

D2.5.1: Report on Natural Language Discovery and Query Interface

| | |
|------------------------------|--|
| Title | D2.5.1: Report on Natural Language Discovery and Query Interface |
| Creator | Claudia Cialone, Kristin Stock |
| Creation date | 04/11/2010 |
| Date of last revision | 05/04/2011 |
| Subject | Interoperability, Query Interface, multilingual discovery, AOC, alignment, Natural Semantic Metalanguage |
| Status | <input type="checkbox"/> Draft <input checked="" type="checkbox"/> Final |
| Publisher | EuroGEOSS |
| Type | Text |
| Description | Natural Language/Semantics Research and Discovery Interface Implementation |
| Contributor | Claudia Cialone, Kristin Stock (CGS, University of Nottingham) |
| Format | Word document. |
| Source | |
| Rights | <input type="checkbox"/> Restricted <input checked="" type="checkbox"/> Public |
| Identifier | EuroGEOSS_D_2_5_1.doc |
| Language | En |
| Relation | WP2 |
| Coverage | Not applicable |

These are Dublin Core metadata elements. See for more details and examples <http://www.dublincore.org/>

TABLE OF CONTENTS

| | | |
|-----|---|----|
| 1 | INTRODUCTION | 5 |
| 2 | Goals & Objectives | 6 |
| 3 | The Natural Language Discovery Interface | 8 |
| 3.1 | Phase 1: Ontology Querying | 9 |
| 3.2 | Phase 2: Natural Language Spatial Relation Querying | 9 |
| 3.3 | Phase 3: More Generalised Natural Language Querying | 10 |
| 4 | Progress | 10 |
| 4.1 | Thesaurus Management | 10 |
| 4.2 | Implementation | 13 |
| 4.3 | The Multilingual WPS User Interface | 17 |
| 4.4 | Comparison with other AOC Discovery Interfaces | 19 |
| 5 | Remaining Work | 19 |
| 5.1 | Phase 2 Progress | 20 |
| 5.2 | Phase 3 | 22 |
| 6 | Output and dissemination | 24 |
| 7 | CONCLUSIONS | 25 |
| | REFERENCES | 25 |

APPENDIXES

| | |
|---|----|
| Appendix 1: SEMANTIC KNOWLEDGE SCHEMES | 27 |
| Appendix 2: NATURAL SEMANTIC METALANGUAGE (NSM) | 28 |
| Appendix 3: USER INSTRUCTIONS FOR PHASE 1 OF THE WPS | 31 |

FIGURES

| | |
|--|----|
| Figure 1: Overall architecture of the WPS | 14 |
| Figure 2: Screenshot of a prototype MNL WPS basic interface: Phase 1 | 18 |
| Figure 3: Screenshot of a prototype MNL WPS basic interface: Phase 2 | 19 |
| Figure 4: List of the most common types of knowledge schemes | 27 |
| Figure 5: Sample table of the NSM semantic primitives and categories for English | 28 |
| Figure 6: Possible NSM grammatical combinations of categories | 30 |

ACRONYMS AND ABBREVIATIONS

| Abbreviation | Name |
|----------------|--|
| IOC | Initial Operating Capacity |
| GUI | Graphical User Interface |
| AOC | Advanced Operating Capacity |
| CNR | National Research Centre |
| JRC | Joint Research Centre |
| SBA | Societal Benefit Area |
| WPS | Web Processing Service |
| CSW | Catalogue Service Web |
| WP | Work Package |
| WPS | Web Processing Service |
| NSM | Natural Semantic Meta-language |
| EEA | European Environmental Society |
| GEO | Group on Earth Observations |
| GEOSS | Global Earth Observation System of Systems |
| XMLHTTP | Extensible Markup Language Hypertext Transfer Protocol |
| OGC | Open Geospatial Consortium |
| GIS | Geographic Information Systems |
| SKOS | Simple Knowledge Organization System |
| RDF | Resource Descriptor Framework |
| OWL | Ontology Web Language |

1 INTRODUCTION

This deliverable is a progress report on the work conducted towards Task 2.5 (within Work Package 2) of the EuroGEOSS Project. The work commenced at the beginning of the project, in November 2009 (M 7), and will conclude near at the end of the project in February 2012 (M 34). The Task is being led by a team at the Centre of Geospatial Science at the University of Nottingham (Dr Kristin Stock, Ms Claudia Cialone and Dr Amir Pourabdollah), and builds on research already conducted at that institution. Contributions to the work have also been made by a number of other project partners, both in terms of the technical aspects (University of Zaragoza, EU Joint Research Centre and National Research Council, CNR, in Italy) and the thematic aspects (University of Zaragoza, Royal Society for the Protection of Birds as one of the Work Packages of the project- WP4).

This interim report describes in more detail the task that was introduced in the original project proposal, and the work towards Task 2.5 has been driven by the original outline of the task, together with the requirements of the thematic experts within the project team. The task specifically involves the development of a multilingual, natural language query interface to provide tools to assist querying both within and across the EuroGEOSS themes. This interface is being developed with a core web service that conforms to the OGC WPS specification and provides semantic querying (over the core EuroGEOSS broker) as well as multilingual and natural language capabilities, and may also be executed from thematic portals developed by other project partners. The natural language components are based on linguistic theory involving Natural Semantic Metalanguage (NSM).

The current interface takes a step aside from other valuable semantic discovery interfaces that have been developed during the course of the project by other project partners. The current interface in fact although bearing resemblance to the other discovery interfaces as per the use of ontological discovery terms for static geographic information and connections to the same semantic repository has as its ultimate goal quite a different function. Not only in fact does the current query interface use the links among the resources to identify semantically related resources, but it also constitutes an added value in expanding the user's query with their multilingual spatial relations, and in its later phases, will include natural language querying, particularly focussing around natural language spatial relations. This will assist users with a more flexible approach and refined discovery of the resources of interest (see section 4.4 below).

Task 2.5 is being carried out in three phases. The first phase involves a simple ontology term selection with discovery of semantically related resources. The second phase includes natural language expressions that express spatial relations between features represented by ontology concepts, and the final phase further expands the natural language capability, possibly with the inclusion of temporal expressions. These phases are explained in more detail in this report.

This report is structured as follows. Section 2 describes the detailed goals and objectives of the Task. Section 3 explains the task in terms of the three phases through which it is being implemented. Section 4 explains the progress so far, including details of the implementation of the discovery interface. Section 5 describes the work that will be conducted in the future, most of which focuses on the details of the natural language research work. Section 6 summarises the outputs from the task (both completed and in preparation), and Section 7 concludes the report.

Readers who are interested in a non-experts summary of the work being undertaken in Task 2.5 may refer to Appendix 3. It provides a simplified version of some of the technical and theoretical detail within this report. Conversely, readers who are interested in technical specifications of the

data models and protocols adopted during implementation may refer to the specification document that is being prepared along with the demonstrator (Deliverable 2.5.2) by the University of Nottingham.

2 GOALS & OBJECTIVES

This document reports on work undertaken and planned under task 2.5 of the EuroGEOSS project. Investigation is being carried out on the implementation of beyond-the-state-of-the-art methods in semantics and GIS technology to allow users to express spatial and geographic (dynamic or static) queries in a natural language of preference, to discover relevant resources (mainly Web Services) under the thematic areas of Forestry, Biodiversity and Drought. These methods will underpin interoperability among the three thematic areas of concern.

Currently, given the difficulty of using machines to process natural language (due to linguistic issues including vagueness in the queries of users, ambiguity in natural language expressions and cultural specificity), scientific research methods that directly address and successfully solve the linguistic question of the discovery of resources are quite hard to find. Especially in the geographic field there is no accepted and successful approach. Most of the query interfaces tend to adopt different approaches to address this issue in the simplest way such as keyword-only query.

This being a natural language-oriented task, the approach undertaken focuses on three phases of implementation reflecting different aspects in users' queries, progressively adding more functionality:

Phase 1: Multilingual selection of static geographically-related objects (identified by thematic terms selected from a vocabulary);

Phase 2: Multilingual spatial queries (natural language expressions of spatial relations, combined with the thematic terms from Phase 1);

Phase 3: Multilingual general queries (particularly spatio-temporal expressions, including spatial and temporal relations together with the thematic terms from Phase 1);

The first phase is accomplished by allowing users to select terms from existing shared (see Appendix 1)¹. Once selected, the multilingual natural language web processing service (MNL WPS) developed under this task performs dynamic identification of semantically equivalent terms.

The first phase is accomplished by allowing users to select terms from existing shared and mainly SKOS-encoded (reasons for encoding are better explained in section 4.1) ontologies (see Appendix 1)². The ontologies are stored in a semantic repository showing a SPARQL endpoint, the repository is owned by the Joint Research Centre in collaboration with the GENESIS project (more info is illustrated in section 2). Once selected, the multilingual natural language web processing service (MNL WPS) developed under this task performs dynamic discovery and retrieval of semantically equivalent terms. The number of ontologies (or knowledge schemes) included in the semantic infrastructure, is flexible. At present two main knowledge schemes have been have been

¹ The word ontology here is used in its generic connotation to indicating a semantically structured conceptualization of knowledge.

² The word ontology here is used in its generic connotation to indicating a semantically structured conceptualization of knowledge.

chosen by all the partners of the project as the core of the semantic infrastructure of EuroGEOSS.³
These are:

- the General Environmental Multilingual Thesaurus (GEMET) developed and maintained by the European Environmental Agency (EEA), and
- the Societal Benefit Areas, which is a set of related categories and subcategories, drafted by the Group on Earth Observation following the plan of a 'Global Earth Observation System of Systems' (GEOSS) whose multilingual version has been developed by the University of Nottingham with inputs from the EuroGEOSS partners.

The added value of using ontologies is represented by the shared content, and the ability to determine semantic relationships between terms within and between the ontologies.

