Are SKOS concept schemes ready for multilingual retrieval applications?

Diana Tanase
School of Electronics and Computer Science
University of Westminster
101 New Cavendish Street London, W1W 6XH
tanasedi@westminster.ac.uk

Epaminondas Kapetanios
School of Electronics and Computer Science
University of Westminster
101 New Cavendish Street London, W1W 6XH
e.kapetanios@westminster.ac.uk

ABSTRACT

This article describes our approach to accessing Knowledge Organization Systems expressed using the Simple Knowledge Organization System (SKOS) data model. We share the view that the Web is becoming a multilingual lexical resource and a distribution infrastructure for knowledge resources. We aim to tap into this for the particular use case of Cross-Language Information Retrieval systems. The SKOS framework allows the description of monolingual or multilingual thesauri, controlled vocabularies and other classification systems in a simple machine-understandable representation. It has support for decentralized distribution on the Web of any resource described with it and includes mechanisms to interconnect different concept schemes. Yet, when building our prototype CLIR system different processes require more than the existing content of a SKOS resource: concept descriptions, labels and basic inter-concept relations. For example the SKOS concept indexing phase entails identifying potential occurrences of a SKOS concept in a text and to disambiguate based on the semantics referenced to in the overall SKOS scheme. By design, the SKOS data model does not formally define semantics of its concepts thus we have built a set of three algorithms that help generate a multilingual dataset linking to the original SKOS dataset and providing more details about the lexical entities that describe concepts. This new dataset contains specific RDF triples that aid concept identification, disambiguation and translation in CLIR.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models*

General Terms

Algorithms, Languages, Standardization, Design

I-SEMANTICS 2012, 8th Int. Conf. on Semantic Systems, Sept. 5-7, 2012, Graz, Austria

Keywords

Cross-language Information Retrieval, Multilingual Semantic Web, Semantic Search, Conceptual Index, SKOS, W3C

1. INTRODUCTION

The Multilingual Web is the reality of today's online content landscape. Our geographical boundaries have disappeared online with multilingual users meandering through the web in search of facts, of answers to questions, in an attempt to discover new information, or just to keep alert on what is going on throughout the world. There are though, the language boundaries. They restrict access to the Web in its entirety. In this context, the research field of Cross-lingual Information Retrieval (CLIR) brings forward a set of questions that focus on helping users locate and present answers to their queries, regardless of their main language from resources in other languages. The challenge is to build a CLIR system that has a broad access to multilingual language and knowledge resources.

We share the view described by Calzolari [3] in 2008 that the Web is becoming a multilingual lexical resource and a distribution infrastructure for knowledge resources. Our aim is to learn how to tap into it in the particular use case of a CLIR system. The foreseen advantages of connecting CLIR to such resources are: a) organic growth in scale, b) translations in tune with cultural changes of meaning within a community, and c) a self-evolving system. Though an ideal scenario, it can never be achieved without investigating how to use existing knowledge resources published on the Web.

In this article, we focus specifically on accessing Knowledge Organization Systems (KOS) openly available and expressed using the Simple Knowledge Organization System (SKOS) data model. The SKOS framework allows the description of monolingual or multilingual thesauri, controlled vocabularies and other classification systems in a simple machine-understandable representation. It has support for decentralized distribution on the Web of any resource described with it and includes mechanisms to interconnect different concept schemes. Thus, SKOS resources provide a promising mechanism to link a CLIR system to a variety of multilingual domain knowledge.

Yet, while building our prototype CLIR system the question emerged, *are SKOS concept schemes ready for multilingual retrieval applications*? We have uncovered that different processes require more than the existing content of a SKOS resource: concept descriptions, labels and basic inter-concept relations. For example the SKOS concept indexing phase entails identifying potential occurrences of a SKOS concept in a text, but also means to disambiguate based on the semantics referenced to in the overall SKOS scheme. By design, the SKOS data model does not formally define semantics of its concepts. Also, not all SKOS datasets are multilingual and we wanted to be able to add acquired machine translations to an initial dataset.

Thus, we have designed a set of three algorithms that help generate a multilingual dataset linking to the original SKOS dataset and adding more details about the lexical entities that describe its concepts. This new dataset contains specific RDF triples that aid concept identification, disambiguation and translation in our prototype CLIR. Before describing the algorithms and their results we introduce SKOS as a data model of multilingual KOS in Section 2. We walk through a use case scenario in Section 3 that underlines how SKOS Resources can be used at different stages while building a CLIR prototype and the need for additional lexical details for each of its concepts. In Section 4 we detail how a new lexical level can be automatically added to a SKOS resource's conceptual and terminological structure, and discuss our results in Section 5.

