Multiterminology cross-lingual model to create the European Health Terminology/Ontology Portal

Julien Grosjean Tayeb Merabti Nicolas Griffon Badisse Dahamna Stefan Darmoni
CISMeF & TIBS LITIS EA 4108, Rouen University Hospital, Cour Leschevin, Porte 21,
3ème tage, 1 Rue de Germont, 76031 Rouen Cedex, France
julien.grosjean@chu-rouen.fr

Abstract

The European Health Terminology/Ontology Portal (EHTOP) is a repository dedicated to European health professionals and students. Currently, it provides access to thirty two health terminologies and ontologies available mainly in French or in English, but also German, Italian and Dutch. EHTOP can be used by humans and by computers via Web services. The main objective of EHTOP is to provide an access to terminologies and ontologies, allowing dynamic browsing and navigation. Methods: To integrate terminologies and ontologies into EHTOP, three steps are necessary: (1) designing a meta-model into which each terminology and ontology can be integrated, (2) developing a process to include terminologies into EHTOP, (3) building and integrating existing and new inter & intra-terminology semantic harmonization into EHTOP. Results: EHTOP is available freely for the MeSH in French (URL: pts.chu-rouen.fr). The access to other terminologies and to other languages is restricted and available only for the scientific community. A total of 32 terminologies are included into EHTOP, with 980,000 concepts, 2,300,000 synonyms, 222,800 definitions and 4,000,000 relations. Twenty one of these terminologies are not included yet in the UMLS among them, some from the World Health Organization. Since January 2010, EHTOP is daily used by CISMeF librarians to index health resources in the CISMeF catalogue in a multi-terminology mode. Currently, 600 unique machines are using the MeSH version of EHTOP, whereas 200 are already registered for its extended version. More recently, the multilingual version of EHTOP is available (URL: http://cispro.chu-rouen.fr/ehtop_site/) and freely provides access to ICD10 and FMA in five languages. Conclusion: EHTOP is a rich tool, useful for a wide range of applications and users, whatever in education, resources indexing, information retrieval or performing audits in terminology management.

1 Introduction

The Internet is currently the major source of scientific and health information and knowledge. If the generally in English, health information for lay people is available in each language. Nonetheless, people around the world is more and more travelling, health information should transcend borders and should become multilingual and based on several health terminologies and ontologies. Some institutions are already proving health information in several languages: e.g. MEDLINEplus is providing health information for lay people in English and Spanish, whereas Europe Medicine Agency is providing drug information for health professionals and lay people in each European language. Health is, with Law, the main scientific field, where coexist several terminologies and ontologies (T/O). For the English languages, over 150 terminologies and classifications are included in the Unified Medical Language System (UMLS) Bodenreider [2004] meta-thesaurus developed by the US National Library of Medicine since 1986. There is an increasing amount of interest today not only in developing and maintaining healthcare T/O but also in making them interoperable within information technology (IT) systems delivering services to applications. Terminology

Short papers of the 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 119–122
Paris, 8–10 November 2011

server has been defined as a tool to manage and to give access to various terminologies Burgun et al. [1997]. Several terminology servers have already been developed, mostly in English Burgun et al. [1997], Navas et al. [2007]. One was recently developed for French terminologies Darmoni et al. [2009]. The principal aim of this work was to create an Health Multi-Terminology Multi-Lingual Portal mainly based on European languages (EHTOP) and connected to the CISMeF information system. The primary goal of EHTOP was to search concepts among all the health terminologies available in French (or in English and translated in French) included in this portal and to browse it dynamically. The ultimate goal was to use this search: (a) to index resources manually or automatically in the CISMeF quality-controlled health gateway (Catalog and Index of Health Resources in French-Fr) [URL:`http://www.cismef.org`] Darmoni et al. [2000]; (b) to allow multi-terminology automatic indexing and information retrieval; (c) to evaluate the integrity of terminological data (audit); (d) to provide a new tool to train health students.

1.1 Material and methods

1.1.1 List of terminologies included in EHTOP

The following terminologies and classifications were included in the CISMeF Information System (n=32), and therefore in EHTOP. Some terminologies and classifications are included in the UMLS meta-thesaurus (n=11) but most are not (n=21): MeSH thesaurus (including the MeSH Supplementary Concepts and the translation of 15,300 MeSH SC in French and the add-on of over 16,000 synonyms to MeSH terms), CISMeF thesaurus (extension to the MeSH thesaurus, including 130 metaterms), SNOMED International (French version) to describe electronic health records Côté et al. [1993], Terminologies developed by the World Health Organization (WHO): ICD10 (International Classification of Diseases, 10th revision)[1], WHO-ART (Adverse Reactions Terminology), for adverse effects[2], WHO-ATC (Anatomical Therapeutic Chemical Classification System)[3] for drugs, WHO-ICPS (International Classification for Patient Safety)[4], WHO-ICF (International Classification of Functioning, Disability and Health)[5], ICPC2 [6], ORPHANET Aymé et al. [1998], MedlinePlus Topics Miller et al. [2000], IUPAC (International Union of Pure and Applied Chemistry) for chemical sciences[7], LOINC (Logical Observation Identifiers Names and Codes) Cormont et al. [2011], FMA (Foundational Model of Anatomy) Rosse and Mejino [2003]. Some other terminologies and ontologies will be integrated in the coming months, in particular US National Cancer Institute Terminology.

