

International Journal of Spatial Data Infrastructures Research, 2012, Vol.7, 225-248.

The Development and Interlinkage of a Drought Vocabulary in the EuroGEOSS Interoperable Catalogue Infrastructure*

Miguel Ángel Latre¹, Barbara Hofer², Javier Lacasta¹, Javier Nogueras-Iso¹

¹Department of Computer Science and Systems Engineering, Universidad de Zaragoza, Spain, {latre, jlacasta, jnog}@unizar.es

²Institute for Environment and Sustainability, Joint Research Centre, European Commission, Italy, barbara.hofer@jrc.ec.europa.eu

Abstract

Metadata catalogues are used for facilitating the discovery of data and web services in, e.g., growing collections of Earth observation resources. Two conditions need to be met in order to successfully retrieve resources in catalogues: the metadata describing resources have to be complete and accurate and the keywords used in searches semantically related to the keywords contained in the metadata descriptions. One method to increase the rate of successfully retrieved metadata in catalogues is the use of controlled vocabularies. Such vocabularies can be used for annotating metadata with appropriate keywords and then also presented to users of the catalogue for specifying search terms. In the process of preparing metadata for drought-related data and services within the EuroGEOSS project, the need of a drought-specific vocabulary arose. This paper presents this drought vocabulary, the methodology followed for its development, its integration in the EuroGEOSS drought infrastructure and discusses its usefulness for the drought thematic area. The usefulness of the vocabulary is hereby measured by an increased use of search terms coming from an appropriate vocabulary and by an increase in the successful retrieval of resources. In particular, metadata must be annotated with appropriate keywords from a controlled vocabulary, thesaurus or ontology suitable for that particular field.

Keywords: Metadata, annotation, vocabularies, appropriateness, drought, GEOSS, INSPIRE

*This work is licensed under the Creative Commons Attribution-Non commercial Works 3.0 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.

DOI: 10.2902/1725-0463.2012.07.art12

1. INTRODUCTION

Nowadays environmental scientists are challenged with the derivation of new insights from connecting the increasingly available data and services coming from Earth observation. The scientists' tasks start from the discovery of new data sources of interest in a certain subject area.

One system that acts as entry point for the discovery of data and services is the Global Earth Observation System of Systems (GEOSS)¹. GEOSS is an effort to bring together data and services from nine societal benefit areas: disasters, health, energy, climate, water, weather, ecosystems, agriculture and biodiversity. The main objective of the GEOSS initiative is to support decision making in the specific thematic areas by providing access to data and tools required for this task.

The bringing together of data and services from different disciplines on a large scale bears semantic and technological challenges. These challenges include the development of search engines for providing the user with well-founded search results, tools for viewing available web services, tools for performing data analysis across disciplines, and data harmonisation issues. Some of these challenges are approached in the EuroGEOSS project², which is a 7th Framework Program project of the European Commission working on a European approach to GEOSS. The EuroGEOSS project focuses on the three thematic areas of drought, biodiversity and forest with the objective to build interoperable infrastructures for each of the disciplines as well as an infrastructure supporting multi-disciplinary interoperability across disciplines.

This paper is based on the work done in the thematic area of drought. The objectives of the drought working group are to connect drought-related resources on different spatial scales in an interoperable infrastructure; this infrastructure is called European Drought Observatory (EDO)³. The main elements of EDO are a drought metadata catalogue⁴ for the discovery of drought-related data and services and a map viewer for visualising drought indices (drought indices are maps and time series graphs that represent the distribution of a certain quantitative measure related to the severity of drought conditions). EDO contains drought indices for the whole Europe and indices coming from national and regional sources that provide more detailed information. The integration of drought indices from different sources in the interoperable infrastructure is based on open web services of the Open Geospatial Consortium (OGC).

¹ <http://www.earthobservations.org/>

² <http://www.eurogeoss.eu/>

³ <http://edo.jrc.ec.europa.eu/>

⁴ <http://eurogeoss.unizar.es/Search/>

Following the proposal of GEOSS and INSPIRE (European Commission, 2007) the discovery of information is based on searching through metadata descriptions. The drought team built a metadata catalogue that is tailored towards the needs of experts from the drought community. One of the key fields of the metadata describing resources is the field 'keywords' that facilitates the discovery of a resource of interest.

In the process of preparing metadata for drought-related data and services, it turned out that the proposed vocabularies within the EuroGEOSS project for the annotation of metadata did not comply with the needs of the drought community, because they were highly generic to allow for a proper classification of the resources or too large to be a practical tool for annotation.

It was decided to prepare a specialised drought vocabulary to improve the discovery of drought-related data and services in an interoperable infrastructure and to facilitate the task of metadata annotation. The resulting vocabulary, developed through an open and collaborative process, contains 103 concepts organised hierarchically in groups (drought, meteorology, soil, hydrology) and provides preferred and alternative labels in fifteen languages (Latre et al, 2011).

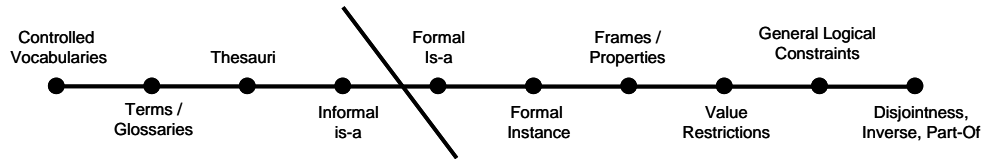
The objective of this work is to describe the methodology followed in the development of this vocabulary and to demonstrate the improvement of search results after the introduction of the drought vocabulary (Lacasta et al, 2007).

The rest of the paper is structured as follows. Section 2 reviews the state of the art in thesaurus related to drought and in thesaurus creation methodologies. Section 3 presents the drought vocabulary which was developed for a detailed annotation of metadata of drought related data and services, focusing on the methodology based on which the vocabulary was derived. The integration of the vocabulary in the existing interoperable infrastructure is presented in section 4 and the appropriateness and usefulness of the established vocabulary is discussed in section 5. The paper finishes with a conclusions and future work section.

2. BACKGROUND AND RELATED WORK

A challenge in information retrieval from metadata catalogues is the provision of search results that are semantically related to the search terms. One approach to meet this challenge is the usage of either controlled vocabularies, glossaries, taxonomies, thesauri or ontologies. These different *knowledge organization systems* (Hodge, 2000) represent different levels of expressiveness. These differences allow a classification of them in what Lassila and McGuinness (2001) called an *ontology spectrum* (see Figure 1).

Figure 1: Ontology Spectrum



Source: Lassila and McGuinness (2001)

An ontology (right part of the spectrum represented in Figure 1) is usually defined as “a formal, explicit specification of a shared conceptualization” (Gruber, 1993) and can be considered composed of a set of concepts that refer to the things of interest in a given domain and some specification of meaning for the concepts by axioms and definitions (Uchold and Gruninger, 2004). When the modelled concepts are just terms, not fully specified by axioms and definitions; and just relationships among these terms (subtype/supertype, part/whole, synonym or relation) are made explicitly, these representations are usually referred to as thesauri or terminological ontologies (represented in the left part of Figure 1) (Lacasta et al, 2010; ISO, 1986; Sowa, 1996). In the case of multilingual thesauri, both terms and relationships are represented in more than one language. When applied to the search of resources, these multilingual thesauri allow the retrieval of resources that may not directly contain the search term among the annotation terms, but a term that is related to the search terms. This can be done by searching not only for the queried term, but also for the terms hierarchically dependant on it, and by the different translations the term may have (Latre et al, 2009). This has the advantage that the user retrieves a richer list of results from his search.