2. MULTILINGUAL SKOS RESOURCES

SKOS, as detailed in [2] and [11], is a mechanism for describing concept schemes in a machine understandable way particularly aimed to be used by semantic technologies. A *concept scheme* is a set of categories of knowledge at different granularity levels. This includes taxonomies, thesauri, and other vocabularies. SKOS itself does not provide solutions for how to create concept schemes, but how to represent them. The value of using SKOS resources is in the lightweight representation of domain specific vocabulary and categorizations. They tend to contain quite large and thoroughly organized concept hierarchies and cross-references.

Specifically, a SKOS description of a concept scheme contains a range of basic information about its concepts and the relations between them. As an example, let us refer to Figure 1 that shows some of the details captured in the GEneral Multilingual Environmental Thesaurus[1] (GEMET) for *climatic change*. For each concept in this dataset, the SKOS concept specification defines a set of multilingual lexical labels: the unique preferred term, a number of alternative terms, and additional documentation such as definitions and optional notes that describe the concept scheme's domain.

The concepts may be related to one another in a variety of ways. In this example, *climate* is a broader concept than *climatic change*. There are no narrower concepts, but there are a number of related terms with which it shares an unspecified association (*climatic alteration, deforestation, man-made climate change*).

The broader and narrower relationships define the hierarchical structure for the concepts, while related is used for associations. It should be noted that the broader/narrower terms do not prescribe a subsumption relationship, but are given the definition that any resource annotated via a given term can be retrieved via its broader term. Note that SKOS allows for a loose specification of facts, and *climatic change* narrower than *climate* for example, does not imply that the former is a specialization of the later.

Another aspect of SKOS resources that can be observed in Figure 1 is its support for interconnecting concept schemes. For example, GEMET specifies mappings between its SKOS concepts and other multilingual datasets such as DBpedia[2], the AGROVOC[3] thesaurus containing specific terms for agricultural digital goods, as well as UMTHES[4], a German-centric thesaurus about environmental protection. These mappings represent connection points to the evolving Linked Data and in the context of an information access system allow for the exploration of concepts and documents across a concept scheme's boundaries. Examples of mappings are *exact match* (equivalent concepts), *close match* (similar but not equivalent concepts), *broad match* (a more general concept), *narrow match* (a more specific concept), and *related match* (an associated concept).

Definition

The long-term fluctuations in temperature, precipitation, wind, and all other aspects of the Earth's climate. External processes, such as solar-irradiance variations, variations of the Earth's orbital parameters (eccentricity, precession, and inclination), lithosphere motions, and volcanic activity, are factors in climatic variation. Internal variations of the climate system, e.g., changes in the abundance of greenhouse gases, also may produce fluctuations of sufficient magnitude and variability to explain observed climate change through the feedback processes interrelating the components of the climate system.

Climatic change	**cambio climático**
broader terms	
climate	clima
related terms	
climatic alteration	alteración climática
deforestation	deforestación
man-made climate change	cambio climático artificial
other relations	
exact match	AGROVOC: Climatic change
Wikipedia article	Climate change
close match	UMTHES: Klimaänderung

Figure 1: *Climatic change* SKOS concept from GEneral Multilingual Environmental Thesaurus

In summary, a SKOS resource has mainly two levels of structure: a *conceptual level*, where concepts are identified and their interrelationships established; and a *terminological correspondence level*, where terms are associated (preferred or non-preferred) to their respective concepts.

A third level, optional level, can be defined using SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL) allowing to define a *lexical level* where lexical relationships are defined to interconnect terms. In the next sections, the focus is on automatically adding this third level to an existing SKOS resource to facilitate specific processes in our prototype CLIR.

3. USING SKOS RESOURCES FOR CLIR

In the CLIR prototype in Figure 2, SKOS datasets are used for several processes: *query construction, processing*