1.1.2 Integration of the terminologies

To integrate the terminologies in the CISMeF Information System (Oracle 11.1g database), three steps are necessary: a) to design a terminology generic model into which each terminology model can be integrated. b) to design a process capable of integrating terminologies into the EHTOP. c) to build and integrate intra & inter-terminology semantic harmonization into EHTOP. Two inter-terminology mappings were performed: one based on UMLS concepts and one based on NLP tools developed by the CISMeF team Merabti [2010]. A generic model was designed for the database in order to fit all the terminologies into one global structure (see Figure 1): this database is the CISMeF BackOffice. Then, a model of each terminology was designed as a specialization of the meta-model. With the specific models, the work consisted to develop a parser for each terminology: the input is the original data (or normalized original data) and output is an OWL file. As data could be in different shapes and structures, in some cases additional processes were performed (temporary databases, files etc.).

The final stage is the integration of the OWL files into the BackOffice. A generic parser was developed to directly insert each terminology into the database. A special model was designed to represent each terminology in a "CISMeF BackOffice view" so that the parser can use this model in input to recognize descriptor classes, definitions, synonyms, relations in order to insert it very

[1] `www.who.int/classifications/icd/en/`
[2] `www.umc-products.com`
[3] `http://www.whocc.no/atcddd/`
[4] `www.who.int/patientsafety/implementation/taxonomy/developement\_site/en/index.html/`
[5] `www.who.int/classifications/icf/en/`
[6] www.who.int/classifications/icd/adaptations/icpc2/
[7] `http://www.iupac.org`

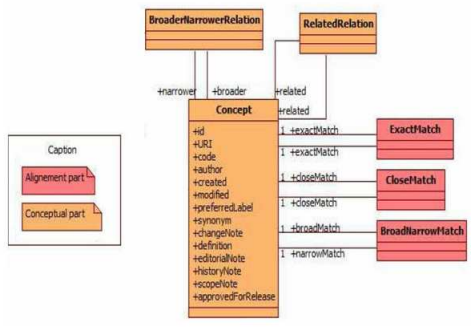

Figure 1: The CISMeF BackOffice database conceptual structure

easily into the database.

1.1.3 Creation of the EHTOP

The EHTOP was designed as a graphic interface of a Web Service, entirely dedicated to information retrieval and associations between terms of several terminologies. Thus, the main objective was to dissociate the substance from the form, in particular the interface. This Web Service was the most important part of the task: retrieve information and major schemes to allow the fullest display in the HTMP interface. The EHTOP Web Service has been developed to respect Web Services Standards with SOAP (Simple Object Access Protocol) and WSDL (Web Service Description Language) signatures. It presents some methods to search terms by descriptor or by database unique identifier. As the EHTOP exploits a SKOS file (RDF), the graphic interface that renders the final HTML was build based on JSP (Java ServerPages) files including XSL (eXtensible Stylesheet Language) functions. The final output (XHTML) deals with W3C standards. For optimal performance, a special AJAX (Asynchronous JavaScript And XML) method is implemented.

2 Results

Currently, the CISMeF team is using one junior engineer (JG) to integrate new terminologies (e.g. SNOMED CT) and one post-doc (TM) to perform semantic harmonization on each terminology to another (more than 900 alignments were already performed -322-) using CISMeF NLP tools. Currently, two versions of EHTOP exist: (a) the first version is mainly bilingual (French and English) specifically devoted to French users. This version is available at the following URL: `http://pts.chu-rouen.fr/`. Only MeSH and CISMeF terminologies are freely available. We provide a restricted access only for the scientific community (click on “Sign in”); (b) the second version is multilingual URL: `http://cispro.chu-rouen.fr/ehtop\_site/`. A total of 32 terminologies are included into EHTOP, with 980,000 concepts, 2,300,000 synonyms, 222,800 definitions and 4,000,000 relations. Since January 2010, EHTOP is daily used by CISMeF librarians to index in multi-terminology mode. Since March 2010, the bilingual version is daily used by 500 unique machines. Three hundred people have already registered to access other T/O, mainly physicians, health students, librarians and translators.

