Applied Ontology 0 (2011) 1–0
IOS Press

Building a Medical Ontology to support Information Retrieval: Terminological and metamodelization issues

Jean Charlet [a,b,*], Gunnar Declerck [a], Ferdinand Dhombres [a,c,d,e,f], Pierre Gayet [g], Patrick Miroux [g], Pierre-Yves Vandenbussche [a,d,h]

[a] Inserm UMRS 872 éq.20, *Ingénierie des connaissances en santé, Paris, France.*
[b] AP-HP – *Assistance Publique Hôpitaux de Paris, Paris, France.*
[c] Inserm SC11, Orphanet, *Plateforme Maladies Rares, Hôpital Broussais, Paris, France.*
[d] UPMC – *Université Pierre et Marie Curie, Paris, France.*
[e] *Faculté de médecine Paris Descartes, Université Paris Descartes, Paris, France.*
[f] *Service de Gynécologie-Obstétrique et Centre de Diagnostic Prénatal de l'Est Parisien, Hôpital Armand Trousseau, AP-HP, Paris, France.*
[g] *Centre hospitalier de Compiègne, France.*
[h] *Société MONDECA, Paris, France.*

Abstract. OntolUrgences is a termino-ontological resource that has been developed for a retrieval information project. This TOR (i) plays the field model role by listing the relevant concepts, and (ii) ensures the link between the concepts and their name in the documents of the electronic Emergency Medical Records. This double function should not only enable an easy annotation and indexation of the patient files, but also facilitate the retrieval of information from the indexed records.

The OntolUrgences development included 6 phases: (i) the building of the TOR ontological skeleton based on a corpus analysis method; (ii) the use of existing terminological and ontological resources to manually complete the TOR concepts system; (iii) the automatic and (iv) the semi-manual TOR enhancement at the terms level; (v) the TOR enhancement of concepts in relation to the medicines; and finally, (vi) the implementation of validation and quality control procedures.

This project shows that: (i) the sustainability of such a resource requires a concise articulation between terms and concepts, and (ii) such a requirement can be met via the implementation of standardized procedures based on a meta-model architecture allowing the modeling of all necessary KOS and other knowledge structures.

Keywords: ontology, terminology, information retrieval, emergency medecine

1. Introduction

The use of terminological systems for the creation of ontologies raises several major issues. Obviously, ontologies and terminologies play a similar normative role. They aim at establishing a common vocabulary and make use of shared representations and concepts to allow the documents interoperability and facilitate knowledge building. However, ontologies and terminologies have clearly a different approach on Semantics. Ontologies are *concepts* architectures and are not organized lists of *terms*. Unlike terms, the concepts are characterized by *formal definitions*. The formal aspect enables the computerized treatment of the information. To use an ontology to normalize a document is, in that sense, to encode it by bringing a characteristic allowing the automated treatment of the information.

However, the use of terminologies, or even more radically the use of corpus of text, for the creation of ontologies is sometimes unavoidable. If the ontology is to be integrated within an

*Corresponding author: Jean Charlet, address of first, jean.charlet@upmc.fr.

automated information treatment system (for instance an information retrieval (IR) system), the concepts should match with the terms appearing on the documents to enable the information treatment. The ontology should ensure the *coverage of the terminological domain*. The conceptual representation would otherwise be unusable.

The Lerudi (emergency services) Project intends to develop an Information System (IS) offering an overview of the Electronic Health Record (EHR) to the health professionals. Additionally, it aims at facilitating the quick reading of the EHR to allow quick medical decisions under tight time constraints. The field experimentation of that project is the reading of hospital files by an emergency regulating physician (Giroud, 2009). The Lerudi IR system is based on a Termino-Ontological Resource (TOR)[1] named OntolUrgences. The TOR (a) plays the field model role by listing the relevant concepts; and (b) ensures the link between the concepts and their name in the EHR documents. This double function should not only enable an easy annotation and indexation of the patient files, but also facilitate the retrieval of information from the indexed records.

The OntolUrgences development included 6 phases: (i) the building of the TOR ontological skeleton based on a corpus analysis method; (ii) the use of existing terminological and ontological resources to manually complete the TOR concepts system; (iii) the automatic and (iv) semi-manual TOR enhancement at the terms level; (v) the TOR enhancement of concepts in relation to the medicines; and finally, (vi) the implementation of validation and quality control procedures.

