Enabling Semantic Search for EO Products: an Ontology Matching Approach*

M. Karpathiotaki[1], K. Dogani[1], and M. Koubarakis[1]

National and Kapodistrian University of Athens, Greece
{mkarpat,kallirroi,koubarak}@di.uoa.gr

Access to Earth Observation (EO) products remains difficult for end-users. To address this, we developed the Prod-Trees platform[1][2], a semantically enabled search engine for EO products. Users guide their search through a number of ontologies related to EO domain. To facilitate users in finding terms that fit better to their needs, we created mappings between these ontologies. In this paper, we present Pythia, an ontology matching system that utilizes and combines various matching techniques [1,3,4] to create mappings between two ontologies.

Pythia is a combination of a string-based technique utilizing Apache Lucene's features, a language-based technique based on WordNet, and a graph-based technique that uses the structure of the ontology and the mappings produced by the two previous techniques. The system supports SKOS ontologies. Therefore, the mappings are also expressed in SKOS using the defined properties for matching concepts: *skos:exactMatch*, *skos:relatedMatch*, *skos:broadMatch*, and *skos:narrowMatch*. Based on these, we create four different types of mappings.

A **terminological matcher** is responsible for implementing the string- and language-based techniques, both applied on the concepts labels (*skos:prefLabel*, *skos:altLabel* and *skos:hiddenLabel*). The mappings created by this component can either be *skos:exactMatch* or *skos:relatedMatch*.

The **string-based technique** uses Lucene for indexing and searching. With Lucene, one can create documents and add fields of a specific type to these documents. When searching the documents, the user can specify which field he wants to search. Taking advantage of Lucene capabilities, the terminological matcher indexes the target ontology. A new document is created for each concept and each available property of the concept is added as a new field. String normalization functions are applied to the field and unnecessary stop words are removed.

When searching for concepts similar to concept A (from the source ontology), the prefLabel, altLabel, and hiddenLabel fields of the indexed ontology are searched using the prefLabel of concept A. The search results fetched back, are ranked according to the string similarity of the compared strings (e.g., *skos:prefLabel* of A and the prefLabel field of a document). This is feasible due to the string similarity functions implemented in Lucene. Also, since each field is indexed, only the index of the specified field is searched, and not all the concepts.

Lucene returns multiple related results. If the two strings are the same, a *skos:exactMatch* is created between A and the corresponding concept from the

* This work was supported by the Prod-Trees project funded by ESA ESRIN.

[1] A video demonstrating the functionalities of the Prod-Trees platform is available at http://bit.ly/ProdTreesPlatform.

target ontology. Otherwise, and only if one string is a substring of the other (e.g., "Elevation" and "Digital Elevation Model"), a *skos:relatedMatch* is created.

The **language-based technique** uses WordNet, a lexical database for English. The technique is optional and can be bypassed, as it adds noise to the results. Putting WordNet to use, a new field, called *relLabel*, is created in the Lucene document of each concept. *relLabel* enhances each concept's labels, by adding synonyms and other related words found in WordNet. During the search, the relLabel fields of the documents are searched, and if a similarity is discovered, a *skos:relatedMatch* relation is created between the corresponding concepts.

In case there are concepts from the source ontology with no *skos:exactMatch* mappings, a **structural matcher** is invoked. This component implements a graph-based technique creating either *skos:narrowMatch* or *skos:broadMatch* mappings. Taking as input a concept A from the source ontology, the matcher finds all the broaders and narrowers of A. Afterwards, it checks whether a *skos:exactMatch* was created by the terminological matcher for one of these concepts. If it did, then a new mapping can be derived. For example, if a *skos:exactMatch* exists between concept B (which is a broader of A) and concept B'(from the target ontology), then it can be derived that B' will be a *skos:broadMatch* of A. Similarly, we can create a *skos:narrowMatch*.

The matcher also checks whether the concepts B and N hold *skos:narrowMatch* or *skos:broadMatch* relations with concepts from the target ontology. If a *skos:broadMatch* exists between B and a concept B", then it is safe to conclude that B" will also be a *skos:broadMatch* of A. This means that when a *skos:broadMatch* exists between a concept B from the source ontology and a concept B" from the target ontology, then this relation can be propagated to concept's B narrowers. Similarly, a skos:narrowMatch between a concept N and a concept N", can be propagated to concept's N broaders. In any other case, no mappings can be derived. When all the concepts are examined, if new mappings were created by the structural matcher, the described process is repeated. Otherwise, Pythia proceeds with the exportation of the mappings to RDF.

Despite the simplicity of the techniques, the results are quite satisfying. Especially, the performance of the language-based technique, which allows tuning WordNet. By stating the types of relations WordNet discovers for a given word, it gives control over the percentage of valid mappings. A higher degree of trust for the final results can be gained with extensions such as a user-evalutation process and the use of domain-specific vocabularies coupled with Wordnet.

References

1. Euzenat, J., Shvaiko, P.: Ontology Matching (2007)
2. Karpathiotaki, M., et. al.: Prod-Trees: Semantic Search for Earth Observation Products. In: ESWC. LNCS, Springer (2014)
3. Nagy, M., Vargas-Vera, M.: Towards an Automatic Semantic Data Integration: Multi-agent Framework Approach. In: Semantic Web. InTech (2010)
4. Pirro, G., Talia, D.: An approach to Ontology Mapping based on the Lucene search engine library. DEXA '07 (2007)