEG	76
4.4	RESTo	80
4.5	Summary	81
5	Semantic Search and Discovery of EO-netCDF Products	83
5.1	EO-netCDF	83
5.1.1	Conventions	83
5.1.2	Libraries	90
5.2	A Semantically Enabled Search Platform	96
5.2.1	Architecture	97
5.2.2	Query Disambiguation	99
5.3	Summary	104

6 Enhancing EO Ontology Services	105
6.1 Ontologies	105
6.1.1 CSCDA	105
6.1.2 GEMET	108
6.1.3 NASA GCMD	111
6.1.4 GEOSS EO Vocabulary	112
6.2 Ontology Matching	113
6.2.1 Ontology Matching Techniques	113
6.2.2 Matching systems	116
6.2.3 Pythia	117
6.3 Ontology Navigation	127
6.3.1 The Cross-Ontology Browser	127
6.3.2 Supported Types of Visualization	127
6.4 From Ontologies to EO-netCDF	129
6.4.1 The EO-netCDF Resources Reasoner	129
6.5 Summary	131
7 Demonstration	133
7.1 Use Cases	133
7.1.1 Envisat	133
7.1.2 SENTINEL-1	136
7.1.3 MyOcean	138
7.2 Search Scenarios	140
7.2.1 Free Text Search	141
7.2.2 Term-based Search	144
7.2.3 EO-related Search	145
7.3 Summary	147
8 Accessing FedEO Clearinghouse	149

8.1	Federated EO Missions Support Environment	149
8.2	Demonstration	150
8.2.1	NUTS dataset	153
8.2.2	Queries and Results	154
8.2.3	Conclusions	172
8.3	Summary	173
9	Conclusions and Future Work	175
9.1	Conclusions	175
9.2	Future Work	175

List of Figures

2.1	The basic concepts of Earth observation	27
2.2	A layered view of O&M EO Products data	37
2.3	Semantic annotation of EO dataset series metadata	41
3.1	A hierarchy example for a taxonomy	44
3.2	An example of a thesaurus	45
3.3	The SKOS data model	47
3.4	The “Classic” netCDF Data Model	51
3.5	The netCDF-4 Data Model	54
3.6	SAFE Information Model	58
3.7	SAFE Logical Model	60
3.8	SAFE Physical Model	60
3.9	An example of a simple OSDD	61
3.10	Supported profilers and accessors by GI-cat	67
4.1	RARE architecture	72
4.2	Resource discovery using ontologies on top of portals	74
4.3	Example of OTE software prototype web page	77
4.4	A little excerpt from GSCDA Multi-Domain Thesaurus	78
4.5	OTEG web client showing results for the term “ocean”	79
4.6	RESto architecture	80
4.7	A number of search results for the query “images of urban area in France acquired in 2013 with less than 25 % of cloud cover”	82
5.1	The Concept of EO-netCDF	84
5.2	Earth Observation Information group	85

5.3	Earth Observation Equipment group	85
5.4	Sensor Information group	85
5.5	SENTINEL-1 EO elements	86
5.6	SENTINEL-1 sensor information elements	86
5.7	EO-netCDF implementation example	89
5.8	Example display of the ncview browser	95
5.9	Plot for air temperature in Panoply	95
5.10	An EO-netCDF file in Panoply	96
5.11	ProdTrees architecture	97
5.12	The lifecycle of the “agriculture brussels 2012” query	98
6.1	CSCDA Multi-Domain Thesaurus (levels 0 and 1)	106
6.2	Multi-Domain Thesaurus structure	107
6.3	The 5-level classification of Earth Science Data in GCMD	112
6.4	Deriving a mapping using a pre-existing <i>skos:exactMatch</i> mapping	122
6.5	Deriving a mapping using pre-existing <i>skos:broadMatch</i> (top) or <i>skos:narrowMatch</i> (bottom) mappings	123
6.6	An example of the creation of a <i>skos:narrowMatch</i> mapping	124
6.7	Deriving a mapping using a pre-existing <i>skos:broadMatch</i> mapping is not possible in this case	124
6.8	Browsing the concepts of CSCDA	128
7.1	Sentinel-1A radar acquisition from 22 April 2014 showing Greece’s Attica region, with mountainous areas and the capital and largest city of Athens near the centre. In the water, different shades of blue indicate different types of sea surface, influenced by currents and waves. <i>Copyright ESA</i>	137
7.2	The 7 areas covered by the MyOcean services	139
7.3	The Web interface of the ProdTrees platform	140
7.4	The default interpretation for the keyword “water”	141
7.5	The different interpretations for the keyword “water”	141

7.6	The results	142
7.7	The filtered results	143
7.8	The details of GEOSS concept AGRICULTURE	144
7.9	Multi-Criteria Search page	145
7.10	The EO-netCDF Model Browser	146
7.11	Search query with bounding box and specified EO-netCDF parameter	146
7.12	The results of the query of Figure 7.11	147
8.1	URL template for collection search in FedEO	150
8.2	Results page for collection search in FedEO	151
8.3	Results page for product search in FedEO	151
8.4	Part from Explain Document of FedEO displaying supported collections of ESA M2CS EO-DAIL	152
8.5	The four levels of NUTS	153

List of Tables

2.1	Observations and Measurements properties mapping within the Earth Observation context	38
3.1	An example of a netCDF dataset in CDL notation	53
3.2	OpenSearch parameters for Geo extension	62
3.3	OpenSearch parameters for Time extension	63
3.4	A number of OpenSearch parameters for collection search	64
3.5	A number of OpenSearch parameters for product search	65
6.1	GEMET Super-groups and Groups	109
6.2	Mappings created by the terminological matcher	120
6.3	Mappings created by the string-based technique	126
6.4	Mappings created by the language-based technique	126
6.5	Mappings created by the graph-based technique	126

Preface

This thesis was conducted as the last assignment of the author towards the completion of the postgraduate program “Advanced Information Systems”, of the Department of Informatics and Telecommunications, of the National and Kapodistrian University of Athens. The work described in this thesis has been motivated, designed, developed and evaluated in the research project ProdTrees¹, in which I was involved since June 2013. In the context of the project ProdTrees, I was cooperating with my colleague Kallirroï Dogani. Therefore, part of her work [10] is also included in this thesis for completeness.

This work has been demonstrated in the demo session of the 14th European Semantic Web Conference. The demonstration was entitled “ProdTrees: Semantic Search for Earth Observation Products”. Also, it was presented in the 8th International Conference On Web Reasoning And Rule Systems with title “Semantic Search for Earth Observation Products using Ontology Services”. Finally, parts of this work will be presented in the poster session of the Ontology Matching Workshop of the 13th International Semantic Web Conference.

¹ProdTrees was funded by European Space Agency.

Chapter 1

Introduction

The demand for aerial and satellite imagery, and products derived from them has been increasing over the years, in parallel with technological advances that allow producing a bigger variety of data with an increasing quality and accuracy. As a consequence of these advances, and the multiplication of deployed sensors, the amount of Earth Observation (EO) data collected and stored has exploded.

However, access to EO products remains difficult for end users in most scientific domains. Various search engines for EO products, generally accessible through Web portals, have been developed. For example, see the interfaces offered by the European Space Agency portal for accessing data of Copernicus, the new satellite programme of the European Union¹ or the EOWEB portal of the German Aerospace Center (DLR)². Typically, these search engines allow searching for EO products by selecting some high level categories (e.g., the mission from which the product was generated, the satellite instrument that was used etc.) and specifying basic geographical and temporal filtering criteria. Although this might suit the needs of very advanced users that know exactly what dataset they are looking for, other scientific communities or the general public require more application-oriented means to find EO products.

The main objective of this thesis was to develop semantic technologies used in the ProdTrees platform, a semantically-enabled search engine for EO products. The implementation of this platform was developed during the project ProdTrees funded by the European Space Agency³. The ProdTrees platform uses semantic technologies to allow users to search for EO products in an application-oriented way using free-text keywords (as in search engines like Google), their own domain terms or both, in conjunction with the well-known interfaces already available for expert users.

The semantic technologies developed in ProdTrees, covered by this thesis in conjunction with [10], belong to three main parts: i) an ontology matching system that creates mappings between ontologies in order to overcome the heterogeneity that arises among

¹<http://gmesdata.esa.int/web/gsc/home>

²<https://centaurus.caf.dlr.de:8443/eoweb-ng/template/default/welcome/entryPage.vm>

³<http://www.esa.int/ESA>

them, ii) a cross-ontology browser that can be used by the users in the query creation phase, as a disambiguation and discovery tool, and, finally, iii) a reasoner, responsible for translating the selected ontology terms to specific EO search criteria.

This thesis provides also a comprehensive overview of the terminology, standards, practices and technologies used in the EO domain. This will help readers to have a better understanding of the concepts of Earth Observation.

In the context of this thesis, we also created a demo to show how we can access the repository of ESA with EO catalogues, named FedEO Clearinghouse, for obtaining EO data. The demo shows also how this data can be combined with linked open data (structured data that can be interlinked and become more useful).

The rest of the thesis is structured as follows. Chapter 2 summarises the basic ideas and standards on which EO technologies are based. Chapter 3 provides the technical background used for the implementation of the ProdTrees platform and Chapter 4 describes previous work related either to EO search or semantic technologies in EO domain. In Chapter 5, we first present a new standard called EO-netCDF⁴ for accessing EO products annotated with netCDF⁵. Next, we describe how the ProdTrees platform implements the EO search based on this standard and what is the role of each component in the system. Chapter 6 contains a detailed description of the semantic technologies we developed during ProdTrees and Chapter 7 demonstrates all the use cases and search scenarios covered by the ProdTrees platform. Chapter 8 is focused on the FedEO demo and, finally, in Chapter 9 you will find some conclusions and thoughts for future work.

⁴EO-netCDF is expected to be submitted to OGC

⁵A well-known standard consisting of set of self-describing, machine-independent data formats and software libraries that support the creation, access, and sharing of array-oriented scientific data. <http://www.unidata.ucar.edu/software/netcdf/>

Chapter 2

A General Overview of Earth Observation Practices

This chapter offers a clear view of the earth observation domain by describing EO terms, initiatives and standards. What is the difference between EO products and EO collections? What is Copernicus and what is the main objective of GEOSS? What standards are used to describe the EO data and why do we need the semantic annotations? All these questions will be answered in the following sections.

2.1 Basic Concepts

The major sources of information in Earth observation systems are satellites. A **satellite** is basically a '[hu]man-made object (such as a spacecraft) placed in orbit around Earth, another planet or the Sun'¹. Because of their orbits, satellites permit repetitive coverage of the Earth's surface on a continuing basis. The satellite itself is also one possible carrier of instruments, called a **platform**. Apart from EO satellite platforms, there are also other types of platforms such as unmanned aircraft vehicles.

An **instrument** is a technical entity that contains detectors, also called sensors. Sensors, in general, are devices that respond to a physical stimulus (as heat, light, sound, pressure, magnetism, or a particular motion) and then react to it in a particular way. In the EO context, **sensors** on board a satellite are typically radiometers and cameras that provide images (here, datasets consisting of a grid of values), but also active sensors like radar sensors or sounders.

There are two large categories of satellites depending on some of the orbit characteristics (size, shape and inclination), sun-synchronous and geostationary satellites. **Sun-synchronous** satellites have a geocentric orbit which combines altitude and inclination in such a way that an object on that orbit ascends or descends over any given Earth latitude at the same local mean solar time. Sun-synchronous orbits permit exploitation of the sun's illumination, and the time to revisit the same point on Earth spans two to four

¹http://www.esa.int/Our_Activities/Space_Science/S

weeks. **Geostationary** satellites have an orbit whose position in sky remains the same for a stationary observer on Earth. They fly at an altitude of 36 000 km above the Earth's equator and follow the direction of the Earth's rotation.

Any observation obtained by satellite instruments can be referred as **EO dataset** or **EO product**. Products may include a wide range of items from single images to huge datasets (e.g. wide coverage, continuous or periodic monitoring, etc.). **EO collections** are collections of datasets sharing the same product specification. These collections are also called **EO data series**. An EO collection typically corresponds to a series of EO datasets derived from data acquired:

- Either from an instrument in a dedicated mode on board a single satellite platform; or
- by a series of instruments, possibly from different satellite platforms, but in this case working in the same instrument mode.

However, other kinds of criteria, such as range of resolution or product quality, can be used to group products into data series. For example, there are snow collections or cloud-free collections.

According to the Oxford dictionary, a mission in general is 'an expedition into space'². When this term concerns satellites (**EO mission**), it is used to describe the whole set of technologies, devices and software components in space and on Earth that accompany a satellite across all of its life-cycle phases. All EO missions play a vital role in systematically generating, preserving and giving access to long- term EO datasets.

Figure 2.1 displays the basic concepts of the EO domain described above.

Another item in the EO domain that is worth mentioning is that EO products can be characterised by the characteristics of the observation payloads that generated the product. The payload carries the instrument that observes Earth. In the following, we limit our description to the EO optical and radar instruments (but there are also other kinds of instruments, e.g altimetric, atmospheric). **Optical** (OPT) imagers are amongst the most common instruments used for Earth observation. In this case, the payload is a passive sensor which detects the electromagnetic radiation originated by the Sun and reflected by Earth. Optical missions are affected by the presence of clouds and therefore important information (metadata) associated with the product is the percentage of cloud cover. Unlike optical systems that rely on reflected solar radiation or thermal radiation

²http://oxforddictionaries.com/view/entry/m_en_gb0525350#m_en_gb0525350

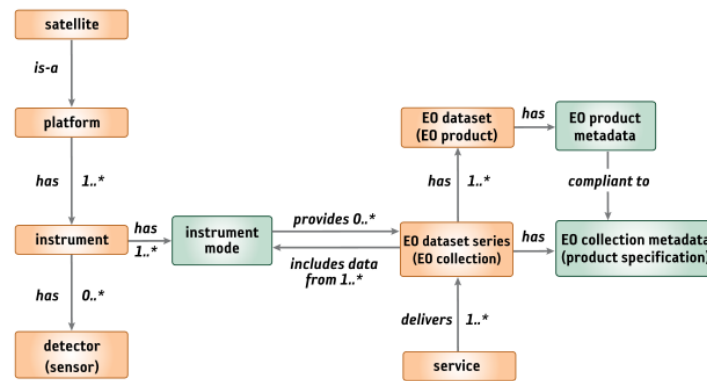


Figure 2.1: The basic concepts of Earth observation

emitted by Earth, imaging **radar** (SAR) instruments work independently of light and heat. Radar is an active system that transmits a beam of radiation in the microwave region of the electromagnetic spectrum. The active nature of the SAR payload enables the development of an EO product that is practically immune from cloud and other atmospheric effects and that works at night as well as in daylight.

2.2 European and Global Initiatives

The following description presents European and global initiatives such as Copernicus, INSPIRE, GEOSS and HMA.

2.2.1 Copernicus

Copernicus³, previously known as GMES (Global Monitoring for Environment and Security), is the European Programme for the establishment of a European capacity for Earth Observation. Its main objective is to provide, on a sustained and operational basis, reliable and timely services related to environmental and security issues in support of public policy needs. It is coordinated and managed by the European Commission (EC). The development of the observation infrastructure is performed under the aegis of the European Space Agency⁴ (ESA) for the space component, whereas the activities about the in situ observation are implemented by the European Environment Agency⁵ (EEA) and the EU Member States.

³<http://www.copernicus.eu/>

⁴<http://www.esa.int/>

⁵<http://www.eea.europa.eu/>

Copernicus consists of a complex set of systems which collect data from multiple sources. More precisely, there are three categories of input data:

- Space observation data provided by satellite missions combined to form a GMES Space Component (GSC). The GSC is co-funded by ESA and the EC under a specific delegation agreement. The GSC also integrates data from other international or national contributing space missions and will provide these data through the “GSC Data Access”TM (GSCDA) component.
- In situ observation data provided by a network of observation infrastructures (in situ sensors such as ground stations, airborne and sea-borne sensors). These networks are typically owned and governed by the EU Member States. The homogeneous and sustainable provision of these data poses a considerable challenge, which is being tackled under the leadership of the European Environment Agency (EAA).
- Reference data, which fulfil a specific and complementary role compared with observation data. These data include topographic data (road networks, hydrography, digital elevation models, etc.) and data such as geological maps.

Copernicus processes these data and provides users with reliable and up-to-date information through a set of services related to environmental and security issues. The services address six thematic areas:

- *Land Monitoring.* This service provides geographical information on land cover and on variables related, for instance, to the vegetation state or the water cycle. It supports applications in a variety of domains such as spatial planning, forest management, water management, agriculture and food security, etc. More information on the Copernicus land monitoring service is available on the `land.copernicus.eu` webpage.
- *Marine Monitoring.* This service provides regular and systematic reference information on the state of the physical oceans and regional seas. The observations and forecasts produced by the service support all marine applications. For instance, the provision of data on currents, winds and sea ice help to improve ship routing services, offshore operations or search and rescue operations, thus contributing to marine safety. The service is currently delivered in a pre-operational mode and is provided through the EU-funded project MyOcean2⁶.

⁶<http://www.myocean.eu.org/>

- *Atmosphere Monitoring.* This service provides continuous data and information on atmospheric composition. The service describes the current situation, forecasts the situation a few days ahead, and analyses consistently retrospective data records for recent years. The service is delivered in a pre-operational mode and its products are provided free of charge through the `atmosphere.copernicus.eu` webportal, which is operated by the EU-funded project MACC-II⁷.
- *Emergency Management.* This service provides all actors involved in the management of natural disasters, man-made emergency situations, and humanitarian crises with timely and accurate geo-spatial information derived from satellite remote sensing and completed by available in situ or open data sources. More information can be found on the EFAS⁸ portal.
- *Security.* This service aims to support the related European Union policies in the following priority: border surveillance, maritime surveillance and support to EU External Action. The service is still in a development phase.
- *Climate Change.* This service will give access to information for monitoring and predicting climate change and will, therefore, help to support adaptation and mitigation. It benefits from a sustained network of in situ and satellite-based observations, re-analysis of the Earth climate and modelling scenarios, based on a variety of climate projections. The pre-operational phase of the Copernicus Climate Change service is supported by a series of projects⁹ launched under the 2013 FP7 Space call related to climate modelling and observation analyses.

2.2.2 GEOSS

The Global Earth Observation System of Systems¹⁰ (GEOSS) is an intergovernmental programme, built by the Group on Earth Observations¹¹ (GEO). It is a 10-year global programme running from 2005 to 2015. GEOSS aims to connect the producers of environmental data and decision-support tools with the end users of these products in order to enhance the relevance of Earth observations to global issues. The result is to be a global

⁷<http://www.gmes-atmosphere.eu/>

⁸<https://www.efas.eu/>

⁹<http://www.copernicus.eu/pages-principales/projects/other-fp7-projects/climate-change/>

¹⁰<https://www.earthobservations.org>

¹¹https://www.earthobservations.org/about_geo.shtml

public infrastructure that generates comprehensive, near-real-time environmental data, information and analyses for a wide range of users.

As of May 2011, 86 countries¹², the EC and 61 organisations¹³ participated in the GEOSS work plan. GEOSS wants to integrate EO systems into a global system, leading to a system-of-systems approach, that can be applied to various areas of environmental science and management. Main enabler of the System of Systems principles is the GEOSS Common Infrastructure (GCI), through which GEOSS resources, including Earth observation data (satellite, airborne, in situ, models), information services, standards and best practices, can be searched, discovered and accessed by scientists, policy leaders, decision makers, and those who develop and provide information services across the entire spectrum of users.

In the context of the GCI, GEO is developing the GEOPortal as a single Internet gateway to the data produced by GEOSS. The purpose of GEOPortal¹⁴ is to make it easier to integrate diverse data sets, identify relevant data and portals of contributing systems, and access models and other decision-support tools. For users without good access to high-speed internet, GEO has established GEONETCast¹⁵, a system of four communications satellites that transmit data to low-cost receiving stations maintained by the users. The GEONETCast toolbox¹⁶ has been made available and contains tools to access some radar altimetry, vegetation, satellite prediction and maritime information.

The GEOSS work plan focuses on the following nine ‘societal benefit areas’ (SBAs), also called GEOSS themes:

1. reduction and prevention of disasters
2. human health and epidemiology
3. energy management
4. climate change
5. water management
6. weather forecasting
7. ecosystems

¹²https://www.earthobservations.org/ag_members.shtml

¹³https://www.earthobservations.org/ag_partorg.shtml

¹⁴http://www.geoportal.org/web/guest/geo_home_stp

¹⁵<https://www.earthobservations.org/geonetcast.shtml>

¹⁶<http://www.itc.nl/Pub/Organization/Geonetcast-Toolbox.html>

8. agriculture

9. biodiversity

Interoperability arrangements ensure that the heterogeneous systems within GEOSS can communicate and operate. Data, information and service providers within GEOSS are guided by technical specifications for collecting, processing, storing and disseminating shared data, metadata and products. Interoperability arrangements in GEOSS are based on open standards, with a preference for formal international standards. The architecture of an Earth observation system refers to the way in which its components are designed so that they function as a whole.

2.2.3 INSPIRE

INSPIRE¹⁷ (Infrastructure for Spatial Information in the European Community) is a legal instrument of the EC. It is driven by Directive 2007/2/EC of the European Parliament and of the Council. The INSPIRE directive came into force on 15 May 2007 and will be implemented in various stages, with full implementation required by 2019. The INSPIRE directive aims to create a European Union (EU) spatial data infrastructure. This will enable the sharing of environmental spatial information among public sector organisations and better facilitate public access to spatial information across Europe.

The motivation for INSPIRE has been that the general situation on spatial information in Europe is one of fragmentation of datasets and sources, gaps in availability, lack of interoperability or harmonisation between datasets at different geographical scales and duplication of information collection. These problems make it difficult to identify, access and use data that are available. In order to avoid these problems, INSPIRE is based on a number of common principles:

- Data should be collected only once and kept where it can be maintained most effectively.
- It should be possible to combine seamless spatial information from different sources across Europe and share it with many users and applications.
- It should be possible for information collected at one level/scale to be shared with all levels/scales; detailed for thorough investigations, general for strategic purposes.

¹⁷<http://inspire.ec.europa.eu/>

- Geographic information needed for good governance at all levels should be readily and transparently available.
- Easy to find what geographic information is available, how it can be used to meet a particular need, and under which conditions it can be acquired and used.

To ensure that the spatial data infrastructures of the Member States are compatible and usable in a Community and transboundary context, the Directive requires that common Implementing Rules (IR) are adopted in a number of specific areas (Metadata, Data Specifications, Network Services, Data and Service Sharing and Monitoring and Reporting). These IRs are technical arrangements that contribute to the maintenance of a common infrastructure.

The INSPIRE Directive requires the Commission to establish a community geo-portal and the Member States shall provide access to their infrastructures through the geo-portal as well as through any access points they themselves decide to operate. The INSPIRE geoportal¹⁸ provides the means to search for spatial data sets and spatial data services, and subject to access restrictions, to view spatial data sets from the EU Member States within the framework of the INSPIRE Directive.

2.2.4 Heterogeneous Missions Accessibility

The initiative of Heterogeneous Missions Accessibility¹⁹ (HMA) is established by national space agencies, satellite or mission owners and operators, and industry in order to provide harmonised access to data of heterogeneous EO missions. These missions range from national missions up to the ESA Sentinel missions developed within the EU co-funded Copernicus Programme.

Heterogeneous EO missions pose the problem that each of them offers its own method and technology to search for, access to and exploit the mission results in terms of software products, i.e. EO datasets or series of datasets, or images derived from these products. Without a coordinated strategy and harmonised development, the ground segment services will all have different interfaces following the needs and business requirements of the individual stakeholders. While this may not be a problem when accessing EO products from just one mission, it becomes difficult and tedious when EO products are required from multiple missions, or, even worse, when EO products from multiple missions have

¹⁸<http://inspire-geoportal.ec.europa.eu/>

¹⁹ESA has published the manual of HMA (it can be found online at <http://esamultimedia.esa.int/multimedia/publications/TM-21/TM-21.pdf>)

to be combined or processed together in order to provide higher-level services. A client application of one mission cannot call the ground segment services of another mission if their interfaces are not agreed upon.

Hence, HMA project came from the need to find and define a common technological foundation in order to harmonise the ground segment interfaces, or, in the language of software architects, to ensure interoperability between the ground segments. This approach of HMA was initiated by the Ground Segment Coordination Body (GSCB, 2009) and driven by ESA to enable the interoperable use of EO products despite the heterogeneous underlying software and system environments of the individual providers.

2.3 Standards Organisations

A standard is a document that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose. Standardisation is key when aiming at open and interoperable solutions. In the next sections a brief overview is given of the basic standards organisations used in the Earth observation domain.

2.3.1 ISO

The International Organization for Standardization, known as ISO, is an international standard-setting body composed of the national standards institutes of 157 countries, on the basis of one member per country, with a Central Secretariat in Geneva, Switzerland. The ISO Technical Committee TC 211²⁰ (ISO/TC 211 Geographic information/Geomatics) is responsible for the ISO geographic information series of standards. Its work aims to establish a structured set of standards for information concerning objects or phenomena (also called features) that are directly or indirectly associated with a location relative to Earth. These standards specify methods, tools and services for the management of geographic data (including definition and description), acquiring, processing, analyzing, accessing, presenting and transferring such data in digital/electronic form between different users, systems and locations.

The work of ISO TC 211 links to other appropriate (ISO) standards for information technology (IT) and data where possible, and provides a framework for the development of sector-specific applications using geographic data. TC211 develops also the ISO family

²⁰<http://www.isotc211.org/>

19xxx of standards for the field of digital information. Below you will find some examples of the series of ISO 19xxx standards:

- ISO 19107 (Spatial Schema)
- ISO 19111 (Spatial Referencing by Coordinates)
- ISO 19115 (Metadata) and ISO 19139 (Metadata - XML Schema Implementation)
- ISO 19116 (Positioning Services)

2.3.2 OGC

The OGC²¹ (Open Geospatial Consortium) is an international consortium of companies, government agencies, research organisations and universities participating in a consensual process to develop publicly available interface specifications. These specifications support interoperable solutions that “geo-enable”™ the web, wireless and location-based services, and mainstream IT. The specifications empower technology developers to make complex spatial information and services accessible and useful with all kinds of applications. The core mission of OGC is to deliver spatial interface and encoding specifications that are openly and publicly available for global use. This mission is achieved through organising interoperability projects, working towards a consensus, formalising OGC specifications, developing strategic business opportunities and standards partnerships, and promoting demand for interoperable products.

The OGC standards baseline comprises more than 30 standards, including:

- GML - Geography Markup Language
- GeoXACML - Geospatial eXtensible Access Control Markup Language
- KML - Keyhole Markup Language
- Observations and Measurements
- SensorML - Sensor Model Language
- WMS - Web Map Service
- WFS - Web Feature Service

²¹<http://www.opengeospatial.org/>

- GeoSPARQL - Geographic SPARQL Protocol and RDF Query Language

The OGC has a close relationship with ISO/TC 211 (Geographic Information/Geomatics). Volumes from the ISO 19100 series under development by this committee progressively replace the OGC abstract specification. Further, the OGC standards Web Map Service, GML, Web Feature Service, Observations and Measurements, and Simple Features Access have become ISO standards.

2.3.3 W3C

The W3C²² (World Wide Web Consortium) is an international community where member organisations, a full-time staff and the public work together to develop web standards. W3C standards define an Open Web Platform for application development that has the unprecedented potential to enable developers to build rich interactive experiences, powered by vast data stores, that are available on any device. Some of the standards developed by W3C are:

- RDF - Resource Description Framework
- SPARQL
- SKOS - Simple Knowledge Organization System
- RIF -Rule Interchange Format
- SOAP - Simple Object Access Protocol
- HTML - HyperText Markup Language
- XML - Extensible Markup Language

2.4 EO Metadata Profile of Observations and Measurements

The OGC Implementation Standard defines a profile (extension) of Observations and Measurements (O&M) (ISO 19156) [8] for describing Earth observation products (EO products). Although this standard has been developed in the context of the Heterogeneous Mission Accessibility (HMA) project initiated by European Space Agency (ESA), the content

²²<http://www.w3.org/>

is generic to Earth observation product description. The metadata model is structured to follow the different types of products (optical, radar, altimetric, ...) which are not HMA specific. The EO profile of O&M provides a standard schema for encoding Earth observation metadata to support the description and cataloguing of EO products [18].

2.4.1 General Concepts

The general mechanism is to create a schema with a dedicated namespace for each level of specificity from a general description (common to each EO product) to a restricted description (specific mission EO products). Each level of specificity is an extension of the previous one.

The General EO product schema is the main application schema for EO Product metadata. It is associated with the “eop” namespace.

Each Thematic EO product schemas extends the “eop” schema:

- The Optical EO Product schema is used to describe optical products. It is associated with the “opt” namespace.
- The SAR EO Product schema is used to describe radar products. It is associated with the “sar” namespace.
- The Atmospheric EO Product schema is used to describe atmospheric products. It is associated with the “atm” namespace.
- The Altimetry EO Product schema is used to describe altimetry products. It is associated with the “alt” namespace.
- The Limb Looking EO Product schema is used to describe limb looking products. It is associated with the “lmb” namespace.
- The Synthesis and Systematic EO Product schema is used to describe “Synthesis and Systematic” products. It is associated with the “ssp” namespace.

The idea behind this layered levels approach is to create an efficient schema set that describes EO product metadata concentrating on the core metadata characteristics that differentiate an EO product within a collection. Figure 2.2 displays the layered view of O&M EO Products data starting from a general layer of O&M and going to mission specific EO products.

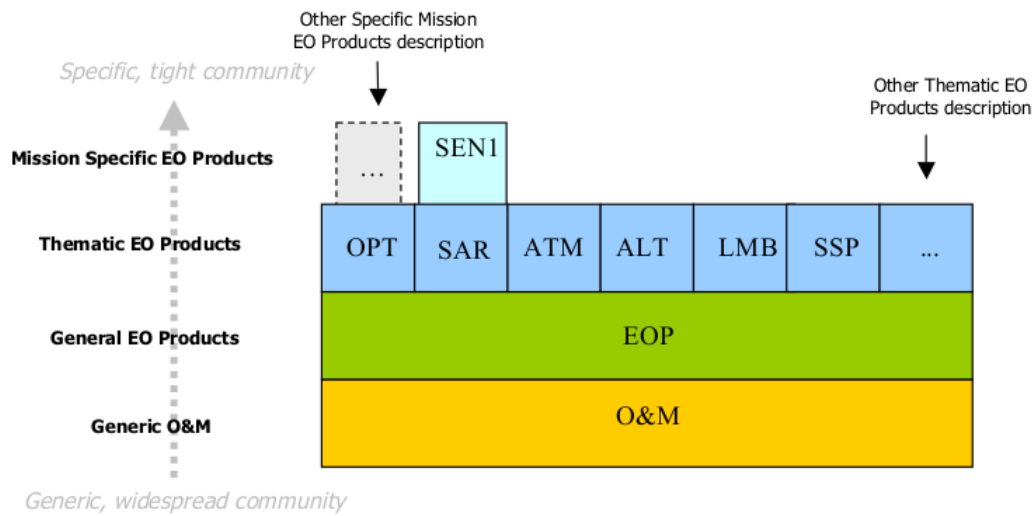


Figure 2.2: A layered view of O&M EO Products data

2.4.2 Observation and Measurements

In this section we will describe the basic characteristics of the O&M model in order to understand how they are extended to satisfy the needs of the EO metadata profile. In natural language, the model states that “An **observation** is an event that estimates an **observed property** of some **feature of interest** using a specified **procedure** and generates a **result**.” Remotely sensed images in the sense of their acquisition can be viewed as observations in which the result of the observation (value of the result property) is a remotely-sensed image product. More precisely, an observation has the following characteristics:

- An **observation** is modelled as a feature type whose instances are created at a specific time point or time period, the **phenomenon time**, i.e. the time when the result applies to the feature of interest. Applied to the EO domain, an observation is the act of acquiring, for example, an image of an observed area on the ground, i.e. the footprint of an acquisition. As this footprint is modelled as a feature of interest, the phenomenon time corresponds to the duration of the acquisition.
- The **observed property** identifies or describes the phenomenon for which the observation result provides an estimated value. It must be a property associated with the type of the feature of interest, e.g. the sea surface temperature if the **feature of interest** is a sea area.
- The **procedure** is the description of a process used to generate the result, i.e. the platform, instrument and detector (sensor) used in the acquisition of the observation,

Table 2.1: Observations and Measurements properties mapping within the Earth Observation context

O&M property	EOP properties	Description
Metadata	eop:EarthObservationMetadata	General properties such as the data identifier, the downlink and archiving information.
phenomenonTime	gml:TimePeriod	The acquisition duration
Procedure	eop:EarthObservationEquipment	The Platform/Instrument/Sensor used for the acquisition and the acquisition parameters (i.e. pointing angles, etc.)
featureOfInterest	eop:Footprint	The observed area (or its projection) on the ground i.e. the footprint of acquisition
Result	eop:EarthObservationResult	The metadata describing the Earth Observation result composed of the browse, mask and product descriptions

or the algorithm applied to a dataset in order to produce a processed result. It must be suitable for the observed property. A **result** of an observation may have been processed after its acquisition and contains the value generated by the procedure.

- The **result time** reflects the time when the result of the observation was produced.

2.4.3 EO Metadata Mapping on Observation and Measurements

To represent Earth Observation metadata, the Observations and Measurements properties are extended with EO specific information. Table 2.1 defines the awaited content of some Observations and Measurements properties.

Thematic extended namespace

In the inheritance mechanism for thematic or mission specific namespaces, existing properties defined in eop are extended or new properties are created in order to fit inside the model.

Thematic extended namespace (opt for example) contains:

1. opt “words”

2. an *opt:EarthObservation* element that inherits from *eop:EarthObservation*. This inheritance is an XML schema extension (to avoid restriction problems) with no element added (because all elements fit inside one of the Observation property *metadata*, *procedure*, *phenomenonTime*, *result* or *featureOfInterest*)
3. one or more extensions of existing eop properties

For example, “opt” thematic EO Products metadata include the cloud cover percentage, named “cloudCoverPercentage”. This property is described within the *opt:EarthObservationResultType* element which extends and acts as a substitution for *eop:EarthObservationResultType*.

Mission specific extended namespace

Mission specific extended namespace (Sentinel-1 for example) contains :

1. sen1 “words”
2. a *sen1:EarthObservation* element that inherits from *sar:EarthObservation*, because Sentinel-1 is a satellite with radar sensors. This inheritance is an XML schema extension (to avoid restriction problems) with no element added (because all elements fit inside one of the Observation property *metadata*, *procedure*, *phenomenonTime*, *result* or *featureOfInterest*)
3. one or more extensions of existing sar properties

2.5 Semantic Annotations

Annotation of Web Services or data compliant to OGC standards refers to the task of attaching meaningful descriptions to the service and the served geospatial data or processes. Without these descriptions, the use of spatial resources is limited to a small group of users. Before publishing a resource in the Web, it has to be annotated with descriptive metadata to make it usable to a broad audience. Otherwise people will neither be able to find the resource using search engines, nor to evaluate if the discovered resource satisfies their current information need.

The OGC standards baseline provides accepted and well thought-out methods to make spatial resources (data and processes) served via Web Services accessible. Service capabilities describe, besides contextual information like contact information, how to access

and invoke the service to retrieve the required geospatial data. The individual name and location of the operations are also listed in the Capabilities document of each OGC-conformal Web Service (as defined in OGC WS-Common). Since such operations and the format to encode the data are predefined in OGC Implementation Standards and OGC Encoding Standards, generic clients can, without knowledge about the nature of the data, display the resulting data on a map.

The OGC standards define how to access, invoke, and finally visualize spatial data, but they lack a well-defined methodology to describe the thematic dimension of a Web Service. They do not tell much about what the served data (or process) represents, and in particular they lack a way to link the resources to external models. For example, the application knows how to load and visualize the data on a map, but the user has no idea how to read the displayed map. With the help of semantic annotations, data providers are able to connect the standardized service descriptions to the modeled knowledge. Such models comprise conceptualized knowledge about the represented geographic phenomena. Having such a link established, reasoning algorithms are able to infer if a Web Service matches an agent's query on the formal level. In addition it allows for extracting valuable contextual information from the knowledge models, making it possible to display thematic information for the displayed data and helping the user to understand.

Semantics can increase the usefulness of geospatial information. As we mentioned above, if two agents agree on how to represent and communicate the data, a seamless and conflict-free integration is established. Hence, semantic interoperability can be achieved if the two agents (both, humans and machines) agree on how to understand the data. The level of understanding can differ and depends mostly on the complexity of the formalized knowledge. The spatial resources include domain-specific knowledge, so the meaning of most terms remain unclear to users. Without a further description, the use of this data is constrained to a very limited user group. Another common issue in service discovery is the different level of expertise between the seeking user and the data provider. The specialist publishing the data is using more specific terms than casual users do, so searching based only on keywords would yield no results. Finally, the different languages spoken by users impair the find-ability of spatial resources as well. Especially in the European region the requirement for multilingual descriptions of geospatial data gains importance. However, all these problems can be resolved with the addition of semantic descriptions linked to the attributes and feature types of data.

The semantic models used to annotate the spatial data are usually expressed as ontologies, controlled vocabularies, taxonomies or thesauri (Section 3.1). All these types of vocabularies can be represented either by the Simple Knowledge Organization System

```

<gmd:descriptiveKeywords>
  <gmd:MD_Keywords>
    <gmd:keyword>
      <gmx:Anchor xlink:href='http://gcmd.gsfc.nasa.gov/skos#spectral_
engineering'>Spectral/Engineering</gmx:Anchor>
    </gmd:keyword>
    <gmd:keyword>
      <gmx:Anchor xlink:href='http://gcmd.gsfc.nasa.gov/skos#visible_
wavelengths'>Visible Wavelengths</gmx:Anchor>
    </gmd:keyword>
    <gmd:keyword>
      <gmx:Anchor xlink:href='http://gcmd.gsfc.nasa.gov/skos#visible_
imagery'>Visible Imagery</gmx:Anchor>
    </gmd:keyword>
    <gmd:type>
      <gmd:MD_KeywordTypeCode codeListValue='theme' codeList='http://www.
isotc211.org/2005/resources/codeList.xml#MD_KeywordTypeCode' />
    </gmd:type>
    <gmd:thesaurusName>
      <gmd:CI_Citation>
        <gmd:title>
          <gmx:Anchor
xlink:href='http://gcmd.gsfc.nasa.gov/skos/'>NASA/Global Change Master Directory (GCMD)
Earth Science Keywords. Version 6.0.0.0.0
          </gmx:Anchor>
        </gmd:title>
        <gmd:date>
          <gmd:CI_Date>
            <gmd:date>
              <gco:Date>2008-02-05</gco:Date>
            </gmd:date>
            <gmd:dateType>
              <gmd:CI_DateTypeCode
codeList='http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO_19139_Schemas/res
ources/Codelist/ML_gmxCodelists.xml#CI_DateTypeCode' codeListValue='publication'>publicat
ion</gmd:CI_DateTypeCode>
              </gmd:dateType>
            </gmd:CI_Date>
          </gmd:date>
        </gmd:CI_Citation>
      </gmd:thesaurusName>
    </gmd:MD_Keywords>
  </gmd:descriptiveKeywords>

```

Figure 2.3: Semantic annotation of EO dataset series metadata

(SKOS) or the Web Ontology Language (OWL). SKOS and OWL are common data models for sharing and linking knowledge via the web and are described in more detail in Subsection 3.1.5. How annotations are inserted into the different types of metadata models is defined in the document “Semantic Annotations in OGC Standards” (OGC 08-167) [23].