3 Discussion

In 2011, the top 3 EHTOP targets are librarians, health professionals and health students to learn how to manipulate health terminologies and to extract knowledge from it, in particular from hierarchies and relations. Via its Web services, EHTOP may also be used by several interactive applications. To the best of our knowledge, the EHTOP is the first of its kind for French. The main added value of EHTOP when compared to any UMLS browser McCray and Razi [1995] is the possibility to access the main health terminologies in French or the multi-lingual terminologies and classification coming from WHO, which are not (yet) included in the UMLS (e.g. ATC for drugs or ICPS for patient safety), as demonstrated in accessing ICD10 in five languages. Currently, the EHTOP is a necessary basic tool to index any document in a multi-terminology multilingual mode. Other portals propose to search and navigate T/O such as NCBO Bioportal Noy et al. [2009] and the EBI Ontology Lookup Service Cote et al. [2006]. Those tools are also very friendly but do not allow users to navigate through terms or search among synonyms in different languages. They are also not adapted to a daily use to index. The HMTP as been evaluated by some medicine student groups and gave 58Even if the HMTP web service does not deal with the HL7/CTS specification, it could be an interesting perspective to implement it in order to be compliant with other

terminological providers. It would be also convenient to get responses from other similar portals such as NCBO Bioportal, UMLS browser or EBI Ontology Lookup Service to enhance our results and to provide the best possible service to users.

4 Conclusion

A health multi-terminology portal is a valuable tool to help to index and retrieve resources from a quality-controlled health gateway. It can also be very useful for teaching or performing audits in terminology management.

Acknowledgements

EHTOP was supported by several grants: PSIP project; (Patient Safety through Intelligent Procedures in medication -FP7-ICT-2007-); URL: `http://www.psip-project.eu/`; InterSTIS project (ANR-07-TECSAN-010); URL: `http://www.interstis.org/`; ALADIN project (ANR-08-TECS-001); URL: `http://www.aladin-project.eu/`; L3IM project (ANR-08-TECS-00); URL: `http://projet4-limbio.smbh.univ-paris13.fr/`; PlaIR project, funded by FEDER; URL: `http://www.plair.org`. The authors thank Richard Medeiros for his advice in the editing of this manuscript and the eight students of the INSA Rouen Engineering School that partially developed the multi-terminology portal.

References

S Aymé, B Urbero, D Oziel, E Lacouturier, and AC Biscarat. Information on rare diseases: the ORPHANET project. *Rev Med Interne*, 19(Suppl 3):376S–377S, 1998.

O Bodenreider. The Unified Medical Language System (umls): Integrating biomedical terminology. *Nucleic Acids Res*, 32:267–270, 2004.

A Burgun, P Denier, O Bodenreider, G Botti, D Delamarre, B Pouliquen, P Oberlin, JM Lévéque, B Lukacs, F Kohler, M Fieschi, and PBergun1 Le Beux. A web terminology server using umls for the description of medical procedures. *J Am Med Inform Assoc*, 4(5): 356–363, 1997.

S Cormont, PY Vandenbussche, A Buemi, J Delahousse, E Lepage, and J Charlet. Implementation of a platform dedicated to biomedical analysis terminologies management. In *AMIA Annu Symp Proc*, 2011.

R A Côté, D J Rothwell, J Patolay, R Beckett, and L Brochu. The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International., 1993.

RG Cote, P Jones, R Apweiler, and H Hermjakob. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7(1):97, 2006.

SJ Darmoni, JP Leroy, B Thirion, F Baudic, M Douyère, and J Piot. CISMeF: a structured health resource guide. *Meth Inf Med*, 39(1):30–5, 2000.

SJ Darmoni, M Joubert, B Dahamna, J Delahousse, and M Fieschi. SMTS: a French Health Multi-Terminology Server. In *Proc. AMIA Symp. 2009*, page 808, 2009.

AT McCray and A Razi. The UMLS knowledge source server. In *MedInfo*, volume 8 Pt 1, pages 144–7, 1995.

T Merabti. *Methods to map health terminologies: contribution to the semantic interoperability between health terminologies*. PhD thesis, University of Rouen, 2010.

N Miller, E.M Lacroix, and J.E Backus. MEDLINEplus: building and maintaining the national library of medecine's consumer health web service. *Bull Med Libr Assoc*, 88(1):11–7, 2000.

H Navas, AL Osornio, A Baum, A Gomez, D Luna, and FG de Quiros. Creation and evaluation of a terminology server for the interactive coding of discharge summaries. *Stud Health Technol Inform*, 129(Pt1):650–654, 2007.

N.F Noy, N.H Shah, P.L Whetzel, B Dai, M Dorf, N Griffith, C Jonquet, D.L Rubin, M.A Storey, C.G Chute, and M.A Musen. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37:170–173, 2009.

C Rosse and J.J Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36:478–500, 2003.