Integrating the Human Phenotype Ontology into HeTOP terminology-ontology server

Julien Grosjean[a], Tayeb Merabti[a], Lina F. Soualmia[a,b], Catherine Letord[a], Jean Charlet[b], Peter N. Robinson[c,d], Stéfan J. Darmoni[a,b]

[a] *CISMeF & TIBS, LITIS EA 4108, Rouen University Hospital, Rouen, France*
[b] *INSERM, Unité Mixte de Recherche en Santé (UMR_S) 872, équipe 20, Paris, France*
[c] *Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany*
[d] *Berlin-Brandenburg Center for Regenerative Therapies, Germany*

Abstract

Background: The Human Phenotype Ontology (HPO) is a controlled vocabulary which provides phenotype data related to genes or diseases. The Health Terminology/Ontology Portal (HeTOP) is a tool dedicated to both human beings and computers to access and browse biomedical terminologies or ontologies (T/O). Objective: The objective of this work was to integrate the HPO into HeTOP in order to enhance both works. Methods: The HPO has been conceived as a formal ontology, interoperable with other vocabularies and databases. The HeTOP multi-terminology and cross-lingual metamodel is a powerful tool to integrate any T/O into its system. Results: The integration of the HPO into HeTOP is a success and allows users to search and browse the HPO with a dedicated interface. Furthermore, the HPO has been enhanced with the addition of content such as new synonyms, translations, mappings. Conclusion: Integrating T/O such as the HPO into HeTOP is a benefit to vocabularies because it allows enrichment of them and it is also a benefit for HeTOP which provides a better service to both and machines.

Keywords: *Terminology; Phenotype; Distributed Systems; Rare Diseases.*

Introduction

Importance of an ontology of human phenotypes

The Human Phenotype Ontology (HPO) is a standardized, controlled vocabulary, which allows phenotypic information to be described in an unambiguous fashion in medical publications and databases [1]. The HPO is freely available at http://www.human-phenotype-ontology.org.

In the field of human genetics, several sources of information and terminologies have already been developed: (a) the Online Mendelian Inheritance in Man (OMIM) database (n=19,794) [2], which was developed at Johns Hopkins University; (b) the London Dysmorphology Database (LDDB) [3]; (c) POSSUM from Australia [4]; and (d) Orphanet originally from France [5], and now extended to Europe with a terminology available in five languages (English, French, Spanish, Italian and Portuguese). Several evaluation studies have been performed on most of these sources of information and terminologies [6].

The more recent developments of terminology and ontology (T/O) browsers or portals (e.g. UMLS browser [7] and BioPortal [8]) led our team to develop a multi-terminology cross-lingual portal: the Health multi-Terminology and Ontology Portal (HeTOP) [8]. The HeTOP is available at http://www.hetop.eu/. The access is restricted for most of terminologies and ontologies, but it is freely provided for academic scientists, health professionals and students. The HeTOP integrates terminologies and ontologies that are not included in UMLS (mostly French T/O, e.g. CCAM for medical procedures but also WHO terminologies such as ATC, ICPS). Further, HeTOP has cross-lingual functionalities, including 23 languages (e.g. French, German, Dutch, Danish but also non-Latin alphabets such as Russian or Chinese).

The objective of this work is to describe the integration of the HPO into the HeTOP terminology-ontology portal and its applications for terminologists, librarians, and medical students and how such integration can contribute to enhance the quality of the information.