The first 2 phases correspond to a usual ontology construction method, widely tested in our team and did not raise major issues. However, the 3 following phases that were specific to this particuliar TOR and its IR supportive purpose were much more problematic. Specifically, the TOR terminological enhancement required external resources: *Knowledge Organization System* (KOS). These external resources are only useable in an architecture supporting a complex modeling of the target TOR. In particular, the architecture should accommodate the terms, the concepts and their interrelation, and simultaneously, the KOS used for the enhancement. The last phase (related to validation and quality control) is also specific to this project and was necessary considering the various participants involved in the TOR construction.

Trough the detailed description of the process guiding this TOR construction and validation within a large team, we aim at showing that: (1) the sustainability of such a resource requires a concise articulation between terms and concepts; and (2) such a requirement can be met via the implementation of standardized procedures based on a meta-model architecture allowing the modeling of all necessary KOS and other knowledge structures.

The rest of the paper is organized as follows: Section 2 briefly presents the advantages of using ontologies for retrieving information. The first two steps of the TOR construction and its specificities are presented in Section 3. Section 4 provides an overview of the UniMoKR model that enables the implementation of the TOR terminological and conceptual enhancement procedures; and its uses. Sections 5 and 6 describe the TOR different enhancement phases. The validation and quality control are detailed in the section 7. Finally, the paper concludes with a summary and a discussion in Section 8.

2. Why using an ontology for information retrieval?

To begin with, we should ask ourselves what the point in using an ontology for an IR is. Specifically, the main advantage of an ontology is to allow an automated reasoning based on the conceptual structure and semantic relations between notions. Consequently, in addition to subsumption relations (*is-a formal relation*), we modeled the semantic relations between signs,

[1] A TOR is an ontology in which the terms are linked to concepts in a systematic and exhaustive way. Several binding ways exist in order to link terms and concepts according to the representation target (Reymonet et al., 2007).

diseases and medical specialties. These relationships enable an interface (*i.e.* a cloud of words) to display the medical specialties that characterize a given EHR.

An ontology for IR has also *de facto*, as any ontology, a structure that depends on the task (Charlet et al., 1996; van Heijst et al., 1997). This structure is not a quality in itself for the IR, but it nevertheless has two advantages: *a*) a well-structured ontology is easier to maintain than a poorly structured ontology, *b*) a well-structured ontology enables valid reasonings. This second point is obviously expected from any ontology, but it is clear that it is not always satisfied. Another important property that has to possess an ontology for IR is the coverage of the terms relevant to express the notions of the target-domain. The following two examples will illustrate these points:

Example of the importance of the formal structure of the TOR. Considering the important following question that asks an emergency physician about a patient: "Has my patient already been infested by an enterobacteria in the past?". Consider that the patient's record contains a document annotated with the concept of "Salmonella". For the system to conclude that Salmonella is an enterobacteria, it is necessary that the TOR specifies that the concept "Salmonella" has a transitive relation of specialization with the concept of "enterobacteria". In this way, the answer to the question of the emergency physician will be positive, *even if the patient record is not directly annotated with the more general concept of enterobacteria.*

Example of the importance of the terminological coverage of the TOR. The annotation of the noun phrases "paracetamol", "Dafalgan" and "paraml." requires that the TOR has a unique concept representing these three syntagms and the availability of the terms related to the chemical molecule (paracetamol) and the proprietary drug[2] or its brand name (Dafalgan).

It appears clearly that, the quality of the information displayed to the final user of the IR system crucially depends on the quality and richness of the TOR. The processes of annotation, indexing and inference rely on the formal structure and the terminological completeness (*i.e.* its capacity to cover the terms of the domain). The figure 1 below illustrates the different uses of the TOR in the Lerudi project.