2.1. Review of Thematic Thesauri Related to drought

Two main issues were identified in the review of thesauri from fields close to drought and of thesauri proposed for metadata annotation in the EuroGEOSS project: they are either highly generic and contain only few terms related to drought or they are so extensive that their use for annotation of resources is impractical.

Thematic thesauri in fields close to drought, such as hydrography, hydrology and meteorology, are generally comprehensive; however, the amount of terms related to drought is limited. The Glossary of Meteorology⁵ of the American Meteorological Society (2000) contains more than 12 000 terms related to meteorology and only a few are related to drought. The International Glossary of

⁵ <http://amsglossary.allenpress.com/glossary>

Hydrology (UNESCO, 1993), which counts with an experimental web version⁶, is available in 11 languages and consists of more than 300 water-related terms, but few of them related to drought. The CUAHSI Water Ontology has the purpose of supporting the discovery of time-series data collected at a fixed point, including physical, chemical, and biological measurements. Again, with more than 5 000 terms, most of them not specific to drought, it is not practical for drought resources annotation and search. Extending the scope of the thesauri does not provide any improvement: AGROVOC⁷, covering subject fields in agriculture, forestry and fisheries, and considered sometimes as a general-themed thesaurus, contains close to 40 000 concepts, but only a dozen of them are drought-related.

General purpose thesauri, such as the thesauri proposed in the EuroGEOSS project for the annotation of metadata (INSPIRE topic categories, GEOSS Societal Benefit Areas categorisation and the General Multilingual Environmental Thesaurus, GEMET) are not appropriate too for its use in the drought field. Below an illustration of the issues related to the reuse of these thesauri:

- They can be highly generic vocabularies: INSPIRE topic categories (European Commission, 2008) and Societal Benefit Areas categorisation⁸ allow a classification of data into general subject areas like 'climatologyMeteorologyAtmosphere' from the INSPIRE topic categories. These categories are too general to establish useful search restrictions by expert drought users when discovering data in a catalogue.
- They can be large collections of terms: GEMET⁹ is a thesaurus containing around 5,244 terms. It is designed to cover a wide range of topic areas and the large amount of terms makes the selection of the right keywords for metadata annotation or restricting a discovery query tedious and cumbersome.

Since the need for a thesaurus for metadata annotation and improvement of searches in the drought metadata catalogue had been identified, this review of existing thesauri led to the preparation of a specialised vocabulary on droughts.

2.2. State of Art in the Process of Construction of Thesauri

The construction of a thesaurus is a complex process in which the terminology used in a knowledge domain is collected, analysed and linked together into a model that can be used for classification of resources in the domain. Along the

⁶ <http://webworld.unesco.org/water/ihp/db/glossary/glu/aglo.htm>

⁷ <http://aims.fao.org/website/AGROVOC-Thesaurus/sub>

⁸ http://en.wikipedia.org/wiki/Societal_Benefit_Areas

⁹ <http://www.eionet.europa.eu/gemet>

years, with the objective of improving the quality of the created models, different thesaurus construction methodologies have been developed. In this field, different standards have been created to normalise the structure and properties of monolingual and multilingual thesauri (ANSI/NISO, 2005; ISO, 2011; BSI, 2007). These standards do not propose a detailed construction methodology, but they describe the general idea of the most common processes used for thesauri construction. In general four steps are usually required:

- A review of similar existent thesauri. This is needed to avoid the creation of a new thesaurus from zero if an existent one can be valid or adapted.
- A modelling stage where the desired structure, format and final display are selected.
- A term selection stage where the set of possible terms to include in the thesaurus are selected and related.
- A validation step in which the candidate terms are reviewed to select only those that fulfil the standards specifications.

Depending on the specific methodology used, each one of these steps can be performed in a different way. For example, the term selection stage can be performed by a committee generating a corpus of terms or they can be extracted from the domain (e.g., other existent knowledge models). And in each of these cases different approaches can be used. In the first case, the corpus can be constructed following a top-down or bottom-up approach. In a top-down approach, the set of general categories the thesaurus is going to be composed of are selected first and then, for each one, the relevant sub-concepts are recursively defined. On the other hand, following a bottom-up approach means that the committee starts defining a list of specific terms and then organizes them in categories. In this second case, the terms can be extracted using an inductive or a deductive method. In the inductive method, new terms are selected as they are found and the hierarchy is built on the fly. The vocabulary control is applied as the terms are selected. In a deductive approach, all the relevant terms are extracted in an earlier stage but no vocabulary control or relationships definition is performed until they are all collected.

Following these general guidelines, De Vorse et al (2006) describe the process used to construct the American Museum of Natural History (AMNH) thesaurus. The process starts with the revision of existent thesauri in the area and then uses a subset of Art & Architecture Thesaurus (AAT) and a set of keywords already used in the AMNH as candidate terms for the thesaurus. The process includes a cleaning phase of the AAT, a second one of harmonisation of AMNH keywords and a third one of definition of relations and scope notes for the selected terms.

Other works provide construction methodologies partially different from the indicated by thesaurus standards. For example, the State Records Authority of

New South Wales (2003) describes a complete thesaurus construction process whose term collection phase is based on DIRKS methodology (Commonwealth of Australia, 2001) for the construction of the organisation classification schemes. The complete methodology has a first stage of preparation (review of thesaurus need and planning); a second one devoted to collecting information that uses the DIRKS methodology and interviews to future users; a third one of analysis, where the thesaurus structure is composed; a fourth one of collation where the model is represented in a final format; a fifth one of revision, where feedback is searched; and a final one of production where the created thesaurus is put into use.

The process described by the Semantic Health Project (2006) to create the Belgian Bilingual Bi-encoded Thesaurus (3BT) is less elaborated. It uses the Amsterdam Thesaurus (AT) as an initial version and then it removes the unrequired concepts. Finally, it applies a set of refinement stages that add concepts, correct linguistic errors and translate the keywords from the German and French teams. The Commonwealth of Australia (2003) also describes a thesaurus construction process in a quite general way. It assumes that the organisation has developed a business classification scheme in accordance with the DIRKS methodology. And then it provides an eight-step guide to convert such scheme into a thesaurus. Another approach is the one indicated by the Working Group on Guidelines for Multilingual Thesauri of the IFLA Classification and Indexing Section (2005), which describes a methodology for the construction of multilingual thesaurus (from scratch and in base to others). However, it focuses on the criteria for selection of symmetrical terms in different languages, not on the basics for selecting terms and identifying relationships.

Finally, there are approaches based on techniques used for the construction of ontologies. In this context, Bechhofer and Goble (2001) describe a construction process that use knowledge representation techniques to facilitate the construction of coherent hierarchies. It does not describe a proper methodology, but it uses the bottom-up approach described by Vickery (1966) and improves it using Description Logics as the scheme to model the relationships between the concepts more precisely.

3. DEVELOPMENT OF THE VOCABULARY

The drought team of the EuroGEOSS project prepared a metadata catalogue tailored towards users from the drought community with data and services linked to the drought field. One of the key fields of the metadata describing resources is the field 'keywords', which facilitates the discovery of a resource of interest. Since neither the proposed general purpose vocabularies within the EuroGEOSS for the annotation of metadata nor the hydrology or meteorology related ones did comply with the needs of the drought community, it was decided to prepare a specialised drought vocabulary to:

- tailor the search to drought specific content,
- support the data providers in the annotation task of the metadata,
- approach the issue of dealing with search terms in various languages.