Methods

Orphanet ontology

Orphanet is a reference information portal on rare diseases and orphan drugs, for healthcare professionals and for general audiences. This unit is led by a European consortium of around 40 countries, coordinated by a French team which is responsible for the infrastructure of Orphanet, management tools, quality control, rare disease inventory, classifications and the production of the encyclopedia. After ten years of evolution, previous Orphanet tools could not support efficiently the edition, update and data sharing processes of a constantly growing of rare diseases knowledge (6,000 rare diseases with annotations and more than a hundred of overlapping classifications). Therefore, in order to improve the edition workflow, the Orphanet team has developed a rare disease knowledge-base founded on an Ontology-based architecture. This architecture complies with the W3C standards and languages of the Semantic Web : OWL,, RDF, SPARQL and SKOS. This ontology design approach is based on both domain expertise (in rare diseases and in knowledge engineering) and knowledge extraction from the relational database. The current version of OntoOrpha comprises over 11,000 classes and 190,000 annotations organized under a Rare Diseases Core Ontology. Relationships between disorders (disease, malformative syndrome,...), groups of disorders (by anatomical system, by physiopathological mechanism,...), clinical signs and genes are therefore represented.

Semi-automatic translation of HPO concepts into French

Because the HPO is not yet included in the UMLS, it was not possible to apply knowledge-based procedures that the

CISMeF team has already applied to semi-automatically translate several heath terminologies or ontologies as described in ref [chapter, Merabti, 2012] (e.g. FMA [9]).

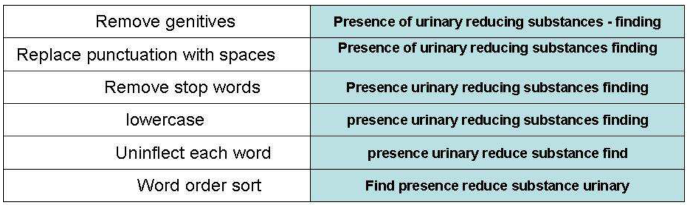
Therefore Natural Language Processing (NLP) techniques were used in this work to propose candidate terms for the translation. In this study, each HPO concept in English was tested versus each concept included in HeTOP in English (n= 1,274,568 concepts).

The HPO ontology contains 10,206 concepts (December 2010 version). Each HPO concept has several links to OMIM rare diseases and to NCBI genes Web site. Unfortunately OMIM rare diseases and NCBI genes are not (yet) included into HeTOP. To our knowledge, neither OMIM diseases nor NCBI genes are organized as a terminology or furthermore as a more formal ontology. Therefore, it was necessary to find semantic harmonization with the Orphanet ontology, both for rare diseases and genes.

Leverage of HPO

Several mappings were used to match the HPO genes and diseases with Orphanet genes and diseases included in the HeTOP. The HPO has links to NCBI genes and to the reference number of the OMIM diseases.

OMIM disease terms were automatically extracted from the OMIM Web site. Then, these OMIM disease labels were automatically mapped to disease terms included in the HeTOP. In this approach, OMIM terms in English were normalized and an algorithm was applied to find target terms which were exactly lexically similar (exact match) from terminologies and ontologies included in HeTOP. This algorithm was exploited in several previously reported studies to map external French and English terminologies to UMLS and the HeTOP. In this method, the Norm module included in the UMLS Specialist Lexicon tools was used in English [11], [12], [13], [14]. The Normalization process involves stripping genitive marks, transforming plural forms into singular, replacing punctuation, removing stop words, lower-casing each word, breaking a string into its constituent words, and sorting the words into alphabetic order, see Figure 1.



Remove genitives	Presence of urinary reducing substances - finding
Replace punctuation with spaces	Presence of urinary reducing substances finding
Remove stop words	Presence urinary reducing substances finding
lowercase	presence urinary reducing substances finding
Uninflect each word	presence urinary reduce substance find
Word order sort	Find presence reduce substance urinary

Figure 1 – Example of NLP Normalization process

In order to map HPO concepts to Orphanet genes, a join was performed between NCBI gene terms (thanks to the Entrez Gene Ids) and Orphanet gene terms. This procedure is a simple exact match without any string normalization.

Finally, this procedure allows a mapping between the HPO concepts and the HeTOP diseases and the Orphanet genes.

The following processes were applied to leverage HPO:

1. Translations using existing automatic mappings:
    - exactMatch with one already translated term. For example, based on this approach, the HPO term “Pelviureteric junction obstruction” is mapped to the English term “obstruction of pelviureteric junction”, which corresponds to the French term “Obstruction de la jonction pyélo-urétérale” and this term was subsequently proposed as a possible translation of the English HPO term;
    - manual translation using a web application connected to the database.
2. New synonyms using mappings: extremely important for Information Retrieval.
3. Pushing CUIs from UMLS: important for interoperability.

