Re-use of terminological and ontological resources for the construction of domain ontologies in medicine: a description of two experimental approaches

Ferdinand Dhombres[1,2,3,4,5], Jean Charlet[1,3,4,5], Jean-Marie Jouannic[2,3,4], Laurent Mazuel[1], Marie-Christine Jaulent[1,4,5]

[1] INSERM U872, Equipe 20, ICS, Paris, France ;
[2] Service de gynécologie-obstétrique et centre pluridisciplinaire de diagnostic prénatal de l'Est Parisien, hôpital Armand Trousseau ;
[3] AP-HP, Paris, France ;
[4] Université Pierre et Marie Curie, Paris, France ;
[5] Université Paris Descartes, Paris, France.
{ferdinand.dhombres, jean.chalet, marie-christine.jaulent}@crc.jussieu.fr

Abstract. By referring to the construction of wo medical domain ontologies, we describe our methodological choices for the re-use of a number of terminological and ontological resources: CCAM, ICD-10, ATC, FMA, MENELAS, ORPHANET and SNOMED 3.5. Good practice was identified and two strategies (pruning versus selective additions) for resource re-use in ontology construction are discussed.

Keywords: knowledge engineering, domain ontology, prenatal diagnosis, emergency medicine.

1 Introduction

The re-use of terminological and ontological resources (TORs) in the construction of medical ontologies is essential, given the number and quality of the available resources and the work that has gone into their development. Our construction of domain ontologies in several fields of medicine (cardiovascular surgery, pneumology, intensive care, prenatal diagnosis and emergency medicine) has prompted us to re-use a certain number of TORs.

In the present article, we describe our strategies for the re-use of a core medical ontology (OntoMenelas[1] [1]), a reference ontology of human anatomy (the Foundational Model of Anatomy[2] [2]), several medical classifications (ICD-10 [3], CCAM [4] and ATC [5]) and a thesaurus (Orphanet [6], a multilingual thesaurus in the field of rare and orphan diseases). We also discuss the case of SNOMED 3.5 [7]. The two projects used to illustrate our approach are OntolUrgences (an ontology for

---

[1] http://purl.oclc.org/NET/spim/ontologies/public/OntoMenelas/
[2] http://sig.biostr.washington.edu/projects/fm/

medical record support in an emergency medicine setting) and OntoDPN (an ontology of prenatal diagnosis).

2 Background

2.1 The construction of domain ontologies in medicine

Our approach to the construction of domain ontologies in medicine starts with an examination of the corpus of documents in the field [8]. It also includes a large number of methodological choices, within which the re-use of TORs is a significant part [9].

OntoDPN is an ontology of prenatal diagnosis. It was developed to represent the relations between the signs on prenatal imaging and the diagnosis of developmental anomalies and use in different applications (help with coding, decision support, etc.). Its construction strategy [9] is both bottom-up (by studying text corpora : 200,000 prenatal examination reports and many reference documents) and top-down (exploitation of existing TORs, whether in the field of medicine in general or prenatal medicine in particular).

OntolUrgences. This ontology was developed as part of the LeRUDI project. The goal is to enable a coordinating physician in an emergency medicine unit to navigate rapidly and efficiently through electronic medical records (EMRs). As with OntoDPN, construction of the ontology [10] comprised both bottom-up techniques (analysis of text corpora) and top-down approaches (re-use of existing TORs).

2.1 Terminological and ontological resources in the medical field

There are many resources in the medical field ; Table 1 presents those which were re-used here. We decided not to use a number of other resources (such as MeSH[3], GALEN[4], DOLCE[5] and UMLS[6]) in our initial work.

[3] http://www.ncbi.nlm.nih.gov/mesh
[4] http://www.opengalen.org/
[5] http://www.loa-cnr.it/DOLCE
[6] http://www.nlm.nih.gov/research/umls/

Table 1. The medical terminological and ontological resources which were re-used for the OntolUrgences and OntoDPN ontologies.