[1] http://www.eionet.europa.eu/gemet/about?langcode=en

[2] http://wiki.dbpedia.org/DBpediaLive

[3] http://aims.fao.org/website/AGROVOC-Thesaurus

[4] http://data.uba.de/umt/de/concepts/_00014452.html

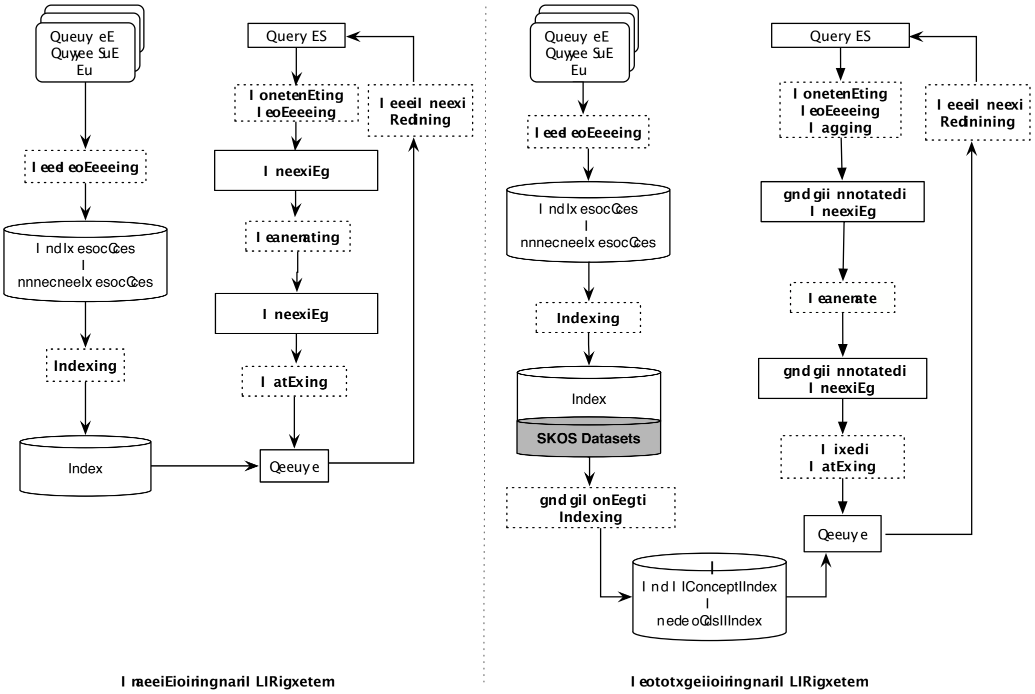

Figure 2: Re-thinking CLIR

and mapping, *translation*, and for *generating a SKOS Concept-based index*. The first two processes are known uses for thesauri, like in the retrieval system created using a SKOS-based astronomical vocabularies by [6] where queries are built using terminology from the SKOS domain vocabulary. Also, there exists some previous work at CLEF in 2002 by [13] that addressed the problem of mapping queries to terms from bilingual thesauri using an exact match (identify the longest matching entry term) or fuzzy match (using a similarity measure find a suitable candidate entry term). These particular experiments were not conclusive on how thesauri influence the operation of a CLIR system in terms of precision and recall. This research question is still debated within the research community and we believe one of the reasons it was hard to investigate in the past is the fact KOS resources were not shared in a standard format like SKOS.

With regards to indexing techniques for mapping queries and documents to concepts from external resources, a recent example is Explicit Semantic Analysis (ESA) [5] that uses Wikipedia as a knowledge resource that allows mapping text to a concept space of Wikipedia articles. Other models, like the *mixed models*, adopt the bag-of-words model, however, they extend it by using taxonomies ([16]), ontologies ([7]), networks of concepts ([9]), thesauri ([14]) or categories derived from WordNet ([4]). In each of these models, a key step is identifying the association between a concept from an external knowledge resource and its textual representation. This entails that there exists a specification of a partial or fit-for-purpose concept's lexicalisation.

3.1 Scenario

Let us start with a comparative look of our prototype in Figure 2. In both sections of the diagram, the classic versus the prototype, a CLIR system starts by processing a query from the source language, Spanish in this case, followed by its translation to the target language, English. Previous research in the field has proven that for large scale applications translating queries is a feasible approach, as opposed to translating all documents in the collection depending on the source language of the query. In our prototype part of the diagram, a query is semantically annotated by matching it to a suitable concept from the multilingual SKOS datasets. The translation step is now performed based on the existing multilingual labels in the dataset. The SKOS resource we will use as a basis for the example below is GEMET.

Let us assume the initial query is asking about information on *cambio climático* and is submitted to a document collection containing D_1, D_2, D_3.

D_1**:** Spanish ratification of the Kyoto Protocol in 2002 implied the commitment of limiting emissions. Since then, Climate Change policies are extremely important for Spanish institutions.

D_2**:** There is a changing climate in today's approach to state welfare in England with the most vulnerable left without the support they need.