The HeTOP and the HPO integration

The HeTOP is based on a generic meta-model of terminologies and ontologies. This meta-model has been developed since 2005 and showed its efficiency because we managed to integrate all terminologies or ontologies we decided to add to the system.

To integrate the terminologies in the HeTOP, three steps were necessary: a) to design a generic terminology model into which each terminology model can be integrated. b) to design a process capable of integrating terminologies into the HeTOP. c) to build and integrate intra and inter-terminology semantic harmonization into the HeTOP. A generic meta-model was designed for the database in order to fit all the terminologies into one global structure. Then, a model of each terminology was designed as a specialization of the meta-model. With the specific models, the work consisted in developing a parser for each terminology: the input is the original data (or normalized original data) and the output is an OWL file. As data exists in different formats 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 CISMeF database. A generic parser was developed to directly insert each terminology into the database. A special model was designed to represent each terminology as a proprietary view so that the parser can use this custom model in input to recognize concept classes, definitions, synonyms, relations in order to insert it very easily into the database.

Since HPO is an ontology and has been defined very cleanly, an OBO (Open Biological and Biomedical Ontologies) parser has been specified for the HPO OBO file, downloaded directly from the HPO web site. The HPO model is also very clear and the integration was very easy as we already performed such ontology integration for FMA [16].

Results

From the 10,206 HPO concepts, for 18% of them at least one potential candidate term was proposed from the different terminologies or ontologies included in the HeTOP (see Table 1).

Tableau 1 - Distribution of candidate terms for the HPO translation into French

Matching T/O included in HeTOP	Number of candidates
MedDRA	1,404
SNOMED int. (3.5)	1,273
Orphanet	816
WHO-ART	788
ICD-10	670
MedlinePlus	309
ICPC-2	80
LOINC	72
FMA	17
Others	29

Each of them was manually reviewed. Then, the CISMeF team (CL & SJD) has manually reviewed 5,458 concepts; for each HPO concept, the CISMeF team has chosen the preferred term among several candidates. The other validated candidates were added as HPO synonyms. The non-validated candidates were discarded. In addition to this method, the CISMeF team has manually translated many other terms not found in other T/O.

Currently, 7,411 HPO concepts are translated into French (72.6%), including 2,984 synonyms. A total of 5,615 semantic mappings were also manually reviewed. These mappings consist of only exact match relations. The main advantage of their use is to perform semantic expansion in the information retrieval process.

This work provides several functionalities that were out of reach before it:

(a) Ontology auditing: for each Orphanet disease which has a semantic exact match with OMIM, it is now possible for the ontologist to confront the Orphanet phenotypes to the HPO phenotypes. For example, for the Marfan disease, the Orphanet ontology provides semantic links to 65 signs, whereas the HPO provides 51 signs, see Figure 3. It is then easy for Peter Robinson (HPO) and Jean Charlet (Orphanet) to review the discrepancies between the two ontologies.

(b) Indexing: since over 78% of HPO is already translated into French, these terms are available to the CISMeF librarians for indexing health Internet resources into the CISMeF catalog (URL: www.cismef.org).

(c) Teaching: HeTOP became a tool to teach rare diseases for Rouen medical students since 2010. For example, in the Orphanet ontology, the Marfan syndrome is linked to two genes: FBN1 and TGFBR2. Students may want to know what are the other rare diseases linked to this gene. In only two clicks it is possible to have the answer (n=6). Similarly, the Marfan syndrome is linked to 65 clinical signs (e.g. ectopia lentis). Students may want to know what are the other rare diseases linked to these clinical signs. In only two clicks, it is also possible to obtain the answer (n=19).

Three qualitative evaluations were performed in the last three years on three successive cohorts of Rouen Medical School students (second year) (76% of satisfaction for the content and 58% satisfaction for the user interface) [17].