Resource	Items/concepts	Type
SNOMED 3.5	106,171	nomenclature
ICD-10	19,856	classification
CCAM	13,468	classification
ATC	5,510	classification
ORPHANET (developmental abnormalities)	1,095	thesaurus
FMA	81,056	ontology
OntoMenelas	1,832	ontology

3 The re-use of TORs

3.1 OntoMenelas: re-use of a core ontology of medicine

Re-use of the OntoMenelas top-level and core ontologies was fruitful in both projects: for conceptualizing the examinations in an emergency medicine setting for OntolUrgences and describing the relationships between the anatomy and morphological signs in OntoDPN.

An example of conceptualization in OntolUrgences (the performance of a creatinine assay) is presented below (Fig. 1); it uses OntoMenelas as a hierarchical basis for the top-level and core ontologies. When modelling medical examinations in the OntolUrgences ontology, the use of OntoMenelas enables us to adopt a distributed conceptualization, as follows:

- *"Substance"*: a general concept representing any substance, which is useful for the domain ontology. In our example, this enables us to classify the concept *creatinine*.
- *"Attribut Physiologique"* : this represents the attributes related to a living entity (in our case, a patient) - the *blood creatinine level* and the *urine creatinine level*, for example.
- *"Examen Paraclinique"*: this corresponds to the act of examining a patient. Three categories are created by specializing this concept (*laboratory test*, *imaging* and *clinical examination*).

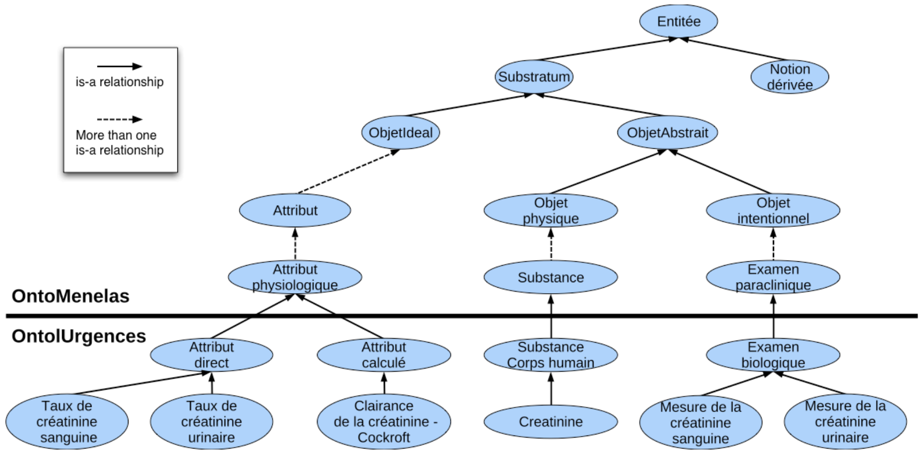

Fig. 1. Re-use of OntoMenelas in OntolUrgences: conceptualization of examinations performed in an emergency medicine setting.

The reader should note that although example in Figure 2 does not feature the concept *blood sample collection* (due to space limitations), the latter naturally fits into the OntoMenelas hierarchy as an *intentional act of exploration*.

It should also be noted that OntoMenelas also features a hierarchic conceptualization of the relations between concepts (although the latter are not yet fully reusable). For example, the relation *attr* links an *attribute* to an *AbstractObject* in order to indicate that an object can possess attributes. In our example, we specialize this relation to create a *has for creatinine clearance* relation between the concepts of *patient* and *creatinine clearance*.

3.2 FMA: re-use of a reference ontology

The OntoDPN ontology's prime objective is to represent prenatal imaging semiology. Since this semiology is mainly morphological (description of foetal anatomical structures), we decided to re-use the Washington University School of Medicine's *Foundational Model of Anatomy* (FMA) reference ontology [2].