ISO 19139 is an example of standard used to annotate keywords contained in dataset series or service metadata. These annotations point to the appropriate concepts defined in a semantic model. Figure 2.3 shows an XML extract that includes ISO 19139 keywords and associated references to the semantic model defined by the GCMD Earth science keywords as a SKOS representation.

2.6 Summary

In this chapter, we introduced the concepts of the EO domain. We started by explaining basic EO terms and continued with the presentation of some of the most important European and Global Initiatives, as well as Standards Organisations in the domain of Earth Observation. Afterwards, we presented the EO Metadata Profile of Observations and Measurements, the OGC Standard for describing EO products and, finally, we discussed about Semantic Annotations the task of attaching meaningful descriptions to the data.

Chapter 3

Technical Background

This chapter concentrates on the technical background with which readers should be informed in order to understand better the terminology and the technologies referred in the next chapters.

3.1 Vocabularies and Ontologies

As the design of the ProdTrees platform is based on the use of semantic technologies, it is important to distinguish the different types of vocabularies (e.g. what is an ontology and how it differs from a thesaurus). In the next sections, you will find a short description of each type of vocabulary, as well as, of the two languages, SKOS and OWL, for defining these vocabularies. The ProdTrees platform is built on the use of thesauri encoded in SKOS.

3.1.1 Controlled Vocabularies

A controlled vocabulary is an arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain. The purpose of controlled vocabularies is to organize information and to provide terminology to catalog and retrieve information.

Controlled vocabularies are nothing more than lists of words (e.g. “Cat, Poodle, Mammal, Collie, Dog, Manx, Bulldog”). Their definition does not include any specific order, although a web form might display long vocabularies alphabetically or short ones in order of popularity to make it easier for people to find the terms that they need. If there was a specific ordering to these lists, that would constitute metadata about their relationships, so we would be moving away from controlled vocabulary territory toward a taxonomy.

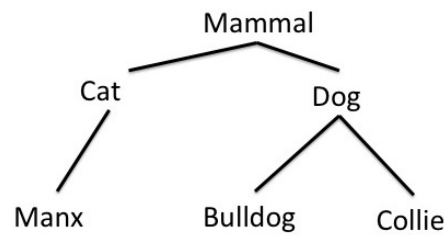


Figure 3.1: A hierarchy example for a taxonomy

3.1.2 Taxonomies

Taxonomies organize controlled vocabulary terms into a hierarchy. Taking as example the controlled vocabulary mentioned above and saying that Cat is a broader term for Manx, that Dog is a broader term for Collie and Bulldog, and that Mammal is a broader term for Dog and Cat, we create a simple taxonomy. The “broader” relationships of a taxonomy are often represented visually as a tree (Figure 3.1).

Each term in a taxonomy is in one or more parent/child (broader/ narrower) relationships to other terms in the taxonomy. There can be different types of parent/child relationships, such as whole/part, genus/ species, or instance relationships.

A taxonomy used for serious business purposes often stores more than just broader-than relationships. These can include alternative terms to assist search (for example, “auto” as an alternative to “car”), translations of the term to foreign languages, metadata about who last edited the term and when, and notes about what exactly the term applies to if there is potential confusion what taxonomists call “scope notes”.

3.1.3 Thesauri

A thesaurus stores even more metadata than a taxonomy. It might store relationship information about opposite terms; for example, that the opposite of Yes is No. It might store what taxonomists call a “Use For” relationship, so that users searching for a particular term that isn’t considered to be the best one can be redirected to the preferred term.

“Broader” relationships are one kind of relationship metadata, and a thesaurus often stores other kinds of relationship metadata as well. These can even connect a term to another term in a different vocabulary. For example, a thesaurus might store metadata indicating that the term Dog in an animal taxonomy is Related To the term Doghouse in a taxonomy of shelter types, or to a particular veterinary product in a pharmaceutical

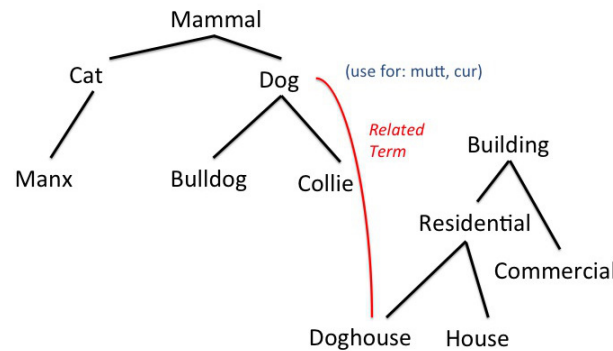


Figure 3.2: An example of a thesaurus

company's taxonomy of products.

If a taxonomy can store more than just broader relationships, and a thesaurus is usually arranged as a taxonomy with additional metadata, it is not clear when a hierarchical controlled vocabulary with metadata is a taxonomy or a thesaurus. These terms are sometimes used interchangeably.

The metadata properties associated with a taxonomy's terms fall into two categories, which we can call relationship properties and attribute properties. Relationship properties indicate a term's relationship with another term, such as that Dog is a broader term than Collie or that Yes is the antonym of No. Attribute properties are typically text entered as metadata about a term, such as the Greek word for that term or the name of the staff member who last edited it.

To be more precise with the definition of the vocabulary used in a thesaurus, the primitive objects are not terms, but abstract notions that are represented by terms and are called *concepts*. A *concept scheme* is a set of concepts, potentially including statements about relationships between those concepts:

1. Broader Terms
2. Narrower Terms
3. Related Terms
4. Synonyms, usage information etc.

Concepts are gathered in concept schemes to provide consistent and structured sets of concepts, representing whole or part of a controlled vocabulary.

3.1.4 Ontologies

In a thesaurus there are various standard, generally applicable relationship and attribute properties that can be used to store more information about the terms in that thesaurus. The difference in an ontology is that we can define our own relationships and attributes, as well as classes of things that are characterized by these relationships and attributes. Where thesauri and taxonomies use generic relationships such as broader, related and “use for” that can be applied to any term, ontologies define relationships and attributes that are specific to a particular area. For example, an ontology might include a relationship to show that one event is a precondition of another event, and a medical ontology might have a relationship to show that one symptom contraindicates a particular treatment.

Ontologies can be used to infer new information, such as class membership. For example, if someone has a “playsInstrument” property value of “guitar” and the ontology says that anyone with a “playsInstrument” value is a musician, we can infer that this person is a musician even if there is no explicit data saying that he is a member of that class.

3.1.5 SKOS and OWL

Simple Knowledge Organization System¹ (SKOS) is a W3C recommendation designed for knowledge organization systems such as thesauri, concept schemes, taxonomies, or any other type of structured controlled vocabulary. SKOS is part of the Semantic Web family of standards built upon RDF and RDFS, and its main objective is to enable easy publication and use of such vocabularies as linked data. The SKOS data model views a knowledge organization system as a concept scheme comprising a set of concepts. These SKOS concept schemes and SKOS concepts are identified by URIs, enabling anyone to refer to them unambiguously from any context, and making them a part of the World Wide Web.

SKOS concepts can be:

1. **labeled** with any number of lexical (UNICODE) strings in any given natural language, such as English or Japanese.
2. **documented** with notes of various types. SKOS provides a basic set of documentation properties, supporting scope notes, definitions and editorial notes, among others.

¹<http://www.w3.org/TR/skos-reference/#intro>

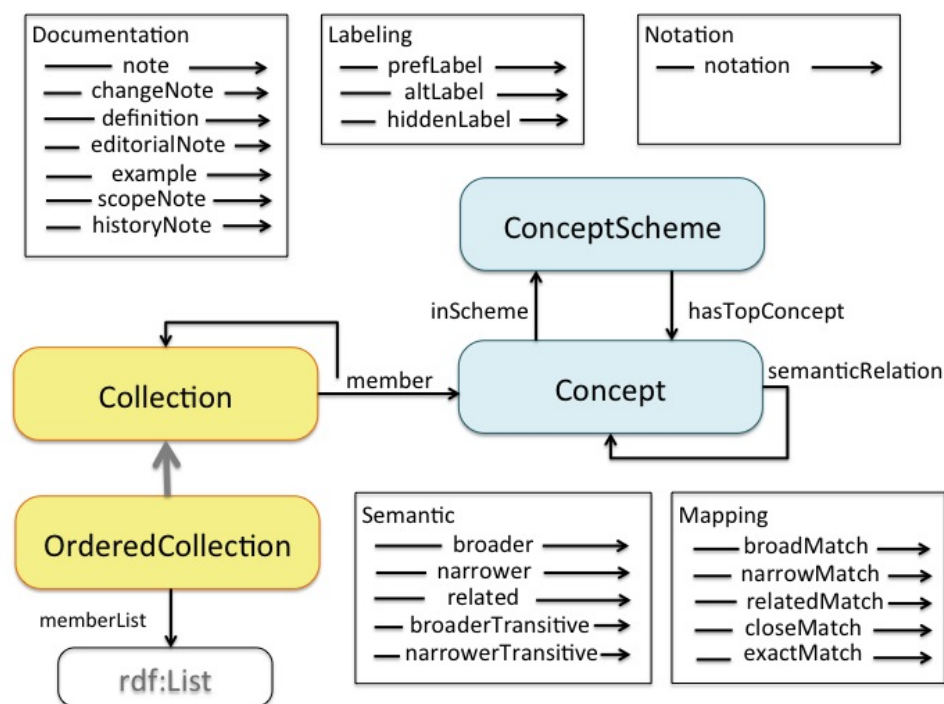


Figure 3.3: The SKOS data model

3. **linked** to other SKOS concepts using hierarchical and associative links.
4. grouped into **collections**, which can be labeled and/or ordered. This feature of the SKOS data model is intended to provide support for node labels within thesauri, and for situations where the ordering of a set of concepts is meaningful or provides some useful information.
5. **mapped** to other SKOS concepts in different concept schemes.

Figure 3.3 displays all the features provided by the SKOS data model. There is also the SKOS eXtension for Labels² (SKOS-XL) that offers additional support for descriptions of labels and links between them (e.g. acronyms, abbreviations).

Whereas SKOS is used for the representation of thesauri, OWL³ (Web Ontology Language) is the W3C standard for defining ontologies. It represents rich and complex knowledge about things, groups of things, and relations between things. OWL is a computational logic-based language such that knowledge expressed in OWL can be exploited by computer programs, e.g., to verify the consistency of that knowledge or to make implicit

²<http://www.w3.org/TR/skos-reference/skos-xl.html>

³<http://www.w3.org/TR/2004/REC-owl-features-20040210/#s1>

knowledge explicit. OWL documents, known as ontologies, can be published in the World Wide Web and may refer to or be referred from other OWL ontologies. OWL is part of the W3CTMs Semantic Web technology stack, which includes RDF, RDFS, SPARQL, etc.

SKOS can be used side-by-side with OWL to express and exchange knowledge about a domain. However, SKOS is not a formal knowledge representation language. The “knowledge” made explicit in a formal ontology is expressed as sets of axioms and facts. A thesaurus or concept scheme is of a completely different nature, and does not assert any axioms or facts. Rather, a thesaurus or concept scheme identifies and describes, through natural language and other informal means, a set of distinct ideas or meanings (concepts). These concepts may also be arranged and organized into various structures (most commonly hierarchies). These structures, however, do not have any formal semantics, and cannot be reliably interpreted as either formal axioms or facts about the world. They serve only to provide a convenient and intuitive map of some subject domain, which can then be used as an aid to organizing and finding objects relevant to that domain.

To make the “knowledge” embedded in a thesaurus or concept scheme explicit in any formal sense requires that the thesaurus or concept scheme be re-engineered as a formal ontology. In other words, the structure and intellectual content of a thesaurus or concept scheme must be transformed into a set of formal axioms and facts. This work of transformation is both intellectually demanding, time consuming and costly. Much can be gained from using thesauri, etc., as-is, as informal, convenient structures for navigation within a subject domain. Using them as-is does not require any re-engineering and is therefore much less costly. In addition, some knowledge organisation systems are, by design, not intended to represent a logical view of their domain. Converting such knowledge organisation systems to a formal logic-based representation may, in practice, involve changes which result in a representation that no longer meets the originally intended purpose. OWL does, however, provide a powerful data modeling language. We can, therefore, use OWL to construct a data model for representing thesauri or concept schemes as-is. This is exactly what SKOS does.

3.2 Data Formats and Models

As we mentioned in Chapter 2, the ProdTrees platform is a search engine for EO products with metadata encoded in EO-netCDF, an extension of NetCDF. For this reason, we need first to explain what kind of information this data format supports. We also describe SAFE data format, as in ProdTrees, EO data in SAFE is used to be translated in EO-netCDF.

3.2.1 NetCDF

Network Common Data Form⁴ [34] (netCDF) is a data model for array-oriented scientific data, a freely distributed collection of access libraries implementing support for that data model, and a machine-independent format. Together, the interfaces, libraries, and format support the creation, access, and sharing of multi-dimensional scientific data. Data in netCDF format is:

- **Self-Describing:** A netCDF file includes information about the data it contains.
- **Portable:** A netCDF file can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.
- **Scalable:** Small subsets of large datasets in various formats may be accessed efficiently through netCDF interfaces, even from remote servers.
- **Appendable:** Data may be appended to a properly structured netCDF file without copying the dataset or redefining its structure.
- **Sharable:** One writer and multiple readers may simultaneously access the same netCDF file.
- **Archivable:** Access to all earlier forms of netCDF data will be supported by current and future versions of the software.

There are four netCDF format variants, two supported data models and two textual representations for netCDF data. In different contexts, “netCDF” may refer to either the data model, or the data format.

The different netCDF formats are: i) the classic format, ii) the 64-bit offset format, iii) the netCDF-4 format, and iv) the netCDF-4 classic model format. The **classic format** was the only format for netCDF data created between 1989 and 2004 by the reference software from Unidata⁵. It is still the default format for new netCDF data files, and the form in which most netCDF data is stored. The **64-bit offset format**, was added in 2004, allows users to create and access far larger datasets than were possible with the original format. In 2008, the **netCDF-4 format** was added to support per-variable compression, multiple unlimited dimensions, more complex data types, and better performance, by layering an enhanced netCDF access interface on top of the HDF5 format [17]. At the same time, a

⁴<http://www.unidata.ucar.edu/netcd>

⁵<http://www.unidata.ucar.edu/>

fourth format variant, **netCDF-4 classic model format**, was added for users who needed the performance benefits of the new format (such as compression) without the complexity of a new programming interface or enhanced data model.

The netCDF data models are the classic model and the enhanced model. The **classic model** is the simpler of the two, and is associated with all versions of netCDF prior to netCDF-4 format. The **enhanced model** (sometimes also referred to as the netCDF-4 data model) is an extension of the classic model that adds more powerful forms of data representation and data types at the expense of some additional complexity. Although data represented with the classic model can also be represented using the enhanced model, datasets that use enhanced model features, such as user-defined data types, cannot be represented with the classic model. Use of the enhanced model requires storage in the netCDF-4 format.

Finally, Common Data Language [33] (CDL) and NetCDF Markup Language [35] (NcML) are the textual representations for the netCDF data. **CDL** provides a convenient way of describing netCDF dataset. A CDL file is an ASCII description of the binary data stored in a netCDF file that is designed to be easily read by humans. CDL files can be generated automatically from netCDF files by using appropriate tools. **NcML** is an XML representation of netCDF metadata. NcML is similar to CDL, except, of course, it uses XML syntax.

A detailed description of the netCDF data models is given in the forthcoming subsections.

The “Classic” netCDF Data Model

The classic netCDF data model [33] contains **dimensions**, **variables**, and **attributes**, which all have both a name and an number by which they are identified. These components can be used together to capture the meaning of data and relations among data fields in an array-oriented dataset.

UML diagrams represent data models visually. Each box contains:

1. the name of a class of objects
2. characteristics of object in the class
3. operations (methods) for that class of objects

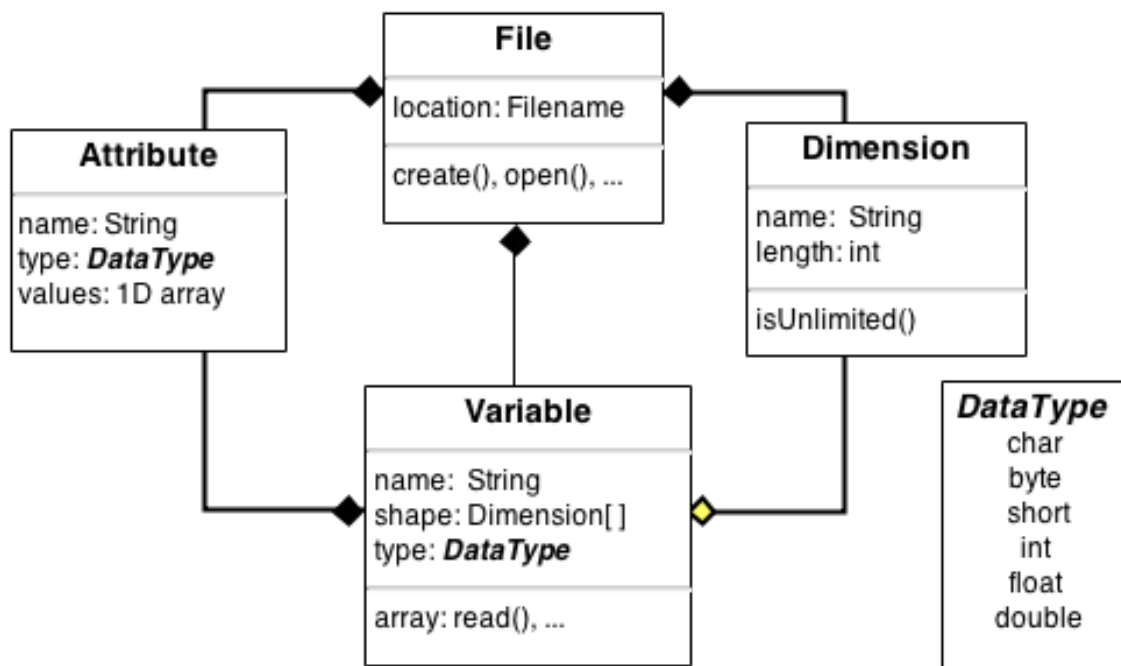


Figure 3.4: The “Classic” netCDF Data Model

Connecting lines identify relationships of containment and use. Figure 3.4 shows a simplified UML diagram of the classic netCDF data model.

A netCDF example [34] illustrating the concepts of the netCDF classic data model can be found in Table 3.1. This includes dimensions, variables and attributes. The notation used to describe this simple netCDF object is CDL. As you can see, all CDL statements are terminated by a semicolon. Spaces, tabs, and newlines can be used freely for readability. Comments may follow the double slash characters ‘//’ on any line.

A CDL description for a classic model file consists of three optional parts: dimensions, variables, and data. The variable part may contain variable declarations and attribute assignments. For the enhanced model supported by netCDF-4, a CDL description may also include groups, subgroups, and user-defined types.

A dimension is used to define the shape of one or more of the multidimensional variables described by the CDL description. A dimension has a name and a length. At most one dimension in a classic CDL description can have the unlimited length, which means a variable using this dimension can grow to any length (like a record number in a file). Any number of dimensions can be declared of unlimited length in CDL for an enhanced model file.

A variable represents a multidimensional array of values of the same type. A variable has a name, a data type, and a shape described by its list of dimensions. Each variable

may also have associated attributes (see below) as well as data values. The name, data type, and shape of a variable are specified by its declaration in the variables section of a CDL description. An attribute contains information about a variable or about the whole netCDF dataset or containing group. Attributes may be used to specify such properties as units, special values, maximum and minimum valid values, and packing parameters. Attribute information is represented by single values or one-dimensional arrays of values. For example, `units` might be an attribute represented by a string such as `celsius`. An attribute has an associated variable, a name, a data type, a length, and a value. In contrast to variables that are intended for data, attributes are intended for ancillary data or metadata (data about data).

In CDL, an attribute is designated by a variable and attribute name, separated by a colon (`:`). It is possible to assign global attributes to the netCDF dataset as a whole by omitting the variable name and beginning the attribute name with a colon (`:`). The data type of an attribute in CDL, if not explicitly specified, is derived from the type of the value assigned to it. The length of an attribute is the number of data values or the number of characters in the character string assigned to it. Multiple values are assigned to non-character attributes by separating the values with commas (`,`). All values assigned to an attribute must be of the same type. In the netCDF-4 enhanced model, attributes may be declared to be of user-defined type, like variables.

The netCDF-4 Data Model

The netCDF-4 data model [33] adds **Groups** and **User-Defined Types** to the classic netCDF data model, but backward compatibility is preserved (Figure 3.5).

Groups, like directories in a Unix file system, are hierarchically organized, to arbitrary depth. They can be used to organize large numbers of variables. Each group acts as an entire netCDF dataset in the classic model. That is, each group may have its own attributes, dimensions, and variables. The default group is the root group, which allows the classic netCDF data model to fit neatly into the new model.

Dimensions are scoped such that they can be seen in all descendant groups. Dimensions can thus be shared between variables in different groups, if they are defined in a parent group. In netCDF-4 files, the user may also define a type. For example a compound type may hold information from an array of C structures, or a variable length type allows the user to read and write arrays of variable length values.

Variables, groups, and types share a namespace. Within the same group, variables,

Table 3.1: An example of a netCDF dataset in CDL notation

```
netcdf example_1 { // example of CDL notation for a netCDF dataset

dimensions:          // dimension names and lengths are declared first
    lat = 5, lon = 10, level = 4, time = unlimited;

variables:            // variable types, names, shapes, attributes
    float    temp(time,level,lat,lon);
                temp:long_name = "temperature";
                temp:units      = "celsius";
    float    rh(time,lat,lon);
                rh:long_name    = "relative humidity";
                rh:valid_range  = 0.0, 1.0;    // min and max
    int      lat(lat), lon(lon), level(level);
                lat:units       = "degrees_north";
                lon:units       = "degrees_east";
                level:units     = "millibars";
    short    time(time);
                time:units      = "hours since 1996-1-1";
    // global attributes
                :source         = "Fictional Model Output";

data: // optional data assignments
    level = 1000, 850, 700, 500;
    lat   = 20, 30, 40, 50, 60;
    lon   = -160,-140,-118,-96,-84,-52,-45,-35,-25,-15;
    time  = 12;
    rh    =.5,.2,.4,.2,.3,.2,.4,.5,.6,.7,
            .1,.3,.1,.1,.1,.1,.5,.7,.8,.8,
            .1,.2,.2,.2,.2,.5,.7,.8,.9,.9,
            .1,.2,.3,.3,.3,.3,.7,.8,.9,.9,
            0,.1,.2,.4,.4,.4,.4,.7,.9,.9;

}
```

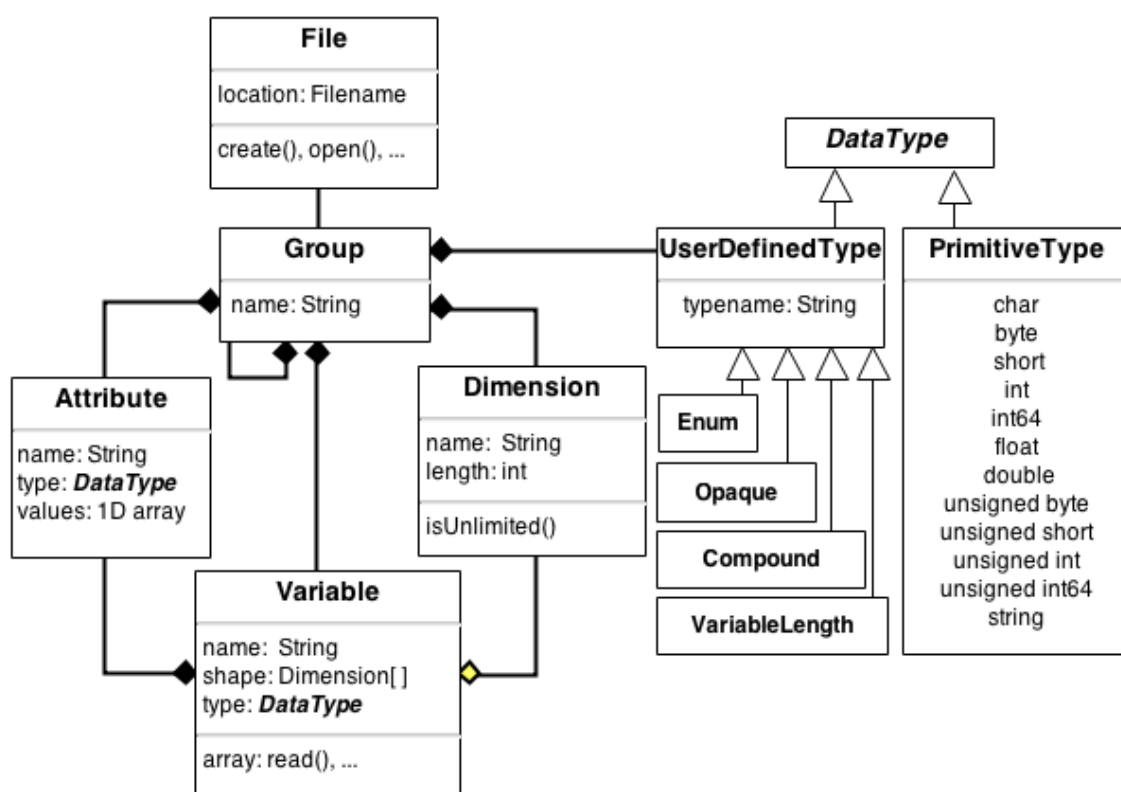


Figure 3.5: The netCDF-4 Data Model

groups, and types must have unique names. In other words, a type and variable may not have the same name within the same group, and similarly for sub-groups of that group.

3.2.2 NetCDF Conventions

While netCDF is designed to read and write data that has been structured according to well-defined rules and so is intended for “self-documenting data”, the netCDF interface enables but does not require the creation of such data. Specific semantics (e.g. applications and/or Community semantics) can be encoded by defining and using conventions, and thus enriching and extending the netCDF data model. These conventions are written up as human readable documents called **netCDF conventions**.

Conventions enhance datasets with sufficient metadata that are self-describing in the sense that each variable in the file has an associated description of what it represents, including physical units if appropriate, and that each value can be located in space (relative to earth-based coordinates) and time. They can enable software tools to display data and perform operations on specified subsets of the data with minimal user intervention. It is, also, possible to provide the metadata describing how a field is located in time and

space in many different ways that a human would immediately recognize as equivalent. The purpose in restricting how the metadata is represented is to make it practical to write software that allows a machine to parse that metadata and to automatically associate each data value with its location in time and space. Finally, it is equally important that the metadata be easy for human users to write and to understand.

The next subsections describe the most frequently used conventions. Other sets of conventions currently available can be found in <http://www.unidata.ucar.edu/software/netcdf/conventions.html>.

CF Conventions

The most used set of conventions is the Climate and Forecast Metadata Conventions [15] (CF-netCDF) which is intended for use with climate and forecast data, for atmosphere, surface and ocean, and was designed with model-generated data particularly in mind. The CF-netCDF encoding format consists in netCDF conforming to the CF conventions.

The conventions define metadata that provide a definitive description of what the data in each variable represents, and of the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, re-gridding, and display capabilities. In addition, version 1.6 of the CF conventions (CF-1.6) introduced the important concepts and relations for managing the *Discrete Sampling geometries* (e.g. point, time series, trajectory, trajectory profile, time series profile, and any ragged and multidimensional array data type). This allows the CF-netCDF to be used, as encoding and data model, not only for the Coverage domain but also for the Feature domain and the Sensor domain (e.g. OGC WFS [36] and SOS [6]).

CF conventions are also well used because they define a set of controlled terms (i.e. vocabularies) for: variable names (i.e. standard names⁶), unit of measures, coordinate reference systems (CRS), grid mappings and projections, cell methods. As a result, CF-netCDF is a *de-facto* standard for many Earth Science Communities, such as: Meteorology, Oceanography, Glaciology, and Climatology. Besides, due to its flexibility and simplicity it is more and more used for addressing multi-disciplinary challenges. For example, the Hydrology community has been using it by leveraging its capacity to bridge the GIS and EO domains.

As to formal descriptions, the CF-netCDF consists of a set of normative specifications,

⁶<http://cfconventions.org/standard-names.html>

recently formalized as OGC standards. These include: the OGC NetCDF Enhanced Data Model Extension Standard [13] and the NetCDF Binary Encoding Extension Standard [11]. Besides, a CF-netCDF encoding extension for WCS v2.0 is in its final draft stage [2] and a discussion paper on CF-netCDF conventions for uncertainty encoding was recently approved by the OGC CF-netCDF SWG [12] (Standardization Working Group).

Finally, this convention is designed to be backward compatible with the COARDS conventions⁷ (a 1995 standard that CF Conventions extends and generalizes), which means that a conforming COARDS dataset also conforms to the CF standard. Thus new applications that implement the CF conventions will be able to process COARDS datasets.

ACDD Conventions

The Attribute Convention for Dataset Discovery⁸ (ACDD) identifies and defines a list of netCDF global attributes recommended for describing a netCDF dataset to discovery systems such as Digital Libraries. Software tools can use these attributes to extract metadata from datasets, and export them to several metadata formats.

For example THREDDS⁹ tools use ACDD for extracting metadata from datasets, and exporting to Dublin Core¹⁰, DIF¹¹, FGDC¹², ISO 19115 etc. metadata formats. These attributes parallel THREDDS catalog specification's digital library metadata. Attributes are used to add information inside the netCDF file, while THREDDS catalog metadata adds information external to the netCDF file.

Where appropriate, ACDD uses attributes described in the netCDF Users Guide as well as some attributes defined in the CF convention. Some are used directly (e.g., "title" and "history"), while others are used unless more detailed attributes defined are given (e.g., "institution" vs. "creator_*").

A metadata mapping between ACDD and CF, THREDDS, Dublin Core, ISO 19115, etc. is available¹³.

⁷http://ferret.wrc.noaa.gov/noaa_coop/coop_cdf_profile.html

⁸http://wiki.esipfed.org/index.php/Category:Attribute_Conventions_Dataset_Discovery

⁹<http://www.unidata.ucar.edu/software/thredds/current/tds/>

¹⁰<http://dublincore.org/>

¹¹<http://gcmd.gsfc.nasa.gov/add/difguide/index.html>

¹²<https://www.fgdc.gov/metadata/geospatial-metadata-standards>

¹³http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery_%28ACDD%29

NetCDF-U Conventions

The NetCDF Uncertainty Conventions [3] (NetCDF-U) introduces a set of conventions for managing uncertainty information within the netCDF classic data model and format. These conventions have the following rationale:

- Compatibility with netCDF-CF Conventions 1.5,
- Human-readability of conforming datasets structure,
- Minimal difference between certain/agnostic and uncertain representations of data (e.g. with respect to dataset structure).

The main mechanism for modelling uncertainty in netCDF-U files consists of annotating the netCDF data variables with uncertainty-related semantics based on the Uncertainty Markup Language¹⁴ (UncertML) dictionary. The netCDF-U Conventions are applicable to data encoded in the netCDF-3 format and are designed to be fully compatible with the netCDF Climate and Forecast Conventions. However, limitations may apply, as regards compliance with conflicting conventions¹⁵.

The NetCDF-U conventions has been proposed as a standard to the OGC.

3.2.3 SAFE

The Standard Archive Format for Europe¹⁶ (SAFE) has been designed to act as a common format for archiving and conveying data within ESA Earth Observation archiving facilities. SAFE aims to preserve the archived data for a long-term, facilitating the conversion into different formats, simplifying the extraction from the archive and enhancing their utilization by end-users and/or processing systems. The format was developed in the context of the HARM¹⁷ (Historical Archives Rationalization and Management) project, which aimed at converting ESA's historical datasets into a new modern format, based on the latest technologies and standards and able to ensure the long-term preservation of its holdings.

During the development of SAFE, particular attention was put to the long-term preservation aspect. To this end, the information model of the generic Archival Information

¹⁴<http://www.uncertml.org/>

¹⁵A netCDF dataset may be compliant with multiple conventions.

¹⁶<http://earth.esa.int/SAFE/>

¹⁷<http://www.werum.de/en/mdm/eo/project/harm/index.jsp>

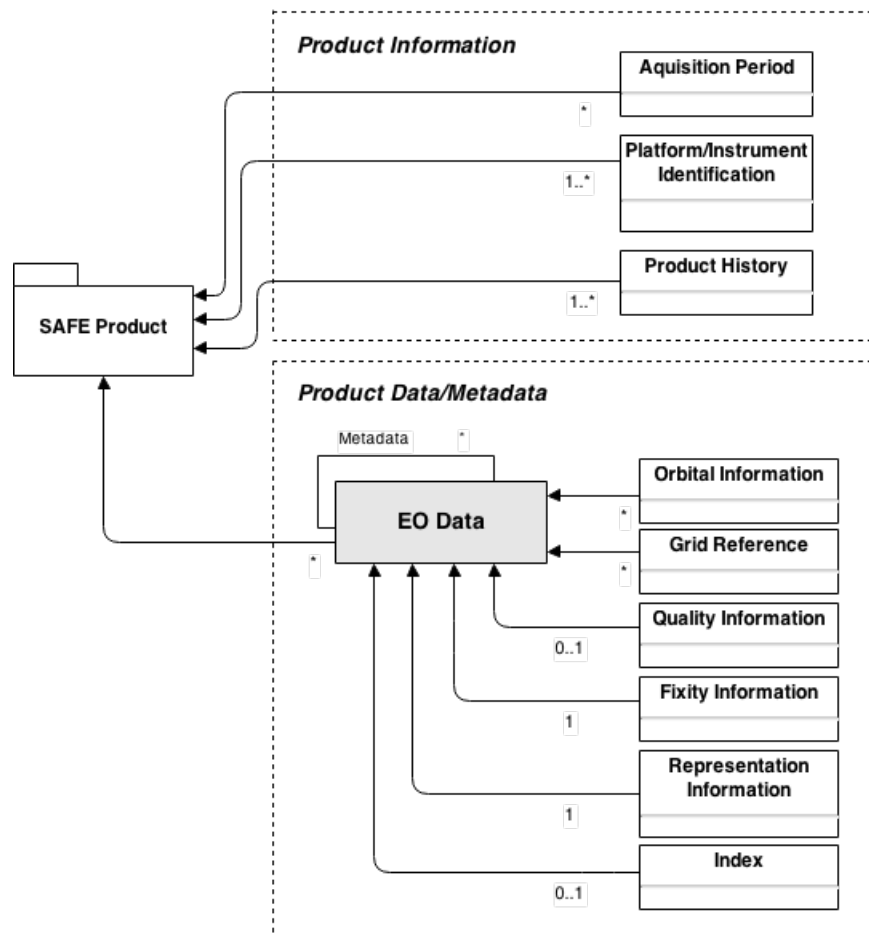


Figure 3.6: SAFE Information Model

Package (AIP), introduced in the ISO 14721:2003¹⁸ OAIS (Open Archival Information System) Reference Model, has been used. Furthermore, SAFE is based on the XFDU¹⁹ (XML Formatted Data Units) standard under development by the Consultative Committee for Space Data Systems²⁰ (CCSDS). In its essence, SAFE is a profile of XFDU, and it restricts the XFDU specifications for the specific utilization in the EO domain.

Although the primary goal of SAFE, in the framework of the HARM project, is to handle EO data with processing levels close to the usually called “*Level-0*” (or “*LO*”), no limitation exists regarding the packaging of higher level products as well as other technical and scientific information. Actually, experience has demonstrated that packaging and archiving higher processing levels or auxiliary data in a common format may be effective in many situations. SAFE embodies this concept by offering a single framework for packaging a large variety of information. SAFE embodies this concept by offering a single framework

¹⁸http://www.iso.org/iso/catalogue_detail?csnumber=24683

¹⁹<http://sindbad.gsfc.nasa.gov/xfdu/>

²⁰<http://public.ccsds.org/default.aspx>

for packaging a large variety of information.

An introduction to the different abstraction models that can be used to describe SAFE products follows in the next subsection.

Abstraction Models

There are three different abstraction models²¹ that can be used to describe SAFE products, which are all based on the XFDU standard:

- SAFE Information Model
- SAFE Logical Model
- SAFE Physical Model

SAFE Information Model wraps or references EO data and associates them with information expressed in EO vocabulary. The primary objective of SAFE is to hold the “LO” data which is close to the telemetry level but it has, moreover, been qualified for the packaging of higher levels products as well. All SAFE products contain the following metadata: i) Acquisition Period, ii) Platform/Sensor identification, and iii) Product History. In addition, a collection of metadata information may be attached. These include: Orbital information, Grid reference, Geological information, Quality/Fixity information, Representation information. Finally, SAFE does not limit the information to the content listed above but supports extensions as far as they preserve the integrity of the mandatory items. The SAFE Information Model is depicted in Figure 3.6.

SAFE Logical Model describes a SAFE product as a logical tree of “Content Units” forming the so-called “Information Package Map”. Conversely to XFDU, only one map is expected per SAFE product. The root Content Unit has predefined associations to the information applicable to the overall product, i.e. at least the “Acquisition Period”, the “Platform/Sensor Identification” and the “Product History”. The structure of the children Content Units is less constrained and depends mainly on the logical view of the wrapped data. In most cases, one Content Unit matches one EO dataset and its accompanying metadata. Several Content Units may, however, share the same metadata. The SAFE Logical Model is depicted in Figure 3.7.