Those results are available online, directly from HeTOP: http://pts.chu-rouen.fr/connexion.html?lang=en (login=hpodemo, password=demo11). One can search HPO terms, or Orphanet and MeSH terms (e.g. coloboma of iris (see Figure 2)) and navigate through the HPO tree and through relations and semantic links.

HeTOP is currently used by about 500 unique machines per day, mainly by librarians, translators, students and physicians, mainly as a French MeSH browser. Furthermore, over 1,000 users are registered to access the extended versions of the terminologies and ontologies with approximately 1.5 new users per working day.

Discussion/Conclusion

This paper has described the integration of the HPO ontology into the HeTOP terminology server, that leads to ontology auditing, indexing and teaching. Auditing terminologies and ontologies using a terminology server is, according to us, of utmost importance. It should be included in the quality process of terminology and ontology building and maintenance, in particular for the two ontologies used in this study: the HPO and the Orphanet ontology.

This paper has several limitations: because HeTOP is originally based on terminologies and ontologies developed in French, it contains much less knowledge resources than ULMS (n=150) and BioPortal (n=303). Nevertheless, HeTOP contains one major functionality that not exists in BioPortal: it is crosslingual, allowing navigation among 23 languages. Currently, this functionality is used for HPO mainly for the French language, but HPO will be translated into other languages (in particular German in the near future). Then, HPO translations will also be integrated into HeTOP terminology server.

Furthermore, the HeTOP interface and relations have been translated in several languages: Arabic, Danish, German, Portuguese, Italian and Hebrew. This allows the use of this portal by health professionals, students and patients that do not speak very well English, which is the case quite often in France.

One other HeTOP feature is very important, compared to UMLS or BioPortal: the web site has been designed for humans. That means that the quality of data matters so as its representation. The HeTOP is currently used by 500 unique machines per day, mainly by librarians, translators, students and physicians, mainly to query MEDLINE/PubMed in his/her native language. The translations of terms and the interoperability between terminologies and ontologies are also a major leverage of the quality of the data and terminologists and ontologists could find in HeTOP a great opportunity to deal with the lexicons quality.

References

[1] Robinson PN, Mundlos S. The Human Phenotype Ontology. Clin Genet 2010;77:525–534.

[2] Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's online mendelian inheritance in man (OMIM). Nucleic Acids Res 2009; 37 (Database issue): D793–D796.

[3] Fryns JP, de Ravel TJL. London Dysmorphology Database, London Neurogenetics Database and Dysmorphology PhotoLibrary on CD-ROM [Version 3] 2001. Winter RM, Baraitser M. Oxford University Press, ISBN 019851-780, pound sterling 1595. [Hum Genet 2002: 111: 113].

[4] POSSUM. Retrieved from http://www.possum.net.au/ (Accessed 2012).

[5] Aymé S. Orphanet, an information site on rare diseases. Soins. 2003 Jan-Feb;(672):46-7.

[6] Pelz J, Arendt V, Kunze J. Computer assisted diagnosis of malformation syndromes: an evaluation of three databases (LDDB, POSSUM, and SYNDROC). Am J Med Genet. 1996 May 3;63(1):257-67.

[7] Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Res 2004;32:267–270.

[8] Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research 2009; Jul;37(Web Server issue):W170-3.

[9] Merabti T, Soualmia LF, Grosjean J, Palombi O, Muller J-M & Darmoni SJ. Translating the Foundational Model of Anatomy into French using knowledge-based and

lexical methods. BMC Medical Informatics and Decision Making 2011 Oct 26;11:65.

[10] Merabti T, Soualmia LF, Grosjean J, Joubert M, Darmoni SJ. Aligning Biomedical Terminologies in French: Towards Semantic Interoperability in Medical Applications. Medical Informatics, March, Pages 41-68, InTech, 2012.