3. Terminological and ontological resources used for designing OntolUrgences

OntolUrgences has been built in several steps and by using different resources, and its target knowledge field has been clarified gradually. From the very beginning of the project, we realized that the knowledge field that had been originally set for the TOR (that is: the repertoire of concepts that had to be present in the TOR) had to evolve. We had left with the idea of building an ontology representing only the specific concepts used by the emergency physicians. But it turned out that, from the perspective of information retrieval in EHRs, such restriction *a priori* of the target knowledge field was a mistake. Indeed, the information system aims to allow the emergency physician to quickly find medically relevant concepts in EHRs. But these concepts can not be reduced to concepts specific to the medical emergency field, they can instead meet *any medical specialty.*

Instead of building an ontology of *emergency* related concepts, we have thus designed an ontology of medicine *in its generality*, but reduced to the only part useful to the emergency physicians. Of course, this decision was not without consequences on the size of the ontology, which currently has no less than 13,000 concepts.

In the paragraphs below, we present the main phases of the development of OntolUrgences and the terminological and ontological resources we have used. We do not discuss the problem

[2]In pharmacology, a proprietary drug means a form of medication offered by a brand *e.g.* "DAFALGAN 1 g, effervescent tablet". "DAFALGAN" is the brand name of the product sold in pharmacies (name which is recommended by the French Agency for Sanitary Safety of Health Products: AFSSAPS). This proprietary drug is made of the chemical active "paracetamol". For more information, see the website of AFSSAPS `http://bit.ly/rhWZtr`

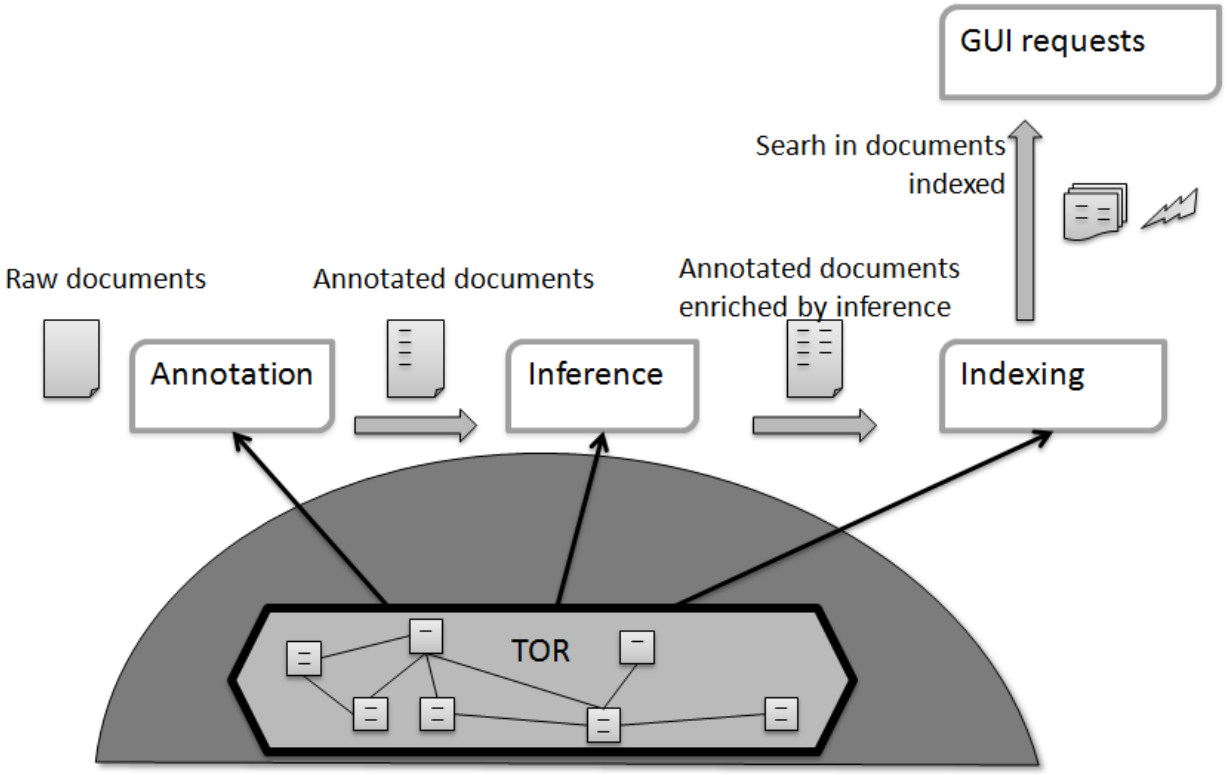

Fig. 1. Uses of the TOR in the Lerudi project. The TOR supports the processes of annotation, indexing and inference.

of the organization of these stages and cycles of development. For this question, we may refer to (Dhombres et al., 2010) and for OntolUrgences to (Charlet et al., 2009a). Suffice it to say that during the development process of OntolUrgences, we followed the Archonte method developed by B. Bachimont (2002).