²¹<http://earth.esa.int/SAFE/models.html>

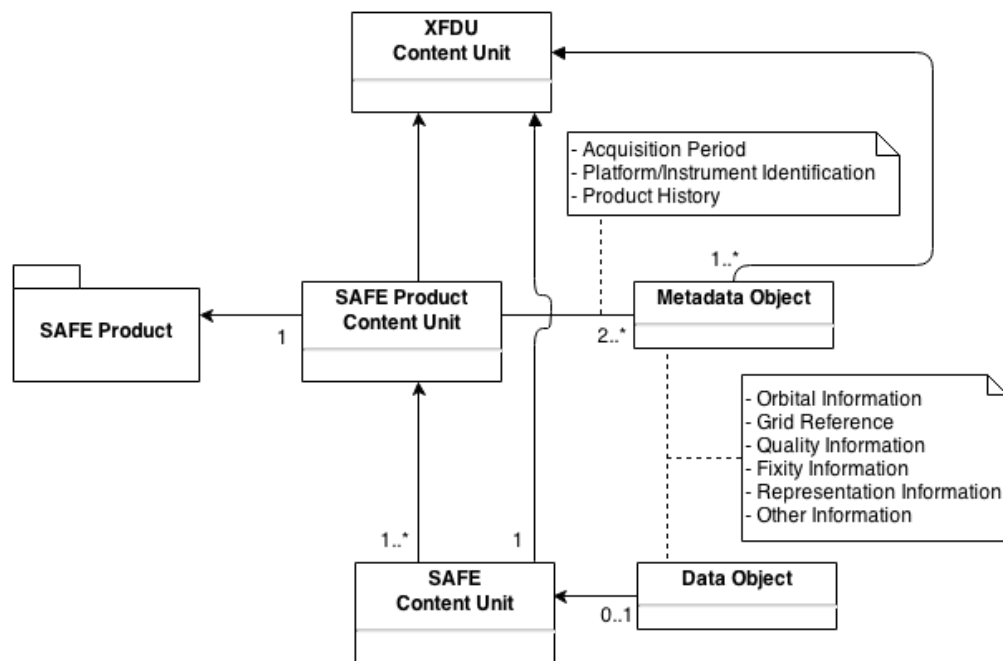


Figure 3.7: SAFE Logical Model

Finally, **SAFE Physical Model** describes a SAFE product physically using the following components: i) a Manifest file, ii) Binary, ASCII or XML files, and iii) XML Schema files. The SAFE Physical Model is depicted in Figure 3.8.

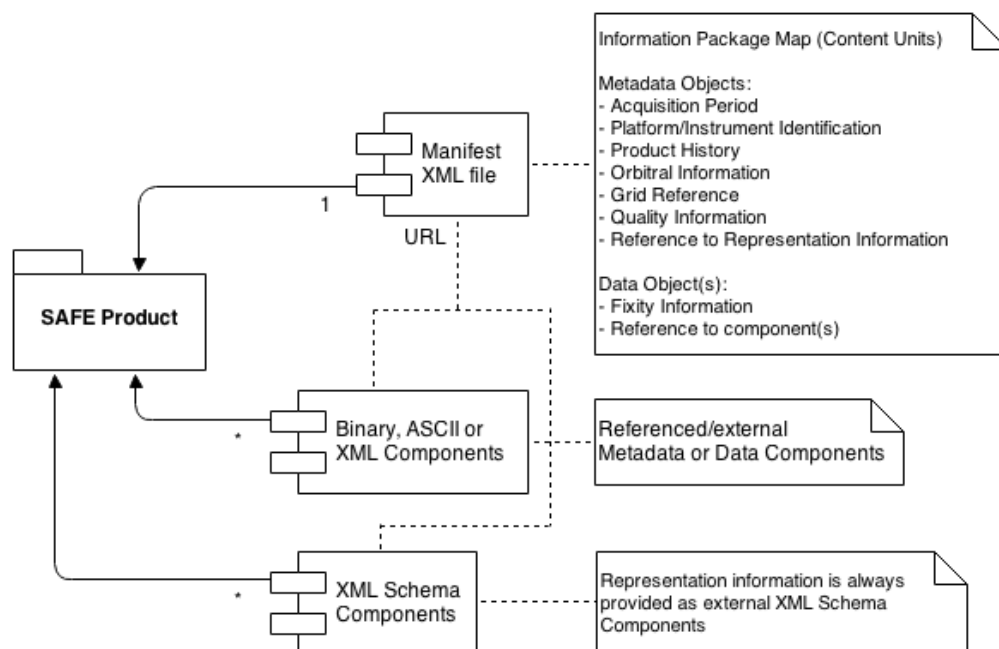


Figure 3.8: SAFE Physical Model

3.3 Related Technologies

Before describing the architecture of ProdTrees in Chapter 5, we should explain what are the technologies on which the ProdTrees components are built. The following sections are focused on these technologies.

3.3.1 OpenSearch in Earth Observation

OpenSearch²² is a collection of technologies and standards for describing search services and publishing of search results in a format suitable for aggregation. OpenSearch helps search engines and search clients communicate by introducing a common set of formats to perform search requests and syndicate search results.

The OpenSearch description document (OSDD) format is used to describe a search engine so that it can be used by search client applications. It is an XML document that provides a set of URL templates which describe the query parameters accepted by the service and the variety of output formats in which results are obtained. For example, an OSDD can answer questions like: “What are the properties described by this service?”, “Who developed this service?”, “What is the license model for this service?”, “What is the URL to call the search service?”, and more. The search results can be returned as Atom, RSS, HTML, RDF, KML, JSON, etc. An example of a simple OSDD is given in Figure 3.9.

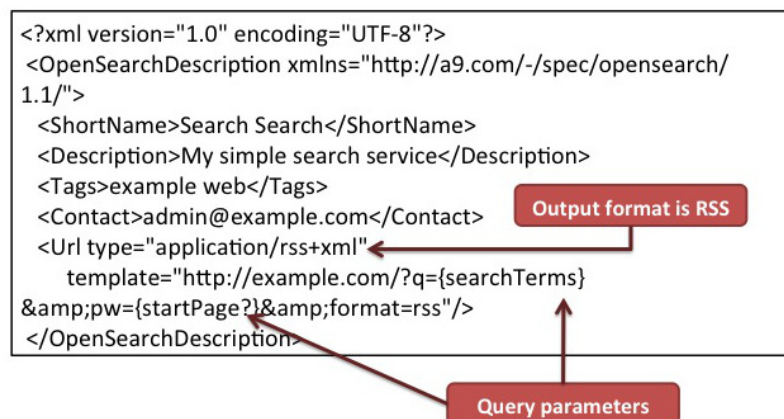


Figure 3.9: An example of a simple OSDD

²²<http://www.opensearch.org/>

OpenSearch GeoSpatial and Temporal Extensions

The OGC 10-032 [20] standard specifies the Geo and Time extensions to the OpenSearch query protocol. The purpose of this standard is to provide a very simple way to make spatial and temporal queries to a repository of geospatial content that contains geographic and temporal properties.

The Geo and Time Extensions specify a series of parameters that can be used to geographically constrain search results. Tables 3.2 and 3.3 describe these parameters.

Table 3.2: OpenSearch parameters for Geo extension

OpenSearch Parameter	Definition	Data types and values
box	Geographic bounding box	The box is defined by “west, south, east, north” coordinates of longitude, latitude, in a EPSG:4326 decimal degrees.
geometry	Geographic area (geometry)	The geometry is defined using the Well Known Text and supports the following 2D geographic shapes: POINT, LINESTRING, POLYGON, MULTIPOINT, MULTILINESTRING, MULTIPOLYGON The Geometry shall be expressed using the EPSG:4326
uid	Local identifier of the record in the repository context	Character String
lat	The latitude of a given point	Latitude in decimal degrees in EPSG:4326.
lon	The longitude of a given point	Longitude in decimal degrees in EPSG:4326.
radius	A search radius from a lat-lon point	The distance in meters along the Earth’s surface.
relation	Spatial relation to result set	Character String. One of “intersects”, “contains”, “disjoint”.
name	A string describing the location (place name) to perform the search	Character String

Table 3.3: OpenSearch parameters for Time extension

OpenSearch Parameter	Definition	Data types and values
start	A string describing the start of the temporal interval to search (bigger or equal to).	Character String, must match the RFC-3339.
end	A string describing the end of the temporal interval to search (smaller or equal to).	Character String, must match the RFC-3339.
relation	A temporal relation to the result set	Character String: One the “intersects”, “contains”, “during”, “dis-joint”.

OpenSearch Extension for Earth Observation

The OGC 13-026 [19] standard is the specification for the OpenSearch extension for Earth Observation collections and products search. It is complementary to the OpenSearch Geo and Time Extensions (OGC 10-032) described in previous section and recommends its use for spatial and temporal queries.

Some of the parameters defined in the Earth Observation Extension are displayed in Tables 3.4 and 3.5. The specification defines also a number of parameters that describe the acquisition process (e.g. acquisitionStation, availabilityTime).

Accessing EO Catalogues with OpenSearch

OpenSearch is used by applications and services in order to make feasible the search and the location of EO data. In the next paragraphs you will find a short description for some of these services.

EuroGeoss Broker is developed in the context of project EuroGeoss²³. EuroGeoss contributes to the GEOSS Common Infrastructure (GCI) in discovery and access and demonstrates applications in Biodiversity, Forest and Drought. The EuroGEOSS Broker is located between the user and the set of datasets and services providers. It is able to interface with existing web services, whatever the interoperability standards used. In technical terms, the Broker takes a request from a user as an entry, translates and dispatches it between the referenced services. Upon return of results from the services, it

²³<http://www.eurogeoss.eu/>

Table 3.4: A number of OpenSearch parameters for collection search

OpenSearch Parameter	Definition	Data types and values
productType	A string identifying the entry type (e.g. ER02_SAR_IM__0P, MER_RR__1P, SM_SLC__1S)	Character String
platformShortName	A string with the platform short name (e.g. Sentinel-1)	Character String
instrument	A string identifying the instrument (e.g. MERIS, AATSR, ASAR, HRVIR, SAR).	Character String
sensorType	A string identifying the sensor type.	Character String. Suggested values are: OPTICAL, RADAR, ALTIMETRIC, ATMOSPHERIC, LIMB
orbitType	A string identifying the platform orbit type (e.g. LEO, GEO)	Character String
resolution	A float number, set or interval requesting the range of sensor resolution given in meters	Integer

Table 3.5: A number of OpenSearch parameters for product search

OpenSearch Parameter	Definition	Data types and values
productionStatus	A string identifying the status of the entry (e.g. ARCHIVED, ACQUIRED, CANCELLED)	Character String
acquisitionType	Used to distinguish at a high level the appropriateness of the acquisition for “general” use, whether the product is a nominal acquisition, special calibration product or other.	Character String. Values: NOMINAL, CALIBRATION, OTHER
orbitNumber	A number with the acquisition orbit.	Integer
orbitDirection	A string identifying the acquisition orbit direction.	Character String. Possible values are: ASCENDING, DESCENDING
processorName	A string identifying the processor software name	Character String
processingCenter	A string identifying the processing center (e.g. PDHS-E, PDHS-K, DPA, F-ACRI)	Character String

merges and displays the results to the user.

GeoNetwork²⁴ is a catalog application to manage spatially referenced resources. It provides metadata editing and search functions as well as an embedded interactive web map viewer. The software offers an easy to use web interface to search geospatial data across multiple catalogs, combine distributed map services in the embedded map viewer, publish geospatial data using the online metadata editing tools and optionally the embedded GeoServer map server.

ECHO OpenSearch²⁵ provides access to Earth Observing System (EOS) Clearing House²⁶ (ECHO) via an OpenSearch interface. ECHO is developed by NASA and is a spatial and temporal metadata registry and order broker. It allows users to more efficiently search and access data and services and increases the potential for interoperability with new tools and services.

Mirador²⁷ is an earth science data search tool developed at the Goddard Earth Sciences (GES) Data and Information Services Center²⁸ (DISC) for data users. It has a simplified, clean interface and employs the Google mini appliance for metadata keyword searches. Other features include quick response, spatial and parameter subsetting, data file hit estimator, Gazetteer (geographic search by feature name capability), and an interactive shopping cart. You can access the OpenSearch interface at http://mirador.gsfc.nasa.gov/mirador_dataset_opensearch.xml.

3.3.2 GI-cat

GI-cat²⁹ is an implementation of a broker catalog service. It allows clients to discover and evaluate geoinformation resources over a federation of data sources. It also publishes different catalog interfaces, allowing different clients to use the service.

GI-cat features caching and mediation capabilities and can act as a broker towards disparate catalog and access services: by implementing metadata harmonization and protocol adaptation, it is able to transform query results to a uniform and consistent interface. GI-cat is based on a service-oriented framework of modular components and can be customized and tailored to support different deployment scenarios.

²⁴<http://geonetwork-opensource.org/>

²⁵<https://api.echo.nasa.gov/opensearch/>

²⁶<https://earthdata.nasa.gov/echo>

²⁷<http://mirador.gsfc.nasa.gov/>

²⁸<http://disc.sci.gsfc.nasa.gov/about-us>

²⁹<http://essi-lab.eu/do/view/GIcat>

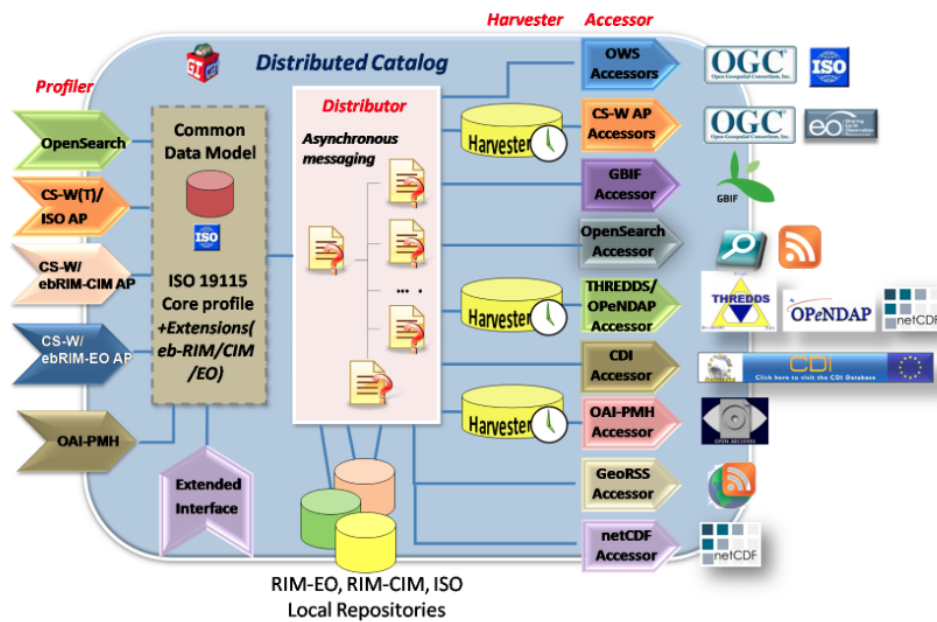


Figure 3.10: Supported profilers and accessors by GI-cat

GI-cat can access a multiplicity of catalogs services, as well as inventory and access services to discover, and possibly access, heterogeneous ESS resources. Specific components implement mediation services for interfacing heterogeneous service providers which expose multiple standard specifications; they are called Accessors.

These mediating components map the heterogeneous providers metadata models into a uniform data model which implements ISO 19115, based on official ISO 19139 schemas and its extensions. Accessors also implement the query protocol mapping; they translate the query requests expressed according to the interface protocols exposed by GI-cat, into the multiple query dialects spoken by the resource service providers. Currently, a number of well-accepted catalog and inventory services are supported, including several OGC Web Services (e.g. WCS, WMS), THREDDS Data Server, SeaDataNet Common Data Index, and GBIF.

GI-cat itself exposes several interfaces, including the OGC CSW interfaces (Core, ISO, ebRIM EO and ebRIM CIM). The query and result mediation is implemented by the Profilers. A distributor component implements the query distribution functionalities (e.g. results aggregation).

3.3.3 The GEOSS Discovery and Access Broker

The GEO Discovery and Access Broker³⁰ (DAB) is a middleware component which is in charge of interconnecting the heterogeneous and distributed capacities contributing to GEOSS; it became part of the GEOSS Common Infrastructure (GCI) since November 2011. The DAB provides three main functionalities:

1. Discovery of resources from brokered sources
2. Semantics-enriched discovery
3. Access of resources

Since it is a middleware component, DAB users are typically software agents, such as web-based or desktop client applications. These can exploit the DAB functionalities implementing the client-side of one (or more) of the protocols published by the DAB for the above functionalities. The available protocols include:

- OGC Catalog Service for the Web (CSW)
- OpenSearch with geo, time and semantic extensions
- Open Archive Initiative (OAI) PMH
- OGC Web Processing Service
- etc.

In order to simplify the development of applications and clients making use of the DAB, this high level client-side Open API (Application Program Interface) has been designed and developed in JavaScript. DAB is also part of the project EuroGeoss mentioned in 3.3.1 and was developed by Consiglio Nazionale delle Ricerche³¹ (CNR).

3.3.4 The Spatiotemporal RDF Store Strabon

Strabon³² has been developed by the UoA team over the years in order to manage linked geospatial data that changes over time.

³⁰<http://api.eurogeoss-broker.eu/docs/index.html>

³¹<http://www.iiia.cnr.it/>

³²<http://strabon.di.uoa.gr>

Strabon is a semantic spatiotemporal RDF store. It may be used to store linked geospatial data that changes over time and pose queries using two popular extensions of SPARQL. Strabon supports spatial datatypes enabling the serialization of geometric objects in OGC standards WKT and GML. It also offers spatial and temporal selections, spatial and temporal joins, a rich set of spatial functions similar to those offered by geospatial relational database systems, and support for multiple Coordinate Reference Systems. Strabon can be used to model temporal domains and concepts such as events, facts that change over time etc. through its support for valid time of triples, and a rich set of temporal functions. Strabon is built by extending the well-known RDF store Sesame and extends Sesame's components to manage thematic, spatial and temporal data that is stored in the backend relational database (RDBMS).

The first query language supported by Strabon is stSPARQL, a spatiotemporal extension of SPARQL 1.1 developed by the UoA group [22]. stSPARQL can be used to query data represented in an extension of RDF called stRDF. stRDF and stSPARQL have been designed for representing and querying geospatial data that changes over time, e.g., the growth of a city over the years due to new developments can be represented and queried using the valid time dimension of stRDF and stSPARQL respectively. The expressive power of stSPARQL makes Strabon the only fully implemented RDF store with rich spatial and temporal functionalities available today.

Strabon also supports the querying of static geospatial data expressed in RDF using a subset of the recent OGC standard GeoSPARQL [28] (OGC 11 052r4), which consists of the core, geometry extension and geometry topology extension.

3.4 Summary

In this chapter, we focused on the technical background of this thesis. We described the different types of vocabularies, from controlled vocabularies to ontologies, and various data formats and models that are utilized in the EO domain. We also presented related technologies, such as the OpenSearch, a collection of technologies and standards for describing search services and, the spatiotemporal RDF store Strabon.

Chapter 4

Related Activities

The ProdTrees platform was not the first system used for EO search. It re-uses components developed during the RARE project and its architecture follows the same approach as in RARE system, which is also a search engine for EO products. The main difference is the use of EO-netCDF. SMAAD, OTE and OTEG are also related projects, as they use vocabularies to annotate EO data. Finally, RESTo provides a semantic search service for EO data.

4.1 RARE

The objective of RARE¹ (Rapid Response Support Server) project is to make the EO products easily accessible by a larger group of users. The last years the technology has evolved and the deployed sensors have been multiplied. As a result, the amount of EO data collected and stored has exploded making the need for easy access to all these EO data much more intense.

RARE faces this problem by building a distributed software system accessible through a Web-based user interface. This system permits users to search for EO-related resources such as satellite images, maps and geo-localized features (e.g. coverages and points of interests). Although there are several web portals that also allows searching for and obtaining the EO products, they are too complex to be used by users who have limited knowledge of the EO domain. On the other hand, RARE satisfies the needs of non-expert users by offering them an easy to use interface, where they can search for EO products using free-text keywords or selecting ontology application terms. These terms are concepts from the CSCDA ontology (see Section 6.1.1), so the users, after the navigation within this ontology, can select the terms they want to use for their search. The benefit from this process is that the users are accustomed with the terminology used in the ontology, so it is easy for them to find the terms that fit their needs. RARE system offers also an interface for expert users, where they can use domain specific EO criteria for their search (e.g. sensor type, sensor resolution).

¹<http://deepenandlearn.esa.int/tiki-index.php?page=RARE+Project>

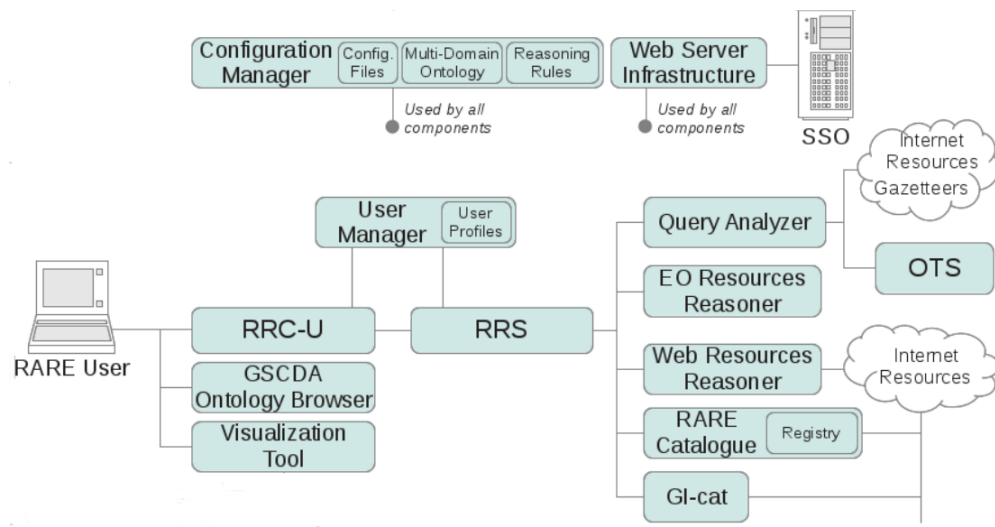


Figure 4.1: RARE architecture

To implement these ways of searching for EO products, RARE built a centralized service that interfaces with a number of on-line resources integrated for first time. These resources include:

1. a terminology service supporting the navigation facility between related application terms (OTS)
2. a query analyzer that augments the knowledge with unforeseen concepts and relations extracted from the Internet
3. gazetteers used to resolve place names (e.g. GeoNames)
4. various reasoners mapping the application terms selected by the users to product categories
5. a centralized catalogue service that collects the properties of on-line resources and other remote catalogues

RARE hides as much as possible the technical information related to the resources themselves: the system determines, based on elaborate mapping rules defined with the help of domain experts, which categories of resources are valuable in regards to the application terms entered by the users. This mechanism greatly reduces the time and the knowledge needed to search for, and obtain, the right resource for the right problem. It brings the advantages of EO-derived resources into the users domain of interest with minimal effort.

Figure 4.1 represents the architecture of RARE platform. RARE users access the platform through its Web interface (Rapid Response Client for Users, or RRC-U). This Web interface is implemented as a Liferay portlet. User authentication is managed at the portal level, possibly integrated with a centralized Single Sign-On infrastructure (e.g. EO-SSO at ESA/ESRIN). RARE receives user information (name, e-mail address) directly from the hosting portal. An ontology browser allows navigating and selecting application terms defined in the CSCDA ontology. A Web-based visualization tool allows displaying certain types of products including WMS and WFS data.

Except for user management, all the interactions with the backend modules go through the Rapid Response Server (RRS). This is in particular the case (1) when a query string entered by the user needs to be disambiguated and (2) when EO and Web resources must be searched for. In the first case, the RRS invokes the Query Analyzer for disambiguating search queries. The Query Analyzer processes the query string, identifying the words that may be mapped to application terms, location names (toponyms), time constraints, or other types of named entities. In the second case, typically when the user agrees with the disambiguated query, the RRS is invoked to obtain the matching (EO and non-EO) resources. The RRS interacts with the EO Resources Reasoner to obtain the filter criteria to be used for querying the GI-cat catalogues broker. The list of matching resources is displayed to the users.

RARE started on December 2010 and will reach its completion in the coming months.

4.2 SMADD

Semantic annotation, as described in Section 2.5, pursues an approach to specify the meaning of elements in a data or metadata element by pointing to the concepts of an ontology, or at least an agreed vocabulary. It is based upon the assumption that an ontology represents the shared knowledge of a community, e.g. a thematic expert group or an international expert initiative, in terms of a ‘conceptualisation’.

SMAAD² (Semantic Web Mediated Across Domains) is a HMA-related research project that focuses on the idea of an Ontology Access Service and a Thesaurus Access Service (OGC 07-097) and interprets them in the context of the HMA service environment and the latest technological developments.

The basic role of the Ontology Access Service is to ‘support the read access to the specification of a logical ontology and to export or import a complete specification of a

²<https://wiki.services.eoportal.org/tiki-index.php?page=SMAAD>

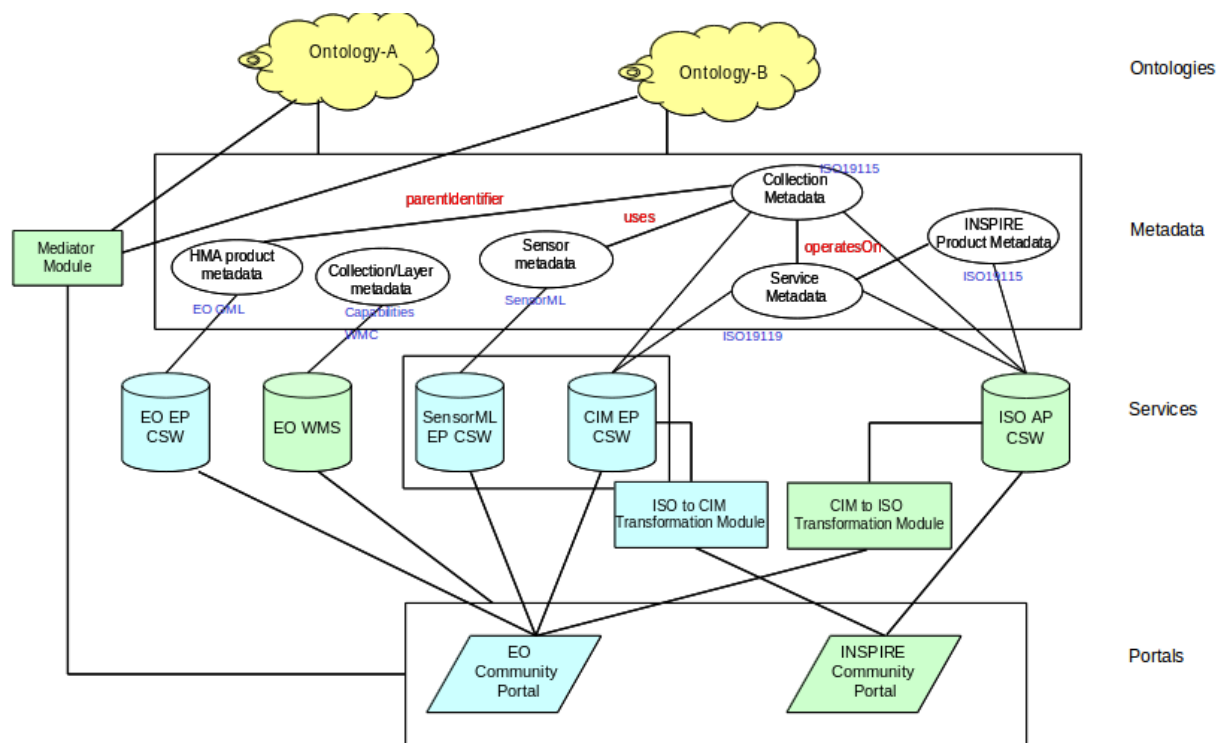


Figure 4.2: Resource discovery using ontologies on top of portals

logical ontology into an ontology store’, whereas the Thesaurus Access Service ‘supports read and write access to a thesaurus that may be multi-lingual’. Thus, a thesaurus is understood to be ‘a variant of an ontology restricting the relations used to a priori relationships between terms, e.g. questioning whether the meaning of two terms is similar, broader, or narrower’.

SMAAD encompasses both aspects under an HMA Ontology Access Service verifying that the service-oriented architecture (SOA) and standards proposed within HMA permit easy linking of metadata keywords to more than one domain ontology and thesaurus.

Figure 4.2 shows in an abstract way how we can achieve the discovery of resources (services, collections, products, sensors) based on the use of ontologies. The ontologies can describe appropriately the metadata used in services built on top of EO portals.

SMAAD addressed the following scenarios:

1. Ontology Based Resource Discovery

- Keywords List from Controlled Vocabulary
- Browsing Ontology for Keywords
- Keywords Narrowing, Widening, or Relating

- Keywords as Terms in Other Language
- 2. Ontology Based Resource Metadata Explanation
 - Following links in metadata records
 - Translate keywords in metadata records
- 3. Ontology Based Service Capabilities Discovery
 - Similar to Scenario 1 (metadata in WxS Capabilities document)
- 4. Ontology Based Layer Discovery
 - Similar to Scenario 1 and 3 (metadata in WMC document)
- 5. Mediation
 - Search using different controlled vocabulary
 - Show keywords from mapped ontology in metadata records
- 6. Discovery of Ontologies and Mediators
 - Discovery of Ontologies Used by a Service
 - Discovery of Mediators
- 7. Ontology Access
 - Ontology Capabilities Discovery
 - Textual Ontology Search
 - Direct Access
 - Results Ranking

SMAAD followed in all use-cases a standards-based approach (HMA, INSPIRE, OGC, ISO, W3C, OpenSearch). Another achievement of SMAAD was the amendment and promotion of the following standards:

1. OGC 08-167r2, Semantic Annotations in OGC Standards
2. OGC 11-035, EO Collection and Service Discovery
3. OGC 08-197r4: INSPIRE Conformance Class of OGC Cataloguing of ISO Metadata following SMAAD implementation feedback

4. CIM EP 07-038 following SMAAD implementation feedback being formalised in HMA-S

SMAAD started in June of 2010 and ended in March of 2013.

4.3 OTE/OTEG

OTE and OTEG are two more related projects based on the idea that the semantic terms can be used to make the EO data easily accessible, homogenize it with auxiliary information and remove any confusion (naming conflicts, unit conflicts, similar but not identical meaning).

A shared Ontology (a semantic representation of knowledge by means of a hierarchy of concepts and their relations) and Terminology (description of meaning of all terms used with synonyms and related words) are essential for the correct exchange of information between human beings and between computer programs. Ontology and Terminology can help within the EO domain, for easing the work among partners, and towards non-EO domains, in simplifying the identification of EO products relevant for specific applications.

The objective of Ontology and Terminology for Earth Observation³ (OTE) project was to: 1) Enlarge and spread the use of EO satellite data, 2) Identify relevant / useful EO data for users belonging to different application domains (non-EO) and 3) simplify the interaction among partners in the context of Ground Segment infrastructures. These needs were supported through the identification of the minimum number of concepts, relations and terms necessary to semantically describe relations among EO resources and non-EO domains.

During the OTE project two services were developed. The first one is a terminology search that searches for terms that match provided words, retrieves information on terms definition and information on Terminology / Ontology mapping, i.e. concepts associated to terms. The second web service implements an Ontology Navigation: search for concepts that match provided words and navigate into the selected sub-tree, inspect a selected domain and retrieve related EO Products and retrieve the ontology structure.

A simple web based interface⁴ allows users accessing search and navigation functionalities of the two web services described above. The application permits non-EO experts to inspect and retrieve information about terms and concepts, identify and access rele-

³<http://deepenandlearn.esa.int/tiki-index.php?page=OTE+Project>

⁴<http://gmesdata.esa.int/OTE/navigateInfoDomain>

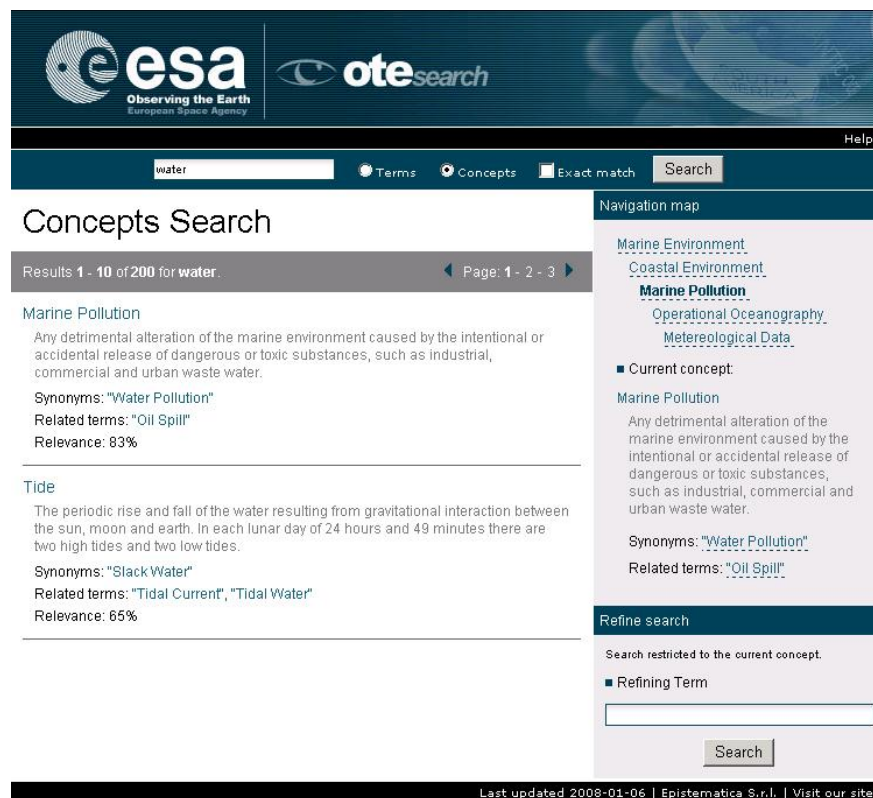


Figure 4.3: Example of OTE software prototype web page

vant EO resources starting from their domain of interest or expertise and using terms familiar to them and, finally, simplify the interactions among the partners of the GMES Space Component (GSC) Ground Segment through the use of a shared ontology for ground segment components.

Figure 4.3 shows an example of the web application developed for accessing Ontologies and Terminologies is reported.

The general purpose of the Open Access Ontology/Terminology for the GMES Space Component (OTEG) project (ended on May 2009) was to revise and expand the results of the OTE project in order to design, implement and validate an openly available GSCDA Semantics, including Multi-domain Thesaurus, Multi-domain Vocabulary and GSCDA Taxonomy.

Like the OTE, the OTEG⁵ project provides a software system for easily accessing the defined semantic information (ontologies and terminologies). The system is composed of two web services and three web applications. The web services provide functions to access the GSCDA Semantics and the web applications call these functions to create a graphical

⁵<http://deepenandlearn.esa.int/tiki-index.php?page=OTEG+Project>

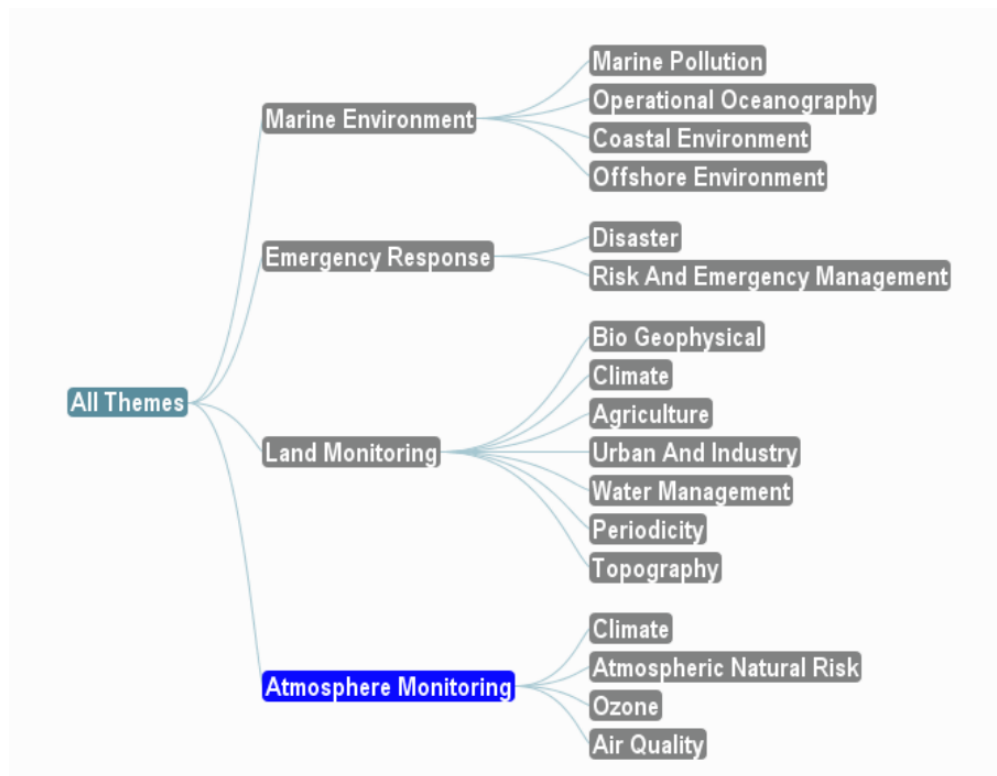


Figure 4.4: A little excerpt from GSCDA Multi-Domain Thesaurus

interface for the end user. The two web services are responsible for terminology search and ontology navigation. Through the web based interface the users can enter search keywords and retrieve information about the results (e.g., term definitions). It is possible to navigate the hierarchical structure of the taxonomy, thus exploring different domains (Figure 4.4). For each concept in the taxonomy, the system will provide a list of all the relevant EO products, showing the related information (missions, sensor, datasets).

The results of the OTEG project is that the knowledge ESA has on its products has been described through the formalisms of logic. This made the use of automated reasoning possible leading to the development of a software able to use the ESA knowledge to drive customers towards products logically, not statistically, more pertinent to the search criteria. This solution enables the customers to perform searches using all the knowledge ESA has on its products, rather than using only their knowledge. Figure 4.5 shows the OTEG web client. The client is available online at <http://gmesdata.esa.int/OTE/navigateInfoDomain>.

The screenshot displays the OTEG web client interface. At the top, the ESA logo and 'GMES Space Component Data Access' are visible. A search bar contains the term 'ocean', and a 'Submit' button is present. Below the search bar, the results are categorized into several sections:

- Search Results for "ocean"**: Shows 'Results: 1 - 10 of 12' and 'Page: 1 - 2'.
- Ocean colour**: Describes water colour as an indicator of water quality, caused by absorption and scattering of sunlight. Includes a 'Focus on' button.
- Ocean surface monitoring**: Describes currents, waves, temperatures, and tides. Includes a 'Navigate' button.
- Ocean level**: Describes changes in sea level connected to tides, air pressures, winds, and global warming. Includes a 'Navigate' button.
- Marine environment**: Describes marine environments including estuaries, coastal marine, and nearshore zones. Includes a 'Navigate' button.
- Water quality monitoring**: Describes water quality monitoring involving chemical conditions of water (sediments, fish tissue, dissolved oxygen, nutrients, metals, oils, pests). Includes a 'Navigate' button.

On the right side, there is a **Navigation Map** showing a map of the world with a focus on 'Ocean colour'. Below the map, there is a **Focus on** section for 'Ocean colour' with a detailed description and synonyms. A **Refine search** section is also present, showing the search is restricted to the current focused concept: 'Ocean colour'.