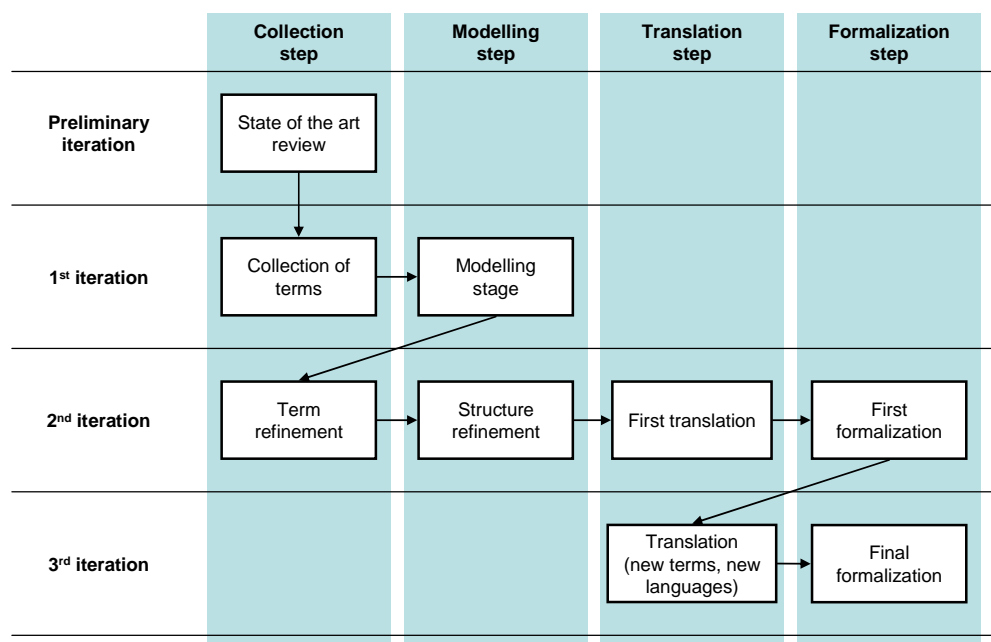
The resulting vocabulary, developed through an open and collaborative process, contains 103 concepts organised hierarchically in groups of concepts (drought, meteorology, soil, hydrology) and provides preferred and alternative labels in fifteen languages (Latre et al, 2011). The rest of this section presents the methodology followed to develop the drought vocabulary.

3.1. Overview of the methodology

The methodology followed for the development of the drought vocabulary is quite similar to the one described in the ISO standards (see section 2.2). It includes a review, modelling and structure refinement stages. Like in the case of the DIRKS methodology, an additional formalisation step was included; in our case, in order to be able to use the thesaurus in an information retrieval environment. These steps were applied in an iterative way, allowing for an increase of the level of refinement, in a similar way to the revision stage of the DIRKS methodology or the refinement stage of the 3BT, but shared among the different steps of this methodology. In the context of the drought work package of EuroGEOSS, this iterative process allowed for a rapid integration in the technological infrastructure being developed as part of the first tasks of the project and it also facilitated quick feedback and flexible collaboration of the different partners in the development and refinement of the vocabulary. A total of three iterations were needed to create the vocabulary. Each iteration can include steps of information collection, modelling, translation and formalisation with a different degree of emphasis. Previously to the first iteration, a state of the art review was performed to ensure that no other thesaurus or vocabulary fitted our purposes. Most thesauri proved to be too general or too big to be used in the drought field, as explained in section 2.

The main part of the effort devoted to the creation of the vocabulary was made during the first iteration. The second one was dedicated to refine the vocabulary based on the results of the first iteration and to provide a first translation of the terms and a draft formalisation in a knowledge representation language. The third and final iteration was devoted to finish the translations and to obtain the final formalised version of the vocabulary. Figure 2 shows a schema of the different steps and iterations followed to create the thesaurus, which are explained in the following sections.

Figure 2: Steps of the Methodology Followed in the Drought Vocabulary Creation



3.2. Collection of Information

The initial step of the first and second iterations was the selection of terms, following a bottom-up, deductive approach. During the first iteration, information and terms were collected: all partners of the drought work package contributed a list of keywords in a common language (English) that described their data and services. In most cases, the submitted terms have already been used informally to tag the created metadata. Apart from the knowledge from partner experts, related terms in well-known sources were also searched, as proposed by the standards (ANSI/NISO, 2005; ISO, 2011; BSI, 2007) and the AMNH thesaurus development methodology (De Vorse et al, 2006). In this case, GEMET and AGROVOC thesauri were searched for terms related to drought missing from the submitted words. The final set of keywords contained terms that allow characterising drought events, drought data, general topics related to droughts, drought indices, etc. The initial list was refined in the second iteration in order to add missing terms, (such as 'drought risk', 'drought management plan', 'discharge', 'drought resilience') or prune not very related ones (such as terms describing time and spatial scale, since there are more specific thesauri to cover that).

3.3. Modelling

After the collection of terms, a modelling stage took place. An identification of synonyms or conflation of different submitted keywords referring to the same term was made in the first iteration. Preferred and alternate labels were then chosen among the keywords for the conflated terms. GEMET and AGROVOC were used to select preferred labels when the term was also present in one of these thesauri. For instance, 'arid climate', 'desert climate' and 'dry climate' were different keywords individually proposed by different partners, all of them referring to the same term. 'Arid climate' was chosen as preferred label (as it is also in AGROVOC) and the other two were maintained as alternate labels. In the case of the terms referred by acronyms or the pairs 'acronym-complete name', terms were split in two in order to separate acronyms (preferred label) from their complete name (alternate label). Finally, a draft structure or hierarchy for the terms was proposed.