3.1. The use of a top-ontology

In Inserm UMR_S 872, Éq. 20, we have developed several ontologies for medical applications, most often of medical coding assistance, but also for modeling issues related to usage studies (Charlet et al., 2009b). The first of these ontologies, OntoMénélas, was developed by B. Bachimont *et al.* (Charlet et al., 1994) in the context of the project Ménélas (Zweigenbaum et al., 1995) and covers the field of cardiac surgery. More recently, the development of three ontologies has been initiated: an ontology of Emergency Medicine – OntolUrgences, which is the subject of this paper –; an ontology of prenatal diagnosis – OntoDPN – and an ontology of rare diseases (Dhombres et al., 2011).

These projects have faced us with the question of what is generic in the building of our ontologies, and may therefore be reused from an ontology to another. Answering this question requires a modularization view of the construction of ontologies. Three distinct parts or levels can usually be distinguished in an ontology:

1. the foundational ontology or *top*-ontology, which corresponds to the most general concepts, which therefore does not belong to a particular field of knowledge (medical or otherwise), and whose organization largely relies on reflections of philosophical nature (for instance reflections on the most abstract and universal categories and conceptual distinctions);
2. the *core*-ontology, which provides the structural concepts of the field and describes the relations between these core concepts – in medicine, It includes concepts such as *diagnosis*, *sign*,

anatomical structure, and relationships such as those related to the location of a pathology on an anatomical structure;
3. the *domain* ontology, which describes the domain concepts as they are manipulated by professionals.

OntoMénélas being an ontology consisting of those three levels, we reused some of its parts for the development of our other ontologies. The upper part of OntoMénélas (its top-ontology) and its core level have been reused for the particular design of OntolUrgences. More generally, we conducted a study to determine under what conditions OntoMénélas would provide a top-ontology as well as a core ontology for medicine. These questions are the subject of another paper (Charlet et al., 2009a).

3.2. The processing of textual data

In the Archonte method, the domain ontology is built on the analysis of documents generated during the activity to be modeled. In our case, we have encountered great difficulties in accessing a corpus that could perform this function. The emergency services being not computerized, and the paper documents shorter and less numerous than in other services, it was difficult to find documents in sufficient numbers to make up the corpus in question.

Consequently, we used two other kinds of documents: the acts of the *Urgences* conference of the discipline and the *Guides to Good Practice*. Besides the difficulty we had to preprocess the corpus, the main problem was the coverage capacity of the corpus compared to the target. Indeed, the corpus of the conference proceedings, that was fully processed, has shown its limits in terms of scope. Conference papers are in many cases concerned with the "rare bird", that is with questions that are not representative of the problems that emergency physicians are confronted daily. A specific work has shown this clearly by comparing the terms most frequently detected in the corpus with the actual incidence of the emergency diseases (Gayet et al., 2010).

This issue of availability of the corpus should not be underestimated: in the areas where we can base the construction of the ontology on a corpus analyzed by tools of natural language processing (NLP), resorting to existing terminologies operates in the validation process of the work having been done. In the case of interest here, they occur much earlier in the development process.

3.3. Reusing the specialty thesaurus

For the PMSI[3] coding, the emergency physicians make use of a CIM-10 extract which contains about 1,000 terms. These terms covering an important part of the terminological repertoire used by emergency physicians for coding, it appeared necessary to incorporate them in the ontology. Consequently, a concept was created and defined for each of them.

One of the major limitation of this work is the fact that the CIM-10 terms are suitable for coding, but some of them are difficult to manage in an ontology because they encompass several heterogeneous concepts. For example, one can find terms such as « subject waiting to be admitted elsewhere, in a suitable establishment » or « Symptoms and signs involving cognitive functions and consciousness, other and unspecified ». The concepts associated with such terms, because they articulate in a complex way a multitude of heterogeneous concepts, are difficult to model.