However, we chose not to use the FMA 3.0 "as is" and did not introduce its very high-level distinction (at the *Substratum* concept level in OntoMenelas) between anatomical entities (corresponding to the concept *AnatomicalEntity*) and other concepts. Given that FMA's vocation is (according to its authors) to be partly re-used and adapted for a specific purpose, we did not include all of the FMA in OntoDPN but used it as a foundation for building our new ontology.

By examining the corpus, we selected descriptions of signs in foetal imaging reports. For each description, the corresponding anatomical element or elements were sought and identified in the FMA so that they could be named, organized and partly annotated in an analogous way. A morphological sign in OntoDPN was thus linked to an existing anatomical entity in the FMA or (if necessary) a new anatomical concept.

Different modelling choices were made according to the circumstances and are specified below.

Concepts not present in the FMA. Some anatomical concepts identified by our method (such as *PortionOfAmnioticFluid*) are quite specific to prenatal diagnosis. In this situation, we create a new concept (Fig. 2C) which fits with the FMA's logic: *PortionOfAmnioticFluid* is thus positioned as a child of *PortionOfBodySubstance* under the OntoDPN-specific concept *GestationalSubstance*, which is itself created in compliance with the FMA's logic.

Concepts present in the FMA but absent in the foetus. Some anatomical structures may exist in both the mother and the foetus; it must be possible to model either without any doubt concerning to whom the structure belongs. Under the FMA's concept of *GestationalStructure*, there are foetus- and embryo-specific concepts. However, there are far too few of these for our ontology[7] and so we decided to proceed in two steps. The first involves moving the FMA's original *GestationalStructure* concept to the same level as *AnatomicalStructure* (Fig. 2B), which is renamed as *NonGestationalStructure* (Fig. 2A). These two concepts are siblings and share *AnatomicalStructure* as a parent. This interfaces with the *Inanimate* concept in the lower part of the OntoMenelas core ontology. We call this type of concept an "interface concept" (elements marked with a * in Fig. 2). The second step consists in re-using the FMA's concepts for the mother under the *NonGestationalStructure* concept, with the same labels. For concepts in foetal anatomy, labels are annotated with the suffix "_F" and the hierarchy is exactly the same as that used by the FMA in the adult (i.e. these are "FMA-like" concepts, Fig. 2). We thus distinguished between prenatal anatomical structures (under *GestationalStructure*) and postnatal anatomical structures (under *NonGestationalStructure*) by ensuring the high level of granularity required for our model in prenatal anatomy.

Concepts present in the FMA but whose parent(s) correspond(s) to a concept in OntoMenelas. By way of an example, the anatomical areas defined in the FMA correspond to children of the *PseudoObject* concept in OntoMenelas. In terms of the link between the two ontologies, we wanted to be able to take advantage of other useful children of OntoMenelas concepts, such as *SystemicObject* and *SociologicObject* (Fig. 2D). Since the concepts of anatomical areas had already been specialized for the embryo and the foetus under *GestationalSpace*, we re-used them "as is".

[7] In the FMA, only the heart and the uterus are specifically described for the foetus.

Concepts present in the FMA but whose position varies in OntoMenelas. In the FMA, the concepts of organ systems (*OrganSystem* and *OrganSystemSubdivision*) cannot be integrated into the concept of *AnatomicalStructure*, since the latter does not include the notion of organisation around a particular function. This notion is present in the definition of a "system" in physiology and is useful in OntoDPN. In order to retain this semantic, these concepts are defined as children of the OntoMenelas concept *PhysicalSystemicObject* (Fig. 2E). Moreover, a distinction between the gestational and non-gestational systems is provided by two interface concepts (*GestationalPhysicalSystemicObject* and *NonGestationalPhysicalSystemicObject*), by analogy with the concept *AnatomicalStructure*.

Our approach clearly differs from that used in the construction of RadLex [11], which consisted in selection of FMA concepts by expert consensus. Here, we selected FMA concepts which were relevant for our field, after studying the corpus of texts. Moreover, the FMA was not able to fully represent the prenatal anatomy, which prompted us to enrich this resource.