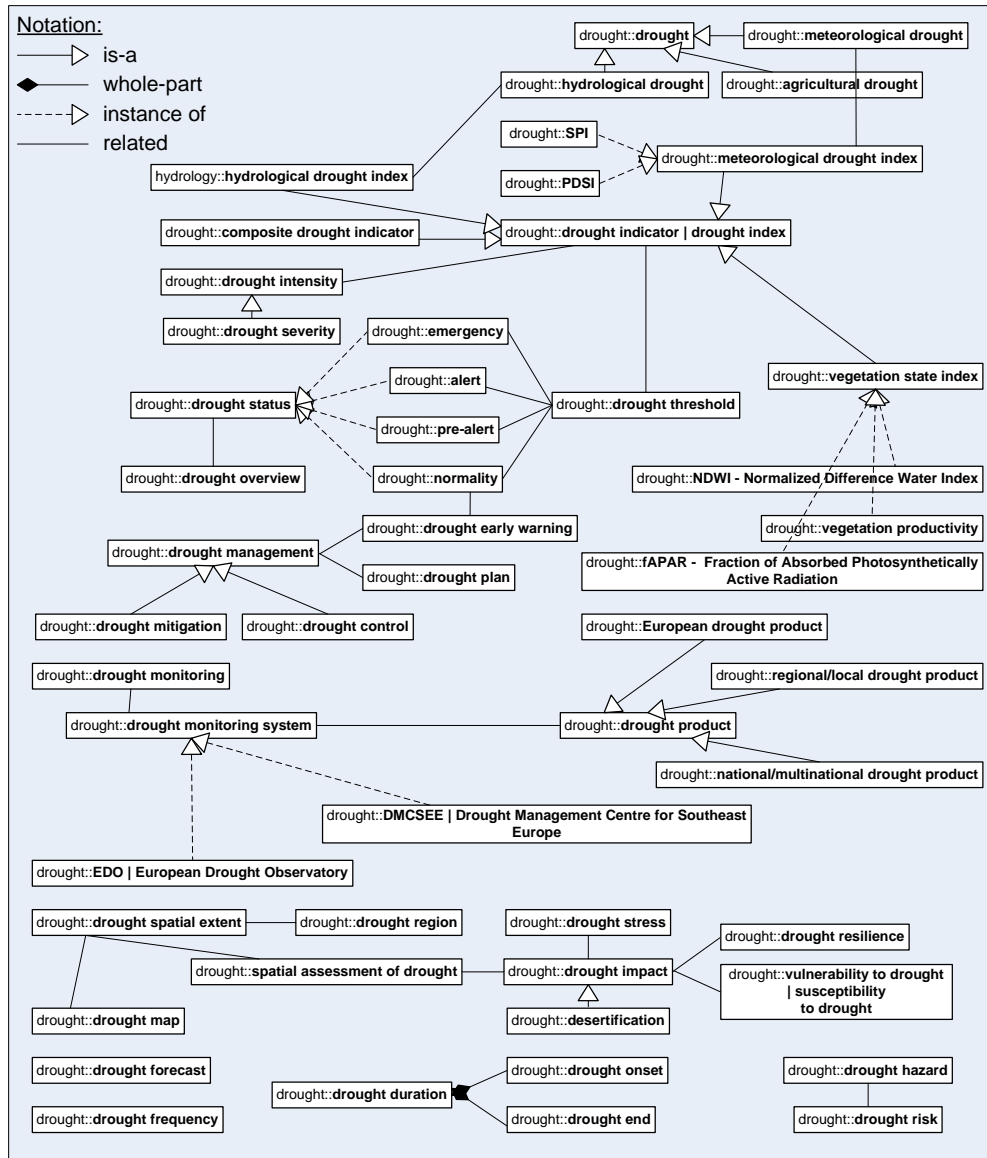
In the second iteration, this modelling was refined: different categories (groups of concepts) were identified and the hierarchical relationships of the first version were refined, maintaining only as purely hierarchical those that could be classified into *is-a* relationships (a 'rainfall anomaly' is a kind of 'precipitation anomaly'), *whole-part* ('drought duration' has an 'onset' and an 'end') or *instance-of* ('EDO' is an instance/individual/particular case of a 'drought monitoring system'). The rest were considered as non-hierarchical and maintained as simple *related* relationships ('soil' is related to 'soil moisture'). This analysis, which was performed just to better identify the hierarchy of the thesaurus, could be the basis for the engineering of this thesaurus into an ontology.

Some of the terms and relations of the vocabulary are shown in figure 3 with English labels.

3.4. Translation

The third step in the creation of the thesaurus was the translation of the terms, following the recommendations of the Working Group on Guidelines for Multilingual Thesauri of the IFLA Classification and Indexing Section (2005). For the first version of the vocabulary, obtained at the end of the second iteration, translations into Slovenian, Spanish, French and German, besides the original English version, could be quickly provided by the partners and, thus, available for testing multilingualism aspects of the use of the thesaurus. In the third iteration, apart from the translation of the newly added terms, translations into Bosnian, Turkish, Italian, Portuguese, Croatian, Serbian, Albanian, Macedonian, Greek and Montenegrin were integrated into the thesaurus, to sum up a total of 15 languages.

Figure 3: Graphical Representation of Part of the Drought Vocabulary



3.5. Formalisation

The final step in the creation of the thesaurus was its formalisation. SKOS was chosen to create a representation of this vocabulary. SKOS (Simple Knowledge Organization System) is a formal language for representing controlled structured

vocabularies, including thesauri, classification schemes, taxonomies and subject-heading systems (Miles and Pérez-Agüera, 2007). Its main objective is to enable easy publication of controlled structured vocabularies for the Semantic Web.

Unique URIs (Uniform Resource Identifier) were constructed for the terms (such as '<http://eurogeoss.eu/DroughtVocabulary/15>' for 'drought') in order to allow referring to a term in a language-independent manner. The *is-a*, *whole-part* and *instance-of* relationships were mapped to the *skos:broader* and *skos:narrower* relationships while the *related* relationship has been maintained too. There exists also the possibility of grouping concepts using the *skos:collections* construction to provide a more consistent grouping of the terms into the different categories identified during the modelling stage: meteorology, drought, soil, hydrology, statistics.

The first version of the thesaurus was represented in SKOS for testing purposes. The final version, together with other resources related to the vocabulary (spreadsheet with labels in different languages, full graphical representation and metadata) can be downloaded from the EuroGEOSS drought catalogue home page¹⁰.

4. INTEGRATION OF THE THESAURUS IN THE EUROGEOSS FRAMEWORK FOR DROUGHT MONITORING

The drought vocabulary has been integrated in the infrastructure of the European Drought Observatory in three ways. Firstly, it has been incorporated into the CatMDEdit tool, the metadata editor tool used in the drought working group of the EuroGEOSS project. Secondly, it has been integrated within the web application used for searching and updating online the metadata records. And thirdly, it has been aligned with the other two thesauri used in the EuroGEOSS framework: GEMET and the GEOSS Societal Benefit Areas categories.

4.1. Integration into the EuroGEOSS Drought Metadata Editor Tool for Resource Annotation

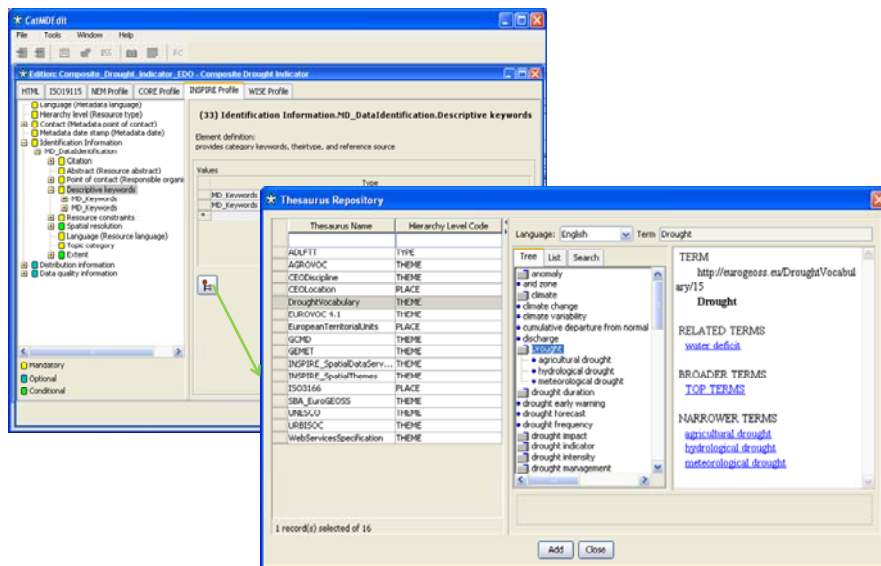
The metadata editor tool used in the EuroGEOSS drought work package, CatMDEdit¹¹ (Nogueras-Iso et al, 2008), uses a serialised version of the vocabulary to browse its content and allows users to create or update metadata to select terms from the vocabulary to tag the resources. Figure 4 shows the drought vocabulary in the CatMDEdit thesaurus browser. The vocabulary can be browsed in any of the fifteen languages available through the thesaurus tree-like structure or through an alphabetical term list. Additionally, there is a third tab that

¹⁰ <http://eurogeoss.unizar.es/home/>

¹¹ <http://catmdedit.sourceforge.net/>

allows searching terms contained in the thesaurus. The thesaurus browser, loaded with the drought vocabulary, allows users to semantically annotate the resources with selected concepts from the vocabulary. Once a term is selected in a particular language, the section “*Identification Information Descriptive keywords*” of the metadata record is updated with a new set of keywords, each of them consisting of a string representing any of the terms in the selected language present in the hierarchy from the root to the selected term, and references to the drought vocabulary.