At the bottom, there is a **Related EO Products** section showing 7 entries related to 'Ocean colour':

Product Type	Dataset	Sensor	Mission
OrbView.SS1.SWF_L2B	DAP_MG3_03	SeaWiFS	OrbView-2
OrbView.SS1.SWF_L2C	DAP_MG3_03	SeaWiFS	OrbView-2
OrbView.SS1.SWF_L2A	DAP_MG3_03	SeaWiFS	OrbView-2
MER_RR__2P	DAP_MG3_03	MERIS	ENVISAT

Figure 4.5: OTEG web client showing results for the term "ocean"

4.4 RESTo

RESTo⁶ (REstful Semantic search Tool for geOspatial) is a framework that provides a semantic search service on Earth Observation data. It implements OpenSearch OGC 13-026 standard (OpenSearch Extension for Earth Observation). RESTo is written in PHP, uses PostgreSQL and PostGIS and follows a RESTful⁷ approach to manage resources.

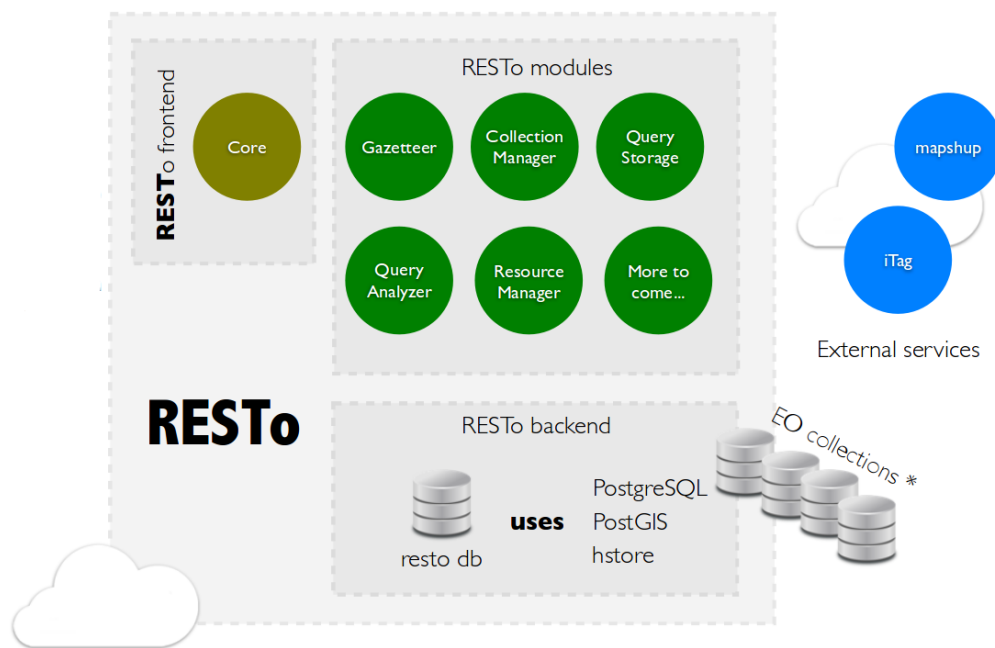


Figure 4.6: RESTo architecture

Figure 4.6 displays the architecture of RESTo. EO Collections can be stored within RESTo database or in external databases. When a resource is inserted to the system, iTag⁸ automatically tags the resource with location and land use. iTag is used as an external service to automatically tag a geographical footprint against location, land cover, population count, etc. The results of the search can be also viewed in mapshup⁹, another external service used by RESTo.

RESTo uses the Query Analyzer to translate natural language query into a set of EO OpenSearch parameters. Query string analysis algorithm is based on simple recognition

⁶<http://mapshup.info/resto/>

⁷REST (Representational State Transfer) is an architectural style consisting of a coordinated set of architectural constraints applied to components, connectors, and data elements, within a distributed hypermedia system. For more information, visit http://en.wikipedia.org/wiki/Representational_state_transfer

⁸<https://github.com/jjrom/itag>

⁹<http://mapshup.info/>

of words and pattern. The basic steps are:

1. Split query string into list of unitary words
2. Extract “key=value” strings (e.g. orbitNumber=4)
3. Extract platforms and instruments. Platforms and instruments list are stored within common dictionary.
4. Remove excluded words and non dictionary words with length less than 4 characters (e.g. “area **of** Mexico **in** 2012”)
5. Extract patterns and dated (e.g. “acquired in the **last 2 days**”)
6. Extract keywords (e.g. “**urban** area in **France**”)
7. Extract location on remaining words (e.g. “images acquired in **Toulouse**”)

All detectable words are stored within a dictionary. RESTo supports also other languages (french, italian, german), synonyms (e.g. unit “M” is “m”, “meter” or “meters”) and automatic typing error correction using a similarity function.

RESTo embeds a Gazetteer service to detect location. The Gazetteer is based on Geonames and contains more than 9.000.000 toponyms.

Figure 4.7 shows a number of results for the following query “images of **urban** area in **France** acquired in **2013** with **less** than **25 %** of **cloud** cover”. The keywords with bold are the search parameters used in the OpenSearch query. Each search result has a “human readable url” that can be index by a web crawler (i.e. google robots). In our example, the url is `http://mapshup.info/resto/Spot/?format=html&lang=en&q=+images+of+urban+area+in+France+acquired+in+2013+with+less+than+25+%25+of+cloud+cover`. Finally, keywords on resources are links to search requests, thus they can be indexed by web crawler and so on.

4.5 Summary

In this chapter, we surveyed related activities to EO search and semantic technologies in the EO domain.

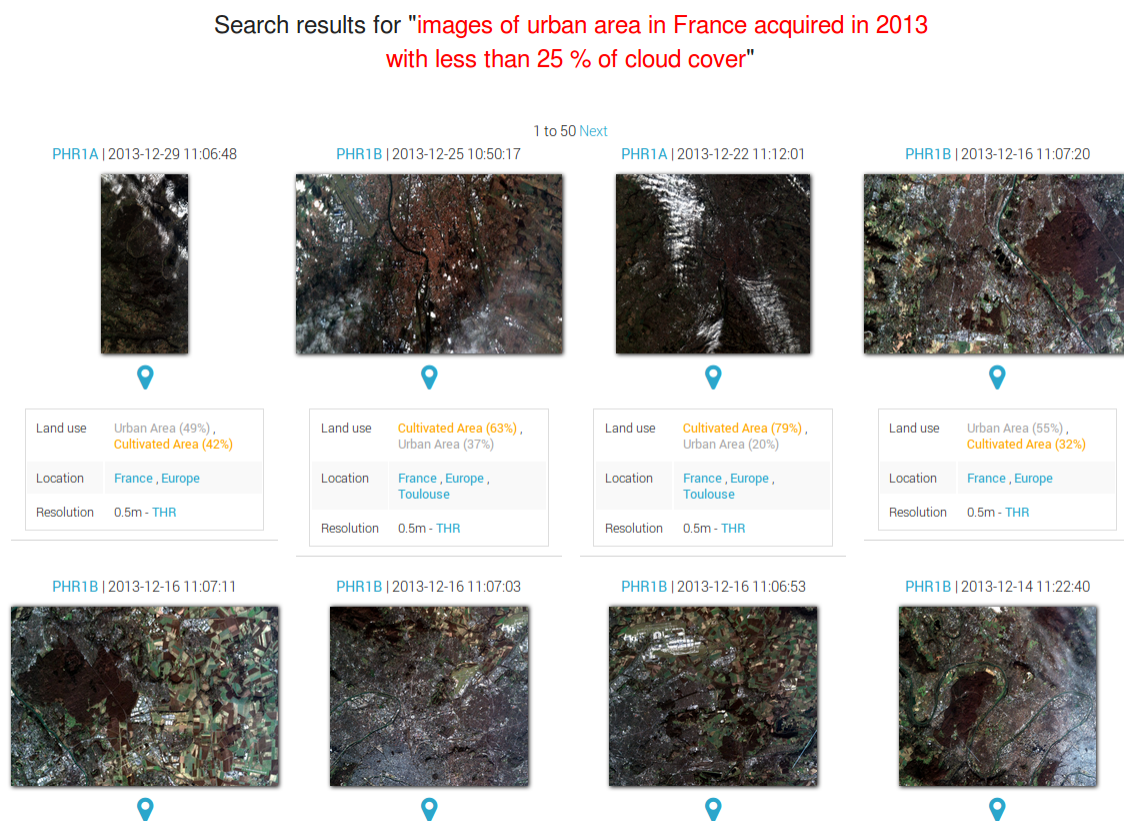


Figure 4.7: A number of search results for the query “images of urban area in France acquired in 2013 with less than 25 % of cloud cover”

Chapter 5

Semantic Search and Discovery of EO-netCDF Products

A typical problem that occurs with systems such as the ones described in the previous chapter, is that they focus on satisfying either casual or advanced users, and not both types of users. Another issue is the lack of a common vocabulary for accessing EO products. This limitation prevents other organizations (scientific communities, commercial companies) from implementing their own compatible solutions, adapted to their own requirements and applications domain.

In this chapter, we present the ProdTrees platform, a semantically-enabled search engine for EO products developed by the project ProdTrees. The system uses semantic technologies to allow users to search for EO products in an application-oriented way using free-text keywords (as in search engines like Google), their own domain terms or both, in conjunction with the well-known interfaces already available for expert users. A specific innovation of the presented system is the use of a new netCDF convention, called EO-netCDF, for accessing EO products annotated with netCDF. In the next sections, we firstly introduce this new convention, and then we give the overview of the ProdTrees system.

5.1 EO-netCDF

5.1.1 Conventions

The EO metadata convention for the netCDF standard, called EO-netCDF, was defined and developed in order to address the issue of discovery, evaluation, access and use of EO products in a standard way. In particular, EO-netCDF provide a standardized solution that permits annotating EO products in such a manner that official and third-party software libraries and tools are able to search for products using advanced tags and controlled parameter names. Annotated EO products can be automatically supported by all the compatible software. Because the entire product's information come from the

annotations and the standards, there is no need for integrating extra components and data structures that have not been standardized.

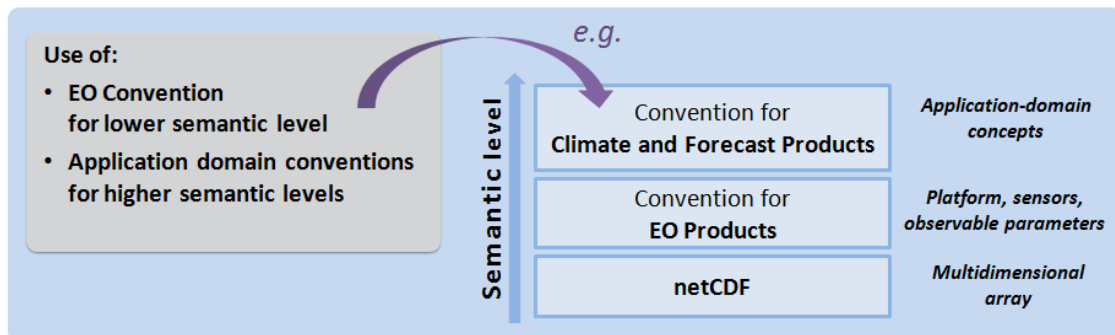


Figure 5.1: The Concept of EO-netCDF

The EO-netCDF are hierarchically structured, similarly to the EOP O&M (see Section 2.4, Figure 2.2). The metadata elements belonging to the more generic products are inherited by the specific products. The concept of the EO-netCDF is that they can be used for describing lower semantic levels, while application domain conventions (like CF-netCDF) can be used for higher semantic levels (Figure 5.1). The current version of the EO-netCDF convention [4] contains a full mapping of EOP O&M to netCDF, SENTINEL-1 metadata profile (see 7.1.2), and an EO vocabulary. The EO vocabulary is a set of standard names that originates from the EOP specification and was increased and extended with concepts connected with EO activities and considering other relevant EO specifications. A version of the EO vocabulary in RDF is also available.

A fraction of the metadata elements that are part of the EO convention are presented below. They are listed using tables, which consist of “group of elements”. A group link another (nested) group by means of special elements having the type “Group”(formatted in red). Metadata elements usually map to netCDF attributes, unless when “Group” or “Variable” is specified. Each element is defined by a name, a format type, a description and an obligation for the element to be documented (mandatory / optional / conditional). Cardinality is documented in case it is greater than 1.

EO Products elements

The main group of metadata elements, called *earth_observation_information* is displayed in Figure 5.2. One of the included groups, the *earth_observation_equipment*, is expanded in Figure 5.3.

Element name	Format	Description	Obligation / Cardinality
phenomenon_time_begin_position	String (ISO 8601)	Acquisition start date time dateTime in ISO 8601 format (CCYY-MM-DDThh:mm[:ss[.cc]]Z)	M
phenomenon_time_end_position	String (ISO 8601)	Acquisition end date time dateTime in ISO 8601 format (CCYY-MM-DDThh:mm[:ss[.cc]]Z)	M
result_time_time_position	String (ISO 8601)	The time when the result becomes available. DateTime in ISO 8601 format (CCYY-MM-DDThh:mm[:ss[.cc]]Z)	M
earth_observation_equipment	Group	Platform/Instrument/Sensor used for the acquisition and the acquisition parameters.	M
observed_property	String array	An xlink to the observed property definition	M
footprint	Group	Observed area on the ground or its projection i.e. the footprint of acquisition.	O
earth_observation_result	Group	Earth Observation result metadata composed of the browse, mask and product description.	O
earth_observation_metadata	Group	Additional external metadata about the data acquisition.	M
result_quality	Group	Result quality information	O [0...n]

Figure 5.2: Earth Observation Information group

Attribute	Format	Description	Obligation
platform_information	Group	Platform information.	O
instrument_information	Group	Instrument information.	O
sensor_information	Group	Sensor information.	O
acquisition_information	Group	Acquisition parameter information.	O

Figure 5.3: Earth Observation Equipment group

Attribute	Format	Description	Obligation / Cardinality
sensor_type	Enumeration	Sensor type. Valid values: <ul style="list-style-type: none"> • OPTICAL • RADAR • ALTIMETRIC • ATMOSPHERIC • LIMB 	O
sensor_operational_mode	Variable (String)	Sensor mode. Possible values are mission specific and should be retrieved using the optional "code_space" attribute.	O
sensor_resolution	Variable (Double)	Sensor resolution. UDUNITS values are suggested for the "units" attribute.	O
sensor_swath_identifier	Variable (String)	Swath identifier (e.g. Envisat ASAR has 7 distinct swaths (I1,I2,I3,...I7) that correspond to precise incidence angles for the sensor). Value list can be retrieved with codeSpace.	O
wavelength_information	Group	List of discrete wavelengths observed in the product	O [0..n]

Figure 5.4: Sensor Information group

Another expansion of a group, which is included in the *earth_observation_equipment* group and called *sensor_information*, is shown in Figure 5.4.

SENTINEL-1 Products attributes

The thematic and mission specific products add new attributes and attribute groups or modify obligation and cardinality of existing attributes.

SENTINEL-1 EO elements set extends the *EO elements* set, displayed in Figure 5.2, by adding the attribute in Figure 5.5.

Element name	Format	Description	Obligation / Cardinality
quality_disclaimer	Group	Quality disclaimer information	0

Figure 5.5: SENTINEL-1 EO elements

Similarly, the *SENTINEL-1 sensor information attributes* set extends the *EO sensor information* set by adding the following attributes in Figure 5.6.

Attribute	Format	Description	Obligation / Cardinality
sensor_type	Enumeration	Sensor type. Allowed values: <ul style="list-style-type: none"> • RADAR 	0
sensor_operational_mode	Enumeration	Restricted enumeration values Sensor mode. Possible values are mission specific and should be retrieved using the optional "code_space" attribute. Allowed values: <ul style="list-style-type: none"> • SM_SP • SM_DP • IW_SP • IW_DP • EW_SP • EW_DP • WV_SP 	0
sensor_swath_identifier	Variable (String)	Restricted String to Enumeration Swath identifier (e.g. IW, EW, WV or from S1 to S7) Added max length constraint	0 Max length: 100

Figure 5.6: SENTINEL-1 sensor information elements

Implementation

The implementation of the convention takes into account both netCDF 4 and netCDF 3 data models. The later implementation includes workarounds, in order to handle the new features added to the netCDF 4 data model, such as Attribute Groups, String Arrays, etc.

An implementation example [4] in NetCDF 4 (NcML) is included beneath. The example is a translation from the eop_example.xml document included in [18].

```
<?xml version="1.0" encoding="UTF-8"?>
<netcdf xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.unidata.ucar.edu/namespaces/netcdf/ncml-2.2
http://www.unidata.ucar.edu/schemas/netcdf/ncml-2.2.xsd"
xmlns="http://www.unidata.ucar.edu/namespaces/netcdf/ncml -2.2">

  <attribute name="Conventions" value="E0"/>
  <attribute name="title" value="Sample NetCDF-E0"/>
  <attribute name="history" value="Manually edited"/>

  <dimension name="track" length="800"/>
  <dimension name="xtrack" length="100"/>

  <group name="earth_observation_information">
    <attribute name="phenomenon_time_begin_position">2001-08-22T11:02:47.000
    </attribute>
    <attribute name="phenomenon_time_end_position">2001-08-22T11:02:47.999
    </attribute>
    <attribute name="result_time_position">2001-08-22T11:02:47.999</attribute>
    <group name="earth_observation_equipment" >
      <group name="platform_information">
        <attribute name="short_name">PHR</attribute>
        <attribute name="serial_identifiser">1A</attribute>
      </group>
      <group name="instrument_information">
        <attribute name="short_name">PHR</attribute>
      </group>
      <group name="sensor_information">
        <attribute name="sensor_type">OPTICAL</attribute>
        <variable name="sensor_operational_mode">
          <attribute name="code_space">urn:eop:PHR:sensorMode
          </attribute>
          <values>PX</values>
        </variable>
        <variable name="sensor_resolution">
          <attribute name="units">meters</attribute>
        </variable>
      </group>
      <group name="acquisition_information">
        <variable name="orbit_number">
          <values>12</values>
        </variable>
        <variable name="last_orbit_number">
          <values>12</values>
        </variable>
      </group>
    </group>
  </group>
```

```

<attribute name="orbit_direction">ASCENDING</attribute>
<variable name="wrs_longitude_grid">
  <attribute name="code_space">EPSG</attribute>
  <values>12</values>
</variable>
<variable name="wrs_latitude_grid">
  <attribute name="code_space">EPSG</attribute>
  <values>12</values>
</variable>
<variable name="wrs_latitude_grid">
  <attribute name="code_space">EPSG</attribute>
  <values>12</values>
</variable>
<variable name="across_track_incidence_angle" >
  <attribute name="units">degree</attribute>
  <values>-14.0</values>
</variable>
<variable name="along_track_incidence_angle">
  <attribute name="units">degree</attribute>
  <values>-13.9</values>
</variable>
<variable name="pitch">
  <attribute name="units">degree</attribute>
  <values>0</values>
</variable>
<variable name="roll">
  <attribute name="units">degree</attribute>
  <values>0</values>
</variable>
<variable name="yaw">
  <attribute name="units">degree</attribute>
  <values>0</values>
</variable>
</group>
</group>
<attribute name="observed_property">#phenom1</attribute>
<group name="footprint">
  <variable name="multi_extent_of">
    <attribute name="gml_type">LinearRing</attribute>
    <attribute name="srs_name">EPSG:4326</attribute>
    <values>2.1025 43.516667 2.861667 43.381667 2.65 42.862778
    1.896944 42.996389 2.1025 43.516667 </values>
  </variable>
  <variable name="center_of">
    <attribute name="gml_type">Point</attribute>
    <attribute name="srs_name">EPSG:4326</attribute>
    <values>2.374167 43.190833</values>
  </variable>
</group>
<group name="earth_observation_result">
  <group name="browse_information">
    <attribute name="type">QUICKLOOK</attribute>
    <variable name="reference_system_identifier">
      <attribute name="code_space">EPSG</attribute>
      <values>epsg:4326</values>
    </variable>
    <attribute name="filename">http://etc</attribute>
  </group>

```

```

<group name="mask_information">
  <attribute name="type">CLOUD</attribute>
  <attribute name="format">VECTOR</attribute>
  <variable name="reference_system_identifier">
    <attribute name="code_space">EPSG</attribute>
    <values>epsg:4326</values>
  </variable>
  <attribute name="filename">http://etc</attribute>
</group>
<group name="parameter_information">
  <attribute name="units">meters</attribute>
  <variable name="phenomenon">
    <values>xyzdef</values>
  </variable>
</group>
</group>
<group name="earth_observation_metadata">
  <attribute name="identifier">DS_PHR1A_20010822110247_TLS_PX_E123N45_0101_01234
  </attribute>
  <attribute name="acquisition_type">NOMINAL</attribute>
  <attribute name="product_type">TBD</attribute>
  <attribute name="status">ARCHIVED</attribute>
  <group name="downlink_information">
    <variable name="downlink_acquisition_station" >
      <attribute name="code_space">urn:eop:PHR:stationCode</attribute>
      <values>TLS</values>
    </variable>
  </group>
  <group name="archiving_information">
    <variable name="downlink_acquisition_station" >
      <attribute name="code_space">urn:eop:PHR:stationCode
      </attribute>
      <values>TLS</values>
    </variable>
    <attribute name="archiving_date">2001-08-22T11:02:47.999
    </attribute>
  </group>
</group>
</group>
<variable name="swathData" shape="track xtrack" type="float">
  <attribute name="_FillValue" type="float" value="NaN"/>
  <attribute name="units" value="meters"/>
</variable>
<variable name="xtrack" shape="x" type="double">
  <attribute name="long_name" value="xtrack"/>
  <attribute name="units" value="meters"/>
</variable>
<variable name="track" shape="y" type="double">
  <attribute name="long_name" value="track"/>
  <attribute name="units" value="meters"/>
</variable>
</netcdf>

```

Figure 5.7: EO-netCDF implementation example

Finally, the EO-netCDF convention is compliant with CF-netCDF and netCDF-U conventions, and is expected to be submitted as a standard to the OGC.

5.1.2 Libraries

The EO-netCDF data model is defined as a set of NetCDF conventions for annotating datasets with EO metadata and is based on NetCDF 3 and NetCDF 4 data models. These EO conventions are also compatible and recommend the use of CF-netCDF conventions. As a result, the software libraries and tools for netCDF and CF-netCDF support also the EO-netCDF data model and can be used to create, read and modify EO-netCDF files. In the following subsections, we described popular netCDF libraries that can be used also for handling EO-netCDF files.

NetCDF Java API

The NetCDF-Java Library¹ is a Java interface to NetCDF files, as well as to many other types of scientific data formats. The library is freely available and the source code is released under the (MIT style) netCDF C library license. Previous versions used the GNU LGPL.

The NetCDF-Java library implements the Common Data Model [7] (CDM), a generalization of the NetCDF, OpenDAP and HDF5 data models. The library is a prototype for the NetCDF-4 project, which provides a C language API for the “data access layer” of the CDM, on top of the HDF5 file format. The NetCDF Java library is a Java framework for reading netCDF and other file formats into the CDM, as well as writing to the netCDF-3 file format. The NetCDF-Java library also implements NcML, which allows adding metadata to CDM datasets and creating virtual datasets through aggregation.

In order to demonstrate the API, an example² is included below. The example writes a two-dimensional array of sample data that look like:

```
netcdf simple_xy {
    dimensions:
        x = 6 ;
        y = 12 ;
```

¹<http://www.unidata.ucar.edu/software/thredds/current/netcdf-java/documentation.htm>

²**Sources:** <http://www.unidata.ucar.edu/software/netcdf/examples/programs/>, <http://www.unidata.ucar.edu/software/netcdf/docs/netcdf-tutorial.html>

```
variables:
  int data(x, y) ;
data:
  data =
    0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,
    12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23,
    24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35,
    36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47,
    48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59,
    60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71 ;
}
```

The Java code is shown below:

```
/* This is a very simple example which writes a 2D array of
sample data. To handle this in netCDF we create two shared
dimensions, "x" and "y", and a netCDF variable, called "data".
*/

import ucar.nc2.Dimension;
import ucar.nc2.NetcdfFileWriteable;
import ucar.ma2.*;

import java.io.IOException;
import java.util.ArrayList;

public class Simple_xy_wr {

    public static void main(String args[]) {
        // We are writing 2D data, a 6 x 12 grid.
        final int NX = 6;
        final int NY = 12;

        // Create the file.
        String filename = "simple_xy.nc";
        NetcdfFileWriteable dataFile = null;
```

```
try {
    dataFile = NetcdfFileWriteable.createNew(filename, false);

    // Create netCDF dimensions,
    Dimension xDim = dataFile.addDimension("x", NX );
    Dimension yDim = dataFile.addDimension("y", NY );

    ArrayList dims = new ArrayList();

    // define dimensions
    dims.add( xDim);
    dims.add( yDim);

    // Define a netCDF variable.
    //The type of the variable in this case
    // is ncInt (32-bit integer).
    dataFile.addVariable("data", DataType.INT, dims);

    // This is the data array we will write.
    //It will just be filled with
    // a progression of numbers for this example.
    ArrayInt.D2 dataOut = new ArrayInt.D2(
        xDim.getLength(), yDim.getLength());

    // Create some pretend data.
    //If this wasn't an example program, we
    // would have some real data to write,
    //for example, model output.
    int i,j;

    for (i=0; i<xDim.getLength(); i++) {
        for (j=0; j<yDim.getLength(); j++) {
            dataOut.set(i,j, i * NY + j);
        }
    }

    // create the file
    dataFile.create();
}
```



```
// Write the pretend data to the file.
//Although netCDF supports
// reading and writing subsets of data,
//in this case we write all
// the data in one operation.
dataFile.write("data", dataOut);

} catch (IOException e) {
    e.printStackTrace();
} catch (InvalidRangeException e) {
    e.printStackTrace();
} finally {
    if (null != dataFile)
        try {
            dataFile.close();

            {\small      } catch (IOException ioe) {
                ioe.printStackTrace();
            }
        }
    }

    System.out.println( "SUCCESS writing example file simple_xy.nc!");
}
}
```

NetCDF C/C++ API

The NetCDF C/C++ library provides an application and machine-independent interface to self-describing, array-oriented data. It supports an abstract view of such data as a collection of named variables and their attributes, and provides high-level access to data that is faithful to the abstraction.

There are two interfaces: the NetCDF C++ Interface³ and the NetCDF C Interface⁴.

³<http://www.unidata.ucar.edu/software/netcdf/docs/netcdf-cxx/index.html#Top>

⁴<http://www.unidata.ucar.edu/software/netcdf/docs/netcdf-c/index.html#Top>

ncgen

The `ncgen`⁵ tool reads a textual representation of a netCDF dataset and generates the corresponding binary netCDF file or a program to create the netCDF dataset. `ncgen` is a utility that generates either a netCDF-3 (i.e. classic) binary `.nc` file, a netCDF-4 (i.e. enhanced) binary `.nc` file or a file in some source language that when executed will construct the corresponding binary `.nc` file. The CDL (network Common Data form Language) is the language that is used as the input to `ncgen`. `ncgen` checks the syntax of the input CDL file. Options may be specified, for example, to create the corresponding netCDF file, or to generate a C program that uses the netCDF C interface to create the netCDF file.

`ncgen` may be used with the program `ncdump` to perform some simple operations on netCDF files. For example, to rename a dimension in a netCDF file, use `ncdump` to get a CDL version of the netCDF file, edit the CDL file to change the name of the dimension, and use `ncgen` to generate the binary netCDF file from the edited CDL file.

ncdump

The `ncdump`⁶ utility generates a textual representation of a specified netCDF file on standard output, optionally excluding some or all of the variable data present in the input. The text representation in CDL (network Common Data form Language) can be viewed, edited, or served as input to `ncgen`. Hence `ncgen` and `ncdump` can be used as inverses to transform the data representation between binary and text representations. `ncgen` documentation contains a description of CDL and netCDF representations.

`ncdump` may also be used to determine what kind of netCDF file is used (which variant of the netCDF file format) with the `-k` option. This utility may also be used as a simple browser for netCDF data files, to display the dimension names and lengths; variable names, types, and shapes; attribute names and values; and optionally, the values of data for all variables or selected variables in a netCDF file.

ncview

The netCDF visual browser is a utility called `ncview`. It provides an extremely quick and easy way to visualize data. It is possible to make line plots by clicking a few buttons

⁵<https://www.unidata.ucar.edu/software/netcdf/docs/netcdf/ncgen.html>

⁶<http://www.unidata.ucar.edu/software/netcdf/docs/netcdf/ncdump.html>

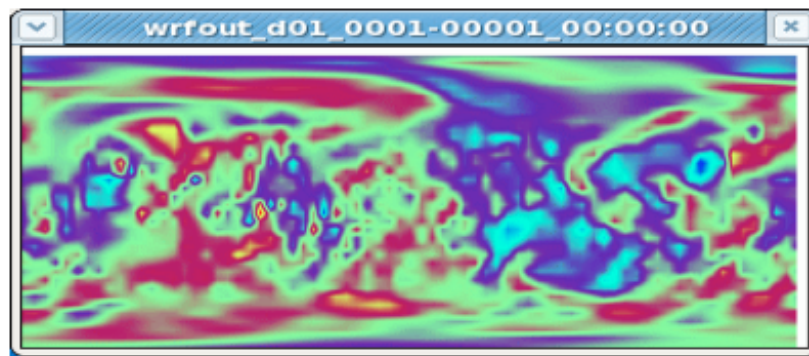


Figure 5.8: Example display of the ncview browser

and simple commands. `ncview` displays a 2-dimensional, color representation of data in a netCDF file (Figure 5.8). The data can be animated in time (making simple movies), flip or enlarge the picture, scan through various axes, change colormaps, etc. `ncview` is not an analysis package. Rather, its purpose in life is to view movies or simple plots of data stored in netCDF format files quickly, easily, and simply.

Panoply

Panoply⁷ is a JAVA application developed by NASA for viewing netCDF files. It plots geogridded and other arrays from netCDF, HDF, GRIB, and other datasets.

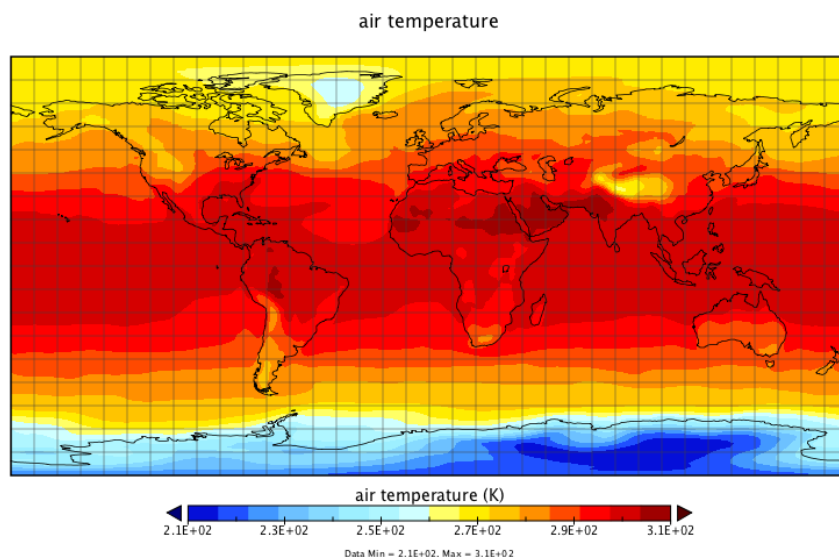


Figure 5.9: Plot for air temperature in Panoply

⁷<http://www.giss.nasa.gov/tools/panoply/>

Figure 5.9 displays a plot that visualizes a variable for air temperature. Finally, Figure 5.10 shows an EO-netCDF file loaded in Panoply. As you can see, it contains variables that describe EO metadata (e.g. platform_information, sensor_information etc.)

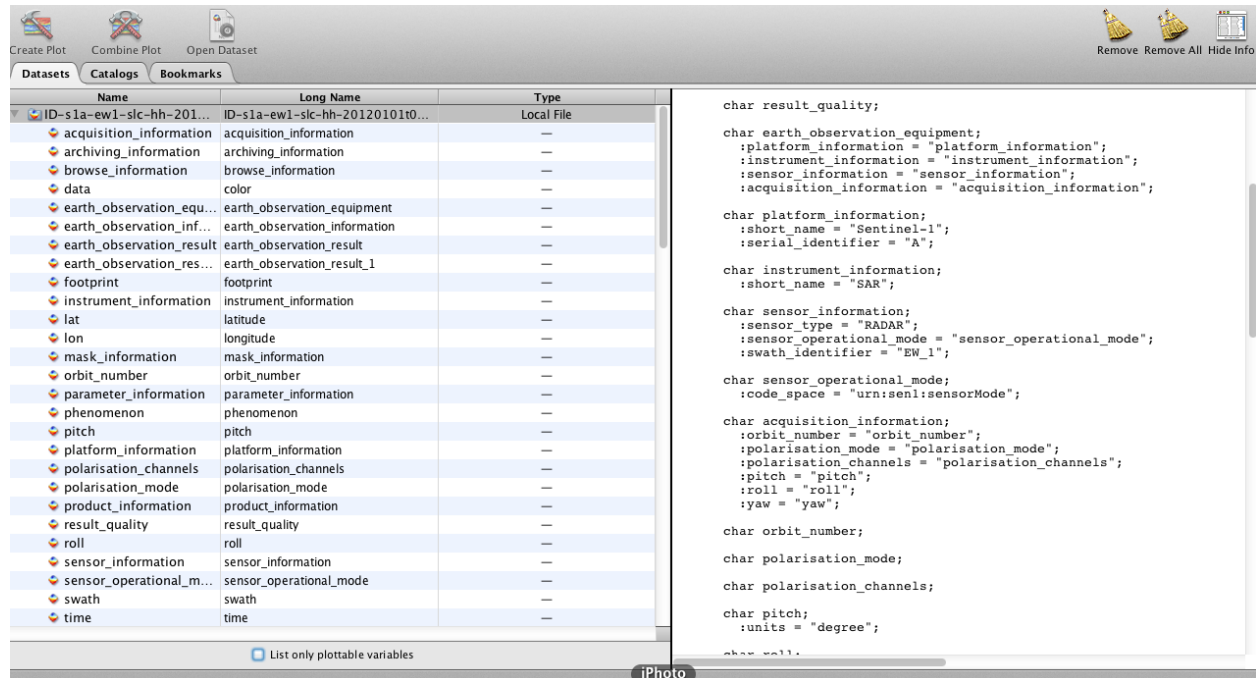


Figure 5.10: An EO-netCDF file in Panoply

5.2 A Semantically Enabled Search Platform

The ProdTrees platform is a semantically-enabled EO products search engine. It allows end-users to search for EO products using filtering criteria provided by the EO-netCDF specification and the EO vocabulary. In particular, the web interface of the ProdTrees platform allows the users to submit free-text queries, navigate to an ontology browser, select applications terms defined in the supported ontologies and finally, search for EO product by specifying EO-netCDF parameters and controlled (bounding box, time, range) search criteria.

When the user has filled the search form, a Query Analyzer is responsible for displaying a number of different interpretations for the inserted free-text. After the user has selected the semantics she wants to be used for the search, the backend service is called, generates one or more queries and sends them to GI-cat through its OpenSearch Enhanced API. GI-cat searches for the matching EO products and returns back the metadata. Depending on the nature of each product (JPG, XML, HDF, etc.), this may be either visualized on-line

or downloaded on the local system.

5.2.1 Architecture

Figure 5.11 depicts the architecture of the platform, which partially re-uses components from the RARE platform. The **Rapid Response⁸ Client (RRC)** provides the user interface to the ProdTrees platform and communicates with several backend services. It displays a search form, where a user can give as input EO-specific search criteria or free text and can navigate to the supported ontologies through the **Cross-Ontology Browser**. This component is a browser for ontologies expressed in SKOS that allows the users to exploit the knowledge contained in the supported ontologies. It provides relevant information for each concept and highlights the connections between different (but related) concepts belonging to the same or other ontologies. Its role is to support the user in the query creation phase, as a disambiguation and discovery tool. The browser is accessed via the RRC search page.

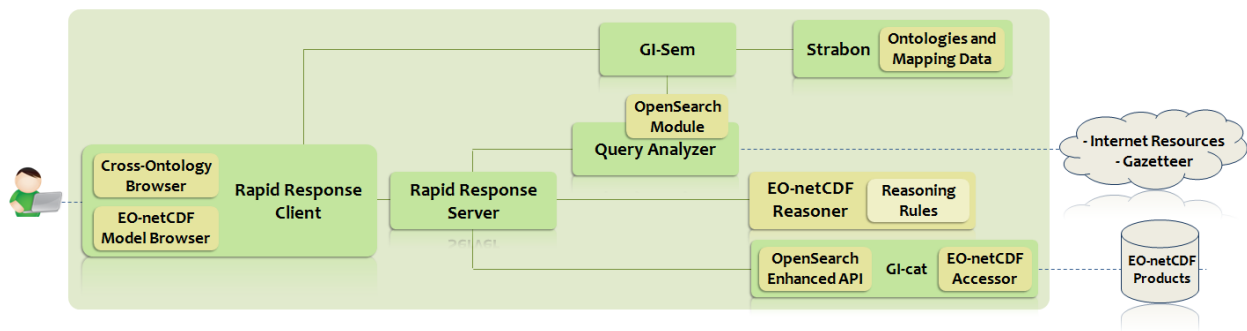


Figure 5.11: ProdTrees architecture

GI-Sem [31] is a middleware which is in charge of interconnecting heterogeneous and distributed components. Its main role in the ProdTrees platform is to create a connection between the Cross-Ontology Browser and the supported ontologies. GI-Sem performs remote queries to Strabon and returns the results to the Cross-Ontology Browser.

Strabon [22], as described in Subsection 3.3.4, is a well-known spatiotemporal RDF store. It holds the supported ontologies and the cross-ontology mappings appropriately encoded in RDF. The supported SKOS ontologies are the GSCDA, GEOSS, GEMET and NASA GCMD. The mappings between these ontologies were created using Pythia, a system developed in the scope of ProdTrees.

⁸The name “Rapid Response” comes from project RARE where the main application of the developed system was rapid response for various emergencies (e.g., humanitarian or environmental). Similarly, for the Rapid Response Server mentioned below.

All the interactions with the backend modules go through the **Rapid Response Server (RRS)**. In case a query string entered by the user needs to be disambiguated, the RRS invokes the **Query Analyzer (QA)**. The QA processes the query string, identifying the words that may be mapped to application terms, location names (toponyms), time constraints, or other types of named entities. In order to carry out this task, the QA interacts with GI-Sem (using an OpenSearch⁹ interface), Internet Resources such as gazetteers, as well as external databases such as Wordnet.

After the disambiguation process, if the user has selected an ontology concept, the RRS interacts with the **EO-netCDF Resources Reasoner** to obtain the filter criteria for the search. The reasoner uses reasoning rules to map an ontology concept to EO-netCDF search criteria. These rules have been built manually with the consultation of experts in the context of the project ProdTrees and the previous project RARE. RSS uses the returned results to build an appropriate query that is sent to GI-cat.

GI-cat [5], as described in Subsection 3.3.2, is an implementation of a catalogue service, which can be used to access various distributed sources of Earth Observation products. In ProdTrees, it has been extended to support products compliant with the EO-netCDF convention. Thus, it provides an EO-netCDF enabled discovery and access engine, so that products annotated with EO-netCDF are searchable and accessible to the users.