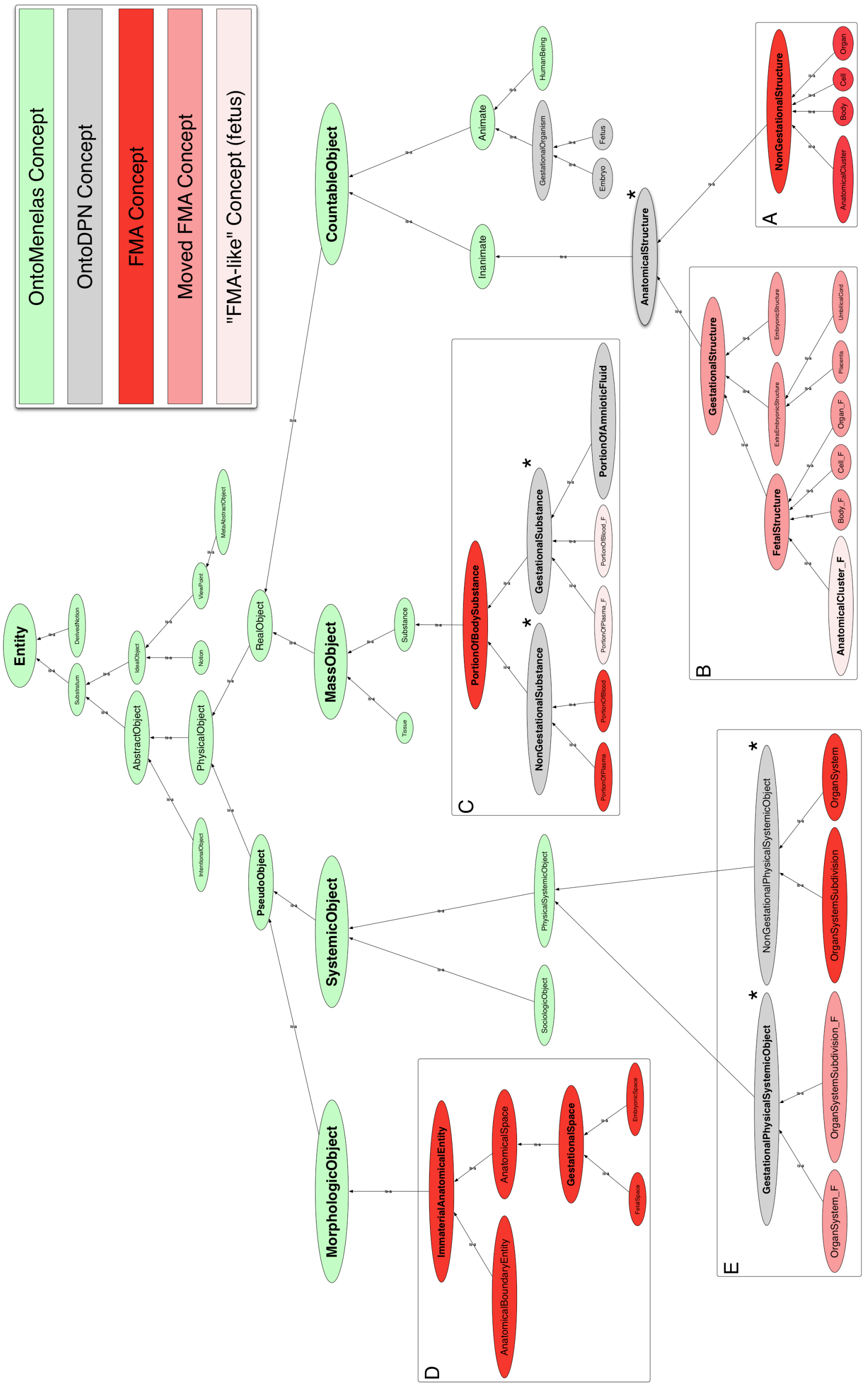

Fig. 2. Re-use of the FMA and OntoMenelas: bridging the two ontologies within OntoDPN.

