# A Thesaurus and an Application Ontology for the Juvenile Arthritis Domain

**Ernesto Jiménez-Ruiz,**
Rafael Berlanga-Llavori,
Victoria Nebot
Universitat Jaume I
Castellon, Spain
ejimenez,berlanga,romerom@lsi.uji.es

**Antonio Jimeno-Yepes,**
Dietrich Rebholz-Schuhmann
European Bioinformatics Institute
Hinxton, Cambridge, UK
yepes,rebholz@ebi.ac.uk

## Abstract

This paper is intended to present our experiences in the creation of a light weight thesaurus for the Arthritis domain and the reuse of this terminological resource to create an ontology to classify patients suffering different subtypes of Rheumatoid Arthritis, which is part of the Health-e Child project.

## 1 Introduction

Ontologies are intended to formally represent knowledge, therefore one of their goals is to provide a way to facilitate inferencing, that is, to get new conclusions when facing new facts. On the other hand, thesauri should be focused on the organization of terms within a narrower-broader taxonomy (e.g. hypernymy and hyponymy) or part-whole taxonomy (e.g. meronymy and holonymy), and additionally the collection of synonyms for the entries (e.g. synsets) contained in it.

Thesauri and ontologies, although different, may live together and complement to each other. In (Jimeno-Yepes et al., 2009b) we analysed the potential benefits of having a reference thesaurus within the ontology lifecycle, moreover, we presented a method to reuse and adapt current terminological resources to build a thesaurus for the Arthritis domain, namely **HeCTh**. In this paper we focus on the creation of **HeCTh** and an application ontology (**JIAO**) that makes use of this thesaurus. **HeCTh** aims at providing the necessary terms, their synonyms, and their basic taxonomic relationships. Instead, **JIAO** is intended to characterize and classify patients suffering different subtypes of Juvenile Idiopathic Arthritis (JIA), a disease being studied in the Health-e Child (HeC) project[1].

## 2 Method Overview

In this section we present the steps we followed to create **HeCTh** and **JIAO**. We have reused and filtered the Unified Medical Language System Metathesaurus (UMLS-Meta)[2], Swissprot[3] and DrugBank[4] in order to extract the necessary terms and relationships to populate **HeCTh**. Next, we briefly comment the phases of our method.

- *Vocabulary extraction*. Domain terms were extracted from a set of medical protocols (Berlanga et al., 2008) and text resources from the literature (Jimeno-Yepes et al., 2008). UMLS-Meta , Swissprot and DrugBank were used to annotate terms within protocols and text. Together with the automatic annotation techniques, manual intervention was also necessary. As a result a flat vocabulary (Jimeno-Yepes et al., 2009a) linked to the domain thesauri was obtained.

- *Thesaurus Conformation*. This step is intended to give an organization to the extracted flat vocabulary. UMLS-Meta has been used as basis to provide such organization. Due to the overload of information and consistency issues, we have adopted a *fragment extraction method* (Nebot and Berlanga, 2009) in order to retrieve only the desired taxonomic-relationships. Additionally, information about synonyms, origin resource, and semantic scope has been added.

[1] http://www.health-e-child.org

[2] http://www.nlm.nih.gov/research/umls

[3] http://www.expasy.ch/sprot/

[4] http://www.drugbank.ca/

Proceedings of the 3rd International Symposium on Languages in Biology and Medicine,
Posters, pages 120-122, Jeju Island, South Korea, 8-10 November 2009.

- *Ontology Creation*. Ontology languages such as OWL provide operators to give a formal conceptualization of the domain, but they still require to associate a proper label to each concept. Thesauri in general, and **HeCTh** in particular, will serve as label provider for ontologies. Moreover they will provide a potential organization for the concepts.

## 3 HeCTh SKOS Thesaurus

We have created **HeCTh** (Jimeno-Yepes et al., 2009a), a thesaurusfor the Arthritis domain in the HeC project. **HeCTh** aims at containing a clear and not overloaded organization of terms, with lexical information (i.e. synonyms) and scope information (semantic group). We have adopted SKOS[5], an RDF-like language, to represent **HeCTh**.

UMLS-Meta relationships were reused and filtered to build our customized term organization. The process to obtain **HeCTh** was split in these phases:

- *Extraction*. Given a set of terms satisfying the requirements (e.g. medical protocols, internal reports) of the application that is intended to use the thesaurus, we apply the fragment extraction method described in (Nebot and Berlanga, 2009). Briefly, this method encodes the whole UMLS-Meta with an interval index scheme that allows fast reconstructions of the hierarchies where the selected terms participate. Additionally, a selection of common and representative ancestors of the required terms is performed in order to organize these terms within the fragment.

- *Fragment Repair*. The obtained fragment can contain undesired narrower-broader relationships. For example *Cell* has *Disease* as broader concept, among others. This is mainly due to wrong matchings between the thesauri integrated within UMLS-Meta, which can imply cycles or undesired classifications. Fortunately, UMLS-Meta associates a *semantic type* to each term, such that terms can be grouped within *semantic groups* (Bodenreider and McCray, 2003). We have used these groups to establish a compatibility criteria between terms and semantic types, that is, two terms can keep a broader-narrower relationship provided that both terms belong to the same semantic group (i.e. they have compatible semantic types).