In addition to these two generic ontologies, it was expected that the EuroGEOSS infrastructure would require the addition of further ontologies for more specialised purposes. For example, a drought vocabulary is being developed by Work Package 5 to provide specialised terms that relate to the drought theme. Thus it was necessary for an approach to be developed that would allow additional ontologies to be included as required, and for those additional ontologies to interoperate semantically to support the types of multi-thematic queries that might be useful in the EuroGEOSS context.

Currently there are two main practices involving the management of ontologies. One is represented by alignment of the ontologies (establishing useful links among ontology terms thus leaving the knowledge schemes unaltered) and the other one is represented by merging (melting the two knowledge schemes into a single scheme thus eliminating overlapping information). The approach that we have developed for EuroGEOSS is based on aligning the different knowledge schemes in a flexible and open structure with no loss of potentially useful information, but with the opportunity to explore the different ontologies/thesauri individually. The advantage in exploiting a semantic set of aligned ontologies stands in the possibility to maintain the original conceptual structure and semantics of the considered ontologies thus avoiding loss of semantic content by converging into a unique conceptualization. This approach also leaves the ontologies re-usable.

The semantic infrastructure can support queries that return not only resources that are directly annotated with the URIs corresponding to the terms that a user queries, but also other resources that are semantically related, either in the same language or another supported language.

As far as the second phase is concerned, the approach adopted involves incorporating spatial relations combined with thematic terms selected from the ontologies. These objectives are expressed by users in a restricted natural language called Natural Semantic Metalanguage⁴ (NSM -see Appendix 2) and are mapped into a formal machine-readable representation of these by an application being developed as part of this work. This is accomplished in two stages:

In the first stage, multilingual NSM expressions are analysed to ensure that they contain valid NSM grammar. For this purpose a first sample of languages are being tested: Italian, Russian and English. A grammar is being developed that is appropriate to validating NSM in the three languages, together with translations of the NSM semantic primitives. These will be used in combination to validate expression in all three languages (and hopefully will be extendable to other Indo-European languages as well).

³ Although the semantic repository under consideration also includes the ISO 19119 categorization of Spatial Data Services and an INSPIRE registry (including the INSPIRE Glossary the INSPIRE Feature Concept Dictionary) that could be potentially used as additional conceptual support.

⁴ <http://www.une.edu.au/bcss/linguistics/nsm/>,
http://en.wikipedia.org/wiki/Natural_semantic_metalanguage

In the second stage, the informal NSM-mapped result will be converted into a formal machine-readable representation to allow comparison between user geospatial objectives and web services (resources) and provide a result in response to the user query. That is, the linguistic spatial expressions (in NSM) will be converted into OGC filter spatial expressions feasible to be used in a Catalogue Service (CSW) request and will thus retrieve the appropriate web services (see section 5.1). The advantage in using the NSM is its possibility to dynamically and freely express spatial relations in different languages that are understandable to and processed by the machine too, but that do not require the user to think about space using any preconceived model. In this way, the user will be relieved from the burden of constraining his or her search to a standard technical language.

Phase 3 of the work is the possibility for the machine to process more general user objectives other than the solely spatial ones. An example could be the inclusion of temporal semantics that are quite often related to spatial semantics.

To carry out the aims within these three Phases, the University of Nottingham has been implementing a Natural Language Discovery Interface whose engine is a multilingual, natural language Web Processing Service (MNL WPS). The interface provides a multilingual Graphical User Interface, which calls the WPS, which in turn accesses semantic information from the ontologies stored in the repository hosted by the EU JRC in collaboration with the GENESIS project and registry information from the EuroGEOSS broker. This architecture also provides the option for different Graphic User interfaces to access the WPS, so that different facades can be adopted within the thematic areas' GUI portals, provided they show a map, the ontologies and the set of primitives from the NSM.

The overall technological added value is the possibility to assist end users to express their own queries with flexibility stemming from semantic universals embedded in almost all the languages of the World. The advantage lies behind the fact that the user query expressions in different languages, which are more difficult to parse, will be formalized, processed and understood by the machines more easily thus returning a higher number of results.

3 THE NATURAL LANGUAGE DISCOVERY INTERFACE

In order to accomplish Task 2.5, a web service that conforms to the OGC Web Processing Service (WPS) Specification, is being developed (Schut P., 2007). This web service is referred to in this document as the multilingual natural language WPS (MNL WPS). In combination with this, a minimal Javascript graphic user interface is also being created to demonstrate the use of the MNL WPS by a user. The user interface contains an ontology box to allow selection of ontology terms, an NSM box to allow the user to enter his or her query using NSM, and a Google map to allow selection of a geographical area. Together, these selections compose the users' query. The interface takes the user selections and constructs a request on the MNL WPS, which in turn performs requests over a semantic repository and the EuroGEOSS service Broker to discover the stored resources (see section 4.3 for the application architecture).

The WPS, endorsing the approaches undertaken by this task to manage the different aspects of user queries (multilingual queries based on simple terms; multilingual spatial queries and multilingual general queries), follows three phases of implementation. Each of these phases and relative implementation progress are described in the sections below.

3.1 Phase 1: Ontology Querying

Phase 1 of the WPS Query Interface is in its last stage of implementation and will allow the multilingual selection of resources based on ontologies shared by project members. The idea is to allow the users to select a term from a knowledge scheme (ontology/thesaurus) by browsing or keyword search, in any supported language, from the existing ontology/thesaurus loaded on screen and stored in the EuroGEOSS semantic repository. The terms selected will be then semantically extended thanks to the use of an algorithm that permits the WPS engine to access and loop through the JRC ontology repository to retrieve semantically related terms and retrieve a wider number of resources (web services) in multiple languages. The ontologies/thesauri have been previously manually aligned by EuroGEOSS project partners (University of Nottingham with contributions from the thematic experts from WP4 and WP5).

As part of this first phase of implementation a simple user query interface that allows selection of ontology terms together with a Google map search that allows users to specify a specific bounding box (a set of coordinates) in their search is provided. The latter narrows down the geographic range within which the search for a specific resource is performed (refer to appendix 3 for the WPS Interface user instructions for Phase 1).

This first phase is significant as it focuses the users' search around some shared and multilingual geographic terminology that will ease the semantic discovery of the resources tagged with metadata from the same ontological sources.

3.2 Phase 2: Natural Language Spatial Relation Querying

Phase 2 of the Query Interface implementation has just started following a period of research on cognitive-linguistic differences of spatial relations, analysis of existing linguistic theories addressing the issue, development of methods to map NSM expressions into spatial queries and actual human experimentation at Nottingham University.

This phase will complement Phase 1, which helped to discover resources through the use of ontological semantics. Phase 2 will instead assist users in the discovery of resources through natural language expressions for spatial relations in different languages⁵, thus avoiding the need to use standard spatial descriptions such as the OGC spatial operators.

The method the University of Nottingham is applying to solve this issue involves developing a natural language extended geospatial query, and uses Natural Semantic Metalanguage (NSM) primitives and their syntactic compositionality (NSM combinatory grammar). NSM is a restricted natural language that has been soundly established as the core of all studied full natural languages, work on which began in the 1960s/70s (Refer to Appendix 2 for a brief overview of NSM).

Phase 2 is divided into two stages.

Stage One of Phase 2 involves developing methods to validate the NSM expressions that users may enter into the user interface. For this purpose, work is currently being undertaken to test the English-based grammar for NSM that was formalised at the University of Nottingham to ensure that

⁵ In June, for the deliverable 2.5.2 we intend to demonstrate some of these natural language functions. This does not exclude the possibility to add other languages given the 'universal' potentialities of the metalanguage in use. However, to implement other languages a more thorough cross-linguistic grammatical analysis needs to be accomplished that might be more suitable for a phase 3 that already represents an additional goal to the achievements originally established for task 2.5 in EuroGEOSS.

it is also valid for other languages. For the purposes of EuroGEOSS, this focuses on Indo-European languages (since the project covers the European Union): specifically Italian and Russian. These languages have been selected to verify the grammar's suitability over a range of linguistic families (respectively the Germanic, Romance and Slavic branches).

Once the formalised grammar for NSM has been verified, it will be used to validate user input into the query interface. In this way, in this phase of the WPS interface, the user will be allowed to select the NSM primitives related to space (for example, NEAR, CLOSE etc) and combine them with terms selected from the ontologies and the geographical area search to create a simple natural language spatial query.

Stage Two of Phase 2 will develop and implement a formalization of the so formed NSM query to be read by the machine adopting methods from formal and modal logic. The NSM query will be compared with and converted into a OGC Filter query and embedded in a CSW request to be sent to the EuroGEOSS Broker (see section 5.1).

This second phase is essential for users while expressing spatial relations since it enhances a more dynamic and flexible search over a wider range of resources.

3.3 Phase 3: More Generalised Natural Language Querying

Phase 3 is planned to be undertaken in the latter part of 2011 and early 2012, when the EuroGEOSS project ends.

This Phase is concerned with the further semantic augmentation of the NSM query, meaning that the user can now chose among all the primes at will to express a wider range of natural language expressions to be sent as queries to the WPS. Initially, this will focus on temporal semantic primitives, as these are mostly closely related to the spatial primitives and often inter-related with them. However the development of this phase will also depend on the users' demand while formulating their queries (what they need at most to express).

Phases 2 and 3 of task 2.5 will provide a more advanced multilingual discovery of resources incorporating natural language, and contributing to the EuroGEOSS Advanced Operating Capacity (AOC).

4 PROGRESS

4.1 Thesaurus Management

For Phase 1 of the work concerning the user querying through ontology terms, an appropriate set of concepts had to be assessed and selected out of a few that covered the three thematic areas proposed by EuroGEOSS. For this task to be fulfilled three main steps were required: **1) An Inventory and Assessment** of the best knowledge schemes to be used for the project; **2) Technical Preparation** of one of these; **3) Mapping** (semantic cross-ontological alignment of the terms) of the ontologies defined.