[3] The french Information System Medicalization Program (PMSI) is intended to introduce concepts of analytic accounting in the administrative management of hospitals: diagnosis and procedures performed in a health facility are coded and recorded, reported to a patient and to the various costs in the structure. This makes it possible to build cost indices for homogeneous group of patients. The PMSI uses an international coding system, the CIM-10, for diagnostics, and a French system, developed through an Anglo-Saxon ontological approach, the CCAM, for the medical acts.

3.4. Reusing the CCAM

The french CCAM classification (commune classification of medical acts) has the benefit of having been designed by teams familiar with ontologies. Which means *a priori* that each concept of this classification has been validated by a formal representation (Rodrigues et al., 1999). The reuse of the CCAM thus enabled us to incorporate a classification made up in accordance with consistent principles to our TOR.

The problems rather came from the way the CCAM is organized and designations used for the acts, which are built for specified accounting policies and not at all suitable for their expression in medical documents - our target. Much work has thus consisted in renaming the terms associated with concepts (*cf.* § 3.6).

These first two examples of reuse confirm, if it was still needed, that one essential difficulty in the reuse of resources comes from the fact that terminologies of all kinds - even ontologies - are each time developed according to specific purposes.

3.5. Reusing the SNOMED V3.5

The creation of the branch of diseases concepts is always a major part in the constitution of medical ontologies. As the necessary corpus for the design of such a branch were not available or did not cover the whole area, we decided to complete the work by integrating in OntolUrgences the diagnoses branch of the Snomed v3.5[4]. This procedure was mainly carried out by physicisians and required more than 100 hours of work: The Snomed v3.5 was notoriously too specific - what could be expected - but appeared also very badly organized - which was quite surprising. From the 25,000 diseases present in the The Snomed v3.5, 6,500 have been preserved.

3.6. Additional methodological comments

To complete the description of the construction of OntolUrgences, a few points of clarification are further needed:

1. OntolUrgences was developed with the OWL2 description logic (DL) language and with the Protégé ontology editor;
2. The SKOS[5] language was used for the formalization of the terms. The SKOS language is a representation language for knowledge organization systems such as thesauri, taxonomies, or any other type of controlled or structured vocabularies. This standard provides some primitives dedicated to the terminology with for each language, a preferred term `skos:prefLabel`, synonyms `skos:altLabel` and a definition `skos:definition`. Those primitives belonging to a standard commonly used are suitable for the representation of names and synonyms of the concepts of the ontology and can be perfectly mobilized within an ontology described in OWL. To facilitate the editing of these labels within the Protégé editor, we use a specific *plugin*[6].

[4]The Snomed v3.5 is a multiaxial classification whose development has been initialized by Canadian anatomopathologists. Its aim is to represent the whole domain of medicine and related notions of society. It contains 105,000 concepts. Snomed v3.5 exists in French and was chosen as the *reference terminology* by the French government (Rosenbloom et al., 2006). An ontology, the SNOMED-CT, has been derived from this classification by successive reorganizations and integrations of other terminologies. SNOMED-CT is not entirely available in French.

[5]The *Simple Knowledge Organization System* (SKOS) is developed within the W3C since 2003.

[6]The Archonte *plugin* was developed in our research unit by L. Mazuel, it corresponds to the integration of some of the features of the Doe software in Protégé associated with an annotation interface managing the multilingual character of terms and the SKOS (*definition*, *prefLabel* et *altLabel*) labels.

3. The resources used in the construction of ontology are diverse. As far as possible, we memorize the origin of the concepts with an annotation that specifies the identifier of the concept in the original resource, *SnomedId* for SNOMED v3.5 or *FmaID* pour the FMA (*Foundational Model of Anatomy*).
4. The concepts of the ontology can be distinguished among those used for IR and the others. The latter are either high-level structuring concepts – *e.g.* ObjetIntentionnel – or medical concepts too general to be discriminating – *e.g.* PhysicalExamination. This feature is described via a boolean annotation – terminologicalConcept – which specifies if the concept has a "terminological" character (it is potentially useful for IR) or not.

4. Meta-modelize to support enhancements

The ONTOLURGENCES ontology provides a conceptualization of the emergency field with terms to designate its concepts. This conceptualization can benefit from (i) the terms present in the KOS of Health to increase the detection of concepts in documents processed and from (ii) specific concepts about drug molecules in the ATC classification. To develop this new resource, you must be able to represent the KOS and ontology at the same level of description. Indeed, these resources are available in different formats and languages.