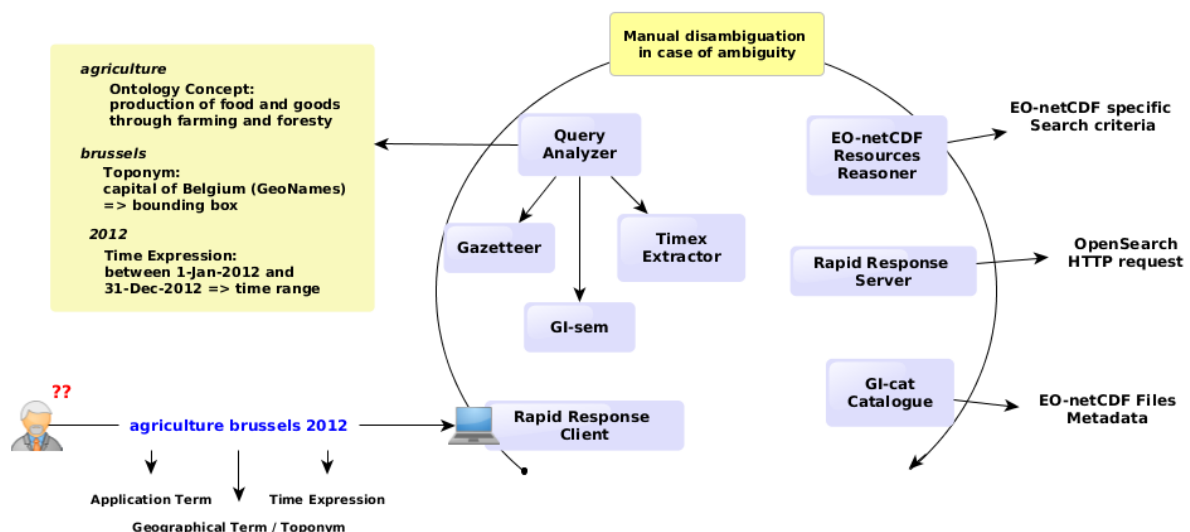


Figure 5.12: The lifecycle of the “agriculture brussels 2012” query

⁹<http://www.opensearch.org/Home>

Figure 5.12 illustrates with an example the various concepts introduced in the previous paragraphs.

The following section will describe in more detail how the query disambiguation phase takes place, while the next chapter will cover the components of the system that were developed in the scope of this thesis.

5.2.2 Query Disambiguation

When the user enters a search query, the search form transmits it to the server-side RCC in order to obtain its default interpretation. The RCC queries the RRS, which in turn invokes the Query Analyzer. The Query Analyzer disambiguates the query and returns the interpretation information to the caller. The interpretation information is displayed in a human readable form to the user and contains the most probable meaning of the terms that compose the query. It may however happen that a term has no detected meaning.

If the user accepts the displayed interpretation of the search query, she may proceed with the search of available resources, otherwise, she has the possibility to control the manner her query must be interpreted. In particular, when the user enters the disambiguation process, the system displays not only the most probable meaning for each (combination of) search term(s) but also all the identified alternate meanings.

The components involved in this activity are the same as for the basic interpretation. The difference is in the returned information: the RCC receives a normalized query that contains the most probable meaning for each search term (or combination of terms), the meaning already selected by the user, if any, and all the identified alternate meanings. The RCC is then able to display lists of meanings and allow the user to choose the most appropriate ones.

An example of the data retrieved for the query “flood” is presented below. The example shows the Disambiguation Data JSON document that is generated and retrieved from the Query Analyzer.

```
{ "DisambiguationData": {  
  "t0": [  
    "flood",  
    [  
      "sid11",  
      "sid12",  
      "sid13",
```

```

"sid14",
"sid1",
"sid2",
"sid3",
"sid4",
"sid5",
"sid6",
"sid7",
"sid8",
"sid9",
"sid10"
],
{
  "sid3": {
    "WNGloss": "light that is a source of artificial illumination
    having a broad beam; used in photography",
    "WNWords": "flood, floodlight, flood_lamp, photoflood",
    "Type": "WNConcept",
    "Id": "urn:wn3.0:flood:n:3"
  },
  "sid2": {
    "WNGloss": "an overwhelming number or amount; \"a flood of
    requests\"; \"a torrent of abuse\"",
    "WNWords": "flood, inundation, deluge, torrent",
    "Type": "WNConcept",
    "Id": "urn:wn3.0:flood:n:2"
  },
  "sid5": {
    "WNGloss": "the act of flooding; filling to overflowing",
    "WNWords": "flood, flowage",
    "Type": "WNConcept",
    "Id": "urn:wn3.0:flood:n:5"
  },
  "sid4": {
    "WNGloss": "a large flow",
    "WNWords": "flood, overflow, outpouring",
    "Type": "WNConcept",
    "Id": "urn:wn3.0:flood:n:4"
  },

```



```

"sid1": {
  "WNGloss": "the rising of a body of water and its overflowing
onto normally dry land; \"plains fertilized by annual
inundations\"",
  "WNWords": "flood, inundation, deluge, alluvion",
  "Type": "WNConcept",
  "Id": "urn:wn3.0:flood:n:1"
},
"sid7": {
  "WNGloss": "fill quickly beyond capacity; as with a liquid;
\"the basement was inundated after the storm\"; \"The images
flooded his mind\"",
  "WNWords": "deluge, flood, inundate, swamp",
  "Type": "WNConcept",
  "Id": "urn:wn3.0:flood:v:1"
},
"sid6": {
  "WNGloss": "the occurrence of incoming water (between a low
tide and the following high tide); \"a tide in the affairs of
men which, taken at the flood, leads on to fortune\"
-Shakespeare",
  "WNWords": "flood_tide, flood, rising_tide",
  "Type": "WNConcept",
  "Id": "urn:wn3.0:flood:n:6"
},
"sid13": {
  "ScopeNote": "http://www.oas.org/DSD/publications/Unit/oea66e/
ch08.htm",
  "Type": "OTSApplicationTerm",
  "HitType": "NarrowMatch",
  "Narrowers": "",
  "Id": "urn:ots3.1:applicationterm:Flood_Plain",
  "PrefLabel": "Flood Plain",
  "Note": "",
  "Broaders": "Risk_Area_Mapping",
  "AltLabels": "Floodplain",
  "Definition": "Floodplains are land areas adjacent to rivers
and streams that are subject to recurring inundation."
},

```

```

"sid9": {
  "WNGloss": "supply with an excess of; \"flood the market with
  tennis shoes\"; \"Glut the country with cheap imports from the
  Orient\"",
  "WNWords": "flood, oversupply, glut",
  "Type": "WNConcept",
  "Id": "urn:wn3.0:flood:v:3"
},
"sid14": {
  "Source": "EOHandbook-Instruments",
  "Type": "RAREVocTerm",
  "data": {
    "Data Format": "",
    "Spatial Resolution Best": "",
    "Swath Width": "",
    "Accuracy": "",
    "Instrument Name Short": "MVIRS",
    "Instrument Type": "Imaging multi-spectral radiometers
    (vis/IR)",
    "Measurements & applications": "Measures surface temperature
    and cloud and ice cover. Used for snow and flood monitoring
    and surface temperature.",
    "Spatial Resolution": "",
    "Instrument Agencies": "NRSCC (CNSA, CAST)",
    "Instrument Status": "Approved",
    "Data Access": "",
    "Wavebands": "VIS - TIR: 0.47 - 12.5 m (20 channels)",
    "Instrument Name Full": "Moderate Resolution Visible and
    Infrared Imaging Spectroradiometer",
    "Waveband Categories": "VIS, SWIR, MWIR, TIR"
  },
  "Id": "CEOSEOHandbook:CEOS_MIMDB-InstrumentTableExport_20121023
  -083841:231"
},
"sid8": {
  "WNGloss": "cover with liquid, usually water; \"The swollen
  river flooded the village\"; \"The broken vein had flooded
  blood in her eyes\"",
  "WNWords": "flood",

```

```

    "Type": "WNConcept",
    "Id": "urn:wn3.0:flood:v:2"
  },
  "sid12": {
    "ScopeNote": "http://en.wikipedia.org/wiki/Flood_alert",
    "Type": "OTSApplicationTerm",
    "HitType": "NarrowMatch",
    "Narrowers": "",
    "Id": "urn:ots3.1:applicationterm:Flood_Alert",
    "PrefLabel": "Flood Alert",
    "Note": "flood, flash flood",
    "Broaders": "Meteorological_Alert",
    "AltLabels": "Flood Watch, Flood Warning",
    "Definition": "A flood watch (or flash flood watch) is issued
when weather conditions are favorable for very heavy rain and
flash flooding. A flood warning (or flash flood warning) is
issued when flooding in a certain area is imminent or
occurring."
  },
  "sid11": {
    "ScopeNote": "http://www.eionet.europa.eu/gemet/concept?
cp=3298",
    "Type": "OTSApplicationTerm",
    "HitType": "ExactMatch",
    "Narrowers": "",
    "Id": "urn:ots3.1:applicationterm:Flood",
    "PrefLabel": "Flood",
    "Note": "early warning",
    "Broaders": "Hydrogeologic_Disaster, Meteorological_Disaster",
    "AltLabels": "Flash Flood",
    "Definition": "An unusual accumulation of water above the
ground caused by high tide, heavy rain, melting snow or rapid
runoff from paved areas."
  },
  "sid10": {
    "WNGloss": "become filled to overflowing; \"Our basement flooded
during the heavy rains\"",
    "WNWords": "flood",
    "Type": "WNConcept",

```

```
        "Id": "urn:wn3.0:flood:v:4"
      }
    ],
    "Interpretation": [[
      "t0",
      "sid11"
    ]]
  }}
}
```

5.3 Summary

This chapter covered the overview of the ProdTrees project. At first, we discussed about the creation of the EO-netCDF standard and then, we described how the ProdTrees platform implements the EO search based on this standard.

Chapter 6

Enhancing EO Ontology Services

Having presented the overview of the ProdTrees system, we can now focus on specific components which were enabled with the use of ontology services. In particular, this chapter firstly introduces the various ontologies used by the system and later covers the descriptions of: i) Pythia, an ontology matching system that creates mappings between the supported ontologies, ii) the Cross-Ontology Browser that can be used by the users in the query creation phase, as a disambiguation and discovery tool, and iii) the EO-netCDF Resources Reasoner, which is responsible for translating ontology terms to specific EO search criteria.

6.1 Ontologies

The ontologies¹ used in the ProdTrees platform are all high level environmental ontologies that contain semantic terms related to EO products. These are: i) CSCDA, ii) GEMET, iii) GEOSS, and iv) NASA GCMD, and they all share the fact that they are represented in SKOS.

6.1.1 CSCDA

The Copernicus Space Component Data Access² (CSCDA) Multi-Domain Thesaurus is a logic-based, hierarchical representation of end-users' applications for EO products. It was developed in the OTEG project (see Section 4.3), which aimed to design and implement an innovative tool intended to link available EO products to semantic terms familiar to specific application domains.

The Multi-Domain Thesaurus (Figure 6.1) contains all the knowledge needed to help the end-users find the relevant EO products. It covers four domains:

- the Marine Environment domain,

¹These are actually thesauri, but we will refer to them as ontologies for simplicity.

²Previously known as GMES Space Component Data Access (GSCDA)

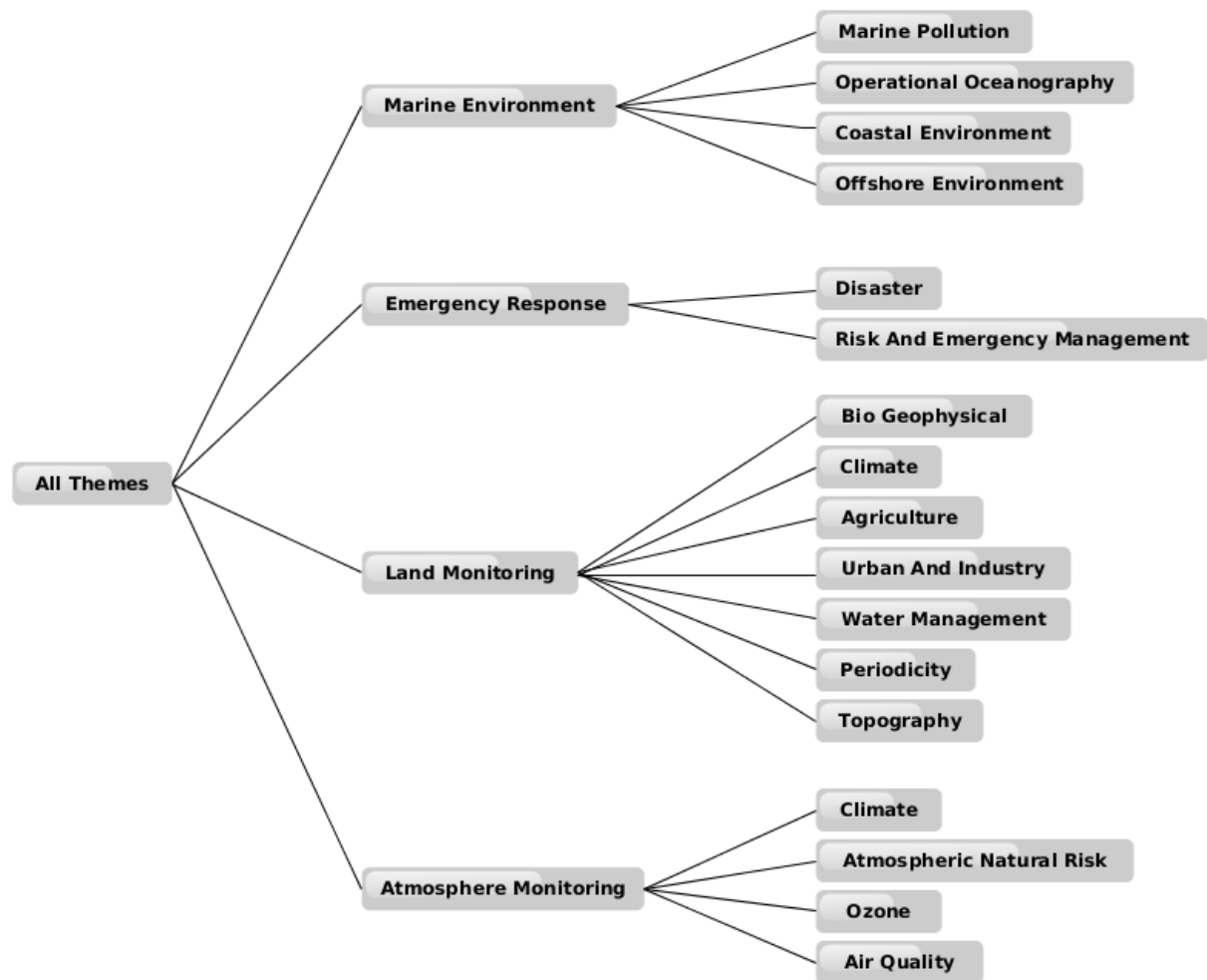


Figure 6.1: CSCDA Multi-Domain Thesaurus (levels 0 and 1)

- the Land Monitoring domain,
- the Emergency Response domain and
- the Atmosphere Monitoring domain.

These domains represent four of the main thematic areas identified in Copernicus (see 2.2.1). The Security domain has not been kept, and the Climate Change domain has been included in the main four top-level domains.

The Multi-Domain Thesaurus is composed of **Application Terms** that are used to represent each application, as well as a full text description, a set of synonyms, and a set of related terms for each one of them. As a result, this information provides a precise description of each Application Term.

Moreover, the Multi-Domain Thesaurus has a graph-like structure. Each Application Term is connected to other more general (*broader*) as well as more specific (*narrower*) Application Terms, thus forming a *logical hierarchy*. Also, each Application Term can be connected with more than one general Application Terms. Therefore, the Thesaurus can be depicted as an acyclic oriented graph (Figure 6.2).

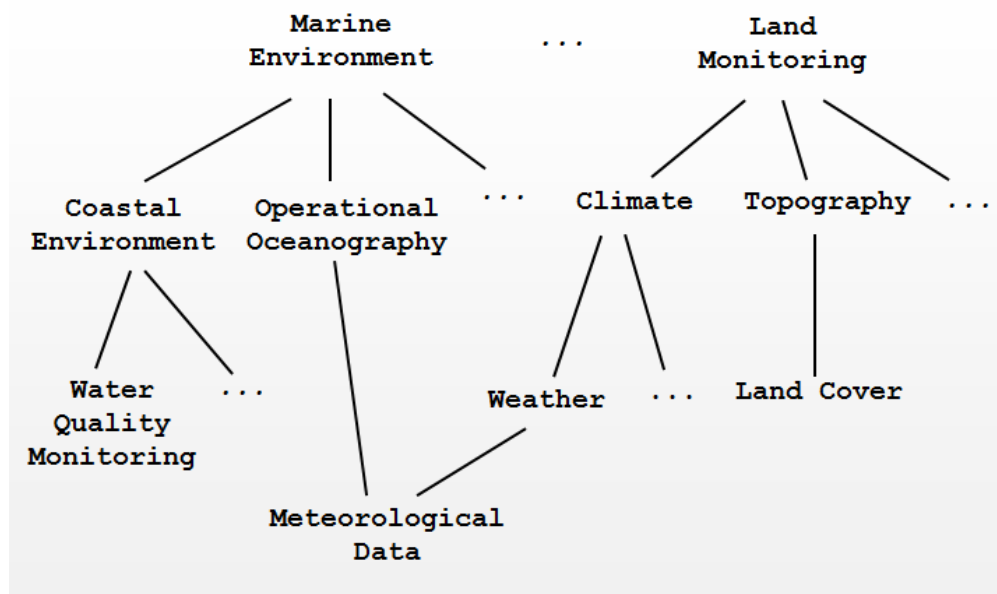


Figure 6.2: Multi-Domain Thesaurus structure

Finally, since the Thesaurus is implemented in SKOS, each Application Term is represented as a SKOS Concept, as shown in the following example:

```
<skos:Concept rdf:about="Water_Management">
  <skos:prefLabel xml:lang="en">Water Management</skos:prefLabel>
  <skos:altLabel xml:lang="en">Water Use</skos:altLabel>
  <skos:altLabel xml:lang="en">Water Monitoring</skos:altLabel>
  <skos:altLabel xml:lang="en">Water Usage</skos:altLabel>
  <skos:altLabel xml:lang="en">Irrigation</skos:altLabel>
  <skos:altLabel xml:lang="en">Farming</skos:altLabel>
  <skos:definition xml:lang="en">Water management is the practice of
    planning, developing, distributing and optimum utilizing of water
    resources under defined water polices and regulations.
  </skos:definition>
  <skos:scopeNote xml:lang="en">
    http://en.wikipedia.org/wiki/Water_management
  </skos:scopeNote>
  <skos:note xml:lang="en">water quality, water chemistry, fertilizers,
    pesticides</skos:note>
  <skos:broader rdf:resource="Land_Monitoring"/>
  <skos:narrower rdf:resource="Surface_Water"/>
  <skos:narrower rdf:resource="Ground_Water"/>
  <skos:narrower rdf:resource="Water_Use"/>
</skos:Concept>
```

6.1.2 GEMET

The General Multilingual Environmental Thesaurus³ (GEMET), was conceived as a “general” thesaurus, aimed to define a common general language, a core of *general terminology* for the environment. The basic idea was to use the best of the presently available excellent multilingual thesauri, in order to save time, energy and funds. Specific thesauri and descriptor systems (e.g. on Nature Conservation, on Wastes, on Energy, etc.) have been excluded and have been taken into account only for their structure and upper level terminology. The present version of GEMET contains about 5200 terms in 33 languages.

GEMET is using a classification scheme, made of 3 super-groups containing 30 groups (Table 6.1); there are in addition 5 accessory groups of terms, instrumental to the thesaurus use. The **super-groups** have been adopted to approach an environmental management perspective and to help the hierarchical structuring of GEMET. The **groups** reflect a systematic, category- or discipline-oriented perspective. Within the groups, the descrip-

³<http://www.eionet.europa.eu/gemet/>

Table 6.1: GEMET Super-groups and Groups

**HUMAN ACTIVITIES AND PRODUCTS,
& EFFECTS ON THE ENVIRONMENT**

- | | |
|----|--|
| 1 | AGRICULTURE, FORESTRY; ANIMAL HUSBANDRY; FISHERY |
| 2 | CHEMISTRY, SUBSTANCES, PROCESSES |
| 3 | EFFECTS, IMPACTS |
| 4 | ENERGY |
| 5 | INDUSTRY, CRAFTS; TECHNOLOGY; EQUIPMENTS |
| 6 | PHYSICAL ASPECTS, NOISE, VIBRATIONS, RADIATIONS |
| 7 | PRODUCTS, MATERIALS |
| 8 | RECREATION, TOURISM |
| 9 | RESOURCES (utilisation of resources) |
| 10 | TRADE, SERVICES |
| 11 | TRAFFIC, TRANSPORTATION |
| 12 | WASTES, POLLUTANTS, POLLUTION |
-

NATURAL ENVIRONMENT, ANTROPIC ENVIRONMENT

- | | |
|----|---|
| 13 | ANTHROPOSPHERE (built environment, human settlements, land setup) |
| 14 | ATMOSPHERE (air, climate) |
| 15 | BIOSPHERE (organisms, ecosystems) |
| 16 | ENVIRONMENT (natural environment, anthropic environment) |
| 17 | HYDROSPHERE (freshwater, marine water, waters) |
| 18 | LAND (landscape, geography) |
| 19 | LITHOSPHERE (soil, geological processes) |
| 20 | SPACE |
| 21 | TIME (chronology) |
-

SOCIAL ASPECTS, ENVIRONMENTAL POLICY MEASURES

- | | |
|----|---|
| 22 | ADMINISTRATION, MANAGEMENT, POLICY, POLITICS,
INSTITUTIONS, PLANNING |
| 23 | ECONOMICS, FINANCE |
| 24 | ENVIRONMENTAL POLICY |
| 25 | HEALTH, NUTRITION |
| 26 | INFORMATION, EDUCATION, CULTURE,
ENVIRONMENTAL AWARENESS |
| 27 | LEGISLATION, NORMS, CONVENTIONS |
| 28 | RESEARCH, SCIENCES |
| 29 | RISKS, SAFETY |
| 30 | SOCIETY |
-

tors are basically allocated in a mono-hierarchical order, but several descriptors needed to be allocated to more than one group or to more than a broader term inside the same group, thus creating a condition of *poly-hierarchy*.

In order to allow a thematic retrieval of terms thematically related but scattered in different groups, a set of 40 themes have been agreed upon with the European Environment Agency⁴ (EEA) and each descriptor has been assigned to as many themes as necessary. These themes have been established according to practical considerations, corresponding to the information needs. They have been developed to reflect the EEA activities in order to support the thematic elements of the EEA DPSIR Dataflow Scheme⁵. The list of themes has taken into account all the main topics of the Scheme, of The Dobris Assessment and of other sources, like ETCs (European Topic Centres) and Eionet⁶ (Environmental Information and Observation Network). They can be used as checklists when dealing with environmental matters. The themes, being complementary to the groups, confer to the thesaurus a *matrix structure*.

Finally, like in other multilingual thesauri, a neutral alphanumerical notation allows the identification of a concept independently to the user's language. Thus, each concept's URI is the union of the GEMET thesaurus URI and a concept number:

```
<skos:Concept rdf:about="concept/57">
  <skos:inScheme rdf:resource="gemetThesaurus"/>
  <skos:narrower rdf:resource="concept/4917" />
  <skos:narrower rdf:resource="concept/5644" />
  <skos:narrower rdf:resource="concept/7938" />
  <skos:narrower rdf:resource="concept/11889" />
  <skos:related rdf:resource="concept/7933" />
  <skos:related rdf:resource="concept/7937" />
  <skos:broader rdf:resource="concept/6033" />
  <skos:closeMatch rdf:resource="http://data.uba.de/umt/_00011022" />
</skos:Concept>
```

⁴<http://www.eea.europa.eu/>

⁵http://root-devel.ew.eea.europa.eu/ia2dec/knowledge_base/Frameworks/doc101182

⁶<http://www.eionet.europa.eu/>

6.1.3 NASA GCMD

The Global Change Master Directory⁷ (GCMD) is a comprehensive directory of Earth Science data sets of relevance to global change research. The GCMD database covers climate change, agriculture, the atmosphere, biosphere, hydrosphere and oceans, geology, geography, and human dimensions of global change. The directory is part of NASA's Earth Observing System Data and Information System (EOSDIS) and also serves as NASA's contribution to the Committee on Earth Observation Satellites (CEOS), through which it is also known as the International Directory Network (IDN).

The project's mission is to assist researchers, policy makers, and the public in the discovery of and access to data, related services, and ancillary information (which includes descriptions of instruments and platforms) relevant to global change and Earth science research. Within this mission, the directory also offers online authoring tools to providers of data and services, facilitating the capability to make their products available to the Earth science community. In addition, citation information to properly credit data set contributions is offered, along with direct links to data and services. The GCMD's primary responsibility is to maintain a complete catalogue of all NASA's Earth science data sets and services.

As an integral part of the project, keyword vocabularies have been developed and are being refined and expanded [27]. Users may perform searches through the Directory's website using controlled keywords, free-text searches, map/date searches or any combination of these. Users may also search or refine a search by data centre, location, instrument, platform, project, or temporal/spatial resolution.

The GCMD includes both controlled and uncontrolled keywords. The controlled keywords include approximately 1000 Earth science terms represented in a subject taxonomy. In particular, there are seven sets of controlled keywords:

- Earth science,
- Data services,
- Data centers,
- Locations,
- Instrument/sensors,

⁷<http://gcmd.nasa.gov/>

- Platforms/sources, and
- Projects

Earth science data are maintained in a 5-level broad keyword classification, as seen in Figure 6, whereas data services are maintained in a 3-level keyword hierarchy. The Climate Diagnostic descriptions include two unique keyword sets: visualization type and analysis type. Several hundred additional controlled keywords are defined for ancillary support, and have been submitted by data providers. These terms tend to be more general than or synonymous with the controlled terms.

Earth Science Data	Example
Controlled (5 levels): <ul style="list-style-type: none"> • Topic • Term • Variable_Level_1 • Variable_Level_2 • Variable_Level_3 Uncontrolled (free text): <ul style="list-style-type: none"> • Detailed Variables • Ancillary 	<ul style="list-style-type: none"> • Topic: Biological Classification • Term: Animals/Invertebrates • Variable_Level_1: Cnidarians • Variable_Level_2: Anthozoans/Octocorals • Variable_Level_3: Sea Fans/Sea Whips Uncontrolled (free text): Detailed Variables: Gorgonacea <ul style="list-style-type: none"> • Ancillary: Gorgonian Colony

Figure 6.3: The 5-level classification of Earth Science Data in GCMD

Moreover, an uncontrolled (i.e., free-text) level of keywords is used for “Detailed Variables” beyond the controlled levels. However, not all of the keywords have data set descriptions (i.e., metadata) behind them.

Finally, additional free-text keywords are also specified in the “Ancillary Keyword” field.

6.1.4 GEOSS EO Vocabulary

As mentioned in Section 2.2.2, GCI is the main enabler of the System of Systems principles and capabilities of GEOSS. It is able to interface with external systems to facilitate end users in discovering and accessing their services and resources. This requires making these systems and components interoperable, so that the data and information they produce can be pooled and combined.

One of the components of the GCI that enables semantic interoperability is the GEOSS Earth Observation Vocabulary (GEOSS EO Vocabulary). The EO Vocabulary was defined using existing glossaries and is associated with other thesauri. More precisely, EO Vocabulary is a selection of 142 “critical observation parameters” that are categorized in a three-level hierarchy according to 80 Global Change Master Directory topics and terms. These terms, also, hold relations linking the EO Vocabulary with other thesauri, and so they can efficiently bridge between the different thematic and application domains. In this way, the EO Vocabulary enables multidisciplinary access to resources by coupling terminologies from different application domains.

For instance, discovery of GEOSS resources is likely to respond to a policy-making need in one of the SBAs defined by GEOSS. Therefore, terms from the EO Vocabulary have been related to the corresponding SBAs so that they can be retrieved by non-scientific users, such as decision makers.

6.2 Ontology Matching

In order to facilitate the search of EO products by end-users, several ontologies in use in the EO domains are mapped to the enhanced netCDF vocabulary. In particular, the various ontologies are first mapped to a main ontology, and this ontology is then mapped to the enhanced netCDF vocabulary. As a result, users can interact with ontologies in their application domains, i.e. ontologies they are familiar with. In this section, we start by surveying the different ontology matching techniques and describing related work in ontology matching. Afterwards, we present Pythia, an ontology matching system that we developed in order to produce various types of mappings that interconnect the ontologies of the ProdTrees platform. We conclude the section with the performance of our system, tested with these ontologies.

6.2.1 Ontology Matching Techniques

Ontology matching is the process of finding relationships or correspondences between entities of different ontologies [16] and is considered as a main factor for enabling interoperability across heterogeneous systems and semantic web applications. This process is essential, as there may be a certain degree of heterogeneity between different ontologies. Usual types of heterogeneity are:

- **Syntactic:** when 2 ontologies are not expressed in the same ontology language,

- **Terminological:** when the same entities have different names in different ontologies,
- **Conceptual:** when different modelling is used for the same domain of interest, and
- **Semiotic:** when an entity is interpreted in regard to the context, and so the same entity might have different interpretations in different ontologies.

Several types of heterogeneity may also occur together. In order to deal with heterogeneity, a variety of matching techniques exist. These either focus on a specific entity in each ontology, or they take into account the various relationships between the entities of the ontology. The rest of this section, introduces to various matching techniques. A complete analysis and different classifications of ontology matching techniques can be found in [16].

Element-level techniques

This category includes techniques that consider ontology entities (or their instances) in isolation from their relations with other entities or instances. The most common techniques are:

String-based techniques, which are used in order to associate names and descriptions of ontology entities. Strings are considered as sequences of letters in an alphabet. These techniques are typically based on the intuition that the more similar the strings are, the more likely they are to denote the same concepts. Usually, distance functions map a pair of strings to a real number, where a smaller value indicates a greater similarity between the strings.

Language-based techniques, which consider names as words in some natural language. They are based on natural language processing techniques exploiting morphological properties of the input word. In most cases, they are applied to names of entities before running string-based or lexicon-based techniques in order to improve their results.

Instance-based techniques or extensional ontology mapping, which depends on measuring the similarity between sets of instances. The idea behind such techniques is that similarity between the extensions of two concepts, i.e. their instances, reflects the semantic similarity of these concepts.

Constraint-based techniques, which are algorithms dealing with the internal constraints being applied to the definitions of entities, such as types, cardinality of attributes, and keys.

Linguistic resources, such as lexicons or domain specific thesauri, which are used in order to match words based on linguistic relations between them.

Alignment reuse techniques, which represent an alternative way of exploiting external resources. These resources record alignment of previously matched ontologies.

Structure-level techniques

Unlike element-level techniques, techniques in this category consider the ontology entities or their instances to compare their relations with other entities or their instances. In particular, this category includes:

Graph-based techniques, which are graph algorithms which consider the input ontologies as labelled graphs. Usually, the similarity comparison between a pair of nodes from the two ontologies is based on the analysis of their positions within the graphs. This technique is motivated by the intuition that if two nodes from two ontologies are similar, their neighbors must also be somehow similar.

Taxonomy-based techniques, which are also graph algorithms which consider only the specialisation relation. The intuition behind taxonomic techniques is that is-a links connect terms that are already similar (being interpreted as a subset or superset of each other), therefore their neighbors may be also somehow similar.

Repositories of structures, which store ontologies and their fragments together with pair wise similarity measures. Unlike alignment reuse, repositories of structures store only similarities between ontologies not alignments. When new structures are to be matched, they are first checked for similarity against the structures which are already available in the repository. The goal is to identify structures which are sufficiently similar to be worth matching in more detail, or reusing already existing alignments, thus avoiding the match operation over the dissimilar structures.

Model-based techniques, which are algorithms that handle the input based on its semantic interpretation. The intuition is that if two entities are the same, then they share the same interpretations. Thus, they are well grounded deductive methods.

Data analysis and statistics techniques, which take advantage of a representative sample of a population in order to find regularities and discrepancies. This helps in grouping together items or computing distances between them.

6.2.2 Matching systems

The basic techniques presented in the previous section are the building blocks on which a matching solution is built. Generally, matching systems are comprised by one or more of the following components: a terminological matcher, a structure-based matcher, and a semantics-based matcher accompanied by a mapping selection. A *terminological matcher* compares string similarities, but also annotations (labels, comments, etc.) and synonyms. A *structure-based matcher* takes into account the structure of the ontologies and usually depends on initial mappings provided by the terminological matcher. A *semantic matcher* refines candidate mappings based on the semantics of a specific knowledge domain. Finally, a mapping selection module is used in combination with one or more of the above techniques and filters out the best mapping candidates.

After more than a decade of research and practice, the field of ontology matching has made a considerable improvement. However, there are still some challenges the ontology matching community has to address (e.g. large-scale matching evaluation, matching with background knowledge, user involvement, explanation of matching results) [32]. The next paragraphs give a brief overview of the basic ontology matching techniques and cite ontology matching systems that use similar background with Pythia.

[26] studies how the previous techniques can be combined within a matching system and how these interactions affect the overall quality of the produced mappings. [32] presents the state of the art systems in ontology matching providing also analytical and empirical comparisons. DSSim [25] and ASMOV [21] are two of these systems. DSSim follows a multi-agent approach that makes use of uncertain reasoning through multi-agent beliefs and conflict resolution. It uses WordNet⁸ to expand ontology concepts and properties and various terminological similarity measures, such as Monger-Elkan and Jaccard distances. The methodology of ASMOV is summarized in two steps: (i) similarity calculation and (ii) semantic verification. In the first step, it uses lexical, structural and extensional matchers to compute similarity measures between two ontologies. Then, it derives an $n:m$ alignment between ontology entities and checks it for consistency.

Another study close enough to Pythia is described in [30] and exploits the features of

⁸<http://wordnet.princeton.edu/>

the Lucene search engine library. Each ontology entity is treated as a Lucene Document (LD). Similarity between entities is computed by constructing an index of LDs derived from a target ontology and using values of entities of a source ontology as search arguments to the index. Finally, many ontology matching systems use machine learning techniques [14, 9].

6.2.3 Pythia

Pythia is an ontology matching system that combines a terminological with a structural matcher. It was designed to satisfy the constraints and the special characteristics of the ProdTrees platform and of the ontologies used by the platform. Firstly, these ontologies are expressed in SKOS, so the mapping technique should be adjusted to the needs of a structured vocabulary. Secondly, during the core search operation, the platform uses one main ontology. This means that the final queries used to access the EO catalogues are constructed according to this ontology. Thus, what is needed is only mappings from the various ontologies to the core ontology, and not an alignment.

Pythia is implemented in Java and it uses the openRDF Sesame⁹ and the Apache Lucene¹⁰ frameworks to handle its main operations. The system takes as input two ontologies and it produces the mappings from the first ontology (source ontology) to the second one (target ontology). The next subsections present how Pythia works.

Mapping Language

Taking into account that Pythia is targeted for SKOS ontologies, the mappings we create are also expressed in SKOS using the defined vocabulary for matching concepts:

- *skos¹¹:exactMatch*: indicates a high degree of confidence that the two linked concepts can be used interchangeably
- *skos:relatedMatch*: states an associative mapping link between two conceptual resources
- *skos:broadMatch*, *skos:narrowMatch*: inverse properties that state a hierarchical mapping link between two conceptual resources

⁹<http://www.openrdf.org/index.jsp>

¹⁰<http://lucene.apache.org/>

¹¹SKOS core namespace: [<http://www.w3.org/2004/02/skos/core#>](http://www.w3.org/2004/02/skos/core#)

Based on these SKOS properties, we can create four different types of mappings. As we will describe below, each mapping technique can produce only specific types of mappings. For example, the structural matcher creates mappings described by the *skos:broadMatch* and the *skos:narrowMatch* properties.

Setting the Properties

Pythia is a flexible system that offers to users self-configuration capabilities. First of all, the users give information about the two ontologies, such as their source (file or URL), location (path to the file or url) and RDF format (N3, RDFXML, etc.). Moreover, they can specify whether a dictionary will be used or not, the path of the dictionary, as well as various parameters that optimize the use of the dictionary. They can also decide if pre-existing mappings will be kept or ignored by the system. Finally, the users define the format of the exported mappings and the type of mappings they are interested in. For instance, they can choose to export: i) one of the *skos:exactMatch*, *skos:relatedMatch*, and *skos:broadMatch* and *skos:narrowMatch* mappings, ii) a combination of them, or iii) all of them.

Using the settings defined by the user, the two ontologies are stored in an openRDF Sesame repository. The repository is then queried in order to retrieve the available information for each concept (different types of labels, narrower/broader/related relations with other concepts, definition, notes).

Terminological Matcher

Pythia deals first with the terminological heterogeneity of the ontologies. A terminological matcher is responsible for implementing a string-based and a language-based technique. The mappings created by this component can either be *skos:exactMatch* (at most one for each concept) or *skos:relatedMatch* (more than one per concept).

The terminological matcher starts by using the string-based technique and if no mappings are found, it continues with the language-based technique. Both techniques are applied on the concepts labels (*skos:prefLabel*, *skos:altLabel* and *skos:hiddenLabel*). Therefore, in order to find a mapping between concept *A* from the source ontology and concept *B* from the target ontology, the technique starts with the *skos:prefLabel* of *A* and searches for a concept *B* in the target ontology with a similar label (*skos:prefLabel*, *skos:altLabel* or *skos:hiddenLabel*). The type of similarity depends on the type of the technique, as it will be explained in the forthcoming paragraphs. If no mapping is found, then the tech-

nique continues with the *skos:altLabel* of *A*, if it exists, and so on. If again no mapping is created, the matcher will try now to find a mapping by applying the language-based technique in a similar way.

The two techniques are explained in more detail in the following paragraphs:

i) String-based technique

In the string-based technique, the terminological matcher uses Apache Lucene which offers to Pythia indexing and search technology, as well as spellchecking, hit highlighting and advanced analysis and tokenization capabilities. Using Lucene, one can create documents and add fields of a specific type to these documents. When adding a field, there is an option on whether this field will be indexed or not. As a result, later, when searching the document, the user can specify which field he wants to search.

Taking advantage of Lucene capabilities, the terminological matcher indexes the target ontology. It creates a new document for each concept and adds each available property of the concept as a new field. When a new field is added, the matcher utilizes a feature of Lucene called Analyzer. This feature removes any unnecessary stop words, apply case normalization to the field, etc.

In order to demonstrate how the terminological matcher works after indexing the target ontology, we will focus again on the example described above. When searching for concepts similar to concept *A*, the *prefLabel*, the *altLabel* and the *hiddenLabel* fields (of the indexed ontology) are searched using the *prefLabel* of concept *A*. The search results that are fetched back are ranked according to the string similarity of the strings that were compared (for example *skos:prefLabel* of *A* and the *prefLabel* field of a document). This is feasible due to the string similarity functions implemented in Lucene. Also, since each field is indexed, the terminological matcher does not search one by one the concepts documents. It only checks the relevant index of the specified field.