1. **An initial inventory and assessment** of concept schemes (controlled vocabularies, thesauri, proper ontologies, controlled lists, taxonomies) was conducted taking into consideration inputs coming from the partner members. The assessment was based on a set of criteria concerning mainly:

- Technical format (RDF, OWL, SKOS);

- Linguistic coverage (able to cover at least a 1/3 of the languages spoken in Europe);
- Thematic coverage (including at least Drought, Biodiversity and Forestry as content);
- Semantic content (semantic relationships, including hierarchical or associative relations among terms in the concept scheme);
- Current use (for example if this is used for environmental purposes, cultural purposes etc.);
- GEOSS requirements (GCI consolidated search requirements demanded by the GEOSS project);
- Availability (free online or not);

As a result of the assessment two ontologies/thesauri that best met the criteria were selected out of those in the inventory. The ontologies chosen were the General Multilingual Environmental Thesaurus (GEMET) created and managed by the EEA and the Societal Benefit Area categories and subcategories created and managed by the Group of Earth Observation (GEO) for meeting the requirements of the GEOSS project.

GEMET is one of the largest environmental ontologies/thesauri in terms of environmental and linguistic coverage, including more than 6,500 descriptors in 28 different languages. It has been flexibly designed by using SKOS, OWL, RDF schema constructs. Its content, although general, is sufficient to provide a high level thematic background for the semantic structure of EuroGEOSS.

The Societal Benefit Areas are a set of 9 Environmental categories and 58 subcategories created and managed by GEO for a total of 63 areas of interest dedicated to specific general aspects of Earth Observation. This has been considered valuable mainly as it fulfils the search requirements specified in the GEOSS clearinghouse, and also because, together with GEMET it provides a high level semantic hierarchy to support the EuroGEOSS project.

However, a careful analysis of this set of categories underlined a number of principal pitfalls mainly related to its linguistic (poor coverage) and its formal (schema constructs adopted) compilation for which some technical preparation was needed prior to any practical usage of the vocabulary. Some minor adjustments still remain to be suggested to and accomplished by the authors of the set of terms.

2. As concerns **the technical preparation** of the SBAs, this involved covering in part some of the shortcomings the classification presented originally in order to accommodate the GEOSS clearinghouse search requirements followed by EuroGEOSS.

First and foremost the SBAs had a poor linguistic range (the original version was drafted only in English). Translations of the SBAs in French, Italian and Spanish were carried out by members of the project team at the University of Nottingham (UK) and thanks to the publication of these on a Wikipedia page, the University of Ljubljana (SL) contributed with a Slovenian version.

As concerns the schema constructs, the SBAs had not been originally written in any technical language (as the classification has not been conceived as an ontology). For this purpose the University of Zaragoza provided a SKOS version of the original English SBA categories and subcategories with these being classified in a hierarchical structure of broader (more general categories) and narrower (more specific subcategories) concepts. The University of Nottingham completed the SKOS file with the addition of foreign languages. However, if considered in these terms, the SKOS semantic classification of the

SBA terms is not entirely correct as this is not consistent with the SKOS logics behind broader and narrower terms. Broader and narrower terms in SKOS mean that the narrower term is a type or subclass of the broader term. For example, in the SBAs, the term 'global biogeochemistry' is classified under the category 'water' and defined as its narrow term in SKOS. However, the former is not a 'type of' water (mineral water, raw water or even river, lake depending on how to semantically consider water in its geographical or chemical connotation) rather a term addressing an associative subject that deals with water. Therefore, it is suggested that the terminology used for the SBA categories and subcategories requires further refinement into a hierarchical structure.

Suggested solutions to overcome this drawback could be multiple. One possibility is to redefine the categories and subcategories so that their relations could fit into the logical schema of SKOS, but the final responsibility for the definition of the SBAs is the remit of the GEOSS members who first created the vocabulary. A possible alternative is to re-define the SKOS (Simple Knowledge Organization System) file, whose creation has been defined by members of EuroGEOSS project, establishing not hierarchical (broader-narrower) but rather more loose or associative (related) links among its categories and subcategories. In this case, the SBAs would need to be conceived as a thesaurus with relative terms but no hierarchical structure, which imposes some limitations on semantic inference.

In the case that the former solution is adopted, the SBAs could upgrade to the level of a thesaurus. However, in this case some additional modifications to the total reformulations of the descriptors, and based on the very formal definition of the terms should be included. First, at present the SBAs as presented on the GEO portal⁶ do not seem to be completely consistent with the standard for the creation of monolingual thesauri (ISO 2788-1986). For example it is NOT recommended that capitalised initial letters be used frequently, as it actually is, for common terms (to distinguish these from specific terms when necessary), and it is NOT recommended that nouns instead of verbs be used to describe processes (eg 'prediction' is recommended instead of 'predicting'), but these are present in the SBA. Finally, if the SBAs are to be considered a thesaurus, a more careful and attentive definition of the subcategories should be pursued. These in fact present different versions and it is not clear whether they correspond to a more nested level of hierarchy or to a better definition of the descriptor (e.g., the subcategory 'Accidentals', presents another version that is 'Accidental Health and Injury').

3. The semantic **matching (mapping or alignment) of the two knowledge schemes** is a crucial part of the approach that has been undertaken. This in fact permits a more flexible and open discovery of the resources in the infrastructure, which have been tagged mostly with metadata whose definitions derive from terms (and URLs) in the two above mentioned knowledge systems (GEMET, SBAs). The flexibility of this approach derives from the possibility to add to the framework any other ontology, provided they are encoded in the technical language of implementation (all should be in SKOS to be aligned). The process of aligning terms from the GEMET and the SBAs has been accomplished manually in two steps. First, a human supervised semantic mapping between the terms in GEMET and the SBAs has been accomplished by members of WP2 and amended thanks to the contribution of thematic experts from WP3, WP4 and WP5. In a second stage, to upload these alignments and to automatically generate SKOS files for any future mapping, a tool developed at the EU-JRC, SKOSMatcher, was used.

The SKOS alignment of the ontologies through the use of the JRC SKOSMatcher tool is suitable for establishing quick manual alignments with an automatic generation of SKOS

⁶ http://www.geoportal.org/web/guest/geo_home

files (while the user makes his/her mapping). This is acceptable when dealing with two knowledge systems such as the ones selected for EuroGEOSS, but the process may not scale up when additional thesauri are added to the semantic infrastructure. For these needs, research into automatic or semi-automatic matching will be of great help (see for example section 4.3 of Euzenat, 2007). The SKOS rationale in fact does not imply an automatically derived 'transitivity' of its relations when the different knowledge systems come from different sources. The transitivity has to be defined in advance (by choosing transitive relations). For example if **A in ontology α** maintains some sort of relation with **B in ontology β** and **B** maintains some sort of relation with **C in ontology γ** , unless the relation under consideration is an exact match (the only one to be defined as transitive by the W3C), it is NOT NECESSARILY IMPLIED that **A** maintains a specific relation with **C**; the SKOS model does not say anything about this relation. The logic underneath this stems from the fact that every ontology can follow a separate path of semantic organization (the same term can mean different things in different contexts, e.g. 'branch' in Business and in Natural Sciences) so, any automatically generated semantic inference is discouraged.

However, this means that any new ontology added to the semantic framework has to be aligned to all the remaining ontologies already stored in the semantic infrastructure. This procedure might require some additional effort to be exerted by thematic experts who will establish these relations thus updating the JRC semantic repository, to make the system functional. Currently a new vocabulary is being developed by members of WP5 to cover the drought theme, which is only sparsely covered in many thesauri. It is hoped that this may be added into the framework by project partners from that WP, and used to extend the semantic framework as designed.

4.2 Implementation

The work described in this report is being implemented in the form of a web service (the MNL WPS), combined with a simple user interface to demonstrate how the web service can be used and to allow users to conduct simple natural language queries. Figure 1 illustrates the interactions between these components and the JRC semantic repository and the CNR EuroGEOSS Broker.

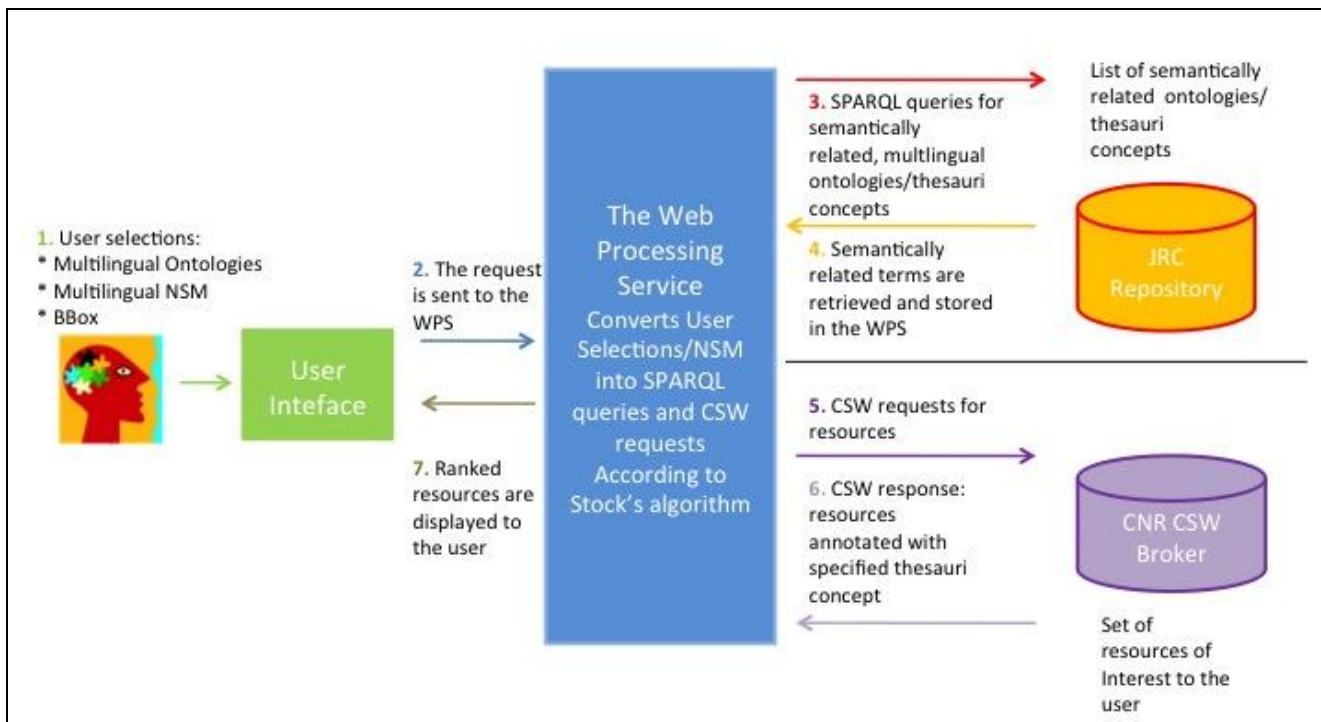


Figure 1: Overall architecture of the WPS

4.2.1 The Web Processing Service

In the natural language and semantic scenario envisioned by Task 2.5 of the project, the architecture proposed includes the implementation of a web service that conforms to the Web Processing Service Specification (Schut P., 2007). The WPS adopted is a customized version of the 52° North open source implementation. This component takes the user selections from the (in our case Javascript) Client interface that are passed in as a WPS request. Within the component, they are converted into specific queries that are performed against the JRC semantic repository and the EuroGEOSS service Broker. The result is a more flexible and open discovery of semantically related resources retrievable in a circumscribed geographical area.