Figure 4: Integration of the EuroGEOSS Drought Vocabulary in CatMDEdit



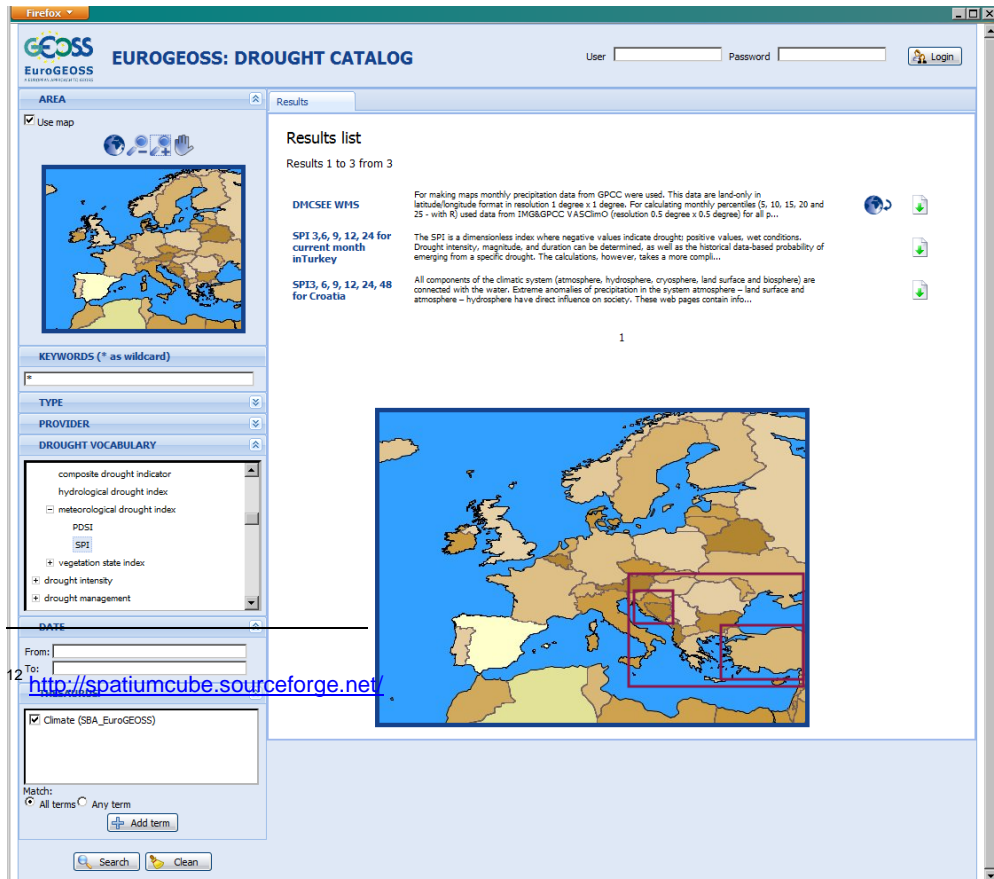
4.2. Integration into EuroGEOSS Drought Catalogue User Interface

The drought vocabulary has also been integrated into the drought catalogue user interface. The metadata records managed by this catalogue are ISO 19115 and INSPIRE compliant and describe 210 datasets and 22 web services submitted by the EuroGEOSS drought partners. About a 58% of the records were written in the original language of the dataset, while the rest were in English, which made difficult the discovery of the described resources when querying by a different language.

The web catalogue is accessible through an OGC compliant catalogue service developed with CatalogCube technology¹² and through a user friendly web application. This web application was designed to take into account the GCI Clearinghouse Requirements (GEO, 2009) about searching criteria: users should be able to search based on location, keywords or text, and temporal extent. In addition to this, a *resource type* and a *provider* criterion were also included as a way to allow users to distinguish the resource type (data or services) or the resource provider in their searches. Figure 5 shows this GUI covering these searching criteria.

Once the first version of the drought vocabulary was developed, it was displayed in the interface in order to facilitate querying the catalogue by using drought-specific related terms. It took the place of the thesauri that were in use at the moment (INSPIRE and SBA categories and sub-categories), since they had proved to be too general to aid the searching. The drought vocabulary, even in its first version and prior to the updating of the metadata with tags from the new vocabulary, proved to be an improvement in the interaction with the user, due to its capability of describing better the resources and the fact that its terms were already used informally in the keywords, abstract or title sections of the resources. Section 5 discusses this in more depth.

Figure 5: Graphical User Interface Web Application for Searching Metadata



4.3. Alignment with Other Thesauri

The drought vocabulary is tailored towards optimising searches in the drought metadata catalogue. To make it useful also in a wider context, such as the multidisciplinary scenarios of EuroGEOSS project, the vocabulary needs to be linked with thesauri that are used for searching through multidisciplinary metadata catalogues. In EuroGEOSS, these thesauri are GEMET and GEOSS Societal Benefit Areas (SBA). The linking of thesauri is referred to as matching: all terms of the drought vocabulary have to be matched to at least one term of another thesaurus. The process needs to be repeated for every thesaurus that needs to be linked to the drought vocabulary. The matching was performed manually with the SKOS matcher of the Semantic Lab¹³ of the Joint Research Centre.

A summary of the alignment activity between the drought vocabulary and either the SBA or GEMET is presented in the following tables. Table 1 shows that 12 concepts out of the 66 of the SBA were mapped to 45 concepts of the drought vocabulary, and that 50 terms out of 5244 from GEMET were mapped to a total of 103 concepts of the drought vocabulary. This validates the authors' claim made in section 2.1 that the general purpose thesauri were too big for its use in metadata annotation and search in the Drought area. Table 2 also justifies the need of a drought vocabulary: most of the mappings between terms of the drought vocabulary and the considered thesauri belong to the category of "related terms". Only 5 out of 125 mappings in the case of the SBA and 23 out of 137 in the case of GEMET are mappings with a more specific meaning: exact match, close match (that is, two similar, but not the same, terms from each thesauri are mapped) and broader match (that is, the term of a thesaurus is mapped to a more generic term of the other). This lack of more specific mappings from the reference thesauri to the drought vocabulary validates the second part of the authors' claim in section 2.1: GEMET and SBA are too general to be useful in the Drought field.

The big advantage of the matching of vocabularies is that the search of the user can be automatically extended to terms of the drought vocabulary that are linked to the selected term of the GEMET or GEOSS SBA.