Considering that Lucene returns multiple related results, the matcher has to verify whether the two strings are exactly the same or not. In the first case, a *skos:exactMatch* is created between *A* and the corresponding concept from the target ontology. An example is displayed in Table 6.2. The **GEOSS¹²:snow depth** has a *skos:exactMatch* with the **GSCDA¹³:Snow_Height**. The mapping was created because the first concept has as *skos:prefLabel* the same string that the second concept has as *skos:altLabel*.

In case one of the two strings is a substring of the other, then a *skos:relatedMatch* is

¹²GEOSS namespace: <http://www.earthobservations.org/GEOSS/EO_Vocabulary/>

¹³GSCDA namespace: <http://thesauri.esa.int/MultiDomain_Thesaurus/>

Table 6.2: Mappings created by the terminological matcher

GEOSS:snow depth	skos:prefLabel	“Snow Depth” @en .
↓ skos:exactMatch		
GSCDA:Snow_Height	skos:prefLabel skos:altLabel	“Snow Height”@en ; “Snow Depth” @en .
-----	-----	-----
GEOSS:elevation	skos:prefLabel	“Elevation” @en .
↓ skos:relatedMatch		
GSCDA:Digital_Elevation_Model	skos:prefLabel skos:altLabel	“Digital Elevation Model”@en ; “Digital Terrain Model”@en , “DEM”@en , “DTM”@en , “Relief Map”@en .
-----	-----	-----
GEMET:1421	skos:prefLabel	“city”@en .
↓ skos:relatedMatch		
GSCDA:Urban_And_Industry	skos:prefLabel	“Urban And Industry”@en .

created between the two concepts. This time (Table 6.2, second example) a *skos:relatedMatch* mapping is created between **GEOSS:elevation** and **GSCDA:Digital_Elevation_Model** because the first concept has as *skos:prefLabel* “Elevation” and the second concept has as *skos:prefLabel* “Digital Elevation Model”.

Finally, if no mappings are discovered for the concept, then the language-based technique is invoked.

ii) Language-based technique

As this technique might add noise to the results (see 6.2.3), its use is optional and can be bypassed.

The language-based technique involves the use of a dictionary API which provides synonyms, related terms, etc. Pythia uses WordNet, which is a lexical database for English. In WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (called synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations, called *pointers*. Thus, the WordNet’s structure makes it a useful tool for computational linguistics and natural language processing. Putting WordNet to use, the terminological matcher creates a new field in the Lucene documents created for each concept. This new field, called *relLabel*, enhances each concept’s labels, by adding synonyms, derived terms and other related words that can be found in WordNet (the type of related words that will enhance the concepts can be specified by the user). Hence, if the matcher does not find any mappings while searching in the *prefLabel*, *altLabel* and *hiddenLabel* fields, then it moves on to the *relLabel* field. If a similarity is discovered, then a *skos:relatedMatch* relation is created between the corresponding concepts.

For example (Table 6.2), a *skos:relatedMatch* can be created between the **GEMET¹⁴:1421** with *skos:prefLabel* “city” and the **GSCDA:Urban_And_Industry** with *skos:prefLabel* “Urban And Industry”. This is a result of the following: When the **GSCDA:Urban_And_Industry** is enhanced, WordNet is invoked in order to retrieve phrases related with “urban and industry”. The retrieved results are: i) citified and industry, ii) cityfied and industry, iii) city-bred and industry, iv) city-born and industry, v) city-like and industry, vi) urbanized and industry and vii) urbanised and industry. Each one of these results is added as a *relLabel* field, in the Lucene document representing **GSCDA:Urban_And_Industry**. Utilizing a Lucene’s Analyzer, when adding the “city-like and industry” string, the text field that is actually added is: “city”, “like”, “industry”.

¹⁴GEMET namespace: <<http://www.eionet.europa.eu/gemet/concept/>>

Therefore, when the *skos:prefLabel* of **GEMET:1421** is used to search the *relLabel* fields, the document of **GSCDA:Urban_And_Industry** will be retrieved and a *skos:relatedMatch* from **GEMET:1421** to **GSCDA:Urban_And_Industry** will be created.

In case there are concepts from the source ontology with no *skos:exactMatch* mappings, the structural matcher is invoked.

Structural Matcher

After the terminological matcher has exhausted all cases, seeking of additional relationships is passed to the structural matcher. This component implements a graph-based technique aiming to enhance the set of the results. The mappings created by the structural matcher are either *skos:narrowMatch* or *skos:broadMatch* and each concept may have multiple mappings of this type.

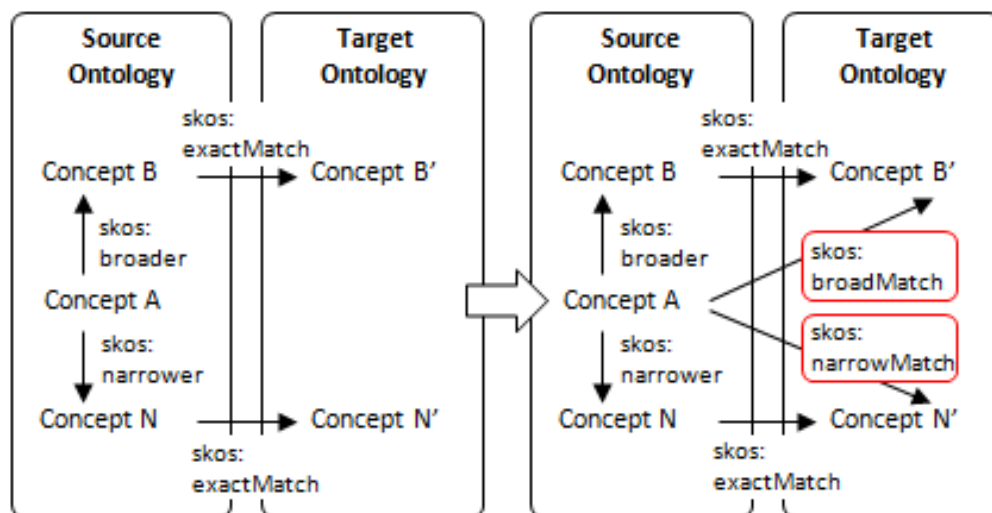


Figure 6.4: Deriving a mapping using a pre-existing *skos:exactMatch* mapping

The structural matcher takes as input a concept *A* from the source ontology and finds all the broaders and narrowers of *A*. Afterwards, it checks whether a *skos:exactMatch* was created by the terminological matcher for one of these concepts. If it did, then the structural matcher can produce a new mapping. For example, if the terminological matcher created a *skos:exactMatch* between concept *B* (which is a broader of *A*) and concept *B'*, then it can be derived that *B'* will be a *skos:broadMatch* of *A*. Similarly, if the terminological matcher created a *skos:exactMatch* between concept *N* (which is narrower of *A*) and concept *N'*, then a *skos:narrowMatch* can be created between *A* and *N'* (Figure 6.4).

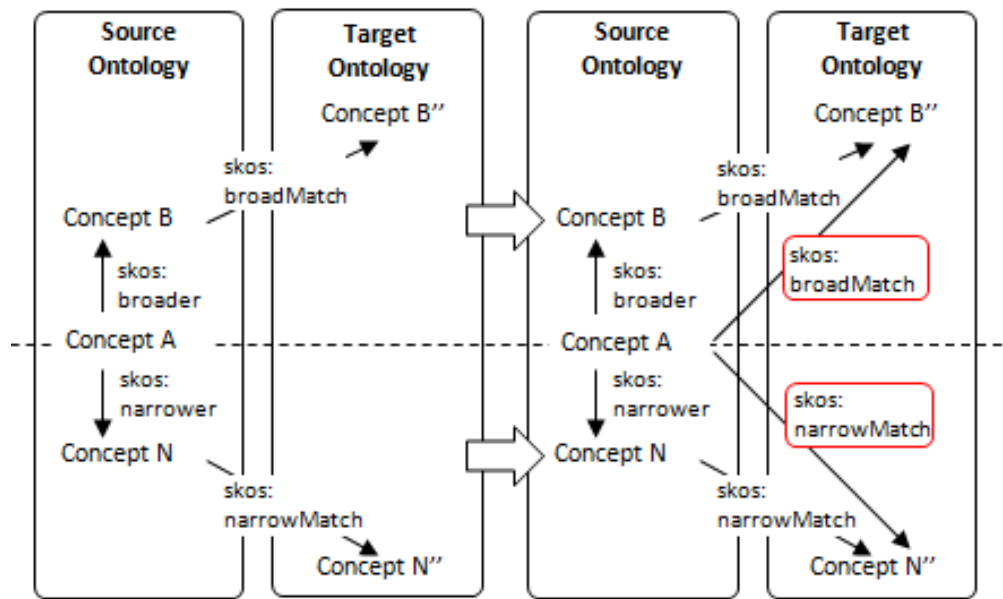


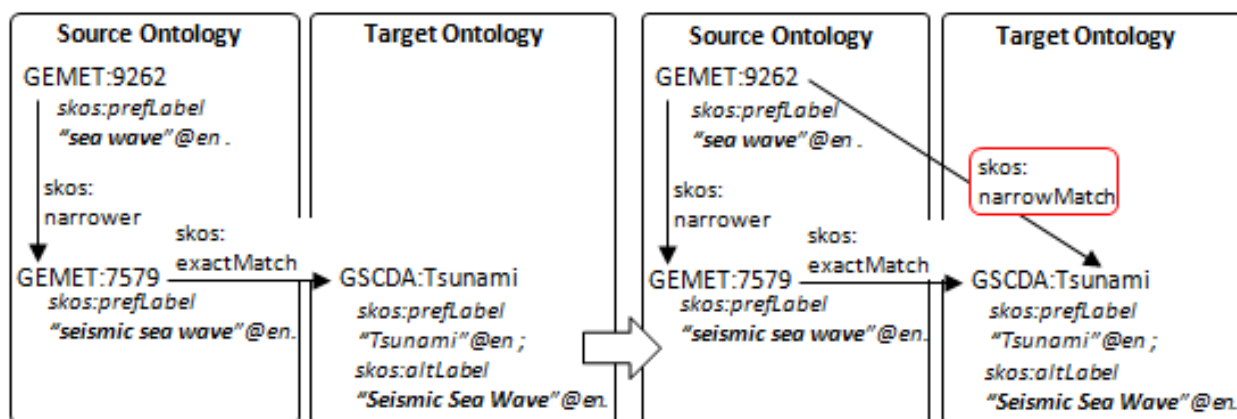
Figure 6.5: Deriving a mapping using pre-existing *skos:broadMatch* (top) or *skos:narrowMatch* (bottom) mappings

When all the concepts are examined, and only if new mappings were created by the structural matcher, the process described in the previous paragraph is repeated. In this case, the matcher will also check whether the concepts *B* and *N* hold a *skos:narrowMatch* or a *skos:broadMatch* relation with concepts included in the target ontology¹⁵. If a *skos:broadMatch* exists between *B* and a concept *B''*, then it is safe to conclude that *B''* will also be a *skos:broadMatch* of *A* (Figure 6.5). This means that when a *skos:broadMatch* exists between a concept *B* from the source ontology and a concept *B''* from the target ontology, then this relation can be propagated to concept's *B* narrowers. Similarly, when a *skos:narrowMatch* exists between a concept *N* and a concept *N''*, then this relation can be propagated to concept's *N* broaders (Figure 6.5).

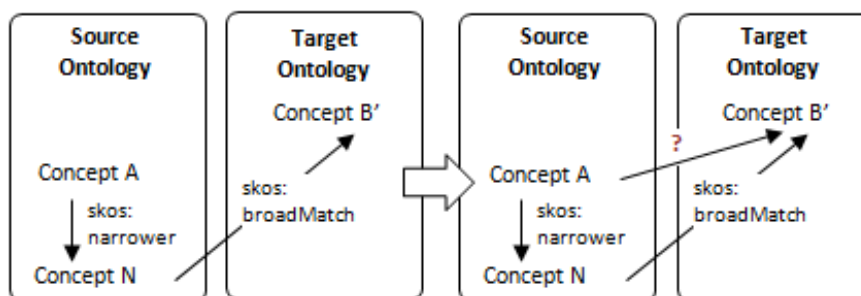
In Figure 6.6, there is an example demonstrating the creation of a *skos:narrowMatch* using a pre-existing *skos:exactMatch*. In particular, a pre-existing *skos:exactMatch* between the **GEMET:7579** and the **GSCDA:Tsunami** is used to derive a *skos:narrowMatch* between the **GEMET:9262** and the **GSCDA:Tsunami**.

It is important to highlight that a pre-existing *skos:narrowMatch* or *skos:broadMatch*, does not necessary conclude that a new mapping will be created. For example, if concept *A* has a narrower concept *N* which holds a *skos:broadMatch* relation with a concept *B'*, it is not safe to conclude a relation between *A* and *B'* (Figure 6.7). It is only safe to con-

¹⁵In case of mapping re-use, *skos:narrowMatch* and *skos:broaderMatch* mappings might be available before the first iteration of the structural matcher. As a result, the matcher will check for these types of mappings during the first iteration.


 Figure 6.6: An example of the creation of a *skos:narrowMatch* mapping

include a *skos:broadMatch* when there is a broader concept which holds this relation, or a *skos:narrowMatch* when a narrower concept holds this relation.


 Figure 6.7: Deriving a mapping using a pre-existing *skos:broadMatch* mapping is not possible in this case

The process terminates when an iteration ends and no new mappings were created. Then, Pythia proceeds with the exportation of the mappings to RDF. An example of mappings produced by Pythia in N3 format is shown below. In the example, the second mapping was created due to the fact that **GEOSS:biomass** has as broader concept **GEOSS:vegetation**, which (as indicated by the first mapping) holds a *skos:exactMatch* relation with **GSCDA:Vegetation**.

```
<http://www.earthobservations.org/GEOSS/EO_Vocabulary/vegetation>
  <http://www.w3.org/2004/02/skos/core#exactMatch>
  <http://thesauri.esa.int/MultiDomain_Thesaurus/Vegetation>.

<http://www.earthobservations.org/GEOSS/EO_Vocabulary/atmosphere>
  <http://www.w3.org/2004/02/skos/core#relatedMatch>
```



```
<http://thesauri.esa.int/MultiDomain_Thesaurus/Atmosphere_Monitoring>.

<http://www.earthobservations.org/GEOSS/EO_Vocabulary/biomass>
<http://www.w3.org/2004/02/skos/core#broadMatch>
<http://thesauri.esa.int/MultiDomain_Thesaurus/Vegetation>.
```

Performance

In order to illustrate the behavior of Pythia, the ontologies used by the ProdTrees platform were employed. More specifically, the GEMET, GEOSS and GCMD ontologies were mapped to the GSCDA ontology. In case of GCMD, only a subset of the ontology was mapped (**Science Keywords** and **Platforms** concept schemes). To demonstrate the performance of the different techniques, we present the mappings produced by each technique separately.

Table 6.3 displays the results of the **string-based technique**. The number of concepts included in each ontology is shown in the second column and the number of the produced *skos:exactMatch* mappings in the third one. Forth column has information about the *skos:relatedMatch* mappings produced by the system. Since each concept from the source ontology may have more than one mappings of this type, alongside with the total number of *skos:relatedMatch* mappings, we present the number of the distinct concepts that were mapped.

Table 6.4 contains the results of mapping GCMD to GSCDA using the **language-based technique**. This technique uses WordNet which utilizes a number of different semantic and lexical pointers in order to connect two words [24]. The technique allows tuning WordNet by stating which types of pointers will be used. Thus, the combination of the pointers can effect the percentage of valid mappings. In Table 6.4, we present the results of the mapping using only one type of pointer each time, as well as, the number of mappings that are not valid. For instance, “Derivationally Related” and “Meronym Part” produce many and correct results, whereas “Hyponym” and “Meronym Substance” produce many mappings, but a lot of them are not valid.

Finally, results of the graph-based technique are displayed in Table 6.5. The technique used as input the *skos:exactMatch* mappings presented in Table 6.3.

Table 6.3: Mappings created by the string-based technique

Ontology	#Concepts	#skos: exactMatch	#skos:relatedMatch Total Mappings/Distinct Concepts
GSCDA	189	-	-
GEMET	5220	114	178/ 85
GEOSS	222	18	20 / 10
GCMD	3264	95	84 / 58

Table 6.4: Mappings created by the language-based technique

WordNet Pointer	Total Mappings/ Distinct Concepts	Non Valid Mappings
Attribute	13/9	0
Cause	10/9	0
Derived from adjective	1/1	0
Derivationally Related	192/122	3
Entailment	3/2	0
Holonym Member	2/2	0
Holonym Part	61/41	23
Holonym Substance	45/35	0
Hypernym	223/136	32
Hypernym Instance	1/1	1
Hyponym	616/236	158
Hyponym Instance	8/8	0
Meronym Part	50/42	3
Meronym Substance	103/43	38
Pertainym	21/17	1
Similar to	21/21	0
Topic	63/37	0
Topic Member	20/17	9
Usage	19/19	0
Verb Group	9/8	1

Table 6.5: Mappings created by the graph-based technique

Ontology	Total Mappings/ Distinct Concepts	#skos:narrowMatch	#skos:broadMatch
GEMET	237/232	56	181
GEOSS	65/65	12	53
GCMD	212/210	52	160

6.3 Ontology Navigation

Ontology navigation is feasible with the use of an ontology browser. The Cross-Ontology browser is a general SKOS Ontology Browser providing all the basic functionality that the common SKOS ontology browsers support. In this section, we will introduce the Cross-Ontology Browser which was developed in ProdTrees. An extended description of the Cross-Ontology Browser can be found in [10].

6.3.1 The Cross-Ontology Browser

The Cross-Ontology Browser was designed in a way that would cover the needs of the users.

At first, the user can select the ontology he wants to browse from a collection of ontologies that are available. In the next step, the user is able to navigate through the concepts of this ontology and the browser visualizes them in a manner that permits the easy selection of the concepts the user is interested in.

The user also sees the values of all the available SKOS properties for each concept, such as *skos:prefLabel*, *skos:altLabel*, *skos:definition*, *skos:narrower*, etc. By providing this functionality, users are able to consult the available properties of various ontology concepts, before they select any of them. Thus, the ontology browser helps non-expert users to disambiguate the meaning of ontology concepts.

The browser is called *cross-ontology*, because it interconnects ontologies with links. These links were created using an ontology matching system, Pythia, described in Section 6.2.3. As mentioned in 6.2.3, the mapping language that is used, is the existing SKOS vocabulary for matching concepts (*skos:exactMatch*, *skos:relatedMatch*, *skos:narrowMatch*, *skos:broadMatch*). So, the user is able to see related concepts not only from the same ontology, but also from other supported ontologies.

Finally, the browser supports keyword search for ontology concepts. This way the users can reach the preferred concepts faster or just check if there are concepts matching a given keyword.

6.3.2 Supported Types of Visualization

The ontology visualization is a main issue regarding the Cross-Ontology Browser, and is implemented in a way that makes ontologies legible and readable by all types of users. In

addition, since GEMET and NASA GCMD, GEOSS and CSCDA are represented in SKOS, implementation of the visualization is mostly SKOS-oriented, in a way that promotes the notable characteristics of this data model.

Hierarchical Concept Browsing

In hierarchical browsing, the ontology is visualized as a tree, showing the hierarchical structure of the concepts. The top concepts are displayed, allowing an overview of the different thematic branches of the ontology. In the next step, after the user has selected a top concept, he is able to navigate through the branch he selected, going from broader to narrower concepts. A demonstration of the hierarchical browsing is shown in Figure 6.8.

Keyword Search

In case the user is unable of finding the preferred concept, the Cross-Ontology Browser provides also a keyword search mechanism. The keyword search enables the users to find concepts in a quicker way, since apart from searching a match between a keyword and a concept's label, it searches also other types of annotations, such as definitions. Pagination is used to group the results in pages of ten, so it is easier for users to navigate within the search results.

The screenshot displays the Cross-Ontology Browser interface. At the top, there is a navigation bar with a dropdown menu for 'Concepts from' set to 'http://thesauri.esa.int/MultiDomain_Thesaurus' and a 'Keyword Search' button. The main content area is divided into two panels. The left panel, titled 'Hierarchical Representation', shows a tree structure of concepts. The 'Disaster' category is expanded, showing sub-categories like 'Man Made Disaster', 'Natural Disaster', 'Geological Disaster', and 'Meteorological Disaster'. Under 'Meteorological Disaster', 'Flood' is highlighted. The right panel, titled 'Flood Concept', has tabs for 'General Info', 'Labels', 'Notes', and 'Concept matching'. The 'General Info' tab is active, showing a 'Broader term' section with 'Meteorological Disaster' and a 'Narrower terms' section which is currently empty. Below these is a 'Definition' section with the text: 'An unusual accumulation of water above the ground caused by high tide, heavy rain, melting snow or rapid runoff from paved areas.'

Figure 6.8: Browsing the concepts of CSCDA

Concept Viewer

Ultimately, after the user has browsed through the ontology and selected the desired concept, an appropriate tool is necessary to display the concept's information.

The concept viewer is a section in the ontology browser, where additional information of a concept appears, when the users selects a specific concept. The viewer, using the concept's properties, shows all the available information about the selected concept. This information includes concept's definition, as well as broader, narrower and related concepts, and so on. Note that through these concepts, the viewer makes possible the navigation to other concepts of the same or another ontology.

6.4 From Ontologies to EO-netCDF

This section describes how the ontology terms selected by the users are finally translated to EO-netCDF criteria used for the search of EO-netCDF products.

In the ProdTrees system, not all ontologies are mapped to EO-netCDF. Instead, the various ontologies are all mapped to the same ontology, and afterwards this ontology is mapped to EO-netCDF. The CSCDA Multi-Domain Thesaurus is selected to play the role of the main ontology. Afterwards, a component called EO-netCDF Reasoner performs the translation of Application Terms contained in CSCDA Multi-Domain Thesaurus to search criteria that can be used to query EO catalogues, and therefore obtain relevant EO products. This translation is based on mapping rules. Below follows an overview of the EO-netCDF Resources Reasoner and the mapping rules used by the component. A detailed description can be found in [10].

6.4.1 The EO-netCDF Resources Reasoner

The EO-netCDF Resources Reasoner, used by the Rapid Response Server (RRS), is responsible for taking an ontology term and returning back the correct EO-netCDF search criteria (rules that set specific values to EO-netCDF terms). The reasoner follows the same approach as the EO Resources Reasoner (EORR) [1] developed in RARE. In particular, the RRS sends an application term from the CSCDA ontology to the reasoning service and the reasoner returns the corresponding EO-netCDF search criteria. The RRS uses these search criteria to build appropriate queries that can be used to query the EO catalogues and fetch back EO resources that satisfy these search criteria.

The EO-netCDF Resources Reasoner performs the translation of an application term to EO-netCDF in two phases:

1. At first, Application (Ontology) Terms are mapped to sets of Parameter-Value. This is implemented by mapping rules expressed as RIF¹⁶ rules, and having the following structure:

```
IF      <application term>
THEN    <parameter-1, value-1> AND
        <parameter-2, value-2> AND ...
```

These parameters, called Application Requirements Parameters (ARP), have been defined by EO experts and reflect common EO products features, such as Product Type, Acquisition Type, Sensor Resolution, etc. An example of a mapping rule is shown below:

```
IF      (Application Term == Precipitation)
THEN
        (Sensor Type = RADAR) AND (Sensor Resolution > 30) AND
        (Sensor Resolution <= 500) AND
        (Polarisation Channels = HH, HV, VH, VV) AND
        (Polarisation Mode = Q)
```

This rule maps the application term “Precipitation” to specific values (or valid ranges) of the parameters “Sensor Type”, “Sensor Resolution”, “Polarisation Channels” and “Polarisation Mode”. These are the criteria on which the search will be based. For example, one of the search criteria is that the ARP “Sensor Type” must be equal with RADAR.

2. ARPs cannot be used to query EO catalogues, so the EO-netCDF Resources Reasoner translates them to related EO-netCDF parameters. In order to do this, a simple XML two-column table is used. Each row contains an Application Requirement Parameter in the first column and an EO-netCDF parameter in the second column.

For example, below is the mapping of the parameter “Sensor Type”:

```
<application_requirement
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  ...
```

¹⁶<http://www.w3.org/2005/rules/wg/draft/ED-rif-primer-20121028/>

```
term="Sensor Type">
  <related_EO_NetCDF_term>
    /eop:earth_observation_information/eop:earth_observation_
      equipment/eop:sensor_information/eop:sensor_type
  </related_EO_NetCDF_term>
</application_requirement>
```

The ARP “Sensor Type” is mapped to the EO-netCDF term `sensor type`, defined as attribute in [4]. This attribute belongs to the group `sensor information` that can be found under the groups `earth_observation_information/earth_observation_equipments`, defined also in [4].

The output is thus available in the form of RIF consequents (THEN), where ARPs are replaced with the equivalent EO-netCDF parameter. This information can be used by the RRS to build appropriate OpenSearch queries and invoke the EO catalogues.

6.5 Summary

In this chapter, we covered the semantic technologies designed and developed in the scope of this thesis. These technologies semantically enhance the ProdTrees platform enabling users that are not familiar with EO concepts to search for EO products.

Chapter 7

Demonstration

The continuous views of Earth supplied by satellite images and data provide scientists and decision makers with the information they need to understand and protect the environment. Among their many applications are monitoring the air, seas and land; providing the basis for accurate weather reports; and supplying national and international relief agencies with data when disasters strike. In order to validate the EO-netCDF conventions and demonstrate the ProdTrees system, we use such data provided by three main use cases.

In particular, the validation process, firstly, includes the annotation of EO Products¹ from these sources with EO-netCDF metadata. Afterwards, the encoded datasets are stored to catalogues accessible by the GI-cat. Finally, the ProdTrees system is used to search for these data.

In this chapter, we start by presenting the three use cases of ProdTrees project. We continue by illustrating the capabilities of the ProdTrees system by querying the EO-netCDF annotated datasets using different features of the system each time.

7.1 Use Cases

7.1.1 Envisat

In March 2002, ESA launched Envisat² (“Environmental Satellite”) global monitoring mission, aiming to endow Europe with an enhanced capability for remote sensing observation of Earth from space. Envisat performed optical, radar and spectroscopic measurements of the atmosphere, ocean, land, and ice, ensuring data continuity with ESA’s pioneering ERS³ missions. With an advanced polar-orbiting Earth observation satellite that included 10 instruments aboard and at eight tons, Envisat was the largest Earth observation space-

¹Both real and test EO products are included.

²<https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat>

³<https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/ers>

craft ever built. The high-tech machine was engineered by a European consortium of companies from 13 countries.

Envisat carried an array of nine Earth-observation instruments that gathered information about the Earth, and a tenth instrument that provided guidance and control. Several of the instruments were advanced versions of instruments that were flown on the earlier ERS 1 and ERS 2 missions and other satellites. The instruments were⁴:

- **ASAR**⁵ (Advanced Synthetic Aperture Radar): the largest single instrument on board. ASAR ensured continuity of data after ERS-2. The radar featured enhanced capability in terms of coverage, range of incidence angles, polarisation and modes of operation.
- **MERIS** (Medium Resolution Imaging Spectrometer): A programmable, medium-spectral resolution, imaging spectrometer operating in the solar reflective spectral range. MERIS allowed global coverage of Earth every three days and had as primary mission the measurement of sea colour in oceans and coastal areas.
- **AATSR** (Advanced Along Track Scanning Radiometer): An infrared radiometer providing high resolution and high accuracy temperature information, for applications such as sea surface temperature or fire observation. AATSR was the successor of ATSR1 and ATSR2, payloads of ERS 1 and ERS 2. It could measure Earth's surface temperature to a precision of 0.3 K (0.54 Â°F), for climate research. Among the secondary objectives of AATSR was the observation of environmental parameters such as water content, biomass, and vegetal health and growth.
- **SCIAMACHY** (SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY): An imaging spectrometer whose primary mission objective was to perform global measurements of trace gases in the troposphere and stratosphere. It compared light coming from the sun to light reflected by the Earth, which provided information on the atmosphere through which the Earth-reflected light had passed.
- **MIPAS**⁶ (Michelson Interferometer for Passive Atmospheric Sounding): A Fourier transform spectrometer for the measurement of high-resolution gaseous emission spectra at the Earth's limb. It complemented SCIAMACHY and operated in the near

⁴http://www.esa.int/Our_Activities/Observing_the_Earth/Envisat/Mission_overview

⁵<https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat/instruments/asar>

⁶<https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat/instruments/mipas>

to mid infrared where many of the atmospheric trace-gases playing a major role in atmospheric chemistry have important emission features.

- **GOMOS** (Global Ozone Monitoring by Occultation of Stars): A medium resolution spectrometer dedicated to atmospheric monitoring, primarily measuring stratospheric ozone. It looked to stars as they descended through the Earth's atmosphere and changed color, which could tell a lot about the presence of gases, and allow for the first time a space-based measurement of the vertical distribution of these trace gases.
- **DORIS** (Doppler Orbitography and Radio-positioning Integrated by Satellite): A microwave tracking system that was used to determine the precise location of the Envisat satellite. It could determine the satellite's orbit to within 10 cm (4 in).
- **RA-2**⁷ (Radar Altimeter 2): An instrument for determining the two-way delay of the radar echo from the Earth's surface to a very high precision: less than a nanosecond. It also measured the power and the shape of the reflected radar pulses, and thus it could be used to define ocean topography, map/monitor sea ice and measure land heights.
- **MWR** (Microwave Radiometer): A microwave radiometer that measured integrated atmospheric water vapour column and cloud liquid water content, as correction terms for the radar altimeter signal. MWR measurement data are useful for the determination of surface emissivity and soil moisture over land, for surface energy budget investigations to support atmospheric studies, and for ice characterisation.
- **LRR**⁸ (Laser Retro Reflector): A passive device used as a reflector by ground-based SLR stations using high-power pulsed lasers.

The data acquired from all these different sensors on the satellite can be used for both scientific studies and a growing number of operational applications. These include studies on atmospheric chemistry, ozone depletion, biological oceanography, ocean temperature and colour, wind waves, hydrology (humidity, floods), agriculture and arboriculture, natural hazards, digital elevation modelling (using interferometry), monitoring of maritime traffic, atmospheric dispersion modelling (pollution), cartography and study of snow and ice.

⁷<https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat/instruments/ra-2>

⁸<https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat/instruments/lrr>

The satellite orbited Earth more than 50.000 times over 10 years - twice its planned lifetime. The mission ended on 08 April 2012, following the unexpected loss of contact with the satellite.

7.1.2 SENTINEL-1

SENTINEL-1 is the first of the five missions which compose the SENTINEL program⁹ that ESA is developing for the Copernicus initiative. Each mission will focus on a different aspect of Earth observation; Atmospheric, Oceanic, and Land monitoring, and the data will be of use in many applications. With the objectives of Land and Ocean monitoring, the SENTINEL-1 mission comprises a constellation of two polar-orbiting satellites, SENTINEL-1A and SENTINEL-1B, sharing the same orbital plane, and operating day and night with the ability to acquire imagery regardless of the weather.

The mission includes C-band imaging operating in four exclusive imaging modes with different resolution (down to 5 m) and coverage (up to 400 km). SENTINEL-1 continues the C-band SAR Earth Observation of ESA's ERS 1, ERS 2 and ENVISAT, and Canada's RADARSAT-1¹⁰ and RADARSAT-2¹¹. It provides dual polarisation capability, very short revisit times and rapid product delivery. For each observation, precise measurements of spacecraft position and attitude are available. Synthetic Aperture Radar (SAR) has the advantage of operating at wavelengths not impeded by cloud cover or a lack of illumination and can acquire data over a site during day or night time under all weather conditions.

The constellation covers the entire world's land masses on a bi-weekly basis, sea-ice zones, Europe's coastal zones and shipping routes on a daily basis and open ocean continuously by wave imageries. Therefore, the SENTINEL-1 mission provides an independent operational capability for continuous radar mapping of the Earth and is designed to provide enhanced revisit frequency, coverage, timeliness and reliability for operational services and applications requiring long time series.

Mission's objectives include:

- Land monitoring of forests, water, soil and agriculture,
- Emergency mapping support in the event of natural disasters,
- Marine monitoring of the maritime environment,

⁹<https://sentinel.esa.int/web/sentinel/home>

¹⁰<http://www.asc-csa.gc.ca/eng/satellites/radarsat1/>

¹¹<http://www.asc-csa.gc.ca/eng/satellites/radarsat2/>

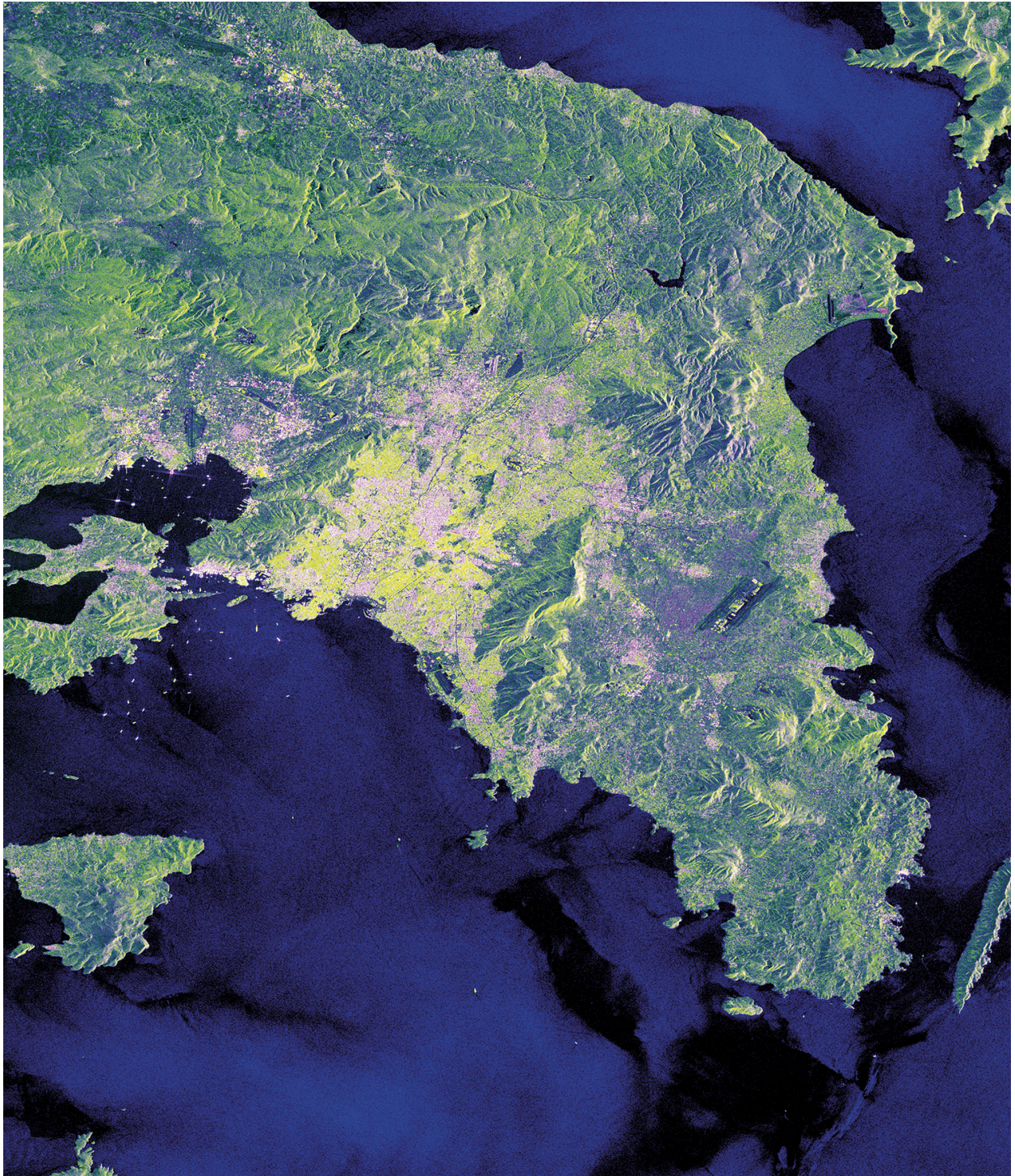


Figure 7.1: Sentinel-1A radar acquisition from 22 April 2014 showing Greece's Attica region, with mountainous areas and the capital and largest city of Athens near the centre. In the water, different shades of blue indicate different types of sea surface, influenced by currents and waves. *Copyright ESA*

- Sea ice observations and iceberg monitoring,
- Production of high resolution ice charts,
- Forecasting ice conditions at sea,
- Mapping oil spills,
- Sea vessel detection, and
- Climate change monitoring

As a result, the mission will benefit numerous services. More specifically, SENTINEL-1 will be the primary source of data for information on the oceans and the Arctic. The mission's ability to provide observation in all weather, and in day or night time conditions, makes it ideal for maritime and Arctic monitoring. The dual polarimetric products will benefit users interested in agriculture, forestry and land cover classification. The enhanced interferometric capabilities will benefit users involved in activities like geohazard monitoring, mining, geology and city planning through subsidence risk assessment. SENTINEL security users will be able to monitor major shipping routes to detect illegal activities, gather prosecution evidence in case of illegal discharges, detect unexpected building in remote areas, monitor deforestation and support search and rescue activities. Finally, the rapid data dissemination and short revisit cycles of SENTINEL-1 together with its interferometric capabilities will also benefit emergency response users, such as the United Nations International Charter on Space and Major Disasters, in emergency situations such as floods, earthquakes, volcanic eruptions and landslides.

Each SENTINEL-1 satellite is expected to transmit Earth observation data for at least 7 years and have fuel on-board for 12 years. The mission is designed to work in a pre-programmed, conflict-free operation mode, imaging all global landmasses, coastal zones and shipping routes at high resolution and covering the global ocean with vignettes. This will ensure the reliability of service required by operational services and a consistent long term data archive built for applications based on long time series.

7.1.3 MyOcean

MyOcean¹² is a series of projects granted by the EC within the Copernicus Program, whose objective is to define and to set up a concerted and integrated pan-European capacity for

¹²<http://www.myocean.eu/>

ocean monitoring and forecasting. The series include MyOcean (2009-2012), MyOcean2 (2012-2014) and MyOcean follow-on (October 2014-March 2015) projects, respectively funded by the EU's Seventh Framework Programme for Research (FP7 2007-2013) and HORIZON 2020 (EU Research and Innovation programme 2014-2020), and they have been designed to prepare and to lead the demonstration phases of the future Copernicus Marine Environment Monitoring Service. Today, the Copernicus Marine Service is yet provided by the MyOcean2 consortium to more than 4000 users worldwide on a pre-operational mode, and the MyOcean follow-on is meant to be full operational from 2015 onwards.