The different performance steps of the WPS follow a specific semantic querying and ranking algorithm as below:

Step 1: When the user query is sent to the WPS, it splits the query and embeds the user's selections into separate requests addressed to two different endpoints: the JRC ontology repository (Steps 3 and 4 below) and to the EuroGEOSS Service Broker (Step 5 below).

Step 2: The first request to be performed by the WPS is sent to the JRC semantic repository. This hosts the semantically related thesauri concepts under the form of SKOS files (ontology files) and exposes a SPARQL endpoint. Therefore, the ontology term selected by the user in the Client interface is embedded in a SPARQL request prior to being sent to the repository. The SPARQL request allows the system to loop through all the terms (mainly the URIs of the terms) which are semantically related to the term selected by the user using the SKOS relations defined both within each thesaurus and between thesauri (in the alignment). The types of relations that are searched are consistent with the SKOS model: related match, broad match, narrow match, exact

match, close match. Once the terms (or URLs) have been retrieved, these are ranked according to their semantic distance (Step 3 below).

Step 3: The retrieved concepts are stored and ranked in the WPS. The ranking of thesauri terms (URLs) follows Kristin Stock's algorithm. This applies a semantic distance (an integer comprised between 0 and 4) to the various types of semantic relations among concepts. In this way:

- if the retrieved concepts (URLs) maintain a **exactMatch** relation with the selected concept (URL), then their semantic distance is set to 0, which means that the meaning of the two concepts corresponds totally.
- if the retrieved concepts (URLs) maintain a **closeMatch** relation with the selected concept (URL), then their semantic distance is set to 1. This means that the two concepts overlap, or that the two concepts might be used interchangeably but only in certain contexts and not indiscriminately.
- if the retrieved concepts (URLs) maintain a **relatedMatch** relation with the selected concept (URL), then their semantic distance is set to 2. This means that the two concepts maintain some sort of loose relation which is not sufficiently based on similarity to be defined as exact or close nor hierarchical to be defined as broad or narrow but on simple conceptual association.
- if the retrieved concepts (URLs) maintain a **narrowMatch** relation with the selected concept (URL), then their semantic distance is set to 3. This means that the relation between the concepts is not sufficiently similar to be defined close or exact, nor only associative to be defined related but hierarchical (for which one is the specification of the other) to be defined as narrow.
- if the retrieved concepts (URLs) maintain a **broadMatch** relation with the selected concept (URL), then their semantic distance is set to 4. This means that the relation between the concepts is not sufficiently similar to be defined close or exact, nor only associative to be defined related but hierarchical (for which one is the generalization of the other) to be defined as broad.

The choice of the distances is quite arbitrary but it is based mainly on the order of usefulness the resources should be searched for. As an example, a close or an exact match is more likely to be useful than a broad match since the latter might contain information that would not be directly related to the concepts under consideration and in any case more general. Although it is arguable whether a broad match is more or less important than a related match it is more likely that a broad match brings about concepts that have nothing to do with the original term, something that the related match does not. For example if a user looks for 'woods', broad matches to this term for example 'green lands', might help retrieving resources that address fields, steppes, paddocks, etc that are clearly not woods and have nothing to do with them, while related matches will at least be related (e.g. wood animals, trees, wood hauling etc).

The algorithm searches recursively and the semantic distance accumulates with each hop away from the originally selected term. The results are ranked according to the semantic distance.

Step 4: Once the related concepts have been stored and ranked in the WPS, a second request addressed to a CSW endpoint and interfacing the EuroGEOSS service broker searches for related resources (web services).

The connection with the broker is realized by sending a CSW request to the broker's endpoint web URL <http://aoc.eurogeoss.eu/>. As a first attempt this has been directed to the resource catalogue developed by the Drought partners (WP5) containing resources concerning drought, and forestry and biodiversity catalogues will be considered for addition at a later stage.⁷

During phase 1 of implementation of the WPS, the selection that is passed to the CSW query to be packaged and sent to the EuroGEOSS broker is composed only of the set of coordinates (the BBox) and an ontology concept as a metadata keyword's URLs selected by the users on the javascript Client GUI. This will be augmented mainly in phase 2 and in phase 3 of the project with the addition of OGC FILTER spatial operators. The request only looks for resources that have previously been tagged with selected and semantically related thesauri terms (and most important with their corresponding URLs). This means that not only will the resources, tagged with the thesaurus term selected by the user, be retrieved but also the resources tagged with the terms previously aligned with the one selected, and those semantically related thereto.

Step 5: The discovered resources are stored locally in the WPS and ranked according to the semantic distance of the terms with which they are tagged, as defined in Step 3 (based on semantic usefulness) from the original selected term.

Step 6: The answers are packaged and embedded in a response that is sent to the user. The final result for Phase 1 of the WPS is a simple page displaying all the resources titles, sorted by the most semantically close to the most general or the broadest, that are related to the user's query in a determined geographical area displaying the language of the metadata retrieved and the link of the resource behind which a GetCapabilities request is performed and an XML file retrieved (see Fig.2).

Resource List

Resources related to <http://eurogeoss.unizar.es/SBA/water>

1. Resource: [MARM SIA: Monthly Drought Index 2010](#)
Language: English
Abstract: WMS service from the Water Information System. The WMS contains: the Drought Index calculated monthly (2010), the Hydrographic Delimitations ("Demarcación Hidrográfica"), and the country towns.

Figure 2: Example of WPS list of resources response (in case one only related resource is found)

One of the main challenges the GIS technology is facing today is to cope with the multiplicity of languages (and meanings) in which most of the resources are described (by geographically distant communities) which renders the communication user-machine or machine-machine even more difficult. Therefore, to circumvent this issue, the overall processing of the information is working with a retrieval of resources based on the Universal Resource Locator (URL) behind their ontology terms through which most of the resources have been tagged. The URL in fact (as a Universal locator of a resource), differently from the simple term (this meant to be the label of a concept), is more easily linked in the infrastructure and parsed by the machines as a unique identifier that overcomes linguistic difficulties especially for static object definitions. For this reason if the

⁷ This choice has been adopted temporarily for problems of format as the Drought Catalogue is the only catalogue where resources have been tagged with URLs, thanks to Metadata Editors such as the University of Zaragoza's CatMDEdit, as required for our multi-lingual, semantic search.

metadata keyword (a label of a concept singled out from the selected thesauri) of a resource is represented in a specific language (e.g., from the SBAs 'landslides, subsidence' in English), there is quite a high probability that the resources tagged with that metadata keyword will not be found if the actual concept is searched in another language (e.g., 'deslizamientos de tierra, hundimientos' in Spanish or 'zemeljski plazovi, udori in usadi' in Slovenian), the reason being that these are treated by the machine as different strings and might not be recognized. On the other hand, the same resource and related resources will be found if the search is based on the linguistically invariable URL (since every language language expression of the term is connected to that URL).

4.3 The Multilingual WPS User Interface

A simple Javascript user interface is being developed to demonstrate the operation of the WPS, to allow project partners to see how the component may be used, to perform user testing and ongoing refinement of the approach.