D_3**:** The fluidity of the global economic climate and evolving immigration enforcement policies lead to a complicated picture of the impact of immigrants to the US economy.

3.1.1 Translation

One of the pivotal aspects of CLIR is translation, which relies heavily on language and knowledge resources to map information encoded in the query-language to information encoded in the document-language. In the example above, we translated both words as a phrase. With external knowledge from GEMET it is straightforward to identify multiword expressions like the one in our query, but such resources for the source language are not always available.

Linguistic resources vary in size, coverage, source style (human or machine readable), form of entries, number of translations alternatives given, translation ambiguity, etc. This set of characteristics directly affects the translations' quality and was at the center of years of research in developing methods that perform well, independently of the pairs of languages the translations are run between. Unfortunately, the devised algorithms and techniques do not have the same efficiency in CLIR tasks irrespective of the language [10].

In summary, according to [8] there are three main groups of translation-related challenges:

- Identifying translation units: what words or phrases should be translated?
- Obtaining translation knowledge: identify suitable resources for machine translation: bilingual dictionaries, corpora, and other knowledge and lexical resources to handle out-of-vocabulary situations.
- Using translation knowledge: words or phrases can have several translations and choosing the most appropriate one based on context is referred to as translation disambiguation; if this is not possible, there should be a weighing process of the translation alternatives.

By representing documents and queries at a higher level of abstraction, namely by using concepts, it is our approach to avoid translation issues arising from translating individual words in the narrow context of a query.

- Identifying translation units: SKOS concepts or phrases; words as a fallback mechanism.
- Obtaining translation knowledge: SKOS resources, Wikipedia and other related Web accessible services or resources
- Using translation knowledge: select the appropriate SKOS resource based on application domain

Therefore the matched concept for the query in our scenario, based-on the preferred label for Spanish described in GEMET, is the SKOS concept *climatic change*[5] and the translated query is the union of preferred and alternative concept labels, in this case just *climatic change*. Then, this gets matched against the document collection in the same language.

3.1.2 Indexing

In the Classic Bilingual CLIR System and for our prototype there is a pre-processing stage when by using Natural Language Processing (NLP) tools such as tokenizers, stemmers, and/or phrase identification each document's text is split into terms that become part of an inverted term index, where each entry points to a vector of frequencies of the term in a document and within the collection. This index is used for tolerance when the knowledges bases are incomplete.

The SKOS Concept Index in the prototype version is generated using the previously created index and maps documents to SKOS concepts from a set of predefined resources. Modeling the relationship between a document and a SKOS concept in the context of the Semantic Web, means defining a corresponding RDF triple within a semantic repository. Determining that a concept is expressed in a document with input from an external knowledge resource, is referred to as *semantic annotation* or *semantic feature extraction*. Each annotation has two components the actual text span and a pointer to the concept from the external resource. An Uniform Resource Identifier (URI) in this case. This is a challenging process that involves identification of a concept's occurrence and disambiguation for the polysemy cases. Both aspects require that the SKOS knowledge resource is lexicalized containing details about a concept's textual representation.

Concepts can occur explicitly or implicitly. For the explicit case, the SKOS specification accommodates this functional requirement as mentioned in the previous section by allowing different types of labels to be added to the dataset. For implicit occurrences, the definition of a concept, other annotations and its relations with other concepts can help build a concept's textual signature, a list of key words and phrases to feed to disambiguation algorithm 2.

For example, with the given version of the GEMET dataset, each concept has a translated label in one of 30 languages, but concepts are only described in English. This means that we can only perform label matching of concepts for non-English cases and miss out any implicit occurrences of a concept. Using the algorithm in Section 4.2, we expand the current resource by creating an RDF graph containing a set of new multilingual annotations derived from the English content of GEMET.

For our running example, the two step concept-indexing process, where we first identify the concepts that occur implicitly in a text using information from a SKOS resource, and then disambiguate, leads to the following result with the annotated concepts italicized.

D_1: Spanish ratification of the Kyoto *Protocol* in 2002 implied the commitment of limiting emissions. Since then, *Climate Change* policies are extremely important for Spanish institutions.

D_2: There is a changing climate in today's approach to state welfare in England with the most vulnerable left without the support they need.

D_3: The fluidity of the global economic climate and evolving immigration *enforcement* policies lead to a complicated picture of the impact of immigrants to the US *economy*.