[11] UMLS® Reference Manual [Internet]. Bethesda (MD): National Library of Medicine (US); 2009 Sep-. 6, SPECIALIST Lexicon and Lexical Tools. Available from: http://www.ncbi.nlm.nih.gov/books/NBK9680/

[12] Browne AC, Divita G, Aronson AR, McCray AT. UMLS language and vocabulary tools. AMIA Annu Symp Proc 2003, :798.

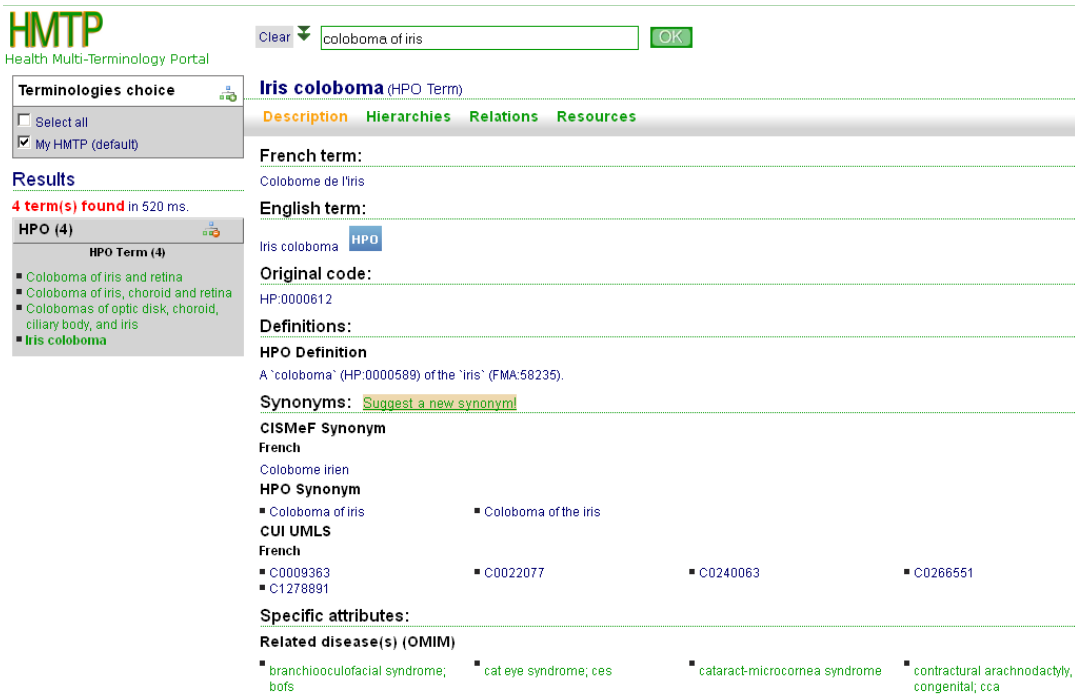
[13] McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care 1994, :235–239.

[14] Peters L, Kapusnik-Uner J, Bodenreider O. Methods for Managing Variation in Clinical Drug Names. In Proc Annu Symp AMIA 2010 2010:637–4.

[15] Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF & Darmoni SJ. Health Multi-Terminology Portal: a semantics added-value for patient safety. Stud Health Technol Inform. 2011;166: 129-138.

[16] Golbreich C; Grosjean J & Darmoni SJ. The FMA in OWL 2. AIM, 2012 (in press).

[17] Grosjean J, Merabti T, Griffon N, Dahamna B, Darmoni SJ. Teaching medicine with a terminology/ontology portal. Stud Health Technol Inform. 2012;180:949-53.



Results

4 term(s) found in 520 ms.

HPO (4)

HPO Term (4)

- Coloboma of iris and retina
- Coloboma of iris, choroid and retina
- Colobomas of optic disk, choroid, ciliary body, and iris
- **Iris coloboma**

Iris coloboma (HPO Term)

French term:

Colobome de l'iris

English term:

Iris coloboma

Original code:

HP:0000612

Definitions:

HPO Definition

A `coloboma` (HP:0000589) of the `iris` (FMA:58235).