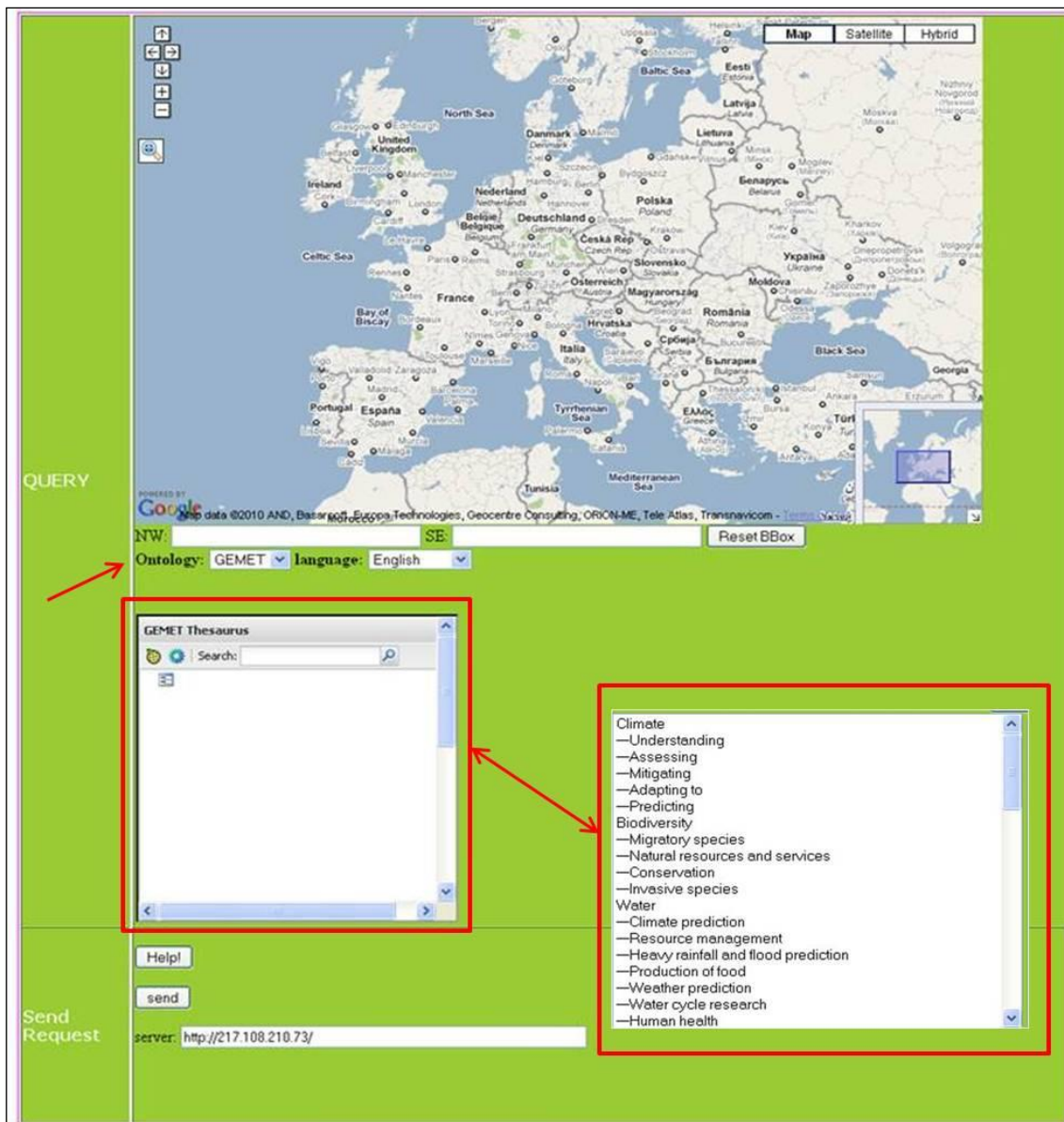


Figure 2: Screenshot of a prototype MNL WPS basic interface: Phase 1

The Client interface in phase 1 (Fig. 2) performs an XML HTTP POST request to the WPS. The request is a simple form containing the selected inputs from the users. At present the only two possible selections are: an ontology term, e.g. agriculture, selected from one of the two ontologies browsed on the Client interface (Fig. 2 in red boxes) and a coordinate bounding box. As concerns the ontology term, when the user selects a term from the list exposed in interface, the information passed to the WPS is not the term in its own right but its URL, thus overcoming linguistic constraints for the machine (to linguistically different versions of the same concept corresponds the same URL). The set of coordinates, these are embedded in a string and passed to the WPS as a string.

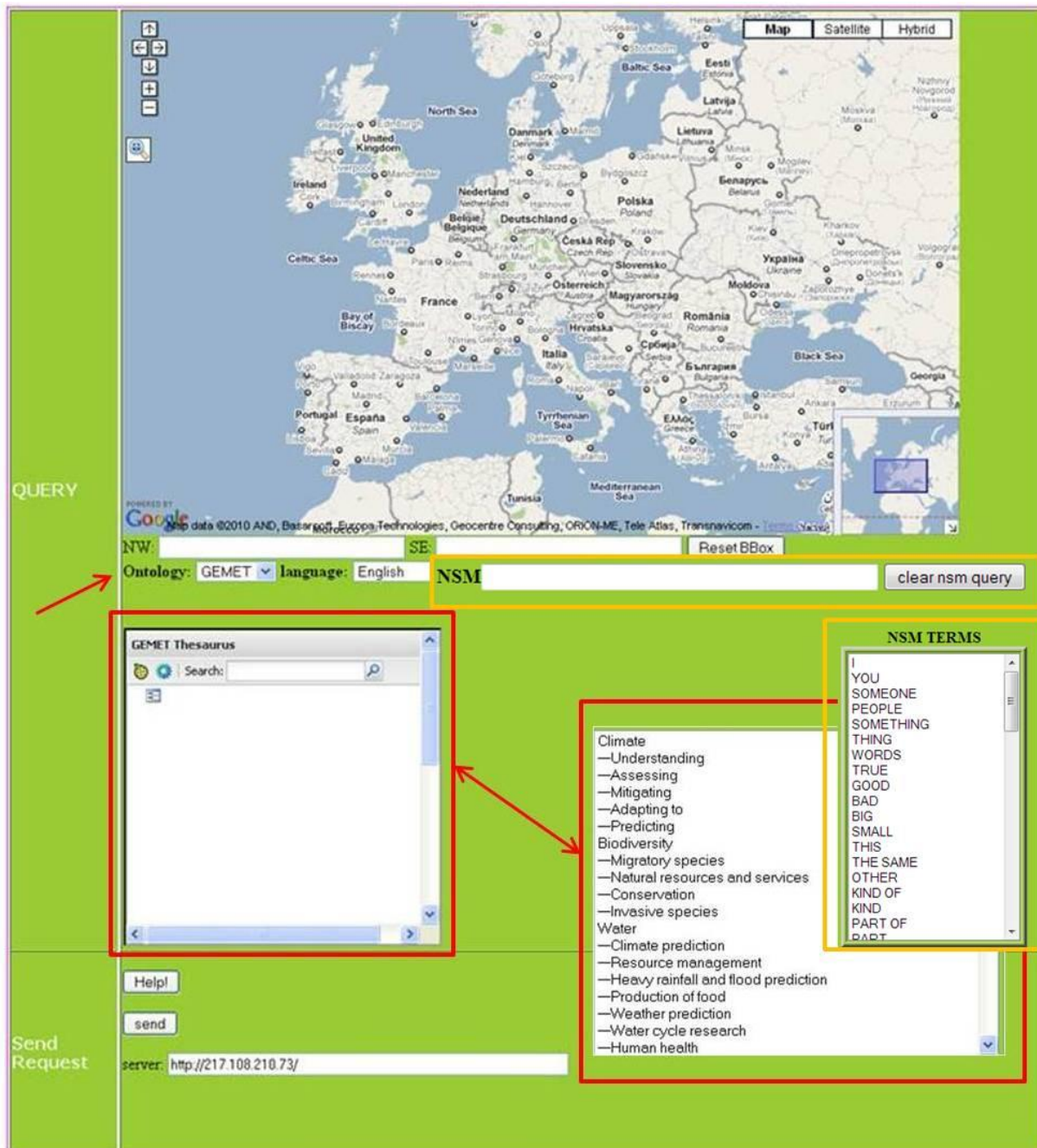


Figure 3: Screenshot of a prototype MNL WPS basic interface: Phase 2

The Client interface in phase 2 (Fig. 3 above) performs the same XML HTTP POST request to the WPS. This time the request is a more complex form including other selected inputs from the users. Alongside the bounding box and the first ontology term, e.g. agriculture from GEMET, another term, e.g. land use planning from the SBAs (Figs. 2 & 3 in red boxes) and a spatial query can be selected for the discovery of resources. This is performed by browsing the NSM box with the words (Fig. 3 in yellow higher box) or typing in inside the NSM query text area (Fig. 3 in yellow long box) the spatial query linking the two terms. The processing of the information by the WPS then is the same but a more complex implementation (an NSM engine for grammatical validation and comparison to the OGC spatial operators on the back-end).⁸

4.4 Comparison with other AOC Discovery Interfaces

Other discovery interfaces have been developed as part of the EuroGEOSS project whose algorithms are similar to the one discussed in this investigation, dealing with the retrieval of broker resources exploiting the semantic relations established in advance among ontology terms (constituting the metadata with which the resources are tagged). An example of semantic user interfaces already in use is the CNR-JRC discovery interface developed to fulfil the aims of the GEOSS AIP-3 e-Habitat.

This is quite a valuable and complementary discovery interface⁹ with an interactive mechanism aimed at providing a clear and user-friendly graphic that assists the user with a visual mapping of his or her semantic search (which in the field of linked data is a powerful strategy). The users are in fact given the possibility to navigate the thesauri concepts, stored in the EU-JRC semantic repository.

Although using a similar semantic mechanism for retrieval of resources based on ontological linked concepts, the current discovery interface does not focus mainly on graphic enhancements but it is a step forward in linguistic research in GIS. The investigation behind the described interface in fact has its focus on a natural language (multilingual) expansion of queries (hence the use of NSM) which in the geospatial field represents a difficult issue to address given the 'geo-cultural' multiplicity of spatial perceptions and the plethora of possible linguistic expressions thereof.

The work described in this report does not exclude a future synergy of the two approaches (the visual and the linguistic) to fulfil a common aim, but this is not one of the primary aims of the current investigation.

5 REMAINING WORK

The outlook for Task 2.5 concerns further natural language query research and implementation of Phases 2 and 3 of the work. Initial research related to the extension of a natural language approach to the query interface planned for Phase 2 has started, as has preliminary work on its implementation.

⁸ This is intended to be part of the deliverable/ demonstrator 2.5.2 in June 2011.

⁹ <http://www.youtube.com/watch?v=NljjmukOFqo>

5.1 Phase 2 Progress

Phase 2 will add querying based on spatial relations to the users' queries that will, together with the described geographic and ontological search, refine dynamically, through the use of natural language expressions, the discovery of web services.

In order for this to be accomplished, some essential work has to be performed both on the client side (on the javascript interface) and on the server side (within the MNL WPS). This will happen as follows.