There were other concepts that were initially identified: *approach* and *lead*, but correctly dropped when disambiguating. In GEMET the concept *approach* refers to a way or means of entry or access, and the scope is *urban settlement*, while in the other case is the metal *lead*. Identifying a concept beyond matching the label of a concept in a text is paramount for a concept-based index to improve a given CLIR system.

Another two annotations that were removed during the disambiguation process are the occurrences of the word *climate* from document D_2 and D_3. By using the classic keyword-based index, all documents will be listed as relevant to the initial query, since several of the words in the

[5] http://www.eionet.europa.eu/gemet/concept/1471

Definition

The long-term fluctuations in temperature, precipitation, wind, and all other aspects of the Earth's climate. External processes, such as solar-irradiance variations, variations of the Earth's orbital parameters (eccentricity, precession, and inclination), lithosphere motions, and volcanic activity, are factors in climatic variation. Internal variations of the climate system, e.g., changes in the abundance of greenhouse gases, also may produce fluctuations of sufficient magnitude and variability to explain observed climate change through the feedback processes interrelating the components of the climate system.

Figure 3: *Climatic change* Semantic Annotation

query appear also in the text of the documents. Yet, D_2 and D_3 do not refer to the topic of *climate change*. By using the concept-based index, only the first document will be listed in response to the query providing better precision. Examples such as this one, justify the cost of adding a new lexical level to an existing SKOS resource.

4. ADDING A LEXICAL LEVEL TO A SKOS RESOURCE

4.1 Development Setup

The following list describes the main components used in implementing and testing the algorithms to be described below.

- Search Engine for CLIR: Terrier IR Platform[6]
- Relevant Java Libraries: skosapi[7],
- Natural Language Processing: GATE[8] Embedded is an object-oriented framework for performing Semantic Annotations tasks; APOLDA a GATE Plugin[9]
- Semantic repository: Virtuoso Universal Server[10]
- Other Resources: English Wikipedia
- Translation Service: GoogleTranslate
- research-esa[11] an implementation for Explicit Semantic Analysis

4.2 Enriching existing SKOS resources using a self-reflection algorithm

This is the description of the algorithm we devised to enrich a given SKOS resource with a set of annotations that are automatically translated with minimal errors in the following algorithm. The minimum requirement for running the algorithm below with a SKOS resource as input is that there exist definitions specifications in the selected dataset for concepts in at least one language. Our test implementations used English as the starting language and examples of the concepts identified for *climatic change* are described in Figure 3, while the phrases identified for the same concept are in Figure 4.

Algorithm 1 Generating Annotations

```
INPUT
KOS expressed using SKOS
for all $c$ SKOS Concept from the resource do
  Index concept definition content
  Create a unique list of the terms in the definition and
  their frequency.
  Semantic annotations
  Using GATE Embedded, tokenize $c$'s definition and
  identify exact occurrences of other concept labels (pre-
  ferred or alternate) in the definition
  Phrase extraction
  Using ESA and EN Wikipedia determine content-
  bearing phrases from $c$'s definition
  Extract groups of words from the definition that have
  a strong association
  The association function is based on the Language
  Model[12] metric
end for
OUTPUT
A set of annotations for each $c$ a SKOS Concept from the
resource
```

The semantic annotation part of the algorithm relies on APOLDA (Automated Processing of Ontologies with lexical Denotations for Annotation) Gate plugin [15] that determines annotations based on the SKOS-converted-to-OWL initial KOS resource. We have used the GEMET expressed in SKOS to OWL format, a representational transformation possible with any SKOS dataset since a SKOS concept is an instance of an owl:Class [2]. This produced 18120 mentions of the 5208 GEMET concepts throughout all its SKOS concept definitions.

The next annotation step, detects short phrases in a text (2, 3, or 4 words), based on the strength of their association computed by determining the semantic relatedness of their English Wikipedia feature vectors. The respective vectors are computed using research-esa implementation of the Explicit Semantic Analysis algorithm on a local instance of the English Wikipedia[13]. The approach provided good results in identifying generic phrases.

We run the above algorithm on GEMET and we have identified 15781 occurrences of 6850 unique content-bearing phrases and single words. These counts are provided after removing any duplicates with the annotations obtained at the previous step. The phrases include multiword-expressions (e.g *toxic chemical*, *oxygen concentration*, *wind velocity*), named entities (e.g *New Zealand*), and other phrases (e.g *pipes supplying water*). The single words that appeared several times within a concept's definition are part of this set of annotations. The automatically selected phrases have all been manually checked as valid atomic groupings of words (in terms of meaning). The semantic annotation sets are fed into the next algorithm for disambiguation.