3.3 CCAM: re-use of the French common classification of medical acts

Version 16 of the French *Classification Commune des Actes Médicaux* (CCAM, Common Classification of Medical Acts) contains over 13,000 names of examinations and treatments and enables medical files to be coded for accounting purposes. Given that there are few CCAM acts in the prenatal field, useful concepts were custom built by an expert in the field. The corresponding CCAM code is added as an annotation to the concept: <owl:AnnotationProperty rdf:about="#ccamId"/>

In contrast, a broad representation of acts is necessary in OntolUrgences (the uses of which include searching for information in medical records from a wide range of sources). Hence, the entire CCAM was included in the version available as part of the project (SKOS). It was transformed in OWL and then linked as a specific branch of acts in OntolUrgences by using a JAVA programme. After this large-scale import from the CCAM, the ontology was pruned by experts so that only relevant acts were conserved and then modelled as concepts. Concepts corresponding to over-detailed acts (which are only useful from an accounting standpoint) were not conserved. Secondary enrichment with data from text corpora and knowledge from medical experts was required. Concepts derived from CCAM acts were annotated in the same way as in OntoDPN when the derived concept and the CCAM concept were truly equivalent. In some cases, more general concepts had to be created and did not correspond to a CCAM act as such.

3.4 ICD-10: re-use of the classification

The ICD-10 was re-used indirectly in both projects: the ICD-code was also added to the corresponding concept as an annotation.

In OntoDPN, the ICD-10 codes were imported at the same time as the 1095 developmental anomalies extracted from one part of the Orphanet database [6]; the annotation process was automated by using the latter's ETL procedure [12].

In OntolUrgences, we manually entered the *specialty thesaurus* used by emergency medicine physicians to code their acts. Since this thesaurus had to be covered by the ontology, all the derived concepts were included in OntolUrgences, with their existing ICD-10 annotations.

3.5 ATC: re-use of the classification

Given the absence of an ontology for drugs, the WHO's ATC classification [5] was re-used in OntolUrgences. At present, drugs are not represented in OntoDPN.

The ATC is a tree structure, which appears a good start to form an ontology. However, this arborescence is misleading because some categories of drugs and compounds are repeated (with different identifiers). This is because the tree is divided up first by organ and then by drug - even though it is possible to factorize some attributes as defined concepts. Furthermore, with a view to information retrieval, the ontology will have to be completed with the drug's brand names, as provided by the AFSSAPS[8].

Hence, the enriched ATC was handled as if it were an ontology. It was then annotated by the emergency medicine physicians, in order to provide the level of descriptive detail required in the EMR interface.

3.6 ORPHANET: re-use of a thesaurus of developmental anomalies

As part of a collaboration with Orphanet [6], we re-used the thesaurus of development anomalies and included this hierarchy in OntoDPN. However, the hierarchy does not always feature “is-a”-type relations. A collaborative modelling project is underway, with a view to building an ontology of Orphanet's domain (rare diseases) ; this will ensure that the disease section in OntoDPN is ontologically correct.

3.7 SNOMED 3.5: re-use of the French classification of medical acts

Given the coverage required for searching within EMRs, the development work was completed via integration of the diagnostic branch of the SNOMED 3.5 terminology into OntolUrgences, as if it was an ontology. Next, over one hundred hours of pruning and reorganizational work by an expert group resulted in the maintenance of 6,000 of the 25,000 imported concepts.

4 Discussion

The re-use of TORs in our research unit's various projects always involves transformation of the resource into SKOS format [13] for direct inclusion into an ontology under construction.