Figure 7.2: The 7 areas covered by the MyOcean services

MyOcean services have been designed to respond to issues emerging in the environmental, business and scientific sectors. Based on the combination of space and in situ observations, MyOcean services provide state-of-the-art information which offers an unprecedented capability to observe, understand and anticipate marine environment events. In particular, MyOcean provides analyses and forecasts of the Global Ocean (worldwide coverage) and of European seas, and their assimilation into 4D models (including the time frame) such as: Temperature, salinity, currents, sea ice, sea level, wind and biogeochemical parameters. These activities benefit several specified areas of use like: Maritime security, oil spill prevention, marine resources management, climate change, seasonal forecasting, coastal activities, ice sheet surveys, water quality and pollution.

The 7 areas covered by the MyOcean services (Figure 7.2) are monitored with an eddy-resolving capacity, based on assimilation of space and in situ data into 3D models, representing the physical state, the ice and the ecosystems of the ocean; in the past (25 years), in real-time and in the future (1-2 weeks). The high-quality products rely on the aggregation of European modelling tools and the scientific methodology is produced through a strong cross-fertilization between operational and research communities.

MyOcean products are available for users of all marine applications in order to add value to their own operational systems or to contribute to their R&D programs. The product are provided through an online catalogue¹³, and users can download them according to their needs in netCDF format and benefit from quality and validation information for most of them.

7.2 Search Scenarios

The ProdTrees platform is accessible through the web interface of the system (Figure 7.3). This interface allows the users to submit free-text search queries, to select application terms that are defined in supported ontologies and to specify multiple search criteria. In order to highlight the functionality of the platform, we will demonstrate three core scenarios. A video demonstrating these capabilities is also available at <http://bit.ly/ProdTreesPlatform>.

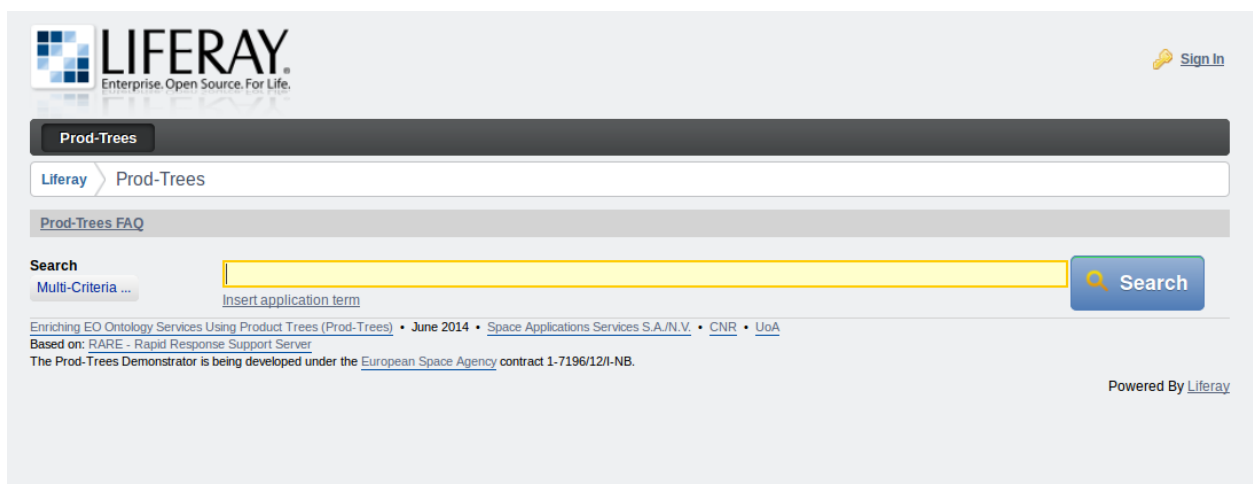


Figure 7.3: The Web interface of the ProdTrees platform

¹³<http://www.myocean.eu/web/69-myocean-interactive-catalogue.php>

7.2.1 Free Text Search

In the first scenario the user inserts a free-text query, for example “water”.

The system replies by presenting a number of different interpretations for the inserted text, which are provided by the Query Analyzer during the disambiguation phase. This way it is clear for the user what are the semantics of the text on which the search will be based. The default interpretation for “water” maps this text to the concept “water” of CSCDA ontology (Figure 7.4).

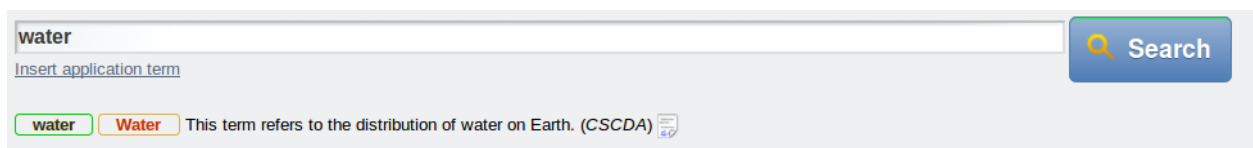


Figure 7.4: The default interpretation for the keyword “water”

In case the user is not satisfied with this interpretation, she can select another one from a proposed list, for example “water level gauges”, “fresh water river discharge”, and more (Figure 7.5). Another option is to use the inserted text without any specific interpretation. In this case, a simple text-based search will be performed.

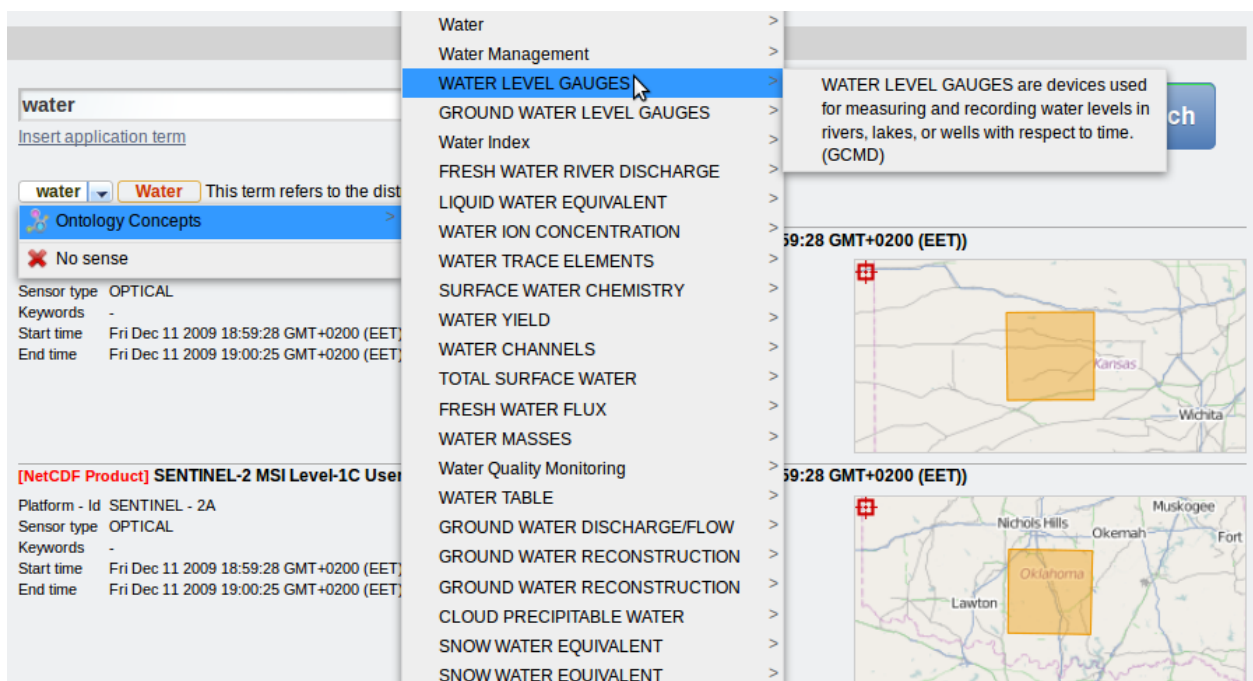


Figure 7.5: The different interpretations for the keyword “water”

After selecting the desired interpretation, for example the concept “water” of CSCDA,

the EO-netCDF Resources Reasoner is invoked in order to map the concept “water” to EO-netCDF parameters with specific values. This is done with the use of appropriate mapping rules which allow us to connect concepts of an ontology (in this case water of CSCDA) to EO-netCDF parameters with specific values. For instance, a mapping rule could specify a combination of Satellite Sensor type, Sensor resolution, Polarisation Channels, and Polarisation Mode:

```
IF (Application Term == Water)
THEN
    (Sensor Type = RADAR) AND
    (Sensor Resolution >= 30) AND
    (Sensor Resolution <= 500) AND
    (Polarisation Channels = HH) AND
    (Polarisation Mode = 5)
```

As a result, GI-cat returns only the EO products that include EO-netCDF parameters with these values. Figure 7.6 displays the first results of the keyword search for “water”.

water

[Insert application term](#)

Search

water

Water

This term refers to the distribution of water on Earth. (CSCDA)

[NetCDF Product] Sentinel-1 SM Level-1 SLC Product (Thu Jan 12 2012 17:08:45 GMT+0200 (EET))

Platform - Id

SENTINEL-1 - A

Sensor type

RADAR

Keywords

-

Start time

Thu Jan 12 2012 17:08:45 GMT+0200 (EET)

End time

Thu Jan 12 2012 17:08:53 GMT+0200 (EET)

[NetCDF Product] Sentinel-1 WV Level-1 SLC Product (Sun Jan 01 2012 04:07:41 GMT+0200 (EET))

Platform - Id

SENTINEL-1 - A

Sensor type

RADAR

Keywords

-

Start time

Sun Jan 01 2012 04:07:41 GMT+0200 (EET)

End time

Sun Jan 01 2012 04:07:43 GMT+0200 (EET)

[NetCDF Product] Sentinel-1 WV Level-1 SLC Product (Sun Jan 01 2012 04:08:10 GMT+0200 (EET))

Platform - Id

SENTINEL-1 - A

Sensor type

RADAR

Keywords

-

Start time

Sun Jan 01 2012 04:08:10 GMT+0200 (EET)

End time

Sun Jan 01 2012 04:08:13 GMT+0200 (EET)

Figure 7.6: The results

In case the user wants to limit the results, she can add more keywords, like toponyms, which are disambiguated using the Geonames gazetteer, or time constraints. Thus, only resources that fulfil the additional search criteria will be retrieved (Figure 7.7).

[Insert application term](#)


This term refers to the distribution of water on Earth. (CSCDA)

is a capital of a political entity located in Belgium (GeoNames).

from 1 January 2012 to 1 January 2013.


[NetCDF Product] Sentinel-1 EW Level-1 SLC Product (Sun Jan 01 2012 18:43:03 GMT+0200 (EET))

Platform - Id SENTINEL-1 - A
 Sensor type RADAR
 Keywords -
 Start time Sun Jan 01 2012 18:43:03 GMT+0200 (EET)
 End time Sun Jan 01 2012 18:43:18 GMT+0200 (EET)



[NetCDF Product] Sentinel-1 EW Level-1 GRD Product (Sun Jan 01 2012 18:43:02 GMT+0200 (EET))

Platform - Id SENTINEL-1 - A
 Sensor type RADAR
 Keywords -
 Start time Sun Jan 01 2012 18:43:02 GMT+0200 (EET)
 End time Sun Jan 01 2012 18:43:20 GMT+0200 (EET)



[NetCDF Product] Sentinel-1 IW Level-1 GRD Product (Mon Jan 09 2012 05:44:06 GMT+0200 (EET))

Platform - Id SENTINEL-1 - A
 Sensor type RADAR
 Keywords -
 Start time Mon Jan 09 2012 05:44:06 GMT+0200 (EET)
 End time Mon Jan 09 2012 05:44:24 GMT+0200 (EET)




Figure 7.7: The filtered results

7.2.2 Term-based Search

Instead of the text queries, the user can also use the ontology browser to select terms she wants. With the browser the user can navigate within and across the supported ontologies, in order to find terms defined in these ontologies that she can use as search criteria.

The screenshot displays the GEOSS ontology browser interface. At the top, there is a navigation bar with 'Concepts' and 'Keyword Search' tabs, and a URL field showing 'http://www.earthobservations.org/GEOSS'. The main content area is titled 'AGRICULTURE Concept' and includes tabs for 'General Info', 'Labels', 'Notes', and 'Concept matching'. On the left, a 'Hierarchical Representation' sidebar lists various concepts, with 'AGRICULTURE' highlighted. The main panel shows the 'AGRICULTURE' concept details, including a 'Broader term' section with 'Top Concept' and a 'Narrower terms' section listing related concepts like 'AGRICULTURAL AQUATIC SCIENCES', 'AGRICULTURAL ENGINEERING', 'AGRICULTURAL PLANT SCIENCE', 'ANIMAL SCIENCE', 'FOREST SCIENCE', 'PLANT COMMODITIES', and 'SOILS'. At the bottom right, there are 'Insert Concept' and 'Close' buttons.

Figure 7.8: The details of GEOSS concept AGRICULTURE

When the user selects a concept, then the selected concept is copied back to the initial text area. Assuming the user has selected the concept “agriculture” of GEOSS ontology to use in the search (Figure 7.8), as before she can add more keywords (toponyms, date etc.) to the text area in order to restrict the search. Afterwards, the work-flow is similar to the one described in the previous scenario.

7.2.3 EO-related Search

Finally, the third scenario will show how search using EO-related search criteria. This case might be more appropriate for expert users, since the user can search for resources using specific metadata values such as bounding box and a more detail date (Figure 7.9). In addition, and more importantly, the user can select one or more EO-netCDF parameters and insert a specific value for each one (Figure 7.10).

The screenshot shows the 'Multi-Criteria Search' interface. At the top, there is a 'Search Terms:' input field. Below it, a 'Where:' section includes a map of Europe and a bounding box input area. A 'When:' section features a date selection dropdown menu with options like 'Today', 'Last 7 days', 'Month to date', 'Year to date', 'The previous Month', 'Specific Date', 'All Dates Before', 'All Dates After', and 'Date Range'. To the right, there are two calendar views for 'August 2014', labeled 'Start date' and 'End date'. The 'Start date' calendar shows the 28th as selected. The 'End date' calendar shows the 28th as selected. At the bottom, there are 'Reset' and 'Find Resources' buttons. A small 'Liferay' logo is visible in the bottom right corner.

Figure 7.9: Multi-Criteria Search page

This is feasible with the use of an EO-netCDF browser that allows the user to navigate through the hierarchy of the NetCDF Earth Observation Metadata Conventions, select the ones that she wants to use in the search, and then specify a value for each one of them. For instance, the user can search for resources where sensor type is “RADAR” and sensor resolution equals to 42.2. (Figure 7.11).

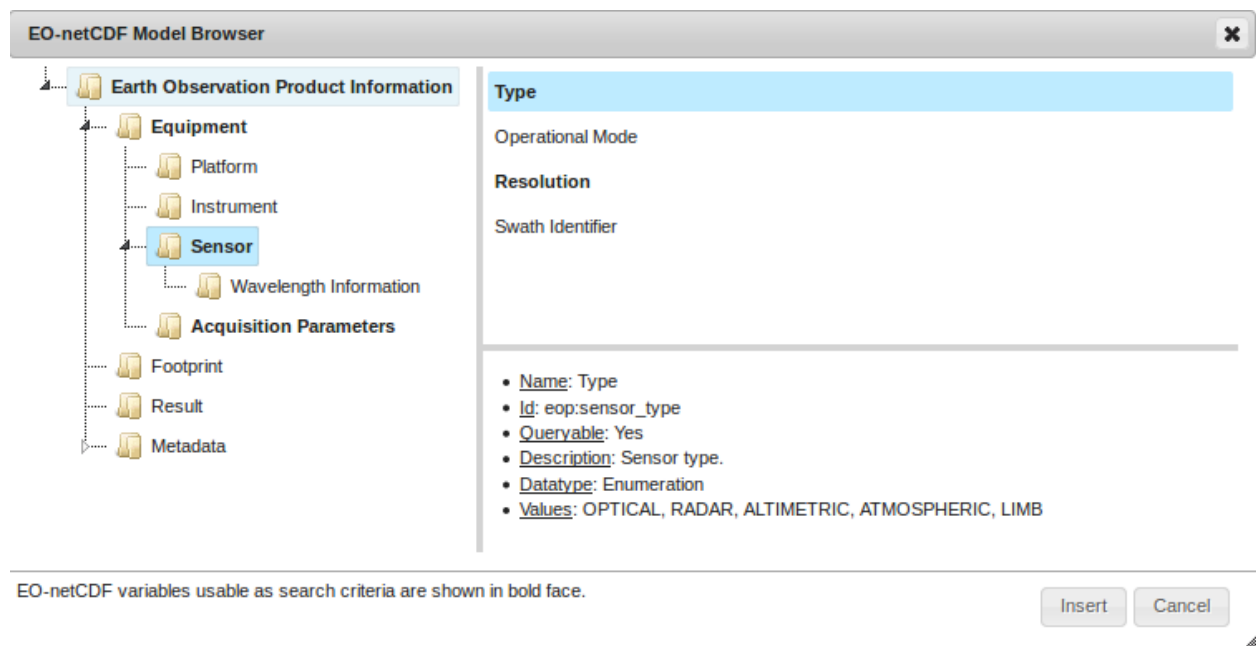


Figure 7.10: The EO-netCDF Model Browser

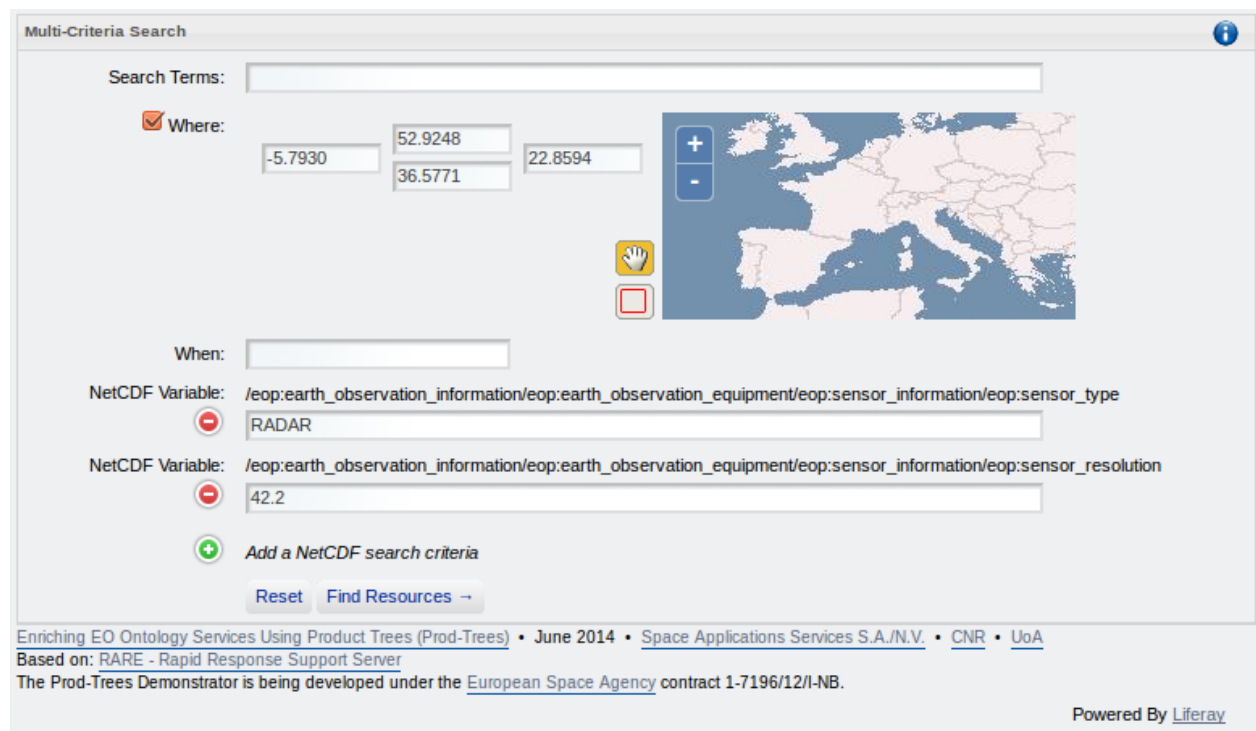


Figure 7.11: Search query with bounding box and specified EO-netCDF parameter

The search will be based on this attribute and will return only EO products that satisfy it. As the EO-netCDF parameter is provided directly by the user, the EO-netCDF

Resources Reasoner will be bypassed and only the GI-cat component will be invoked to return the relevant resources (Figure 7.12).

The screenshot displays the 'Prod-Trees' search interface. At the top, there are tabs for 'Your Query', 'Your Selection', and 'Filter and Sort Criteria'. The 'Your Query' tab is active, showing a list of search criteria:

- Resource type(s): netcdf
- No specific resource providers selected
- Search through the catalogs broker: yes
- Bounding box: (-5.7930, 52.9248), (22.8594, 36.5771)
- /eop:earth_observation_information/eop:earth_observation_equipment/eop:sensor_information/eop:sensor_type = RADAR
- /eop:earth_observation_information/eop:earth_observation_equipment/eop:sensor_information/eop:sensor_resolution = 42.2

Below the query, a section titled 'Matching Resources' shows a single result: '[NetCDF Product] Sentinel-1 → Sentinel-1 EW Level-1 SLC Product (Sun Jan 01 2012 18:43:02 GMT+0200 (EET))'. To the left of this result, a table lists details:

Acquisition platform	Sentinel-1
Sensor type	RADAR
Keywords	-
Start time	Sun Jan 01 2012 18:43:02 GMT+0200 (EET)
End time	Sun Jan 01 2012 18:43:17 GMT+0200 (EET)

To the right of the table is a map of Europe with a red bounding box highlighting the search area in Western Europe, covering parts of the Netherlands, Belgium, and Germany. At the bottom of the interface, there is a footer with the following text:

Enriching EO Ontology Services Using Product Trees (Prod-Trees) • June 2014 • Space Applications Services S.A./N.V. • CNR • UoA
 Based on: RARE - Rapid Response Support Server
 The Prod-Trees Demonstrator is being developed under the European Space Agency contract 1-7196/12/I-NB.
 Powered By Liferay

Figure 7.12: The results of the query of Figure 7.11

7.3 Summary

In this chapter, we first described the various uses cases of the ProdTrees project. Then, we demonstrated the various capabilities of the ProdTrees system by presenting different search scenarios.

Chapter 8

Accessing FedEO Clearinghouse

This chapter describes the demo we created to illustrate how we can access the EO data offered by the FedEO Clearinghouse, how this data can be combined with linked open data and, finally, what kind of knowledge can be extracted. The first section provides information about the FedEO Clearinghouse and the next section follows with the details of the demo.

8.1 Federated EO Missions Support Environment

FedEO Clearinghouse¹ is setup by the HMA-S Evolution Project and provides access to the FedEO Clearinghouse Product and Collections Catalogs via an OpenSearch interface. The implementation is based on the following standards:

- OGC 10-157r3, Earth Observation Metadata profile of Observations & Measurements, Version 1.0.0 [18]
- OGC 10-032r8, OpenSearch GeoSpatial and Temporal Extensions [20]
- OGC 13-026r4, OpenSearch Extension for Earth Observation [19]
- OASIS searchRetrieve specifications²

The OpenSearch description document³ provides the URL templates that can be used to search for products or collections. Figure 8.1 displays a URL template for collection search. As you can see, there are many types of parameters that can be used for this search (e.g. productType, platform, instrument, etc.).

The following URL is a search query for three collections of EOP:ESA:FEDEO:COLLECTIONS that contain in title the word “vegetation”.

¹<http://geo.spacebel.be/opensearch/readme.html>

²<http://docs.oasis-open.org/search-ws/searchRetrieve/v1.0/os/part0-overview/searchRetrieve-v1.0-os-part0-overview.html>

³<http://geo.spacebel.be/opensearch/description.xml>

```
<Url rel="collection"
template="http://geo.spacebel.be:80/openserach/request/?http
Accept=application/atom%2Bxml&parentIdentifier={eo:p
parentIdentifier?}\&subject={dc:subject?}\&query={se
archTerms?}\&startRecord={startIndex?}\&startPage={
startPage?}\&maximumRecords={count?}\&startDate={ti
me:start?}\&endDate={time:end?}\&type={dc:type?}\&a
mp;title={dc:title?}\&publisher={dc:publisher?}\&bb
ox={geo:box?}\&name={geo:name?}\&lat={geo:lat?}\&
lon={geo:lon?}\&radius={geo:radius?}\&uid={geo:ui
d?}\&organizationName={eo:organizationName?}\&produ
ctType={eo:productType?}\&platform={eo:platform?}\&
instrument={eo:instrument?}\&classifiedAs={semantic:cla
ssifiedAs?}\&recordSchema={sru:recordSchema?}"
type="application/atom+xml"/>
```

Figure 8.1: URL template for collection search in FedEO

```
http://geo.spacebel.be/openserach/request/?httpAccept=
application/atom%2Bxml&parentIdentifier=EOP:ESA:FEDEO:
COLLECTIONS&title=VEGETATION&maximumRecords=3
```

The results of this query are shown in Figure 8.2. In the address bar you will also see the query.

Figure 8.3 displays the results page for a product search. As you see, the results now are images. Each image is accompanied with O & M metadata (see Section 2.4). So, there is information about the sensor type, the sensor resolution, orbit direction etc.

FedEO provides also an “Explain Document”⁴ with the names of the supported collections and datasets (see Figure 8.4).

FedEO supports requests for various formats, like ATOM, RSS, SRU. It can also access EO data in RDF format. The main disadvantage with RDF data in FedEO is that there are EO data that are not encoded at all in RDF. As a result, some queries that ask for data in RDF do not return any results, even if there are data (in other formats) that satisfy the constraints.

8.2 Demonstration

In the context of this thesis, we created a demo to show how linked open data can be compined with EO data and how we can extract interesting knowledge (e.g. spatial data analytics). The demo uses a Strabon endpoint where linked open data and RDF EO data

⁴<http://geo.spacebel.be/openserach/request/>

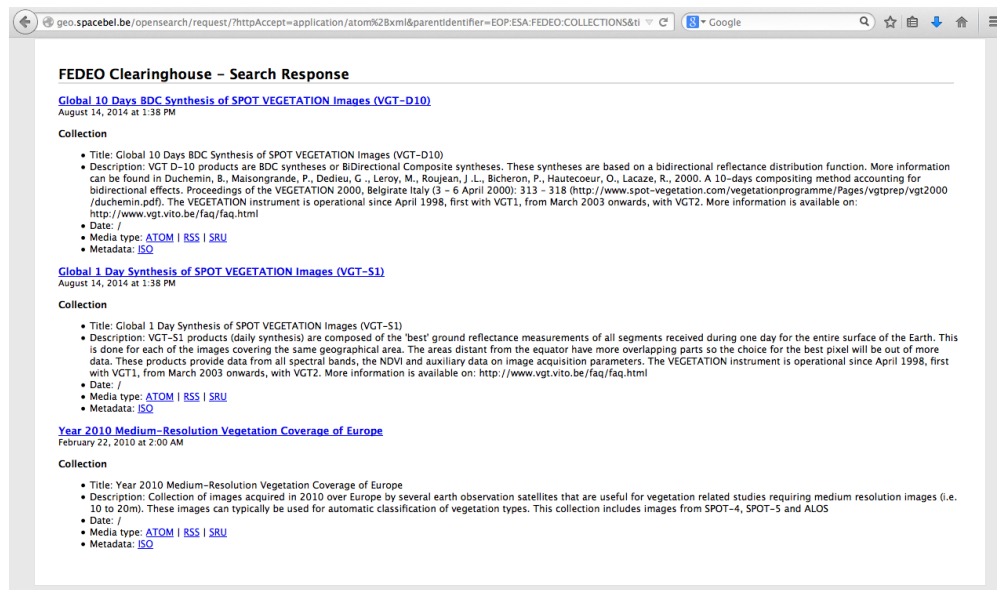


Figure 8.2: Results page for collection search in FedEO

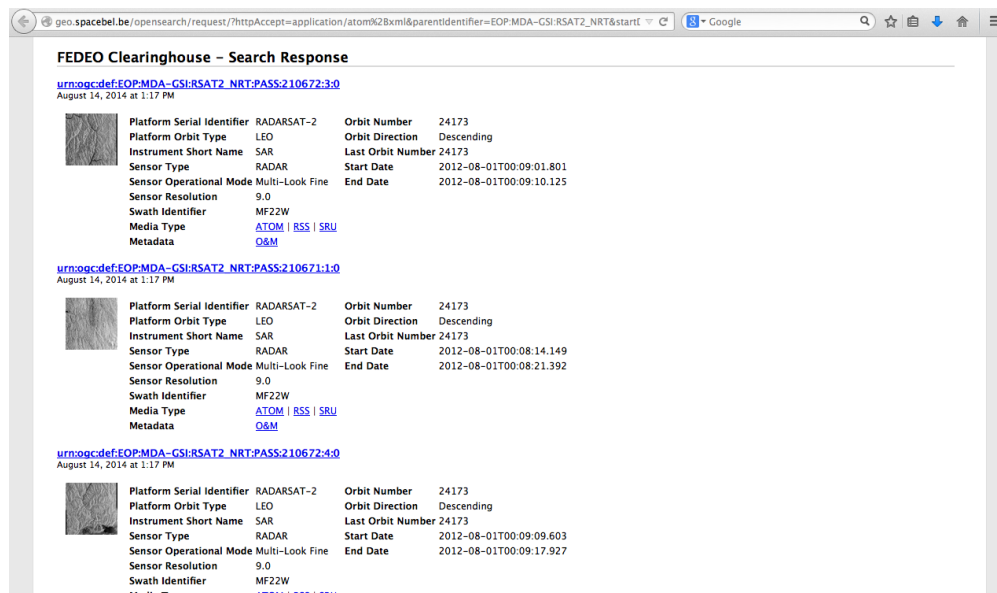


Figure 8.3: Results page for product search in FedEO

```

- <!--
  //////////// ESA M2CS EO-DAIL Collections ////////////
  -->
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_ASA_APC_0S</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_ASA_APH_0S</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_ASA_APV_0S</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_ASA_GMI_1S</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_ASA_IMx_xS</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_ASA_WSx_xS</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_ASA_WV_0S</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_ATS_xxx_xS</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_GOMOS</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_MER_FR_xS</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_MER_RR_xS</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_MIP_NL_xC</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_RA2_MWx_2C</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ENVISAT_SCI_C</supports>
  <supports type="value">EOP:ESA:ESA.EECF.EOS_MOD_xS</supports>
  <supports type="value">EOP:ESA:ESA.EECF.ERS_SAR_xS</supports>
  <supports type="value">EOP:ESA:ESA.EECF.LANDSAT_MSS_xS</supports>
  <supports type="value">EOP:ESA:ESA.EECF.LANDSAT_TM_ETM_P</supports>
  <supports type="value">EOP:ESA:ESA.EECF.LANDSAT_TM_ETM_xS</supports>
  <supports type="value">EOP:ESA:ESA.EECF.PROBA_HRC_xS</supports>

```

Figure 8.4: Part from Explain Document of FedEO displaying supported collections of ESA M2CS EO-DAIL

of FedEO are stored.

For the purpose of this demo, we used as linked open data the NUTS dataset (described below). We harvested RDF EO data from FedEO by using a script with OpenSearch queries. To harvest all the available collections and datasets, the OpenSearch queries use the names included in the Explain Document, one name for each request. The default maximum number of results for each query was 10, so we downloaded locally only a part of the existed products and collections in RDF.

Before we store the EO data to Strabon, we needed to do some preprocessing to modify the triples with geometries according to the GeoSPARQL standard. This means that we added the Well Known Text (WKT) literal and the reference system⁵ EPSG:4326 in each geometry. Also, we modified the triples with geometries to add a geo:Feature and a geo:SpatialObject according to the GeoSPARQL representation [29].

⁵<http://en.wikipedia.org/wiki/SRID>

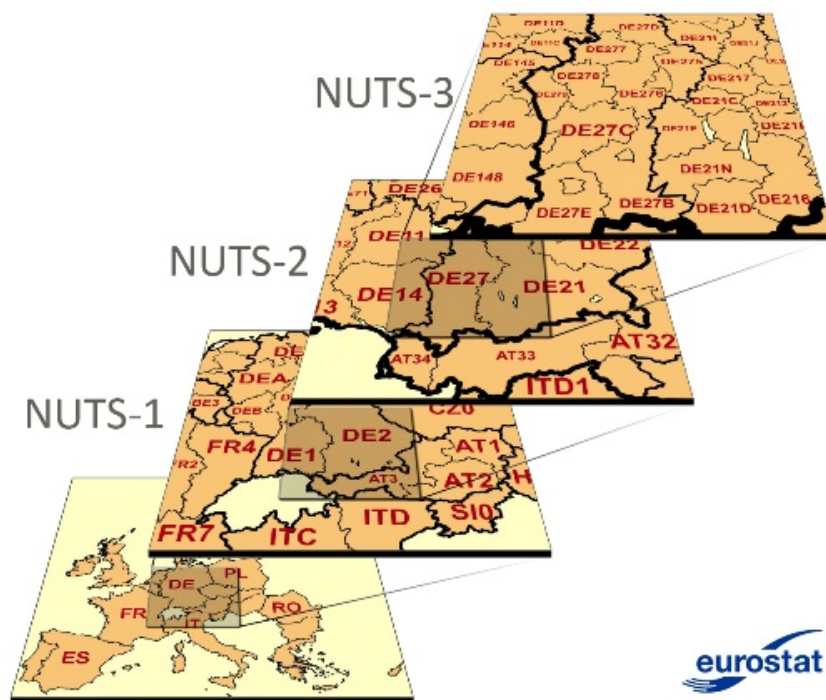


Figure 8.5: The four levels of NUTS

8.2.1 NUTS dataset

The NUTS⁶ (Nomenclature of Territorial Units for Statistics) is a classification defined by the Eurostat⁷ office of the European Union. The NUTS classification is a hierarchical system for dividing up the economic territory of the EU, mainly for statistical and policy purposes. The four level of division are:

- NUTS 0: countries
- NUTS 1: major socio-economic regions
- NUTS 2: basic regions for the application of regional policies
- NUTS 3: as small regions for specific diagnoses

All the information about the four levels is available in RDF. For the demo we used only the level 0, the countries of EU. NUTS is also linked with other datasets, like DBpedia⁸.

⁶<http://nuts.geovocab.org>

⁷<http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/>

⁸<http://dbpedia.org/>

8.2.2 Queries and Results

The demo contains three types of queries:

- Discovery queries, to explore what kind of information is available for the EO data of FedEO
- NUTS based queries, that combine EU countries with EO data
- Statistics using NUTS, that produce charts as result

The prefixes used in the following queries are:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dclite: <http://xmlns.com/2008/dclite4g#>
PREFIX purl: <http://purl.org/dc/elements/1.1/>
PREFIX eop:
    <http://www.genesi-dr.eu/spec/openserch/extensions/eop/1.0/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX xmlns: <http://xmlns.com/2008/dclite4g#>
PREFIX gn:
    <http://www.genesi-dr.eu/spec/openserch/extensions/eop/1.0/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
```

Collections Discovery Queries

1. Total number of collections.

```
select distinct (COUNT(?x) AS ?totalCollections)
where
{
    ?x rdf:type dclite:Series
}
```

Result:

The number of collections in the demo is 288.

2. Select all collection properties.

```
select distinct ?prop
where
{
    ?x rdf:type dclite:Series ;
    ?prop      ?propValue .
}
```

Result:

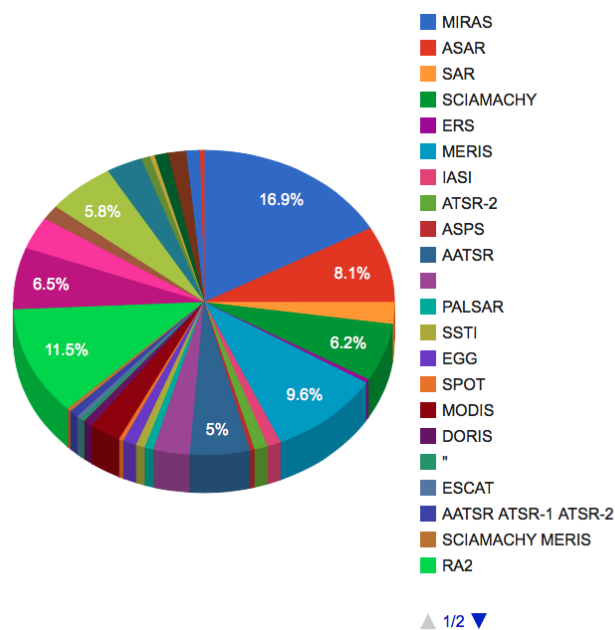
prop
http://www.w3.org/2005/Atomlink
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://xmlns.com/2008/dclite4g#Series
http://purl.org/dc/elements/1.1/identifier
http://purl.org/dc/elements/1.1/description
http://purl.org/dc/elements/1.1/title
http://purl.org/dc/elements/1.1/subject
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/sensor
http://www.w3.org/2002/12/cal/ical#dtstart
http://www.w3.org/2002/12/cal/ical#dtend
http://purl.org/dc/terms/spatial
http://purl.org/dc/elements/1.1/format
http://purl.org/dc/terms/created
http://purl.org/dc/terms/modified
http://purl.org/dc/terms/extent
http://purl.org/dc/terms/abstract
http://purl.org/dc/elements/1.1/rights
http://xmlns.com/2008/dclite4g#resolution
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/processingLevel
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/platform
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/mission
http://www.opengis.net/ont/geosparql#hasGeometry

The collections can contain information about their title, sensor, platform, mission, resolution, start and end time, format.

3. Collections per sensor.

```
select distinct ?propValue ( COUNT(?x) AS ?incCollections )
where
{
    ?x rdf:type dclite:Series ;
    eop:sensor ?propValue .
} GROUP BY ?propValue
```

Result:



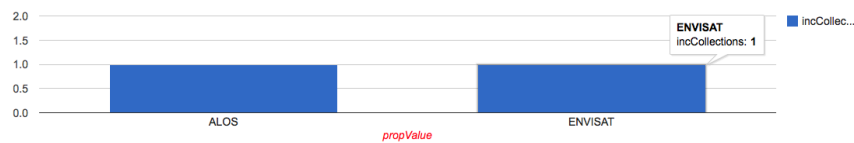
The most used sensor is “MIRAS”.

4. Collections per platform.

```
select distinct ?propValue ( COUNT(?x) AS ?incCollections )
where
{
    ?x rdf:type dclite:Series ;
    eop:platform ?propValue .
} GROUP BY ?propValue
```

Result:

Only two collections have information about their platform. The one platform is “ALOS” and the other one is “ENVISAT”.



5. Collections per mission.

```
select distinct ?propValue ( COUNT(?x) AS ?incCollections )
where
{
    ?x rdf:type dclite:Series ;
        eop:mission ?propValue .
} GROUP BY ?propValue
```

Result:

propValue	incCollections
"Earth Explorer, Soil Moisture and Ocean Salinity Satellite"	"11"^^<http://www.w3.org/2001/XMLSchema#integer>
"Earth Explorer, Ocean Salinity and Ocean Salinity Satellite"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>

Two available missions for 12 collections.

6. Collections per format.