¹³ <http://semanticlab.jrc.ec.europa.eu/>

Table 1: Number of Mapped Concepts

| | Thesaurus size | # of mapped concepts | |
|--------------|----------------|-------------------------|---------------------------|
| | | from original thesaurus | to the drought vocabulary |
| SBA | 66 | 12 | 45 |
| GEMET | 5244 | 50 | 103 |

Table 2: Number of Mapping Relations to the Drought Vocabulary

| | Broad | Close | Exact | Related | Total |
|--------------|-------|-------|-------|---------|-------|
| SBA | 2 | 1 | 2 | 120 | 125 |
| GEMET | 8 | 1 | 14 | 114 | 137 |

5. ANALYSIS OF THE USEFULNESS OF THE DROUGHT VOCABULARY

The previous sections discussed the motivation for establishing a drought vocabulary and the procedure of defining it. This section begins with an example to illustrate how the drought vocabulary improves searches in the drought metadata catalogue in practice, and follows with a quantitative analysis of its usefulness.

To illustrate the usefulness of the vocabulary, a comparison of search results before and after the introduction of the drought vocabulary and the annotation of the metadata descriptions with the drought keywords is going to be made. The inclusion of the drought vocabulary in the search web application was done on 27th December 2010. Before that date, a first version of the drought resources search application without integrating the drought vocabulary was used. Users were able to perform searches based on location, temporal extent, resource type and provider, and by selecting keywords. Although the drought vocabulary was not yet available, users could search using keywords from the SBAs categories and subcategories list (a panel was open in the GUI to show this vocabulary), ISO 19115 topic categories or by writing free text. From the 27th December 2010 on, and using the current version of the search application, the drought vocabulary, already developed, is displayed in a panel in the user interface. To restrict a query with a term from the drought vocabulary, users just need to click on it in the panel. Users are still enabled to search for drought resources using location, temporal extent, resource type, provider and other thesauri and vocabularies (SBAs, ISO 19115 topic categories, INSPIRE spatial data themes and GEMET), selecting terms alone or combined with others and by writing free text. Both previously to and afterwards the introduction of the vocabulary, the title,

abstract and keywords of the metadata descriptions are searched when providing keywords from the interface or writing free text.

Querying for drought indicators through the graphic user interface previous to the introduction of the drought vocabulary had to be done by providing free text to the search query. Specifying the term 'drought indicator' in the free text textbox would have provided a result list that consisted, due to an implementation choice, of all the resources in the drought catalogue containing the words 'drought' or 'indicator' in English. With the current resources of the catalogue, this search provides a total of 137 records (a 58% of the resources contained in it). In the case of searching using only the term 'indicator', 65 results would appear, including resources related not only to drought indicators but also 9 other related to other types of environmental indicators. The results of the search could be restricted to avoid getting extra results by providing the text 'drought?indicator' (even if it is very unintuitive for an end user). This last query would have provided 20 results strictly related to drought indicators: those tagged with the words "drought indicator" in English, but leaving out a total of 37 drought indicators whose metadata was written in Spanish.

With the introduction of the Drought Vocabulary, there are 57 results provided by a query where the term 'drought indicator' is selected from the Drought Vocabulary panel (Figure 6), and those resources with metadata in other languages than English are also retrieved (third, fifth and sixth result in Figure 6, for example, tagged in Spanish with the term '*indicador de sequía*').





















The inclusion of the drought vocabulary in the user interface and its use in the metadata records facilitates the retrieval of more accurate results than just writing free text, for example. Additionally, the annotation of the metadata with elements of the drought vocabulary increases the identification of the metadata descriptions relevant to the query. Without the vocabulary, these resources are not tagged with this term and cannot be found, even if the term is written as free text.

In order to validate in a quantitative way the developed drought vocabulary in the context of the EuroGEOSS drought framework, the logs of the drought web catalogue service from 30th September 2010 to 23rd June 2011 have been analysed. Two different periods have been considered and can be compared while analysing these logs. These periods are divided by the event of the inclusion of the drought vocabulary in the application interface on 27th December 2010. Logs included IP address, date and time, type of request and, in the case of query requests, all the constraints contained in the query.

Figure 6: Search Results for Keyword ‘Drought Indicator’ with the Drought Vocabulary.

Results list

Results 1 to 10 from 57

| | | | |
|--|--|---|---|
| CHE Hydro Index WFS | This Web feature service provides data related to the Ebro River Basin District Hydrological Index in accordance with the OGC WFS Implementation Specification 1.1 standard... |  |  |
| Vegetation Water Content NDWI EDO | The Normalized Difference Water Index (NDWI) assesses the water content of vegetation, which can be affected by drought. It is a remote sensing product that is derived from the Near-Infrared (NIR) and Short Wave Infrared (SWIR) channels of the MODIS (Moderate Resolution Image Spectroradiometer) sens... |  |  |
| Reservas medias mensuales de los últimos 5 años en los embalses pertenecientes a la Cuenca del Ebro | Cobertura de los embalses más significativos de la cuenca del Ebro con el dato de volumen medio mensual embalsado de los últimos 5 años y su porcentaje respecto al volumen máximo, de acuerdo con los valores suministrados por el Sistema Automático de Información Hidrológica (SAIH). Reproduce de forma... |  |  |
| Forecasted Soil Moisture Anomaly for Europe, EDO | In the forecasting mode the European Flood Alert System produces information on the development of soil moisture in Europe for up to ten days ahead. The forecasted soil moisture seven days ahead is compared to the long-term average conditions of this date as derived from re-analysis data of... |  |  |
| Reservas mensuales de embalses pertenecientes a la Cuenca del Ebro | Cobertura de los embalses más significativos de la cuenca del Ebro con el dato de volumen embalsado en la última fecha del mes y su porcentaje respecto al volumen máximo, de acuerdo con los valores suministrados por el Sistema Automático de Información Hidrológica (SAIH). Reproduce de forma cartográ... |  |  |
| Reservas medias de los últimos 12 meses en los embalses pertenecientes a la Cuenca del Ebro | Cobertura de los embalses más significativos de la cuenca del Ebro con el dato de volumen medio embalsado los últimos 12 meses y su porcentaje respecto al volumen máximo, de acuerdo con los valores suministrados por el Sistema Automático de Información Hidrológica (SAIH). Reproduce de forma cartográ... |  |  |
| Daily Soil Moisture for Europe, EDO | Daily soil moisture information provides an image of the current situation of the water content in the top soil layer. Information on soil moisture is presented in form of soil suction (pF) values of the top soil layer that commonly range between 1.5 for very wet conditions up to 5.0 for very dry so... |  |  |
| CHE Hydro Index WMS | This Web Map Service provides geographic visualization of the data relative to the Ebro River Basin District Hydrological Index. This service implements the OGC WMS Implementation 1.3 Specification... |  |  |
| Ebro river basin hydrological status index | Hydrological status index which is derived from several indicators (water flow, water storage, snow package, piezometric levels). The Confederación Hidrográfica del Ebro (CHE) publishes a monthly drought index based on a combination of the following indicators: three-month water flow, water stored i... |  |  |
| JRC EDO (European Drought Observatory) - Draft Drought Products Delivery WMS | The service is a draft version of a WMS service to deliver drought products, i.e. maps of drought indicators as they are displayed into the EDO MapServer Prototype (please see http://edo.jrc.ec.europa.eu/php/index.php?action=view&id=201) and described on the EDO website (http://edo.jrc.ec.europa.eu/ ... |  |  |

1 2 3 4 5 6 Next

The analysis of the logs (Table 3) show a comparable number of uses per day in the both periods considered and a user's preference for searching through terms (each query, in average, uses at least one term); this fact suggests that users of the web application tend to use the other available restriction means (location, date, providers) only marginally. The preference for searching by keywords and terms is maintained with the release of the second version (the average number of terms in each query is in both periods about slightly higher than 1.3).