4.1 Choice of the method: pruning or selective additions?

Choice of the method depends on the availability of experts and the size of the resource for re-use. When the resource was large (FMA and SNOMED), we tested both approaches (Fig. 3); selective additions by an expert proved to be satisfactory for OntoDPN and integrated well into the overall approach for ontology construction. For the re-use of SNOMED in OntolUrgences, "pruning" proved to be too expensive.

Furthermore, pruning or selective addition must be made by someone who understands what an ontology is. Whereas selective addition leaves the ontologist faced with a blank page and requires a good understanding of ontological issues, pruning can be performed without needing to check on the ontology's organizational status. In our approach to re-use of TORs through selective addition, there was a need

[8] *Agence française de sécurité sanitaire des produits de santé*, the French Agency for Healthcare Product Safety (http://www.afssaps.fr/)

for extensive interaction (Fig. 4) between the domain modelling strategies at the ontological commitment step and the selection of the resource's terms for re-use.

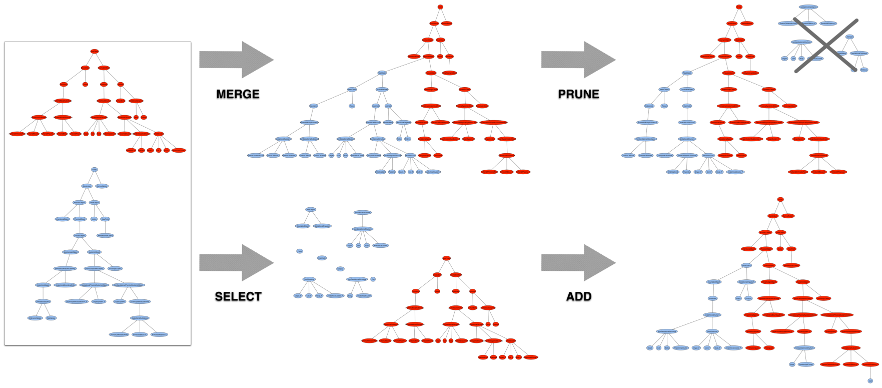

Fig. 3. Strategies in ontology re-use. Top: merge and prune. Bottom: select and add.

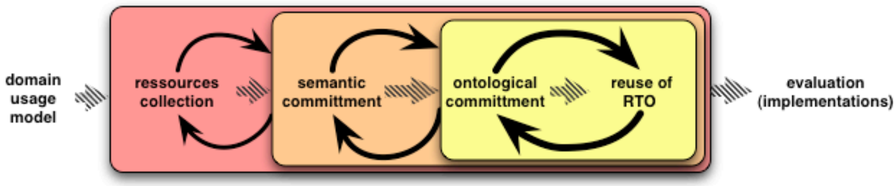

Fig. 4. Use of a nested iteration approach to build a domain ontology in medicine [9].

4.2 The annotation strategy

The source of the re-used concepts is conserved when annotated; this enables retention of the concept's source information and also facilitates automatic alignment between the newly built ontology and the re-used TORs. In fact, secondary alignment in terminological portals like BioPortal [14] is feasible for OntoDPN.

We decided not to keep the re-used concepts' *Uniform Resource Identifiers* - notably because of the structural changes in some parts of the re-used ontologies (the FMA, in particular).

4.3 The representation of a specialty thesaurus

In the case of OntolUrgences (and as performed in the field of pneumology in previous work [15]), we decided to look at a *specialty thesaurus*. In fact, the latter is a part of ICD-10 which specifically concerns the medical specialty in question. This part is identified by an expert in the field as a reference list for the T2A accounting[9]. The legitimacy of this type of list is clearly attractive when one is building an ontology. Firstly, the *specialty thesaurus* is a well-thought-out selection (though not a specialist construction). Secondly, given that one of the possible uses of ontologies is to assist with health economics coding, all the potential concepts in the coding should be available in the ontology [16].

Although the majority of ICD-10 concepts have obvious medical legitimacy, others are merely used to structure the ICD-10 and lack medical legitimacy. These same concepts pose representation problems and prompt the construction of defined concepts which interrelate many of the ontology's primary concepts. Moreover, their "raison d'être" means that some of these concepts are only used for health economics coding, without any medical legitimacy.