4.1. The UniMoKR metamodel

The diversity that exists in the nature, representation, and organization of the knowledge can be explained by different pasts, objectives, and uses. However, these KOS always intend to grasp information, to share it, and to support the human and computerised processing. Thus, it is possible to extract a common model core from this obvious heterogeneity (i.e. a model common to all knowledge structuring). In the field of knowledge organization system representation, some norms and standards are in place and facilitate the interoperability (Miles, 2006; BS8723, 2008; Clarke, 2008). Although SKOS and BS 8723 allow terminologies representation, none of them adress the issue of concepts group in a satisfactory manner[7]. We reuse in this project, the UniMoKR model designed in our previous work (Vandenbussche and Charlet, 2009) and illustrated in figure 2. This model uses and extends modeling elements from SKOS, BS 8723 and is already used by research and commercial projects (Joubert et al., 2011; Cormont et al., 2011).

The Termino-Conceptual part of UniMoKR model describes the relation between a *Concept* and its related *Preferred Term* and *Simple non preferred Terms* (aka synonyms) in each language. The Group Part enables not only the representation of a whole terminology, but also the representation of a terminology subset. It allows two different ways to characterize membership: by intension (concepts have to meet the restriction requirement to be part of a group; all concepts answering this request are implicitly members of the group) and by extension (concepts have to explicitly refer to this group via the relationship *inGroup*; this kind of definition is also available in SKOS and BS 8723). Our modeling reified the SKOS original alignement relations and allows alignments representation generated by various sources as well as the representation of the associated metadata information. Finally, meta-classes intend to guaranty the UniMoKR model extensibility and to facilitate its re-use and adaptation: some artifacts particular to some terminologies are not taken into account in UniMoKR; however, they need to be represented to avoid the loss of information. The terminology CIM-10-specific extension is illustrated in figure 3.

This modeling complies with the KOS construction best practices. As illustrated in the figure 4, the knowledge that constitutes the CIM-10 terminology is represented by instances. Note that the hierarchical relation between concepts is represented by an instance from the reified relationship

[7]For instance SKOS and BS 8723 models can not cope with SNOMED CT value sets or any concept groups defined in intension.

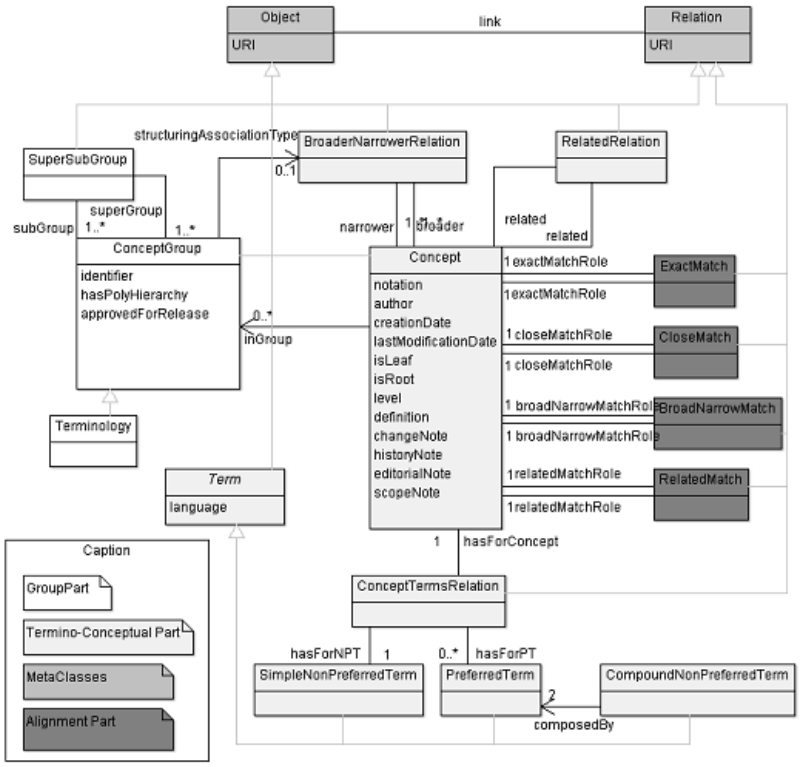

Fig. 2. UniMoKR Simplified UML classes Diagram.