4.3 Disambiguation

The task of disambiguating the semantic annotations from

[6] http://terrier.org/

[7] http://skosapi.sourceforge.net/

[8] http://gate.ac.uk/download/

[9] http://apolda.sourceforge.net/

[10] http://virtuoso.openlinksw.com/

[11] http://code.google.com/p/research-esa/

[13] http://en.wikipedia.org/wiki/Wikipedia:Database_download

Definition

The long-term fluctuations in temperature, precipitation, wind, and all other aspects of the Earth's climate. External processes, such as solar-irradiance variations, variations of the Earth's orbital parameters (eccentricity, precession, and inclination), lithosphere motions, and volcanic activity, are factors in climatic variation. Internal variations of the climate system, e.g., changes in the abundance of greenhouse gases, also may produce fluctuations of sufficient magnitude and variability to explain observed climate change through the feedback processes interrelating the components of the climate system.

Figure 4: *Climatic change* Phrase Identification

the previous algorithm is difficult and the results we have obtained show how the details specified for each SKOS concept impact the ability to determine if a concept is used in a piece of text in the same sense characterized by the SKOS resource. The algorithm involves pre-processing the content of a SKOS resource and extracting concepts' signatures. We than compute a metric of relatedness between a SKOS concept's definition and the concepts' signatures from the annotating concepts set.

Algorithm 2 Disambiguating Semantic Annotations

```
INPUT
KOS expressed using SKOS
for all c a SKOS Concept do
  Build c's concept signature
  for each of c's neighbors, narrower or broader concepts
  do
    Add the union of their annotations' labels to c's concept signature
  end for
end for
for all c a SKOS Concept do
  for all annotation identified for c and a SKOS concept
  do
    Compute the semantic relatedness between the annotation's concept signature and c's textual definition
  end for
end for
OUTPUT
Disambiguated set of semantic annotations
```

For the *climatic change* concept the annotation for *lithosphere* was removed. Judging by the SKOS concept definition this is not necessarily a bad annotation, but its removal has low impact in terms of SKOS Concept Indexing accuracy. We run this algorithm with a low threshold of semantic relatedness and found 530 annotations were removed. This is not surprising considering we have annotated the GEMET resource using concepts from the same resource. A third of the annotations were wrong and this happened consistently when there was no definition for that particular concept or the definition was very short and there were very few neighboring concepts to build the concept's signature. Another third of the annotations were borderline, like in the case of *lithosphere*. This reflected that the definition of the annotated concept just referred to the other concept, but only in this context. The other third of annotations were correct. For example *state* refers to a territory with an organized government, yet, many concepts refer to *state* as a state of matter. Overall, we believe the algorithm needs further refinements and testing on other collections where the nature of the language used is not necessarily similar to the language describing the concepts.

4.4 Multilingual Annotations

The final step in generating a multilingual dataset that links to the original SKOS dataset is described in this section. The annotations obtained after the disambiguation are serialized as RDF triples. The added triples are expressed using SKOS eXtension for Labels (SKOS-XL), which provides additional support for identifying, describing and linking lexical entities.

The SKOS data model described in [1] defines the property skosxl:labelRelation that links instances of skosxl:Label. It is an extension point, for which we define two object sub properties: *literalTranslation* and *domainTranslation*. The *literalTranslation* is used for handling the machine translation of a label using Google Translate Web Service, while *domainTranslation* is intended to link labels from different concept schemes, when there exists the transitive relation *exact match* between the concepts the labels refer to. We think *domainTranslation* is a useful extension to include, since several GEMET concepts have pointers to concepts in the bilingual UMTHES. We do not explore this further for now, but added it to our extensions. These two relations, capturing both translations and context, are in agreement with other work on representing translations for the Semantic Web [12].

We also define a third property, *annotation*. The latter is a sub property of skosxl:hiddenLabel and is used to express a link between the preferred label of a concept and the annotations identified previously from a SKOS concept's definition.

Algorithm 3 Serializing Multilingual Annotations

```
INPUT
KOS expressed using SKOS
for all c a SKOS Concept load its annotation maps do
  Source Language: en
  Target Languages: es, fr, ro
  for all annotation a SKOS based annotation do
    Generate label ID
    if c does not have a label for the target language
    then
      Translate
    end if
    Generate RDF triples description
  end for
  for all annotation a phrase annotation do
    Generate label ID
    Translate phrase
    Generate RDF triples description
  end for
end for
OUTPUT
A new RDF graph of lexical annotations resulted from SPARQL queries
```

Due to the output of algorithm 1 this last algorithm only translates phrases or single words. The machine translation outputs in some instances are words with the wrong inflections or the word order in the translated expression is not correct. In the particular case of CLIR, words get stemmed during indexing (for example, the word *climate* is stemmed to *climat*), thus wrong inflections and mixed word order does not affect the particular case of our application domain. By running this algorithm on GEMET we were able to enrich all concepts with relevant lexical details in Spanish, French and Romanian. All new annotations can be used to disambiguate concepts in other languages than English.