Synonyms: Suggest a new synonym!

CISMeF Synonym

French

Colobome irien

HPO Synonym

- Coloboma of iris
- Coloboma of the iris

CUI UMLS

French

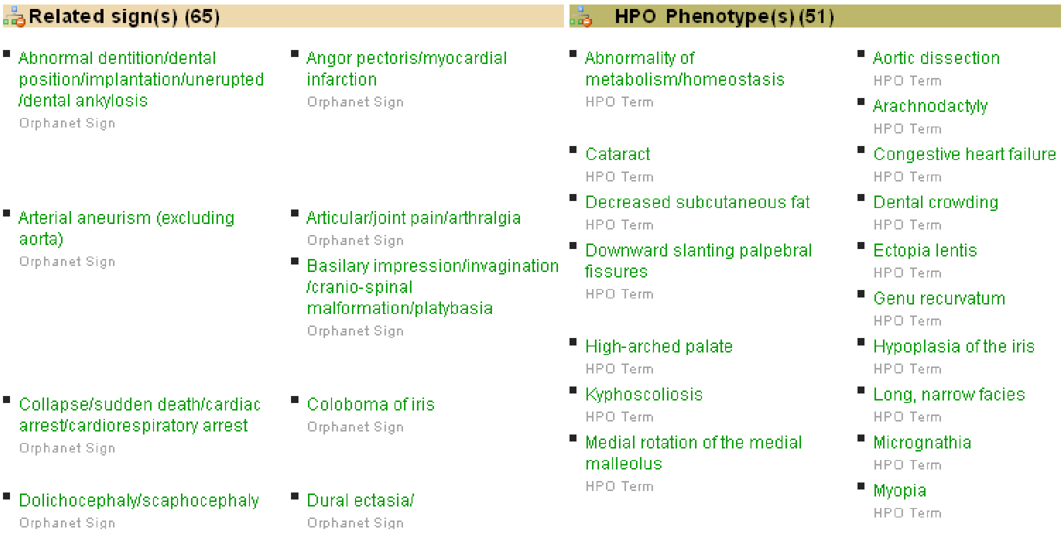
- C0009363
- C0022077
- C0240063
- C0266551
- C1278891

Specific attributes:

Related disease(s) (OMIM)

- branchiooculofacial syndrome; bofs
- cat eye syndrome; ces
- cataract-microcornea syndrome
- contractural arachnodactyly, congenital; cca

Figure 2 - HeTOP screenshot of an HPO term



Related sign(s) (65)

- Abnormal dentition/dental position/implantation/unerupted /dental ankylosis
Orphanet Sign
- Angor pectoris/myocardial infarction
Orphanet Sign
- Arterial aneurism (excluding aorta)
Orphanet Sign
- Articular/joint pain/arthralgia
Orphanet Sign
- Basilary impression/invagination /cranio-spinal malformation/platybasia
Orphanet Sign
- Collapse/sudden death/cardiac arrest/cardiorespiratory arrest
Orphanet Sign
- Coloboma of iris
Orphanet Sign
- Dolichocephaly/scaphocephaly
Orphanet Sign
- Dural ectasia/
Orphanet Sign

HPO Phenotype(s) (51)

- Abnormality of metabolism/homeostasis
HPO Term
- Aortic dissection
HPO Term
- Arachnodactyly
HPO Term
- Cataract
HPO Term
- Congestive heart failure
HPO Term
- Decreased subcutaneous fat
HPO Term
- Dental crowding
HPO Term
- Downward slanting palpebral fissures
HPO Term
- Ectopia lentis
HPO Term
- Genu recurvatum
HPO Term
- High-arched palate
HPO Term
- Hypoplasia of the iris
HPO Term
- Kyphoscoliosis
HPO Term
- Long, narrow facies
HPO Term
- Medial rotation of the medial malleolus
HPO Term
- Micrognathia
HPO Term
- Myopia
HPO Term

Figure 3 - Auditing HPO and Orphanet ontologies (HeTOP screenshots)