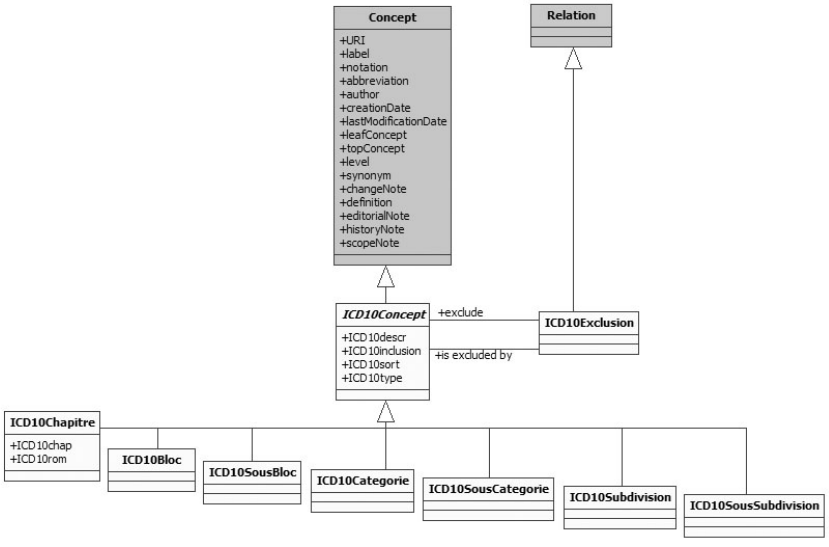

Fig. 3. Model extension for the CIM-10 terminology - UML classes diagram

BroaderNarrowerRelation (that carries a subsumption or partition relationship) and not by a subsumption relation. Specifically, it avoids frequent mistakes linked to the subsumption relation transitivity in cases where concepts must be linked by a partition relation (Aitken et al., 2003).

The UniMoKR model is inspired by, and is compliant with the current norms about KOS representation to avoid reducing the interoperability. UniMoKR is represented in OWL which fits perfectly the formal expressivity requirements necessary to the targeted semantic interoperability between KOS.

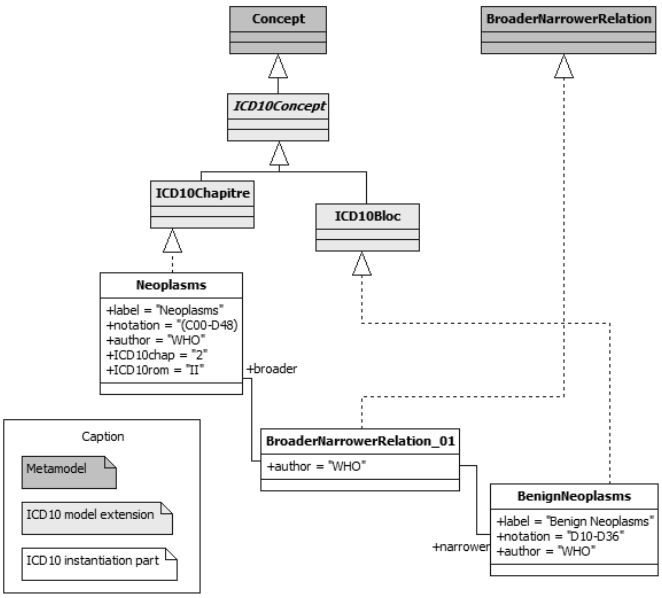

Fig. 4. CIM-10 extended model instantiation. - UML classes diagram

4.2. Process

Semantic repositories provided by INSERM, CISMeF and LERTIM laboratories to the French Shared Healthcare Information Systems Agency (ASIP) are integrated within a single repository server in the MONDECA's ITM tool. We used our UniMoKR model integrated in ITM tool for the representation of the KOS. Most of the terminology used in this project have already been modeled in the framework of the French research project InterSTIS[8] (Joubert et al., 2011). Extensions of the UniMoKR model are necessary to take into account the specificities of each KOS. Ontology OntolUrgences is in OWL DL format. We therefore can apply a model transformation to be compatible with our own model. This transformation consists in adapting the representation of ontology classes into UniMoKR concepts.