5 Conclusion

The re-use of TORs for ontology construction is an essential but complex approach. Resource size and availability of experts in the field appear to be decisive factors in choosing between pruning or selective addition when seeking to integrate existing resources into a new ontology.

When building an ontology, thinking that it may be re-used in the future is a significant factor in facilitating this very process. Hence, since a future extension of OntoDPN will involve indexing diseases on the basis of their prenatal signs, concepts whose labels may correspond to index entries are annotated specifically.

In general, the re-use of TORs enabled us to identify good practice, such as the use of (i) annotations to conserve the original identifier and (ii) exchange formats (such as SKOS).

[9] In T2A system, physician lists all the diagnostic and therapeutic acts performed for a given patient. This method forms the basis of hospital accounting controls in many countries. Schematically, it corresponds to designation of 5 acts from the CCAM and 5 diagnoses from the CIM-10 for each patient.

References

1. Bouaud, J., Bachimont, B., Charlet, J., Zweigenbaum, P.: Methodological principles for structuring an “ontology”. IJCAI'95 : International Joint Conference on AI. Workshop on “Basic Ontological Issues in Knowledge Sharing.”. p. 95-148, Montreal, Quebec (1995).
2. Rosse, C., Mejino, J.L.V.: A reference ontology for biomedical informatics: the Foundational Model of Anatomy. Journal of Biomedical Informatics. 36, 478-500 (2003).
3. CIM 10 - ICD 10, http://www.icd10.ch/.
4. CCAMv16 en ligne sur le site de l'Assurance Maladie en ligne (AMELI), http://www.ameli.fr/accueil-de-la-ccam/.
5. WHO Collaborative Center for Drug Statistics Methodology: Anatomical Therapeutic Chemical (ATC) classification system, Structure and principles, http://www.whocc.no/atc/.
6. Aymé, S.: Orphanet : The portal for rare diseases and orphan drugs, INSERM SC11, http://www.orpha.net/.
7. Présentation de SNOMED en ligne | esante.gouv.fr, le portail de l'ASIP Santé, http://www.esante.gouv.fr/snomed/snomed/.
8. Bachimont, B., Isaac, A., Troncy, R.: Semantic Commitment for Designing Ontologies: A Proposal. EKAW '02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web. p. 114–121Springer-Verlag, London, UK (2002).
9. Dhombres, F., Jouannic, J., Jaulent, M., Charlet, J.: Choix méthodologiques pour la construction d'une ontologie de domaine en médecine prénatale. Actes des 21e Journées Francophones d’Ingénierie des Connaissances. , Nîmes, France (2010).
10. Giroud, M.: L'accès au Dossier Médical Personnel par le médecin régulateur du Samu. 3eme Congrès Urgences. p. 807-816Société Française de Médecine d'Urgence, Paris, France (2009).
11. Mejino Jr, J.L., Rubin, D.L., Brinkley, J.F.: FMA-RadLex: An Application Ontology of Radiological Anatomy derived from the Foundational Model of Anatomy Reference Ontology. AMIA Annual Symposium Proceedings. 2008, 465–469 (2008).
12. Talend Open Studio v3.0. Talend open data solutions.
13. SKOS Simple Knowledge Organization System, http://www.w3.org/2004/02/skos/.
14. NCBO BioPortal: Ontology Listing, http://bioportal.bioontology.org/ontologies.
15. Baneyx, A., Charlet, J., Jaulent, M.: Building medical ontologies based on terminology extraction from texts: an experimentation in pneumology. Studies in Health Technology and Informatics. 116, 659–64 (2005).
16. Blanc, F., Baneyx, A., Charlet, J., Housset, B.: Représentation des connaissances en pneumologie : l'ontologie doit pouvoir aider au codage. Revue des maladies respiratoires. (2010).