5.1.1 Phase 2 Stage 1

Firstly, the ontology querying developed in Phase 1 will be augmented:

- The ability to select a single ontology term will be extended so that two ontology terms can be selected (those that might participate in a spatial relation).
- An NSM list box will be added to the prototype javascript interface. This will allow the end users to browse the 63 semantic primitives of the NSM (see Appendix 2) in order to formulate a more spatially-refined query. At this stage mainly the spatial NSM primes will be taken into consideration.
- An input box that can be used to construct the query combining selected NSM primitives and ontology terms will be added, together with controls to insert appropriate terms and primitives.

Secondly, a component based on NSM grammar will be added (or possibly a separate WPS) that validates the NSM queries composed by the users against some syntactic rules defined in the NSM grammar (based on the possible compositionality of the single primes since not all the possible compositions are possible and this depends on the language). Previous research has been conducted at Nottingham as per the creation and validation of a NSM grammar in English. Current research at Nottingham has started to validate and compare the same grammar over two other languages belonging to two different linguistic branches: Italian (Romance language) and Russian (Slavic language). In this way, provided the three NSM grammars are found to be similar in the syntactic compositionality of their primes, the validity of the NSM for a natural language orientated task will be justified and future uses tested perhaps over a larger set of languages.

The application will be modified to construct a request from the validated user inputs in the augmented user interface and construct a WPS request to send to the Phase 2 MNL WPS.

5.1.2 Phase 2 Stage 2

Stage 2 of Phase 2 involves extending the MNL WPS to handle the more complex request received from Stage 1 of Phase 2 (that is, including the NSM expression consisting of NSM primitives combined with ontology terms). This will involve converting potentially complex, individual NSM expressions based on language-specific spatial cognitive models into OGC FILTER spatial operators that can be included in a CSW request on the EuroGEOSS broker.

Below is the list of spatial operators (Panagiotis A. Vretanos, 2005) currently defined by the OGC Filter Encoding Specification:

- Equals

- Disjoint
- Touches
- Intersects
- Contains
- Within
- Overlaps
- Crosses
- Beyond

While these spatial operators are understood by geospatial information professionals, our experiments and theoretical research have shown that they are not intuitive for non-expert users and that they are not semantically equivalent or universally available across all languages.

For example the meaning of the simple operator BEYOND for a non-expert user, is not straight forward. This depends on the individual's perceptions of space revealed by parameters such as the size of a place and the referential frame (egocentric or absolute). In this reasoning its meaning might be understood in different ways. For example to some people the same concept is understood in Italian as 'on the other side of...something' or in Russian as 'far away' or 'behind' or more simply 'not inside' depending on the referential landmark (BEYOND a forest or BEYOND a fence). Also, some spatial operators' meanings might be quite difficult to distinguish for non-expert users who might be induced to consider for example INTERSECTS and CROSSES as the same concept.

In any case these spatial operators provide a generic as well as vague way of expressing a spatial relation, without the possibility to describe more complex relations. For this reason, many more complex spatial relations may not easily be expressed using these OGC spatial operators, except by construction of complex, multi-nested queries that are beyond the understanding of expert users and the functionality of most simple web-based GIS tools. An example is given by spatial relations such as 'surrounds', 'alongside' or 'through'. However, these might not have a correspondent translation in other languages whereas they could be expressed by means of NSM primes. For example, the concept of 'alongside' in a geospatial field is considered by one of the authors of this report as well as by an English native speaker and other random individuals taken as sample who speak a Slavic language (note that such interpretations are often individual, based on the person's background) as expressing closeness along the perimeter of a surface including the notions of laterality and dynamicity/movement (also it might not be a chance that this term when working as a spatial relation is supported by verbs of movements, e.g. 'runs alongside the river', 'flows alongside the park' and somehow imposes a visual movement all along the side of a surface of an object not on a single point). This could not be expressed by using the previous spatial operators since none of them conveys the image of moving along a surface including the two spatial meanings at once and even combining them would render things far more complicated given their vagueness.

However, if attempting to formulate it in other languages an exact match cannot be found, and the different results are far from universal in meaning:

Italian: <gemet: fiume> costeggia <gemet: foresta>
meaning 'along the limits', (expressing laterality but not dynamicity)

Or

Italian: <gemet: fiume> fiancheggia <gemet: foresta>
meaning 'along one side' (expressing laterality)

Or

Italian: <gemet: fiume> di lato a <gemet: foresta>

meaning 'at one side of' (expressing laterality but not dynamicity)

Spanish: <gemet: ríos> a lado de<gemet: bosques>

Or

Spanish: <gemet: ríos> de costado a <gemet: bosques>

meaning 'on one side' (less used, expressing only laterality)

Russian: <gemet: река > сбоку<gemet: лес>

meaning 'on one side' (laterality but not dynamicity)

Or

Russian: <gemet: река > рядом с<gemet: лес>

meaning 'close to' (expressing closeness but not really laterality or dynamicity)

However, using a simple NSM query it is possible to express movement along the perimeter of a surface in a simple and universal way that encompasses: laterality and dynamicity such as:

English: <gemet: river> MOVES ON ONE SIDE OF<gemet:forest>

Spanish: <gemet: ríos> SE MUEVE AL LADO DE <gemet: bosques>

Italian: <gemet: fiume> SI MUOVE DI LATO A <gemet: foresta>

Russian: <gemet: река> ДВИГАЕТСЯ¹⁰ НА ОДНОЙ СТОРОНЕ<gemet: лес>

The advantage lies in being able to formulate an unambiguous spatial concept that is universally recognized and understood. This means where a single word does not exist, a concept can be constructed.

Users will construct an NSM expression to reflect their own semantics of the spatial relationship between the selected concepts (to reflect the query). The work under this Phase will then develop methods to map a range of such expressions to the OGC spatial operators. A number of methods will be used to perform this mapping, including the application of theoretical work by cognitive scientists and linguists (for example, Talmy) examining the different ways in which spatial relations may be expressed and the application of inference to ensure that combinations of spatial expressions are correctly interpreted. Additionally, semantic noise must be removed to ensure that individual and cultural differences in expression that do not equate to differences in semantics do not affect the interpretation of the NSM expressions. A range of previous research undertaken at the University of Nottingham will contribute to the development of these methods, including user experiments and theoretical investigations.

The intelligent interpretation of these NSM expressions will thus be added to the functionality of the MNL WPS so that it can construct a CSW request that includes appropriate spatial operators and that is then submitted to the EuroGEOSS broker with the goal of discovering a set of resources that meet the user's original NSM expression.

5.2 Phase 3

The third phase is planned to begin in June 2011 and conclude at the end of the project. The implementation of this phase will depend on the progress and success of the previous phases.

¹⁰ ДВИЖЕТСЯ is also possible in Russian.

Phase 3 is concerned with the semantic augmentation of the NSM query. It is anticipated that the mapping developed in Phase 2 will be extended to handle additional types of NSM expressions, including those with temporal semantic primitives. The following is an example of such a query:

English:

<gemet: riversideVegetation> LIVE NEAR A PLACE WHERE **<SBA: migratorySpecies>** EXISTED A LONG TIME BEFORE.

NOW THIS **<SBA: migratorySpecies>** NEAR THIS SAME PLACE DO NOT EXIST ANYMORE.

The approach developed by this work ensures that these types of expression will also be supported multilingually. That is, the existence of a multilingual grammar and translations of the NSM semantic primitives into three languages (at this stage), ensure that when the work is extended in one language, it is easily transferable to other languages. Some examples of the expression above in different languages follow:

Spanish:

<gemet: vegetaciónDeOrilla> VIVE CERCA DE UN SITIO DÓNDE
<SBA: speciesMigratorias > EXISTIERON MUCHO TIEMPO ANTES (DE AHORA).

AHORA ESTE **<SBA: speciesMigratorias >** YA NO EXISTE CERCA DE ESTE MISMO SITIO

Italian:

<gemet: vegetazioneRiparia> VIVE VICINO AD UN POSTO DOVE
<SBA: specieMigratorie> ESISTEVANO MOLTO TEMPO PRIMA.

ADESSO QUESTE **<SBA:SpecieMigratorie>** VICINO QUESTO STESSO POSTO NON ESISTONO PIÙ.

Russian:

<gemet: береговаяРастительность> ЖИВЁТ РЯДОМ С МЕСТОМ ГДЕ
<SBA: мигрирующиеВиды>¹¹ СУЩЕСТВОВАЛИ ДОЛГО ДО ЭТОГО.

СЕЙЧАС ЭТИ **<SBA:мигрирующиеВиды>** В ЭТОМ САМЫМ МЕСТЕ НИКОГДА НЕ СУЩЕСТВУЮТ.

In this more complex example, a user might have wanted to find out not all the names of all the examples of flora growing close to rivers in a certain place, but the riverside vegetation that is present in a geographical area (selected through a map as in phases 1 and 2) that used to be populated by extinct migratory species. This query would be quite difficult to express and parse in different languages by machines since every language has its own peculiar way of expressing spatiality and temporality. On the other hand, NSM is universal. The refined natural language specification of the query will hopefully help any user to formulate requests with any number of combinations possible with the NSM primes.

¹¹ The SBAs do not provide at the moment a Russian version but the term does exist in Russian and a Russian version of this vocabulary could be accomplished easily.

6 Output and dissemination

As concerns the outputs produced by Task 2.5, the major contributions thus far include the alignment of thesauri terms (SBA-GEMET and INSPIRE spatial Data Themes) in the EuroGEOSS infrastructure and the translation of the monolingual (English only) version of the GEOSS SBAs in Italian Spanish and French to accommodate easy multilingual discovery for international users.¹² Also, a number of papers for dissemination in journals and at international conferences have been and are being prepared. Finally, a demonstrator is being developed as the second major deliverable of the work, and will be delivered in June 2011. It is also anticipated that a video will be created to show how this demonstrator works.

Papers in preparation:

'Multilingual Knowledge Systems. The EuroGEOSS' Case Study:GEOSS' Societal Benefit Areas Translations for Italian, Spanish and French' addressing the issue of cross translation of thesauri.

'An Approach to the Management of Multiple Aligned Ontologies for a Geospatial Earth Observation System' (in preparation for journal publication). This paper explains the overall approach of managing the semantic framework through alignment of resources.