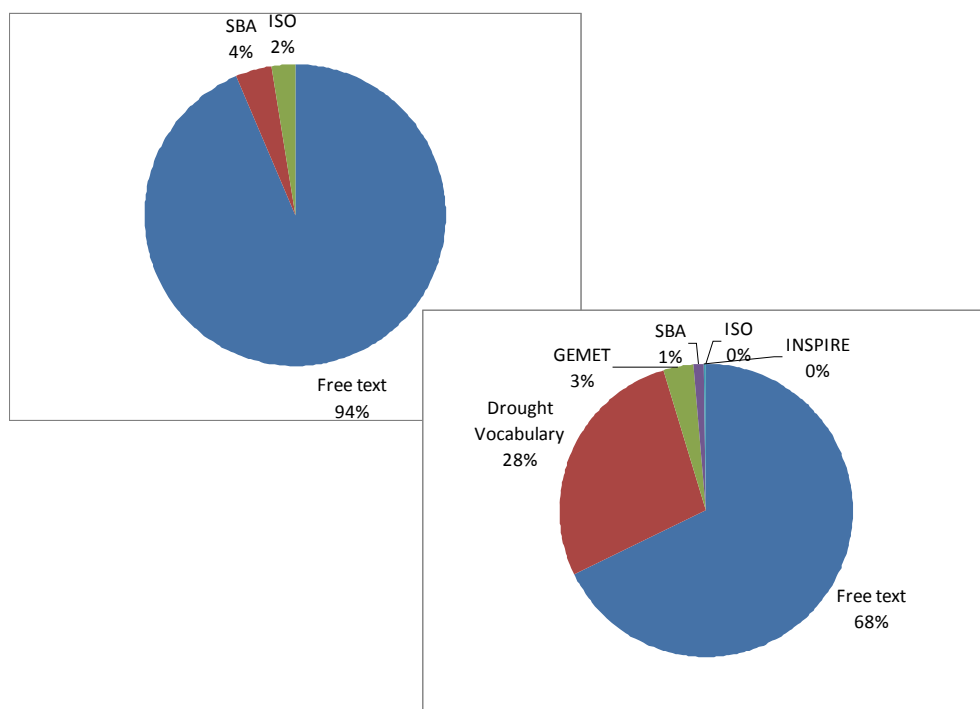
Table 3: Measurements from the Drought Catalogue Queries

| | 1 st period | 2 nd period |
|---|------------------------|------------------------|
| Duration in days | 88 | 178 |
| Number of performed queries | 1295 | 2143 |
| Average number of queries per day | 14.72 | 12.04 |
| Total number of concepts queried | 1745 | 2831 |
| Average number of terms per query | 1.35 | 1.32 |
| Number of queried terms that belong to the drought vocabulary | | 781 |
| Percentage of queried terms that belong to the drought vocabulary | | 28% |
| Number of different queried terms that belong to the drought vocabulary | | 67 |
| Number of concepts in the drought vocabulary | | 103 |
| Percentage of use of the vocabulary | | 65.0% |
| Queried terms from other thesauri | 125 | 135 |

The concepts queried from the user interface were specified either by the controls and panels the different versions of the user interfaces offered: SBAs categories panel, ISO 19115 topic categories panel or by providing free text in the first version; drought vocabulary panel, other thesauri panel (SBAs, ISO 19115 topic categories, INSPIRE spatial data themes and GEMET) and free text in the second version. In order to track the source thesauri of each queried term, we have considered the coincidence of the queried term with one of the concepts of the thesauri relevant to the EuroGEOSS project (SBAs, INSPIRE topic themes, GEMET and the drought vocabulary), with independence of analysing whether the term was provided by a user choosing a term from a thesaurus panel or from the free text field. Figure 7 displays the provenance of the queried terms for the two considered periods, i.e., it indicates whether the source of terms is free text or a specific thesaurus.

It is noticeable that, in the first version of the application, the SBA categories and the ISO 19115 topic categories were hardly used, even if, in the case of the SBAs, these categories were easily available in the graphic user interface. The reason for that is that, as stated before, these categories do not properly classify drought resources. In the case of SBAs, most of the resources could be best classified under the 'drought prediction' subcategory of the 'water' category, being this term not an easy one to be found in the hierarchy and not providing a good classification of the resources. All this leads the users not to find drought resources of interest with their queries when using the proposed thesauri and, usually, force them to browse all the contents of the catalogue.

Figure 7: Source Thesauri of the Queried Terms
 Up left: before the development of the vocabulary; down right: after



However, for the second version of the application, figure 7 shows how the drought vocabulary is the most used vocabulary to specify searching terms. Part of this preference is obviously due to the fact that the vocabulary is easily accessible in the user interface, but it is also true that, after the metadata records were annotated with the terms of the new vocabulary, it is the thesaurus, among the available ones, that best classifies the resources. It is also remarkable that 67 out of 103 terms of the vocabulary have been used at least once in a search, which is an indicative of the appropriateness of vocabulary, meaning by *appropriateness* that the terms chosen to be part of the vocabulary are the relevant ones in the field of drought.

6. CONCLUSIONS AND FUTURE WORK

The proposed vocabularies in the European project EuroGEOSS for the annotation of metadata proven to be, in the thematic area of drought, either too generic to adequately classify drought resources, or too large to be practical for

their annotation. As a consequence, a drought vocabulary has been developed in a collective way, in order to improve the accessibility to appropriate drought resources (datasets and services) to users and experts.

It was thought that a first-guess vocabulary could be prepared for this area based on a collection of terms that would considerably improve the discovery of available resources and, in the end, a 103-term vocabulary, organised into a hierarchy and translated into 15 languages has been developed. The methodology followed for the creation of this specific drought vocabulary has been presented, methodology that could be also applied to other subject areas where the same needs and problems could be identified. The documentation of newly created vocabularies (such as this paper, describing the vocabulary itself and the methodology followed for its construction) is necessary for making them usable to a wider community.

This vocabulary has been first used in the EuroGEOSS drought catalogue in three ways. Firstly, by annotating the resources it holds according to the new vocabulary, since the quality of the search results of a catalogue query depends on the quality of the metadata. The terms of the vocabulary had to be used in the annotation of the metadata, since otherwise resources cannot be properly found. A 51.5% of the drought vocabulary concepts have been used in the annotation of the EuroGEOSS drought catalogue resources. Secondly, it has been included in the user search application interface. Analysis from the catalogue logs shows that it improves the interaction with users, helping them to establish their searching parameters. The fact that the vocabulary has been used both in the metadata and in the user search application helps to improve the search results. Logs also show that the terms selected to be part of the vocabulary are appropriate from a user's point of view, since 65.0% of them have been used in at least a query. Finally, the vocabulary has been aligned with the other two thesauri chosen for metadata annotation in the project: GEMET and GEOSS Societal Benefit Areas categories.