```
select distinct ?propValue ( COUNT(?x) AS ?incCollections )
where
{
    ?x rdf:type dclite:Series ;
        purl:format ?propValue .
} GROUP BY ?propValue
```

Result:

propValue	incCollections
"SMOS"	"43"^^<http://www.w3.org/2001/XMLSchema#integer>
"ERS"	"5"^^<http://www.w3.org/2001/XMLSchema#integer>
"ENVISAT"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"SMOS "	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"NETCDF"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"TIFF"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"GOCE"	"5"^^<http://www.w3.org/2001/XMLSchema#integer>
"SPOT"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"HDF"	"5"^^<http://www.w3.org/2001/XMLSchema#integer>
"NetCDF"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"	"4"^^<http://www.w3.org/2001/XMLSchema#integer>
"GeoTIFF PNG"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"CEOS"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"ESA"	"7"^^<http://www.w3.org/2001/XMLSchema#integer>
"N1 (ENVISAT)"	"4"^^<http://www.w3.org/2001/XMLSchema#integer>

The most common format is “SMOS” (44 collections).

Datasets Discovery Queries

1. Total Datasets.

```
select distinct (COUNT(?x) AS ?totalDatasets)
where
{
    ?x rdf:type dclite:DataSet
}
```

Result:

The total number of datasets in the demo is 2254.

2. Datasets per collection.

```
select distinct ?collection (COUNT(?dataset) AS ?dataSetPerSeries)
where
{
    ?dataset rdf:type dclite:DataSet ;
    dclite:series ?collection .
} GROUP BY ?collection
```

Result:



As you see, the most collections contain only 10 datasets. This is because of the default limit (max 10) in the number of results. So, probably these collections contain more than 10 datasets. As the purpose of this demo was just to illustrate what kind of information is possible to extract from the RDF files of FedEO and how these can be combined with linked open data, the limited number of datasets in the demo is not a problem.

3. Dataset properties.

```
select distinct ?prop
where
{
    ?x rdf:type dclite:DataSet ;
        ?prop ?propValue .
}
```

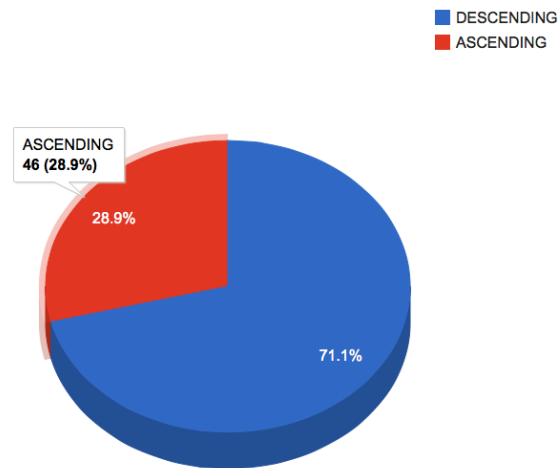
Result:

prop
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://purl.org/dc/elements/1.1/identifier
http://www.w3.org/2002/12/cal/ical#dtstart
http://www.w3.org/2002/12/cal/ical#dtend
http://purl.org/dc/terms/spatial
http://purl.org/dc/terms/created
http://purl.org/dc/terms/modified
http://xmlns.com/2008/dclite4g#series
http://xmlns.com/2008/dclite4g#onlineResource
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/processorVersion
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/orbitNumber
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/acquisitionStation
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/processingCenter
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/size
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/processingDate
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/trackNumber
http://earth.esa.int/sarpolarisationChannels
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/orbitDirection
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/swathIdentifier
http://xmlns.com/2008/dclite4g#quicklook
http://purl.org/dc/elements/1.1/relation
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/wrsLongitudeGrid
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/cycle
http://purl.org/dc/terms/dateSubmitted
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/wrsLatitudeGrid
http://purl.org/dc/elements/1.1/coverage
http://xmlns.com/2008/dclite4g#thumbnail
http://purl.org/dc/elements/1.1/source
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/processorName
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/processingVersion
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/productType
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/processingCenterId
http://purl.org/dc/terms/isPartOf
http://www.genesi-dr.eu/spec/opensearch/extensions/eop/1.0/polarisationMode
http://www.opengis.net/ont/geosparql#hasGeometry

4. Datasets per orbit direction.

```
select distinct ?propValue ( COUNT(?x) AS ?incDatasets )
where
{
    ?x rdf:type dclite:DataSet ;
        eop:orbitDirection ?propValue .
} GROUP BY ?propValue
```

Result:

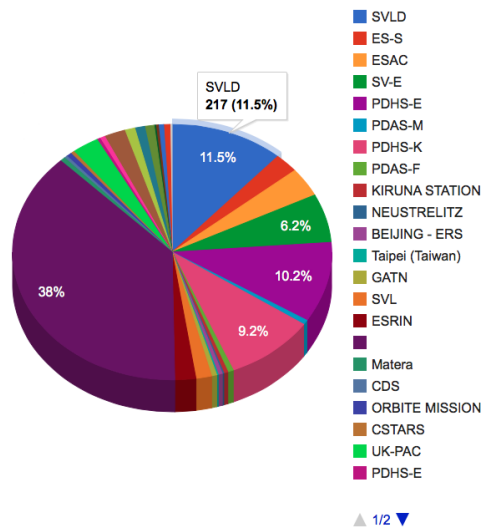


There are 113 datasets with descending orbit direction and 46 datasets with ascending orbit direction.

5. Datasets per acquisition station.

```
select distinct ?propValue ( COUNT(?x) AS ?incDatasets )
where
{
    ?x rdf:type dclite:DataSet ;
        eop:acquisitionStation ?propValue .
} GROUP BY ?propValue
```

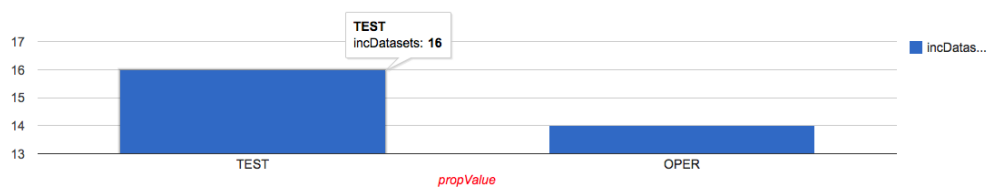
Result:



6. Datasets per product type.

```
select distinct ?propValue ( COUNT(?x) AS ?incDatasets )
where
{
    ?x rdf:type dclite:DataSet ;
        eop:productType ?propValue .
} GROUP BY ?propValue
```

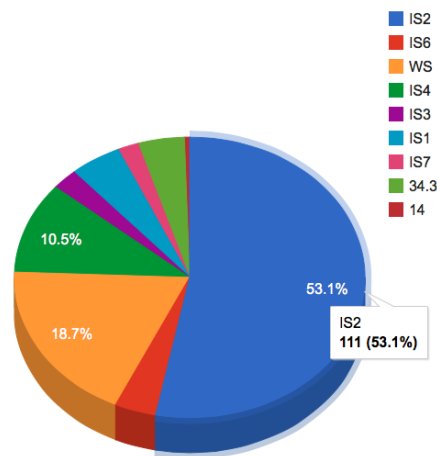
Result:



7. Datasets per swath identifier.

```
select distinct ?propValue ( COUNT(?x) AS ?incDatasets )
where
{
    ?x rdf:type dclite:DataSet ;
        eop:swathIdentifier ?propValue .
} GROUP BY ?propValue
```

Result:

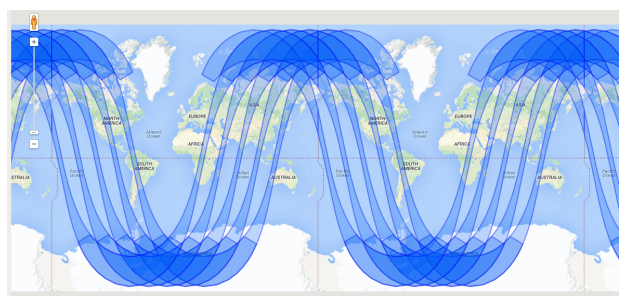


8. Display the geometries of the datasets in collection “MIR_SC_F1A_REPR”.

```
SELECT  ?geos ?dataset_ID
WHERE {

    ?s rdf:type xmlns:Series.
    ?s purl:identifier "MIR_SC_F1A_REPR".
    ?d xmlns:series ?s.
    ?d rdf:type xmlns:DataSet.
    ?d purl:identifier ?dataset_ID.
    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.
}
```

Result:



Because of the limit restriction, only the geometries of 10 datasets are displayed. However, the important thing is to observe the orbit of the datasets in this collection.

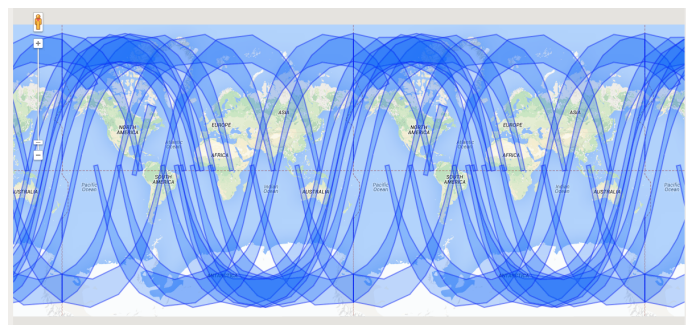
9. Display the geometries of the datasets in collection “AT1_ARD_2P”.

```
SELECT  ?geos ?dataset_ID
WHERE {

    ?s rdf:type xmlns:Series.
    ?s purl:identifier "AT1_ARD_2P".
    ?d xmlns:series ?s.
    ?d rdf:type xmlns:DataSet.
    ?d purl:identifier ?dataset_ID.
    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.

}
```

Result:



10. Display the geometries of the datasets in collection “MIR_SMDAP2”.

```
SELECT  ?geos ?dataset_ID
WHERE {

    ?s rdf:type xmlns:Series.
    ?s purl:identifier "MIR_SMDAP2".
    ?d xmlns:series ?s.
    ?d rdf:type xmlns:DataSet.
    ?d purl:identifier ?dataset_ID.
    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.

}
```

Result:



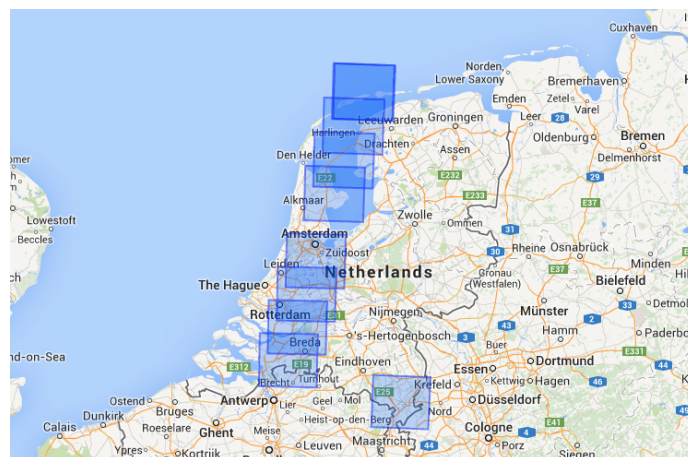
11. Display the geometries of the datasets in collection “ALPSMB_1B2G”.

```
SELECT  ?geos ?dataset_ID
WHERE {

    ?s rdf:type xmlns:Series.
    ?s purl:identifier "ALPSMB_1B2G".
    ?d xmlns:series ?s.
    ?d rdf:type xmlns:DataSet.
    ?d purl:identifier ?dataset_ID.
    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.

}
```

Result:

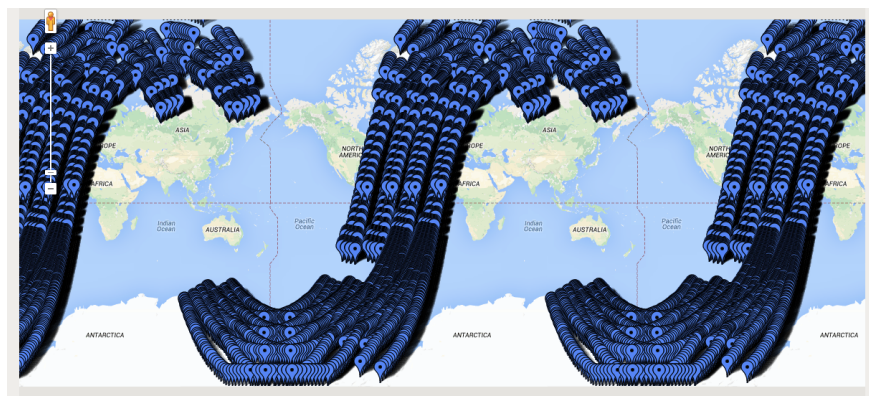


12. Display the geometries of the datasets in collection “SCI_NL__2P”.

```
SELECT  ?geos ?dataset_ID
WHERE {

    ?s rdf:type xmlns:Series.
    ?s purl:identifier "SCI_NL__2P".
    ?d xmlns:series ?s.
    ?d rdf:type xmlns:DataSet.
    ?d purl:identifier ?dataset_ID.
    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.
}
```

Result:



Below you can see the same result after zoom in:



NUTS based queries

1. Which EU countries have an intersection with the dataset “ASPS20_H_040411075415.E2”?

```
SELECT  distinct ?label
WHERE {

    ?d rdf:type xmlns:DataSet.
    ?d purl:identifier "ASPS20_H_040411075415.E2".
    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.

    ?x rdfs:label ?label.
    ?x geo:hasGeometry ?g.
    ?g geo:asWKT  ?geo.
    OPTIONAL{?x <http://geovocab.org/spatial#PP> ?pp}

    FILTER(!bound(?pp))
    FILTER(geof:sfIntersects(?geo, ?geos))
}
```

Result:

label
"CY - ΚΥΠΡΟΣ / KIBRIS"
"BG - БЪЛГАРИЯ / BULGARIA"
"GR - ΕΛΛΑΔΑ / ELLADA"
"TR - TÜRKİYE"
"RO - ROMÂNIA"

2. Which EU countries are included in datasets with processing center “I-PAF”?

```
SELECT  distinct ?label
WHERE {

    ?d rdf:type xmlns:DataSet.
    ?d gn:processingCenter "I-PAF".
    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.
```

```

?x rdfs:label ?label.
?x geo:hasGeometry ?g.
?g geo:asWKT ?geo.
OPTIONAL{?x <http://geovocab.org/spatial#PP> ?pp}

FILTER(!bound(?pp))
FILTER(geof:sfIntersects(?geos, ?geo))
}

```

Result:

label
"IT - ITALIA"
"FR - FRANCE"
"SI - SLOVENIJA"
"HR - HRVATSKA"
"ES - ESPAÑA"

3. Which EU countries are included in collections with sensor “ASAR”?

```

SELECT    distinct ?label
WHERE {

    ?s rdf:type xmlns:Series.
    ?s gn:sensor "ASAR".
    ?d xmlns:series ?s.
    ?d rdf:type xmlns:DataSet.
    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.

    ?x rdfs:label ?label.
    ?x geo:hasGeometry ?g.
    ?g geo:asWKT ?geo.
    OPTIONAL{?x <http://geovocab.org/spatial#PP> ?pp}

    FILTER(!bound(?pp))
    FILTER(geof:sfIntersects(?geos, ?geo))
}

```

Result:

label
"IT - ITALIA"
"LI - LIECHTENSTEIN"
"BE - BELGIQUE-BELGIË"
"LU - LUXEMBOURG (GRAND-DUCHÉ)"
"UK - UNITED KINGDOM"
"ES - ESPAÑA"
"CH - SCHWEIZ/SUISSE/SVIZZERA"
"NO - NORGE"
"FR - FRANCE"
"DE - DEUTSCHLAND"
"PL - POLSKA"
"SE - SVERIGE"
"SI - SLOVENIJA"
"NL - NEDERLAND"
"DK - DANMARK"
"HR - HRVATSKA"
"AT - ÖSTERREICH"

4. Which EU countries are included in collections with “NETCDF” format?

```

SELECT    distinct ?label    (COUNT(distinct ?dataset_ID)
                                as ?NumberOfDatasets)

WHERE {

    ?s rdf:type xmlns:Series.
    ?s purl:format "NETCDF".
    ?d xmlns:series ?s.
    ?d rdf:type xmlns:DataSet.
    ?d purl:identifier ?dataset_ID.
    ?d gn:size ?dataset_size.

    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.

    ?x rdfs:label ?label.
    ?x geo:hasGeometry ?g.
    ?g geo:asWKT    ?geo.
    OPTIONAL{?x <http://geovocab.org/spatial#PP> ?pp}

    FILTER(!bound(?pp))
    FILTER(geof:sfIntersects(?geo, ?geos))
}
GROUP BY ?label

```

Result:

label	NumberOfDatasets
"IS - ÍSLAND"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"IT - ITALIA"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"MT - MALTA"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"AT - ÖSTERREICH"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"BE - BELGIQUE-BELGIË"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"CH - SCHWEIZ/SUISSE/SVIZZERA"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"CZ - ČESKÁ REPUBLIKA"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"DE - DEUTSCHLAND"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"DK - DANMARK"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"HR - HRVATSKA"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"LI - LIECHTENSTEIN"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"LU - LUXEMBOURG (GRAND-DUCHÉ)"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"NL - NEDERLAND"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"NO - NORGE"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"SE - SVERIGE"	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"SI - SLOVENIJA"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"FR - FRANCE"	"3"^^<http://www.w3.org/2001/XMLSchema#integer>
"ES - ESPAÑA"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"FI - SUOMI / FINLAND"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"GR - ΕΛΛΑΔΑ / ELLADA"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"TR - TÜRKİYE"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"CY - ΚΥΠΡΟΣ / KIBRIS"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"RO - ROMÂNIA"	"1"^^<http://www.w3.org/2001/XMLSchema#integer>

5. Display the intersection of EU countries with the geometry of datasets that have processing center "I-PAF".

```

SELECT  (geof:intersection(?geo, ?geos) as ?intersection)  ?id
WHERE {

    ?d rdf:type xmlns:DataSet.
    ?d gn:processingCenter "I-PAF".
    ?d purl:identifier ?id.
    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.

    ?x rdfs:label ?label.
    ?x geo:hasGeometry ?g.
    ?g geo:asWKT ?geo.
    OPTIONAL{?x <http://geovocab.org/spatial#PP> ?pp}

    FILTER(!bound(?pp))
    FILTER(geof:sfIntersects(?geos, ?geo))
}
    
```

Result:



Statistics using NUTS

1. Number of EU countries per dataset with acquisition center “ESRIN”.

```

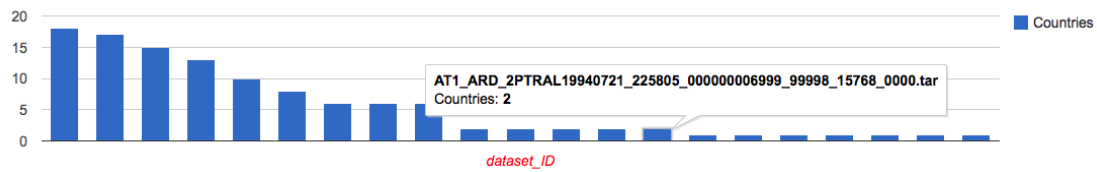
SELECT  ?dataset_ID (COUNT(distinct ?label) as ?Countries)
WHERE {

    ?d rdf:type xmlns:DataSet.
    ?d purl:identifier ?dataset_ID.
    ?d gn:acquisitionStation "ESRIN".
    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.

    ?x rdfs:label ?label.
    ?x geo:hasGeometry ?g.
    ?g geo:asWKT ?geo.
    OPTIONAL{?x <http://geovocab.org/spatial#PP> ?pp}

    FILTER(!bound(?pp))
    FILTER(geof:sfIntersects(?geos, ?geo))
}
GROUP BY ?dataset_ID
ORDER BY DESC(?Countries)
    
```

Result:



2. Number of EU countries per dataset in collection with identifier “SCI_NL__1P”.

```
SELECT  ?dataset_ID (COUNT(distinct ?label) as ?Countries)
WHERE {
```

```
    ?s rdf:type xmlns:Series.
```

```
    ?s purl:identifier "SCI_NL__1P".
```

```
    ?d xmlns:series ?s.
```

```
    ?d rdf:type xmlns:DataSet.
```

```
    ?d purl:identifier ?dataset_ID.
```

```
    ?d gn:size ?size.
```

```
    ?d geo:hasGeometry ?gs.
```

```
    ?gs geo:asWKT ?geos.
```

```
    ?x rdfs:label ?label.
```

```
    ?x geo:hasGeometry ?g.
```

```
    ?g geo:asWKT  ?geo.
```

```
    OPTIONAL{?x <http://geovocab.org/spatial#PP> ?pp}
```

```
    FILTER(!bound(?pp))
```

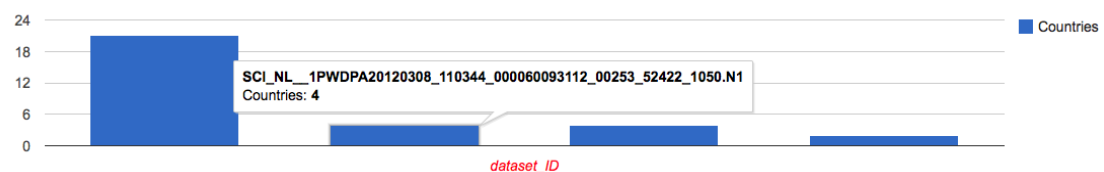
```
    FILTER(geof:sfIntersects(?geos, ?geo))
```

```
}
```

```
GROUP BY ?dataset_ID
```

```
ORDER BY DESC(?Countries)
```

Result:



3. Number of datasets with descending orbit direction per EU country.

```

SELECT ?label (COUNT(distinct ?d) as ?NumberOfDatasets)
WHERE {

    ?d rdf:type xmlns:DataSet.
    ?d purl:identifier ?dataset_ID.
    ?d gn:orbitDirection "DESCENDING".
    ?d geo:hasGeometry ?gs.
    ?gs geo:asWKT ?geos.

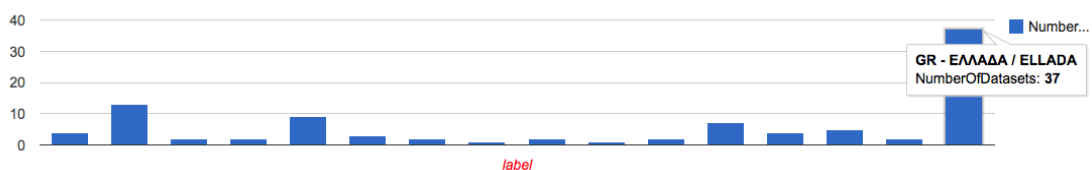
    ?x rdfs:label ?label.
    ?x geo:hasGeometry ?g.
    ?g geo:asWKT ?geo.
    OPTIONAL{?x <http://geovocab.org/spatial#PP> ?pp}

    FILTER(!bound(?pp))
    FILTER(geof:sfIntersects(?geos, ?geo))
}

GROUP BY ?label

```

Result:



8.2.3 Conclusions

All the queries described above show how users can easily extract knowledge about EO data using the query language SPARQL and GeoSPARQL. Also, EO metadata can be compined with linked open data and produce additional information, like the presented charts.

To create this demo, we needed first to download locally all the RDF data, store it to Strabon and then the users make queries directly to Strabon. In a future version, it would

be possible to develop an online service where the users make OpenSearch queries to FedEO for the collections (or datasets) they are interested in and a database is dynamically created to store the RDF results of these queries. Then, the users would be able to make SPARQL queries in this database using Strabon. This way, it is not needed to harvest FedEO and the results are always kept updated.

As we already mentioned, in FedEO Clearinghouse there are O&M metadata available in XML, but not in RDF. The translation from XML to RDF is not accurate, as there is not a standardize RDF vocabulary for HMA metadata. However, if there was an ontology to describe these metadata, it would be easy to make an alignment with EOP O&M, and finally translate the XML metadata in RDF. So, another plan for future work in this field would be to define and standardize RDF vocabulary for publishing O&M metadata as linked data in RDF.

8.3 Summary

In this chapter, we discussed about how someone can access the EO data offered by the FedEO Clearinghouse. Afterwards, we showed how this data can be combined with linked open data and, finally, what kind of knowledge can be extracted.

Chapter 9

Conclusions and Future Work

This chapter presents this thesis's conclusions, as well as our future work.

9.1 Conclusions

In the context of this thesis, we studied the accessing and discovering challenges that end-users face while searching for EO products. At first, in Chapter 2, we introduced the most important elements of the EO domain such as the basic concepts of earth observation, and initiatives and standards organisations of the area. In Chapter 3, we described the technical background of our work, explaining the theoretical distance from vocabularies to ontologies, presenting the data formats and models that we used, as well as related technologies. We continued, in Chapter 4, by surveying related work in the areas most relevant to our work, like the RARE and the SMAAD projects that were the ones that introduced us to this area of interest. Afterwards, in Chapter 5, we presented a new standard called EO-netCDF for accessing EO products annotated with netCDF. We then described how the ProdTrees platform, a semantically-enabled search engine for EO products, implements the EO search based on this standard and what is the role of each component in the system. Chapter 6 covered the semantic technologies of the ProdTrees platform that were developed in the context of this thesis and Chapter 7 illustrated the capabilities of the platform by presenting various search scenarios. Also, Chapter 7, included an introduction to the use cases that were used to validate the EO-netCDF, by providing the EO data that were annotated and then queried by the ProdTrees platform. At last, in Chapter 8, we illustrated an example of accessing open EO dataset repositories and utilizing the provided data. For this, we used the FedEO Clearinghouse. Finally, we conclude in this Chapter by presenting our plans for future work.

9.2 Future Work

Our future work concentrates on enhancing the semantic technologies developed in the scope of this thesis. In particular, we focus on the ontology matching system that we

developed.

The most important of addition to Pythia is a user-evaluation process. As the semantics of each term always depend on the knowledge domain in which it is used, we cannot be totally confident for the correctness of the produced mappings, even if we perform only a strict string similarity algorithm. On the other hand, if users were able to give feedback and proposals regarding the accuracy of the mappings, then we would have a higher degree of trust for the final results. Another useful extension would be to use, not only WordNet, but also other domain-specific vocabularies. This way it would be easier for Pythia to infer which is the most appropriate meaning for a term. For instance, as in our use case we are interested in the EO domain, we should use the GEneral Multilingual Environmental Thesaurus (GEMET) as a base ontology to better describe the content of the other ontologies. At last, adding various ranking methods will eliminate the noise in the final results.

Acronyms - Abbreviations

AATSR	Advanced Along Track Scanning Radiometer (instrument on Envisat)
ACDD	Attribute Convention for Dataset Discovery
API	Application Programming Interface
ARP	Application Requirements Parameters
ASAR	Advanced Synthetic Aperture Radar (instrument on Envisat)
ASCII	American Standard Code for Information Interchange
ATM	Atmospheric (GML extension)
ATSR	Along Track Scanning Radiometer (instrument on ERS)
CCSDS	Consultative Committee for Space Data Systems
CDL	Common Data Language
CEOS	Committee on Earth Observation Satellites
CF	Climate and Forecast
CF-netCDF	Climate and Forecast network Common Data Form
CIM	Cataloguing of ISO Metadata
CDM	Common Data Model
CNR	Consiglio Nazionale delle Ricerche
COARDS	Cooperative Ocean/Atmosphere Research Data Service
COB	Cross-Ontology Browser
CRS	Coordinate Reference Systems
CSCDA	Copernicus Space Component Data Access
CSS	Cascading Style Sheets
CSW	Catalogue Services for the Web
DAB	GEO Discovery and Access Broker
DIF	Directory Interchange Format
DISC	Data and Information Services Center
DLR	Deutsches Zentrum für Luft- und Raumfahrt (German Aerospace Center)
DORIS	Doppler Orbitography and Radio-positioning Integrated by Satellite (instrument on Envisat)
ebRIM	ebXML Registry Information Model

EC	European Commission
ECHO	Earth Observing System (EOS) Clearing House
ECSS	European Cooperation for Space Standardization
EEA	European Environment Agency
EFAS	European Flood Awareness System
Eionet	Environmental Information and Observation Network
Envisat	Environmental Satellite (operated by ESA)
EO	Earth Observation
EO-netCDF	Earth Observation network Common Data Form
EOP	Earth Orientation Parameters/Earth Observation Product (GML extension)
EOP-O&M	Earth Observation Metadata profile of Observations & Measurements
EORR	EO Resources Reasoner
EOSDIS	NASA's Earth Observing System Data and Information System
EPSG	European Petroleum Survey Group
ERS	European Remote Sensing
ESA	European Space Agency
ESRIN	European Space Research Institute
ETCs	European Topic Centres
EU	European Union
FedEO	Federated Earth Observation
FGDC	Federal Geographic Data Committee
GBIF	Global Biodiversity Information Facility
GCI	GEOSS Common Infrastructure
GCMD	Global Change Master Directory
GEMET	General Multilingual Environmental Thesaurus
GEO	Group on Earth Observations
GeoSPARQL	Geographic SPARQL Protocol and RDF Query Language
GEOSS	Global Earth Observation System of Systems
GeoXACML	Geospatial eXtensible Access Control Markup Language
GES	Goddard Earth Sciences
GI	Geographic Information
GIS	Geographic Information System
GMD	Geographic MetaData XML (encoding of ISO 19115)
GMES	Global Monitoring for Environment and Security

GML	Geography Markup Language
GOMOS	Global Ozone Monitoring by Occultation of Stars (instrument on Envisat)
GRIB	GRIdded Binary or General Regularly-distributed Information in Binary form
GSC	GMES Space Component
GSCB	Ground Segment Coordination Body
GSCDA	GSC Data Access
HARM	Historical Archives Rationalization and Management
HDF	Hierarchical Data Format
HMA	Heterogeneous EO Missions Accessibility
HTML	HyperText Markup Language (W3C Standard)
HTTP(S)	(Secured) HyperText Transport Protocol
ICT	Information and Communication Technology
IDN	International Directory Network
INSPIRE	Infrastructure for Spatial Information in Europe
ISO	International Organization for Standardization
JPG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
KML	Keyhole Markup Language
LRR	Laser Retro Reflector (instrument on Envisat)
LTDP	Long-Term Data Preservation
MACC-II	Monitoring Atmospheric Composition & Climate II
MERIS	MEDium-Resolution Imaging Spectrometer (instrument on Envisat)
MIPAS	Michelson Interferometer for Passive Atmospheric Sounding (instrument on Envisat)
MWR	Microwave Radiometer (instrument on Envisat)
NASA	National Aeronautics and Space Administration
NcML	NetCDF Markup Language
netCDF	network Common Data Form
NetCDF-U	NetCDF Uncertainty Conventions
NOAA	National Oceanic and Atmospheric Administration
NUTS	Nomenclature of Territorial Units for Statistics
OAI	Open Archival Information
OAIS	Open Archival Information System

OGC	Open Geospatial Consortium
O&M	Observations and Measurements
OPT	Optical (GML extension)
OSDD	OpenSearch Description Document
OTE	Ontology and Terminology for Earth Observation
OTEG	Open Access Ontology/Terminology for the GMES Space Component
OWL	Web Ontology Language
OWS	OGC Web Service
PDF	Portable Document Format
PPT	PoolParty Thesaurus Server
QA	Query Analyzer
RA	Radar Altimeter (instrument on Envisat)
RARE	Rapid Response Support Server
RSS	Rich Site Summary
RDBMS	Relational Database Management System
RDF	Resource Description Format
RDFS	RDF Schema
RESto	REstful Semantic search Tool for geOspatial
RRC	Rapid Response Client
RRS	Rapid Response Server
RIF	Rule Interchange Format
SAFE	Standard Archive Format for Europe
SAR	Synthetic Aperture Radar
SBA	(GEOSS) Societal Benefit Areas
SCIAMACHY	SCanning Imaging Absorption spectroMeter for Atmospheric CHartography (instrument on Envisat)
SensorML	Sensor Model Language
SKOS	Simple Knowledge Organization System
SKOS-XL	SKOS eXtension for Labels
SMAAD	Semantic Annotation and Mediation
SOA	Service-Oriented Architecture
SOAP	Simple Object Access Protocol
SOS	Sensor Observation Service
SPARQL	SPARQL Protocol And RDF Query Language
SRU	Search/Retrieve via URL

SWE	Sensor Web Enablement
TC	Technical Committee
UDDI	Universal Description, Discovery and Integration
UML	Uniform Modeling Language
UncertML	Uncertainty Markup Language
UoA	National & Kapodistrian University of Athens
URI	Universal Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WCS	Web Coverage Service
WFS	Web Feature Service
WKT	Well Known Text
WMS	Web Map Service
WPS	Web Processing Service
XFDU	XML Formatted Data Units
XML	eXtensible Markup Language

Bibliography

- [1] Space Applications. *RARE Interface Control Document*. RARE-SA-ICD, Version 2.2.0.
- [2] Peter Baumann. *OGC®WCS 2.0 Interface Standard-Core: Corrigendum*, July 2012. OGC 09-110r4, Version 2.0.1.
- [3] Lorenzo Bigagli and Stefano Nativi. *NetCDF Uncertainty Conventions (NetCDF-U) 1.0*, November 2011. OGC 11-163, Version 1.0.
- [4] Enrico Boldrini, Paolo Mazzetti, and Stefano Nativi. *NetCDF Earth Observation (EO) Metadata Convention*, Mar 2014.
- [5] Enrico Boldrini, Stefano Nativi, Fabrizio Papeschi, Mattia Santoro, Lorenzo Bigagli, Fabrizio Vitale, Valerio Angelini, and Paolo Mazzetti. *GI-cat Catalog Service ver. 6.0 Specification*. Draft Specification.
- [6] Ame BrÄ¶ring, Christoph Stasch, and Johannes Echterhoff. *OGC®Sensor Observation Service Interface Standard*, April 2011. OGC 12-006, Version 2.0.
- [7] John Caron. *Unidata’s Common Data Model Version 4*, 2014. [Online; Accessed 25-September-2014].
- [8] Simon Cox. *Observations and Measurements - XML Implementation*, March 2011. OGC 10-025r1, Version 2.0.
- [9] AnHai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. *Ontology matching: A machine learning approach*. In *Handbook on Ontologies*, International Handbooks on Information Systems, pages 385–403. Springer Berlin Heidelberg, 2004.
- [10] Kallirroï Dogani. *Enabling Semantic Search and Discovery for Earth Observation Products*.
- [11] Ben Domenico. *NetCDF Binary Encoding Extension Standard: NetCDF Classic and 64-bit Offset Format*, April 2011. OGC 10-092r3, Version 1.0.
- [12] Ben Domenico. *OGC®Network Common Data Form (NetCDF) Core Encoding Standard version 1.0*, April 2011. OGC 10-090r3, Version 1.0.
- [13] Ben Domenico. *OGC®Network Common Data Form (NetCDF) NetCDF Enhanced Data Model Extension Standard*, October 2012. OGC 11-038r2, Version 1.0.

- [14] Ngo DuyHoa, Zohra Bellahsene, and Remi Colletta. A Flexible System for Ontology Matching. *CAISE'11: International Conference on Advanced Information Systems Engineering*, 2011.
- [15] Brian Eaton, Jonathan Gregory, Bob Drach, Karl Taylor, Steve Hankin, John Caron, Rich Signell, Phil Bentley, Greg Rappa, Heinke HÄ¶ck, Alison Pamment, and Martin Juckes. CF Convention, 2011. [Online; Accessed 25-July-2014].
- [16] Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.
- [17] Mike Folk and Elena Pourmal. *HDF5 1.6 Standard*, January 2007. ESDS-RFC-007v1, Online; Accessed 22-September-2014.
- [18] Jerome Gasperi, Frédéric Houbie, Andrew Woolf, and Steven Smolders. *Earth Observation Metadata profile of Observations & Measurements*, June 2012. OGC 10-157r3, Version 1.0.0.
- [19] Pedro Gonçalves. *OGC®OpenSearch Extension for Earth Observation* , May 2013. OGC 13-026.
- [20] Pedro Gonçalves. *OGC®OpenSearch Geo and Time Extensions*, April 2014. OGC 10-032r8, Version 1.0.0.
- [21] Yves Jean-Mary, E Shironoshita, and Mansur Kabuka. Ontology Matching with Semantic Verification. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3), 2009.
- [22] Kostis Kyzirakos, Manos Karpathiotakis, and Manolis Koubarakis. Strabon: A Semantic Geospatial DBMS. In *ISWC*, volume 7649 of *LNCS*, pages 295–311. Springer, 2012.
- [23] Patrick Maué. *Semantic annotations in OGC standards*, October 2012. OGC 08-167r2, Version 2.0.
- [24] George A. Miller. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, November 1995.
- [25] Miklos Nagy and Maria Vargas-Vera. Towards an Automatic Semantic Data Integration: Multi-agent Framework Approach. In *Semantic Web*. InTech, 2010.

- [26] DuyHoa Ngo, Zohra Bellahsene, and Konstantin Todorov. Opening the Black Box of Ontology Matching. In *The Semantic Web: Semantics and Big Data*, volume 7882 of LNCS, pages 16–30. Springer Berlin Heidelberg, 2013.
- [27] G. Major K. Shein J. Scialdone S. Ritz T. Stevens M. Morahan A. Aleman R. Vogel S. Leicester H. Weir M. Meaux S. Grebas C. Solomon M. Holland T. Northcutt R. A. Restrepo R. Bilodeau Olsen, L.M. *NASA/Global Change Master Directory (GCMD) Earth Science Keywords*, 2013. Version 8.0.0.0.0.
- [28] Matthew Perry and John Herring. *OGC GeoSPARQL - A Geographic Query Language for RDF Data*, September 2012. OGC 11-052r4, Version 1.0.
- [29] Matthew Perry and John Herring. *OGC®GeoSPARQL - A Geographic Query Language for RDF Data*, September 2012. OGC 11-052r4, Version 1.0.
- [30] Giuseppe Pirro and Domenico Talia. An approach to Ontology Mapping based on the Lucene search engine library. *18th International Workshop on Database and Expert Systems Applications*, 2007.
- [31] Mattia Santoro, Paolo Mazzetti, Stefano Nativi, Christiano Fugazza, Carlos Granell, and Laura Diaz. Methodologies for Augmented Discovery of Geospatial Resources. In *Discovery of Geospatial Resources: Methodologies, Technologies and Emergent Applications*, chapter 9.
- [32] Pavel Shvaiko and Jerome Euzenat. Ontology Matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):158–176, 2013.
- [33] Unidata. 2011 Unidata NetCDF Workshop, 2011. [Online; Accessed 20-July-2014].
- [34] Unidata. The NetCDF Users’ Guide, 2011. [Online; Accessed 20-July-2014].
- [35] Unidata. The NetCDF Markup Language (NcML), 2013. [Online; Accessed 22-September-2014].
- [36] Panagiotis Vretanos. *OpenGIS Web Feature Service 2.0 Interface Standard*, November 2010. OGC 09-025r1, Version 2.0.