'Demonstrator on natural language Discovery and Query Interface': creation of a set of instructions for the user interface.

A paper describing the inputs, outputs and technical functionality of the MNL WPS (possibly for submission to the OGC).

Papers and projects already disseminated:

Alignment of GEMET and the GEOSS SBAs.

Translation of the SBAs into French, Spanish and Italian (web dissemination through Wikipedia).¹³

Internal report on 'Domain Ontologies to Support EuroGEOSS'. An inventory of semantic schemes out of a sample selected by experts to be inserted and aligned in the EuroGEOSS infrastructure.

Co-authored paper with Javier Noguera (University of Zaragoza), 'WP 2 SECTION: User-Driven Requirements for Resource Annotation'. A user instruction paper for the users based on how to use the University of Zaragoza's metadata Editor, CatMDEdit, to annotate resources and the EU-JRC SKOS Matcher tool to align thesauri.

'The Semantic Management of Environmental Resources within the Interoperable Context of EuroGEOSS: Alignment of GEMET and the GEOSS SBAs'. EGU meeting in Vienna (02-07 May 2010). Presentation to the EGU of the ontological approach undertaken at CGS on behalf of EuroGEOSS.

¹² Several inputs from thematic experts and professionals have been received for the improvement of both semantic mapping of the thesauri and the SBAs translations.

¹³ http://en.wikipedia.org/wiki/Societal_Benefit_Areas

7 CONCLUSIONS

This report has described the work being undertaken under Task 2.5 of the EuroGEOSS project, including specific details of the work completed so far, and the work planned for the remainder of the project. The Task is aimed at providing a multilingual natural language querying interface, and builds on research already conducted at the University of Nottingham, as well as taking advantage of components developed by a number of other project partners, and thematic inputs from Work Packages 3, 4 and 5.

The report has demonstrated that work towards the final goal of a multilingual, natural language query interface is well underway, and progress towards the demonstrator expected in June 2011 is continuing. The majority of the work so far has focused on providing a semantic framework for later development, both in terms of preparation and management of the thematic thesauri, and the development of the basic multilingual capability of the query interface and accompanying MNL WPS. Work in the remainder of the project will focus on the complex, natural language aspects that will provide additional functionality for users, and this will extend the research work already begun in that area.

As well as conducting the work, we have focused on creating a series of outputs and dissemination products that provide additional technical and theoretical details of the work, and more of these are expected as the project progresses. Furthermore, the demonstrator will be available to members of the project team to test and evaluate as development progresses.

REFERENCES

- Cialone C., Stock K., 'Domain Ontologies to Support EuroGEOSS', internal report (final).
- Cialone C., Stock K., 'Multilingual Knowledge Systems. The EuroGEOSS' Case Study:GEOSS Societal Benefit Areas Translations for Italian, Spanish and French', (forthcoming 2011).
- Goddard C., *Cross-Linguistic Semantics*. John Benjamins Publishing Company. Amsterdam/Philadelphia, 2008.
- Gomez-Perez A., Corcho-Garcia O., Fernandez-Lopez M., *Ontological Engineering: With Examples from the areas of knowledge Management, e-commerce and the Semantic Web*, Springer-Verlag London Limited, 2004.
- Euzenat J., Shvaiko P., *Ontology matching*, Springer-Verlag Berlin Heidelberg, 2007.
- Nebert D., Whiteside A., Panagiotis (Peter) Vretanos. 'OpenGIS® Catalogue Services Specification', OGC 07-006r1, 2007, version 2.0.2
- Panagiotis A. Vretanos, 'Open GIS® Filter Encoding Implementation Specification', OGC 04-095, 2005.
- Peeters B., *Semantic Primes and Universal Grammar: Empirical Evidence from the Romance Languages*. John Benjamins Publishing Company. 2006.
- Schut P. 'OpenGIS® Web Processing Service', OGC 05-007r7, 2007, version 1.0.0

Stock K., Cialone C., 'An Approach to the Management of Multiple Aligned Ontologies for a Geospatial Earth Observation System' (forthcoming 2011).

Talmy L., *Toward a Cognitive Semantics*, volumes I and II, MIT Press, 2000, <http://linguistics.buffalo.edu/people/faculty/talmy/talmy.html>

Wierzbicka A., Goddard C., *Meaning and Universal Grammar: Theory and Empirical Foundings*, vol. I,II. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2002.

Appendix 1: SEMANTIC KNOWLEDGE SCHEMES

| Knowledge Scheme | Definition |
|-----------------------|---|
| Controlled vocabulary | A finite list of concepts, usually without definitions or relationships between the concepts. |
| Data dictionary | A finite list of concepts, usually with definitions but without relationships between the concepts. |
| Taxonomy | A hierarchical classification of concepts, which means a set of generalisations and specialisations of something (e.g., a concept, an animal species and so forth). |
| Thesaurus | A set of defined concepts including additional semantic relationships between concepts (i.e. synonyms, antonyms, hyperonyms, hyponyms, etc.). |
| Ontology | Could be considered as a semantic extension at a large scale of the previously described knowledge schemes. This is a formal, semantically and hierarchically structured collection of concepts reflecting the shared conceptualisation of a community. It can include formal constraints on the ways in which concepts may relate to each other. |

Figure 4: List of the most common types of knowledge schemes

Appendix 2: NATURAL SEMANTIC METALANGUAGE (NSM)

| Category | | Semantic Primitives | |
|-----------------------------|--------------|---|---|
| Substantives | | I, YOU, SOMEONE, PEOPLE, SOMETHING/THING, BODY, WORDS, TRUE | |
| Modifiers | Attributes | GOOD, BAD, BIG, SMALL | |
| | Specifiers | Determiners | THIS, THE SAME, OTHER, KIND (OF), PART (OF) |
| | Quantifiers | ONE, TWO, SOME, ALL, MUCH/MANY, MORE/ANY MORE | |
| Substantive-likes | | KIND, PART, WHERE/PLACE, WHEN/TIME, HERE, NOW, MORE/ANY MORE | |
| Predicates | | THINK, KNOW, WANT, FEEL, SEE, HEAR, SAY, LIVE, DIE, DO, HAPPEN, MOVE, TOUCH, BE(SOMEWHERE), THERE IS/EXIST, HAVE/BELONG, BE(SOMEONE/SOMETHING), WHERE/PLACE, HERE, ABOVE, BELOW, SIDE, INSIDE | |
| Substantive Adjuncts | Location | WHERE/PLACE, HERE, ABOVE, BELOW, SIDE, INSIDE | |
| | Time | WHEN/TIME, NOW, BEFORE, AFTER, FOR SOME TIME, MOMENT | |
| Predicate Adjuncts | | NOT, CAN, LIKE | |
| Intensifiers | | VERY | |
| Scalar Predicates | | FAR, NEAR, LIKE | |
| Scalar Substantive Adjuncts | Location | FAR, NEAR | |
| | Time | A LONG TIME, A SHORT TIME | |
| Clause modifiers | Bi-clausal | IF, BECAUSE | |
| | Mono-clausal | MAYBE | |

Figure 5: Sample table of the NSM semantic primitives and categories for English.

The NSM primes

The NSM is a linguistic set of 63 Universal semantic primitives that has been identified, analysed and tested over a very large number of foreign languages since the late '60s under a school of thought commenced and carried on by the acknowledged linguist/semanticist Anna Wierzbicka.

This means that every language possesses a set of NSM (so there is an English version a French version, a Chinese version, a Yimitirr version etc. of this same NSM) that is composed of the same set of concepts (with some minor exceptions).

Each and every prime can belong to more than one category thus assuming a different linguistic meaning of use.

The reasons behind the use of the NSM are multiple:

- After a period of experimentation at the University of Nottingham, the **NSM primitives** have been found to be very **useful in the human description of spatial and temporal issues**;
- **NSM is a very basic and easy language** that could be used to describe in simple terms the language we use every day (meta-language) with the advantage of overcoming the socio-cultural semantic complexities that plain languages present.
- **NSM** has a long linguistic theoretical and empirical foundation. This means it **has been designed so as to reflect human thinking**. Therefore it is **user-oriented** but also easily processed by the machine for its simplicity.
- **NSM**, being it a basic semantic set of concepts present in almost all the languages in the world, is **suitably cross-translatable**.

The NSM grammar

NSM theory also includes a grammar that stemming from these primes would permit the reiterative formulation (from the very basic to the very complex) of linguistic expressions. This grammar is applicable across the languages examined and so is suitable for use in a multilingual context too. The NSM grammar is delineated, keeping an eye on the constraints each language imposes, from the combinatory character of its primitives (for example, THIS BIG THING IS HERE FOR SOME TIME etc..). An example of the combinatory possibilities of the NSM categories is expressed below (developed by our previous at the University of Nottingham):

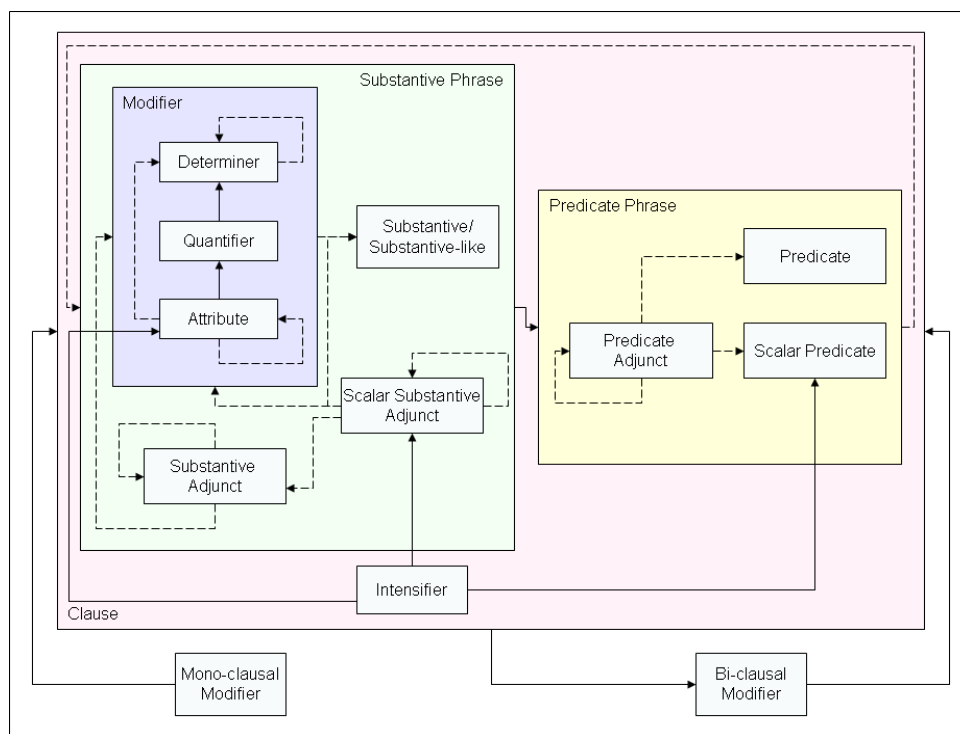


Figure 6: Possible NSM grammatical combinations of categories

- indicate that the connection is mandatory from the point of view of the category at the beginning of the arrow.
- - -→ indicates the connection.