Figure 5 details a partial SPARQL query for creating the new lexicalizations dataset. We precede each skosxl:Label instance with the string *label* followed by a concept id. In the example query, the id number 1471 points to the *climatic change* concept in GEMET, while ids numbers 1462, 8366, 9327 match respectively *climate*, *temperature*, *wind*. We are using the original ids for creating a GEMET annotated dataset. Note, as expected from the two types of annotations we determined in algorithms 1 and 2, we are required to differentiate between the two annotations. For example, a label like *label_1471_1_en* describes the phrase annotation *climatic variation*, while *label_8366_0_en* describes the semantic annotation with concept *temperature*. By using part of algorithm 2 we can easily create multilingual concept signature in Spanish, French and Romanian. The algorithm in this section can easily be extended to support any number of target languages.

5. DISCUSSION

Throughout this article we provided a detailed example of customizing a SKOS resource for our work-in-progress prototype of a CLIR system. Existing examples of SKOS use cases[14] show that most research and development work has focused on creating vocabularies and applications that support editing them. We believe that more diverse applications can be built and we are trying to open the way for more classification systems to be shared in SKOS format. Though, we are not formally evaluating the prototype CLIR system as a whole within the scope of this article, it is important to note that at its core this prototype is a classic CLIR system with more flexibility in accessing language and knowledge resources.

Returning to the initial question *are SKOS concept schemes ready for multilingual retrieval applications*? We have uncovered that different processes require a level of lexicalization of a SKOS resources's content. This can be obtained automatically by applying the algorithms in Section 4. The cost of building the annotations set for a SKOS resource is computationally high, but we would like to publish after further testing the annotated dataset for GEMET under the Open Database License[15].

Proving that connecting a CLIR system to SKOS resources can lead to an improved retrieval system in tune with cultural changes of semantics is the broader research question we are investigating. Thus far, we have focused on understanding the relations between content and structure of a SKOS resource and its applicability to creating a CLIR ap-

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX skosxl: <http://www.w3.org/2008/05skos-xl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX gemet:<http://www.eionet.europa.eu/gemet/gemet-
skoscore.rdf#>

INSERT INTO <http://gemet-annotated> {
gemet:1471 a skos:Concept ;
 skosxl:prefLabel gemet:label_1471_0_en ;
 skosxl:altLabel gemet:label_1471_0_es ;
 skosxl:altLabel gemet:label_1471_0_ro ;
 skosxl:altLabel gemet:label_1471_0_fr .

gemet:label_1471_0_en a skosxl:Label ;
  skosxl:literalForm "climatic change"@en .
gemet:label_1471_0_es a skosxl:Label ;
  skosxl:literalForm "cambio climático"@es .
gemet:label_1471_0_fr a skosxl:Label ;
  skosxl:literalForm "changement climatique"@fr .
gemet:label_1471_0_ro a skosxl:Label ;
      skosxl:literalForm "schimbare climatică"@ro .

gemet:1471 gemet:annotation gemet:label_1462_0_en .
gemet:1471 gemet:annotation gemet:label_8366_0_en .
gemet:1471 gemet:annotation gemet:label_9327_0_en .

gemet:1471 gemet:annotation gemet:label_1471_1_en .
gemet:1471 gemet:annotation gemet:label_1471_1_es .
gemet:1471 gemet:annotation gemet:label_1471_1_ro .
gemet:1471 gemet:annotation gemet:label_1471_1_fr .

gemet:label_1471_1_en  gemet:literalTranslation
gemet:label_1471_1_es .
gemet:label_1471_1_en  gemet:literalTranslation
gemet:label_1471_1_ro .
gemet:label_1471_1_en  gemet:literalTranslation
gemet:label_1471_1_fr .
gemet:label_1471_1_en a skosxl:Label ;
  skosxl:literalForm "climatic variation"@en .
gemet:label_1471_1_es a skosxl:Label ;
  skosxl:literalForm "la variación climática"@es .
gemet:label_1471_1_ro a skosxl:Label ;
  skosxl:literalForm "climatice variaţie"@ro .
gemet:label_1471_1_fr a skosxl:Label ;
skosxl:literalForm "les variations climatiques"@fr .