Future work will deal with the assessment of the completeness of the vocabulary, that is, to evaluate if are there any relevant terms missing from the vocabulary and reassessing the appropriateness of the already selected terms of the vocabulary, by analysing a text corpus composed of documents related to drought. Additionally, the ontological aspects of this drought vocabulary will be explored. This vocabulary will be the basis for a heavyweight ontology, with axioms and constraints that could allow more complex analysis on data sources. The authors plan to approach to define more explicitly the concepts related to drought indicators, their range of values and their semantics to try to define the classification structure of the different drought indexes and to facilitate the establishment of state equivalences between them. From the semantics point of view, it seems highly interesting to explore this ontological approach even if the

topic is sufficiently complex, due to the fact that different drought indices show different impacts of drought (on vegetation, water levels, soil moisture or availability of water).

ACKNOWLEDGEMENTS

This work has been partially supported by the European Commission (FP7 Project nr 226487 - EuroGEOSS) and the Spanish Government (project TIN2009-10971). The authors would like to thank Barbara Medved-Cvikl and Andrej Ceglar (from the Biotechnical Faculty of the University of Ljubljana), Rogelio Galván (from the Ebro River Basin Authority), Carolina de Carvalho (from the Alcalá University's General Foundation - Observatory on Sustainability in Spain), Florian Husson and Jean-Francois Vernoux (from the Bureau de Recherches Géologiques et Minières), José Miguel Rubio (from the Centro Nacional de Información Geográfica) and Stefan Niemeyer and Jürgen Vogt (from the European Commission Joint Research Centre) for their contribution to the development of the drought vocabulary.

REFERENCES

- American Meteorological Society (2000). *Glossary of Meteorology, Second Edition*. Todd S. Glickman (Ed). Boston, United States: AMS Books.
- ANSI/NISO (2005). *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*. American National Standards Institute (ANSI), Z39-19.
- Bechhofer, S. and C. Goble (2001). Thesaurus construction through knowledge representation. *Data & Knowledge engineering*, 37: 25-45.
- BSI (2007). *Structured vocabularies for information retrieval. Guide*. British Standards Institute (BSI), BS 8723.
- Commonwealth of Australia (2001). The DIRKS methodology: A users guide, at: http://www.naa.gov.au/Images/dirks_part1_tcm2-935.pdf [accessed 17 August 2011].
- Commonwealth of Australia (2003). *Developing a Functions Thesaurus. Guidelines for Commonwealth Agencies*, at: http://www.naa.gov.au/images/developing-a-thesaurus_tcm2-916.pdf [accessed 16 August 2011].
- EuroGEOSS (2010). *Deliverable Description of the Initial Operating Capacity*, at: http://www.eurogeoss.eu/wp/wp5/Documents/D.5.5_InitialOperatingCapacity_DeliverableDescription_FINAL.pdf [accessed 12 August 11].

- European Commission (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Official Journal of the European Union*, 50 (L 108) of 25 April 2007: 1-14.
- European Commission (2008). Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata *Official Journal of the European Union*, (L 326), of 4 December 2008: 12–30.
- GEO, Group of Earth Observation (2009). *GCI Consolidated Requirements*, at: http://www.earthobservations.org/documents/gci/gci_requirements_20090312.doc [accessed 9 August 2011].
- Gruber, T. R. (1993). A translation approach to portable ontology specifications, *Knowledge Acquisition*, 5(2): 199–220.
- Hodge, G. (2000). *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. The Digital Library Federation, Washington DC.
- ISO (1986). Documentation: Guidelines for the establishment and development of monolingual thesauri. ISO 2788, International Organization for Standardization.
- ISO (2011). *Thesauri and interoperability with other vocabularies (draft)*. International Organization for Standardization (ISO), ISO 25964.
- Lacasta, J., Nogueras-Iso, J. and F. J. Zarazaga-Soria (2010) *Terminological Ontologies: Design, Management and Practical Applications*. Germany: Springer.
- Lacasta, J., Nogueras-Iso, J., Béjar, R., Muro-Medrano, P. R. and F. J. Zarazaga-Soria (2007). A Web Ontology Service to facilitate interoperability within a Spatial Data Infrastructure: applicability to discovery, *Data & Knowledge Engineering*, 63: 945-969.
- Lassila, O. and D. L. McGuinness (2001). *The Role of Frame-Based Representation on the Semantic Web*. Stanford University.
- Latre, M. Á., Lacasta, J., Mojica, E., Nogueras-Iso, J. and F. J. Zarazaga-Soria (2009). “An Approach to Facilitate the Integration of Hydrological Data by means of Ontologies and Multilingual Thesauri” in Sester M., Bernard L. and Paelke V. (Eds). *Advances in GIScience*. Springer Berlin Heidelberg, pp. 155–171.
- Latre, M. Á., Nogueras-Iso, J. and B. Hofer (2011). “Searching drought-related resources through a specialized vocabulary - testing the interoperable infrastructure of the EuroGEOSS project”. *Proceedings of the INSPIRE*

Conference 2011: INSPIREd by 2020 - Contributing to smart, sustainable and inclusive growth. June 27 - July 1. Edinburgh, Scotland.

- Miles, A. and J. R. Pérez-Agüera (2007). SKOS: Simple Knowledge Organisation for the Web. *Cataloging and Classification Quarterly*, 43(3–4): 69–83.
- Nogueras-Iso, J., Barrera, J., Gracia-Crespo, F., Laiglesia, S. and P. R. Muro-Medrano (2008). “Integrating catalog and GIS tools: access to resources from CatMDEdit thanks to gvSIG”, *Proceedings of the 4th International gvSIG conference: moving forward together*. December 3-5 2008, Valencia, Spain.
- Semantic Health Project (2006). *The creation of the Belgian Bilingual Bi-encoded Thesaurus (3BT)*. World Health Organization. at: http://www.semantichhealth.org/PUBLIC/Belgium_The%20creation%20of%203BT.pdf [accessed 16 August 2011].
- Sowa, J. F. (1996). Ontologies for Knowledge Sharing. In *Manuscript of the invited talk at Terminology and Knowledge Engineering Congress (TKE '96)*, Vienna.
- State Records Authority of New South Wales (2003). *Guidelines for Developing and Implementing a Keyword Thesaurus*, at: <http://www.records.nsw.gov.au/recordkeeping/government-recordkeeping-manual/documents/recordkeeping-guidelines/Developing%20and%20Implementing%20a%20Keyword%20Thesaurus.pdf> [accessed 16 August 2011].
- UNESCO (1993). *International Glossary of Hydrology / Glossaire International D'hydrologie* (Wmo/Omm/Vmo, No. 385), United States: Unipub.
- Uschold, M. and M. Gruninger (2004). Ontologies and semantics for seamless connectivity, *ACM SIGMOD Record*, 33(4): 58–64.
- Vickery, B. C. (1966). *Faceted Classification Schemes*. Rutgers Series on Systems for the Intellectual Organization on Information, Rutgers State University, New Brunswick, NK,
- Vorse, K. L. de, Elson, C., Gregorev, N. P. and J. Hansen (2006). The development of a local thesaurus to improve access to the anthropological collections of the American Museum of Natural History. *D-Lib Magazine*, 12 (4).
- Working Group on Guidelines for Multilingual Thesauri of the IFLA Classification and Indexing Section (2005). *Guidelines for Multilingual Thesauri*, IFLA Professional Reports, No. 115. International Federation of Library Associations and Institutions, at: <http://archive.ifla.org/VII/s29/pubs/Profrep115.pdf> [accessed 16 August 2011].