Appendix 3: USER INSTRUCTIONS FOR PHASE 1 OF THE WPS

Below an informative set of slides explaining the work so far accomplished by task 2.5, of WP2 for EuroGEOSS.¹⁴

| | |
|--|--|
|  <p>EuroGEOSS WP2, task 2.5 Natural Language and Query Interface</p> <p>Kristin Stock & Claudia Cialone Centre for Geospatial Science University of Nottingham, UK</p> |  <p>1. EUROGEOSS TASK 2.5</p> <p>As part of the EuroGEOSS proposal.. one of the tasks (2.5) of WP2 is responsible for:</p> <ul style="list-style-type: none"> • Allowing users (even non-experts) to express their own objectives (geographical, geospatial, etc...) <p>In order to:</p> <ul style="list-style-type: none"> • Discover resources within the EuroGEOSS infrastructure and linked to the outside Web 2.0 <p>By means of:</p> <ul style="list-style-type: none"> A. Simple thesauri querying B. Semantics and Natural Language querying |
|  <p>A. SIMPLE THESAURI QUERYING</p> <p>Allows the user to select a term from the core ontologies GEOSS SBA (Societal Benefit Areas) categories and subcategories and GEMET, or other ontologies aligned later by the users.</p> <p>These ontologies would provide a source of shared and static geographic terms to be used for metadata annotation of Resources (Web Services).</p> <p>B. NATURAL LANGUAGE QUERYING</p> <p>Augments semantically the ontology query allowing the users to define spatial queries using the Natural Semantic Meta-language (or NSM)...</p> |  <p>2. HOW WILL WE MEET THIS REQUIREMENT?</p> <p>A. Allowing discovery of resources with multilingual thesauri (GEMET and the GEOSS SBA categories and subcategories)</p> <ul style="list-style-type: none"> ○ USING an Ontology Multilingual Query Interface (phase 1) ... |
|  <p>A. THE MULTILINGUAL QUERY INTERFACE (ONTOLOGY QUERYING, PHASE 1) MNL WPS Query Service</p> <p>Screen dumps of the multilingual query interface showing the two ontologies.</p>  |  <p>B. Supporting Multilingual Natural Language Spatial Queries over SDI resources to augment the simple ontology discovery from Phase 1</p> <ul style="list-style-type: none"> ○ EXTENDING the Ontology Multilingual Query Interface to include Natural Semantic Metalanguage (NSM) open geospatial queries (phase 2) <p>Example query-</p> <p><code>THIS <SBA:pollutionEvents> EXISTS INSIDE THIS<gemet:forest></code></p> |

¹⁴ The slides above will be available in ppt on the EuroGEOSS website shortly.

D2.5.1: Report on Natural Language Discovery and Query Interface




C. Supporting Multilingual Natural Language Spatial Queries over SDI resources to augment the simple ontology discovery from Phase 1

- o AUGMENTING the semantic search of phase 2 with additional functions, e.g. temporality (probable phase 3).


Example query-

THIS <gemet:EndemicSpecies> LIVED IN THIS <SBA:Ecosystems> FOR A LONG TIME
NOW THIS <gemet:EndemicSpecies> DIES IN THE SAME <SBA:Ecosystems> .




B. WHAT IS THE NATURAL SEMANTIC META-LANGUAGE ?



- o NSM is a 'mini'-Natural Language of semantic universals empirically analysed by linguists such as A. Wierzbicka since the '60s,
- o NSM includes 63 semantic primitives (atomic semantic units) present in all the languages of the world studied so far (eg. Chinese, English, Suomi, Russian, Turkish, Maori, Lao, Malay, Mangaaba-Mbula etc.);
- o NSM is simple in units and syntax and intuitive for non-experts, but mostly it is easy for a machine to parse and process;
- o NSM is suitable to describe objectives, actions, processes, spatio-temporal relationships;

B. WHAT ARE THE SEMANTIC PRIMITIVES?

| Category | Semantic Primitives | | | | | | | | |
|--|--|-------------|---|---|---|-------------------|-------------|--|--|
| Substantives | I, YOU, SOMEONE, PEOPLE, SOMETHING/THING, BODY, WORDS, TRUE | | | | | | | | |
| Modifiers | GOOD, BAD, BIG, SMALL | | | | | | | | |
| <table border="1"> <tr> <th>Specifiers</th> <th>Determiners</th> </tr> <tr> <td>THIS, THE SAME, OTHER, KIND (OF), PART (OF)</td> <td>ONE, TWO, SOME, ALL, MUCH/MANY, MORE/ANY MORE</td> </tr> </table> | Specifiers | Determiners | THIS, THE SAME, OTHER, KIND (OF), PART (OF) | ONE, TWO, SOME, ALL, MUCH/MANY, MORE/ANY MORE | <table border="1"> <tr> <th>Substantive-likes</th> <th>Quantifiers</th> </tr> <tr> <td>KIND, PART, WHERE/PLACE, WHEN/TIME, HERE, NOW, MORE/ANY MORE</td> <td>THINK, KNOW, WANT, FEEL, SEE, HEAR, SAY, LIVE, DIE, DO, HAPPEN, MOVE, TOUCH, BE(SOMEWHERE/HERE), THERE IS/EXIST, HAVE/BELONG, BE(SOMEONE/SOMETHING), WHERE/PLACE, HERE, ABOVE, BELOW, SIDE, INSIDE</td> </tr> </table> | Substantive-likes | Quantifiers | KIND, PART, WHERE/PLACE, WHEN/TIME, HERE, NOW, MORE/ANY MORE | THINK, KNOW, WANT, FEEL, SEE, HEAR, SAY, LIVE, DIE, DO, HAPPEN, MOVE, TOUCH, BE(SOMEWHERE/HERE), THERE IS/EXIST, HAVE/BELONG, BE(SOMEONE/SOMETHING), WHERE/PLACE, HERE, ABOVE, BELOW, SIDE, INSIDE |
| Specifiers | Determiners | | | | | | | | |
| THIS, THE SAME, OTHER, KIND (OF), PART (OF) | ONE, TWO, SOME, ALL, MUCH/MANY, MORE/ANY MORE | | | | | | | | |
| Substantive-likes | Quantifiers | | | | | | | | |
| KIND, PART, WHERE/PLACE, WHEN/TIME, HERE, NOW, MORE/ANY MORE | THINK, KNOW, WANT, FEEL, SEE, HEAR, SAY, LIVE, DIE, DO, HAPPEN, MOVE, TOUCH, BE(SOMEWHERE/HERE), THERE IS/EXIST, HAVE/BELONG, BE(SOMEONE/SOMETHING), WHERE/PLACE, HERE, ABOVE, BELOW, SIDE, INSIDE | | | | | | | | |
| Predicates | WHERE/PLACE, HERE, ABOVE, BELOW, SIDE, INSIDE | | | | | | | | |
| Substantive Adjuncts | WHERE/TIME, NOW, BEFORE, AFTER, FOR SOME TIME, MOMENT | | | | | | | | |
| Predicate Adjuncts | NOT, CAN, LIKE | | | | | | | | |
| Intensifiers | VERY | | | | | | | | |
| Scalar Predicates | FAIR, NEAR, LIKE | | | | | | | | |
| Scalar Substantive Adjuncts | FAIR, NEAR | | | | | | | | |
| Location | A LONG TIME, A SHORT TIME | | | | | | | | |
| Time | IF, BECAUSE | | | | | | | | |
| Clause modifiers | Bi-clausal Mono-clausal MAYBE | | | | | | | | |

Atable with the English semantic primitives as created by Wierzbicka and Goddard and refined by Kristin Stock.






B. WHAT ARE THE SEMANTIC 'SPATIAL' PRIMITIVES?

Out of the group of the 63, some primitives can be singled out as spatial primitives, some examples are:

NEAR, FAR, HERE, WHERE, PLACE, ABOVE/BELOW, ON, SIDE, INSIDE but also the verbs MOVE and TOUCH could express spatial interaction between objects (<gemet: river> TOUCHES <gemet: forest>, that expresses intersection and cross over).

But what is interesting and useful is their capability to combine with other primitives to express more complex spatial concepts.






B. WHY USE NSM FOR GEOSPATIAL QUERY?

NSM is the only system of semantic analysis based on empirically tested linguistic realities valid universally.

NSM is simple for both users and for machines to process.

NSM constitutes the semantic core on which the complex linguistic expressions of our everyday speech build.

B. WHY USE NSM FOR GEOSPATIAL QUERY?

NSM is the only system of semantic analysis based on empirically tested linguistic realities valid universally.

NSM is simple for both users and for machines to process.

NSM constitutes the semantic core on which the complex linguistic expressions of our everyday speech build.