gemet:1471 gemet:annotation gemet:label_1471_2_en .
gemet:1471 gemet:annotation gemet:label_1471_2_es .
gemet:1471 gemet:annotation gemet:label_1471_2_ro .
gemet:1471 gemet:annotation gemet:label_1471_2_fr .


gemet:label_1471_2_en  gemet:literalTranslation
gemet:label_1471_2_es .
gemet:label_1471_2_en  gemet:literalTranslation
gemet:label_1471_2_ro .
gemet:label_1471_2_en  gemet:literalTranslation
gemet:label_1471_2_fr .

gemet:label_1471_2_en a skosxl:Label ;
  skosxl:literalForm "processes"@en .
gemet:label_1471_2_es a skosxl:Label ;
  skosxl:literalForm "los procesos de"@es .
gemet:label_1471_2_ro a skosxl:Label ;
  skosxl:literalForm "procese"@ro .
gemet:label_1471_2_fr a skosxl:Label ;
  skosxl:literalForm "processus"@fr .

}
```

Figure 5: SPARQL query to serialize annotations and translations for the *climatic change*

[14] http://www.w3.org/2006/07/SWD/SKOS/reference/20090315/implementation.html

[15] http://opendatacommons.org/licenses/odbl/

plication. We described a set of algorithms on how to enrich a given SKOS resource with RDF triples that facilitate the concept-indexing stage of the CLIR prototype. Algorithms 1 and 2 can be applied for any text document not necessarily SKOS concepts' definitions. Also, we believe that the disambiguation process can be improved by using also the inter concept schemes relations such as *exact match* or close match to discover further lexical entities and improve a concept's textual signature.

6. REFERENCES

[1] SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL). `http://www.w3.org/TR/skos-reference/skos-xl.html`, March 2009.

[2] SKOS Simple Knowledge Organization System Reference. `http://www.w3.org/TR/skos-reference/`, February 2010.

[3] N. Calzolari. Initiatives, tendencies and driving forces for a lexical web as part of a language infrastructure. In T. Tokunaga and A. Ortega, editors, *Large-Scale Knowledge Resources. Construction and Application*, volume 4938 of *Lecture Notes in Computer Science*, pages 90–105. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-78159-2_10.

[4] C. Fellbaum and P. Vossen. Connecting the universal to the specific: Towards the global grid. In *Intercultural Collaboration I : Lecture Notes in Computer Science, Springer-Verlag*, 2007.

[5] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007.

[6] A. J. G. Gray, N. Gray, and I. Ounis. Searching and exploring controlled vocabularies. In *ESAIR '09: Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 1–5, New York, NY, USA, 2009. ACM.

[7] N. Guarino and P. Giaretta. Ontologies and Knowledge Bases: Towards a Terminological Clarification. *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pages 25–32, 1995.

[8] D. He and J. Wang. *Information Retrieval: Searching in the 21st Century*, chapter Cross-Language Information Retrieval. John Wiley & Sons, 2007.

[9] D. Lenat. *The Dimensions of Context-Space*. Cycorp, 1998.

[10] G.-A. Levow, D. W. Oard, and P. Resnik. Dictionary-based techniques for cross-language information retrieval. *Information Processing Management*, 41(3):523–547, 2005.

[11] A. Miles and D. Brickley. SKOS Core Guide. World Wide Web Consortium, Working Draft WD-swbp-skos-core-guide-20051102, November 2005.

[12] E. Montiel-Ponsoda, J. Gracia, G. Aguado-De-Cea, and A. Gómez-Pérez. Representing translations on the semantic web. In *The 10th International Semantic Web Conference*, October 2011.

[13] V. Petras, N. Perelman, and F. C. Gey. Using thesauri in cross-language retrieval of german and french indexed collections. In *CLEF*, pages 349–362, 2002.

[14] U. P. School and U. Priss. Lattice-based information retrieval. *Knowledge Organization*, 27:132–142, 2000.

[15] C. Wartena, R. Brussee, L. Gazendam, and W.-O. Huijsen. Apolda: A practical tool for semantic annotation. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications*, DEXA '07, pages 288–292, Washington, DC, USA, 2007. IEEE Computer Society.

[16] W. A. Woods. Conceptual indexing: A better way to organize knowledge. Technical report, Mountain View, CA, USA, 1997.