- *Term Grouping*. The terms that have not been classified under another term can be considered as *roots* since they do not have a broader term; however, in some cases they do not represent root terms. This lack of classification may be due to failures in the UMLS-Meta term organization or an incomplete domain vocabulary selection. In order to alleviate these potential problems we have defined a set of preferred *top terms* and we have associated to each UMLS-Meta semantic type one of these top terms. Thus, if a term has not broader terms and it is not included within the set of top terms, then that term is organized under one of the top terms, according to its semantic type.

- *Completion*. Swissprot and Drugbank specialised vocabularies were used to complete, with new entries and additional synonyms, the fragment extracted from UMLS-Meta. These vocabularies lack of a rich classification scheme therefore new terms (not included in UMLS-Meta) were associated a semantic type (e.g. Protein, Drug) and were organized within **HeCTh** using the *term grouping method* commented above.

As a result **HeCTh** thesaurus contains 816 terms (786 coming form UMLS-Meta, 48 from SwissProt and 22 from DrugBank), organized in a nine-roots tree using 1135 broader relationships, and with a maximum depth of 11 levels. Moreover **HeCTh** comprises 6097 term labels, at least one semantic scope label (i.e. semantic type) per term, and links to UMLS-Meta, SwissProt and DrugBank identifiers.

## 4 Towards an Application Ontology

Currently we are developing **JIAO** (an excerpt can be downloaded from (Jimeno-Yepes et al., 2009a)), an application ontology to classify patients suffering a kind of arthritis called *Juvenile Idiopathic Arthritis* (JIA), for which there is not yet a consensus about its classification nor even its name. We have adopted the ILAR classification, since it is the classification

[5] http://www.w3.org/2004/02/skos/

criteria used in the HeC project, which proposes eight subtypes for JIA.

We have adopted the revision 2 of OWL[6] as ontology language to represent **JIAO**. **JIAO** allows us to classify patients given a set of facts and symptoms. The use of rich ontology-like representations have provided us more semantic expressiveness than thesauri in order to characterize complex concepts and express relationships between them. Nevertheless, thesauri play a very important role (Jimeno-Yepes et al., 2009b) in the lexical (e.g. concept labeling) and syntactic (e.g. preliminary concept organization) phases of the ontology life cycle.

**JIAO** benefits from **HeCTh** in several ways: (1) **HeCTh** provides **JIAO** with labels for the concepts. (2) **JIAO** does not need to care about global concept organizations (e.g. joints, drugs, etc.) so that it can focus on specific conceptualizations (e.g. JIA patients and symptoms). (3) Indeed, **JIAO** can propose an alternative organization of concepts and then contrast if such organization is compatible with the thesaurus. On the other hand, **HeCTh** can also profit from **JIAO**'s evolution and incorporate specific terms not considered previously (e.g. specific subtypes of JIA).

**JIAO** concepts are annotated with **HeCTh** terms through their identifiers, moreover the adopted concept label has been reused from **HeCTh**. Notice that, **HeCTh** terms are organized within a hierarchy using explicitly the *broader* property, whereas ontology concepts are defined by means of axioms and the concept hierarchy could be implicitly defined through them. As commented, the thesaurus organization can be reused to complement the ontology concept hierarchy, in this way, *broader properties* in the thesaurus can be translated to *subclass properties* (i.e. *is-a* properties) within the ontology. This reuse of thesaurus knowledge may have side effects which can help the ontology engineer to detect errors or conflicts in the development. For example, in **HeCTh**, *Juvenile spondyloarthropathy* has both *Ankylosing Spondylitis* and *Juvenile idiopathic arthritis* as broader terms; this knowledge could be conflictive and may cause non-intended consequences if the ontology engineer declares *Ankylosing Spondylitis* and *Juvenile idiopathic arthritis* ontology concepts as *disjoint*.

[6] http://www.w3.org/2007/OWL/wiki/Syntax

## Acknowledgments

The authors wish to thank the EU project Health-e-Child (IST 2004-027749) for providing us the application domain. This work has been partially funded by the Spanish National Research Program (contract number TIN2008-01825/TIN), the PhD Fellowship Program of the *Generalitat Valenciana*.

## References

Rafael Berlanga, Ernesto Jimenez-Ruiz, Victoria Nebot, David Manset, Andrew Branson, Tamas Hauer, Richard McClatchey, Dmitry Rogulin, Jetendr Shamdasani, Sonja Zillner, and Joerg Freund. 2008. Medical data integration and the semantic annotation of medical protocols. *IEEE Symposium on Computer-Based Medical Systems*, pages 644–649.

Olivier Bodenreider and Alexa T. McCray. 2003. Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics*, 36(6):414 – 432. Unified Medical Language System.

Antonio Jimeno-Yepes, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(Suppl 3):S3.

A. Jimeno-Yepes, E. Jimenez-Ruiz, R. Berlanga, and D. Rebholz-Schuhmann. 2009a. Health-e-child terminological resources: Vocabulary, Thesaurus (HeCth) and Ontology. http://krono.act.uji.es/people/Ernesto/hec-thesaurus.

A. Jimeno-Yepes, E. Jiménez-Ruiz, R. Berlanga-Llavori, and D. Rebholz-Schuhmann. 2009b. Reuse of terminological resources for efficient ontological engineering in Life Sciences. *BMC Bioinformatics*, 10(Suppl 10):S4.

Victoria Nebot and Rafael Berlanga. 2009. Efficient retrieval of ontology fragments using an interval labeling scheme. *Information Sciences*, August.