These matches can also be imported into the server. This process results in a server containing at the same level of description, the KOS and the mappings between their concepts needed to create the TOR. From this server, it is now possible to perform automatic operations for the ontology linguistic and conceptual enhancements.

5. Linguistic enhancement of the ontology

As mentioned above, for the Lerudi information system to be operational in situation, it is necessary that the TOR OntolUrgences covers almost all linguistic forms under which medical concepts relevant for emergency decisions appear in the EHR the system will have to deal with. Ultimately, the system must also be able to accommodate the « shortcuts » and « imperfections » of the language in which patient records are written, which for instance make use of abbreviations or may simply contain spelling errors.

The overall Lerudi system works as follows: the text of the various documents comprised in the EHR is processed by an algorithm that seeks to establish a correspondence (if necessary, by integrating NLP methods) between the phrases (treated as mere strings) and the system of

[8]The aim of the InterSTIS project, supported by the French National Agency for Research, is to develop a French multi-language medical server. See: `http://www.interstis.org/`

concepts of the TOR. If a string has been matched with a concept, the concept will be used to index the document (semantic interpretation process (Mazuel and Sabouret, 2007)).

Now, medical records are usually written in natural language (or at least in this semi-standardized language suitable for concrete medical activities), for the semantic interpretation process to reach a satisfactory level (or an optimal one: the optimum being set by the performance attained by an emergency physicist), it is often necessary to have available all lexical variations (synonyms, short forms, etc.) that may present the textual form of the concept. If a form encountered in the EHR has not been specified in the ontology, the record will not be indexed with the corresponding concept. The medical term will not be displayed by the interface. The emergency physicist will then have to put up with an incomplete or incorrect information[9].

To overcome this problem, two terminological enhancement processes of the TOR have been performed: (i) an automatic enhancement of the TOR by the adding of terms extracted from various KOS; (ii) a semi-automatic enhancement of the TOR by the adding of noun phrases extracted from the EHRs.

5.1. Enhancement of the TOR through the alignment with KOS

A first version of the enhanced TOR is realized through the aligment of the emergency domain ontology with 10 KOS relevant for the field, including CIM-10, SNOMED 3.5, MedDRA, ATC. By providing a controlled vocabulary, the SOC support the functions of analysis (annotation) of the EHRs. This enhancement process takes place in three main phases plus a phase of export:

Alignment of the ontology with the SNOMED V3.5**.** To align ONTOLURGENCES to other KOS, it was decided to align it initially to a reference ontology (Rosenbloom et al., 2006), the SNOMED V3.5. The advantage of this strategy is that it allows us to build on existing work of alignment of the SNOMED V3.5 to several KOS. The alignment was performed with the alignment software ONAGUI (Mazuel and Charlet, 2010) and by manually validating all the automatic alignments made.

Creating the mappings between SNOMED V3.5 **and other KOS.** CISMeF and Lertim laboratory have proposed a French lexical approach allowing mapping few KOS to the UMLS Metathesaurus to achieve interoperability between these KOS (Merabti et al., 2010).
This approach has been implemented with the UniMoKR model in order to be imported in a server and give us the alignment results betweenn mainly SNOMED V3.5, CISP, MeSH, ATC.

Lexical enhancement of concepts. The third phase is the enhancement of the concepts of the ontology with the terms of the concepts of the SOC matched. The alignments performed make it possible to establish connections of equivalence *exactMatch*[10] between concepts. The ONTOLURGENCES TOR is built automatically during the export phase. A pretreatment process supports the addition of terms coming from the various SOC to the concepts of the ontology (*cf.* figure 5). Each concept of the ontology that has been aligned with the concept of a SOC is enhanced with the terms of the latter. Such operation enables the concepts of the resource to have different lexical forms of designation.

[9] Authors such as (Mohammed and Sahroni, 2010) have suggested that a major reason for the rejection of medical information systems was related to the quality of information delivered by these systems. It is easily understandable within a framework such as the management of a patient in an emergency department: if the system provides incorrect information on the patient's medical history, the consequences can be dramatic. However, the case of incomplete information is more difficult to analyze because emergency physicians are used to work with little information, so that it is easier for them to handle this lack of information.

[10] We have deliberately used only strict equivalence connections to avoid the adding of terms that are too distant and therefore that cause noise in future detections.

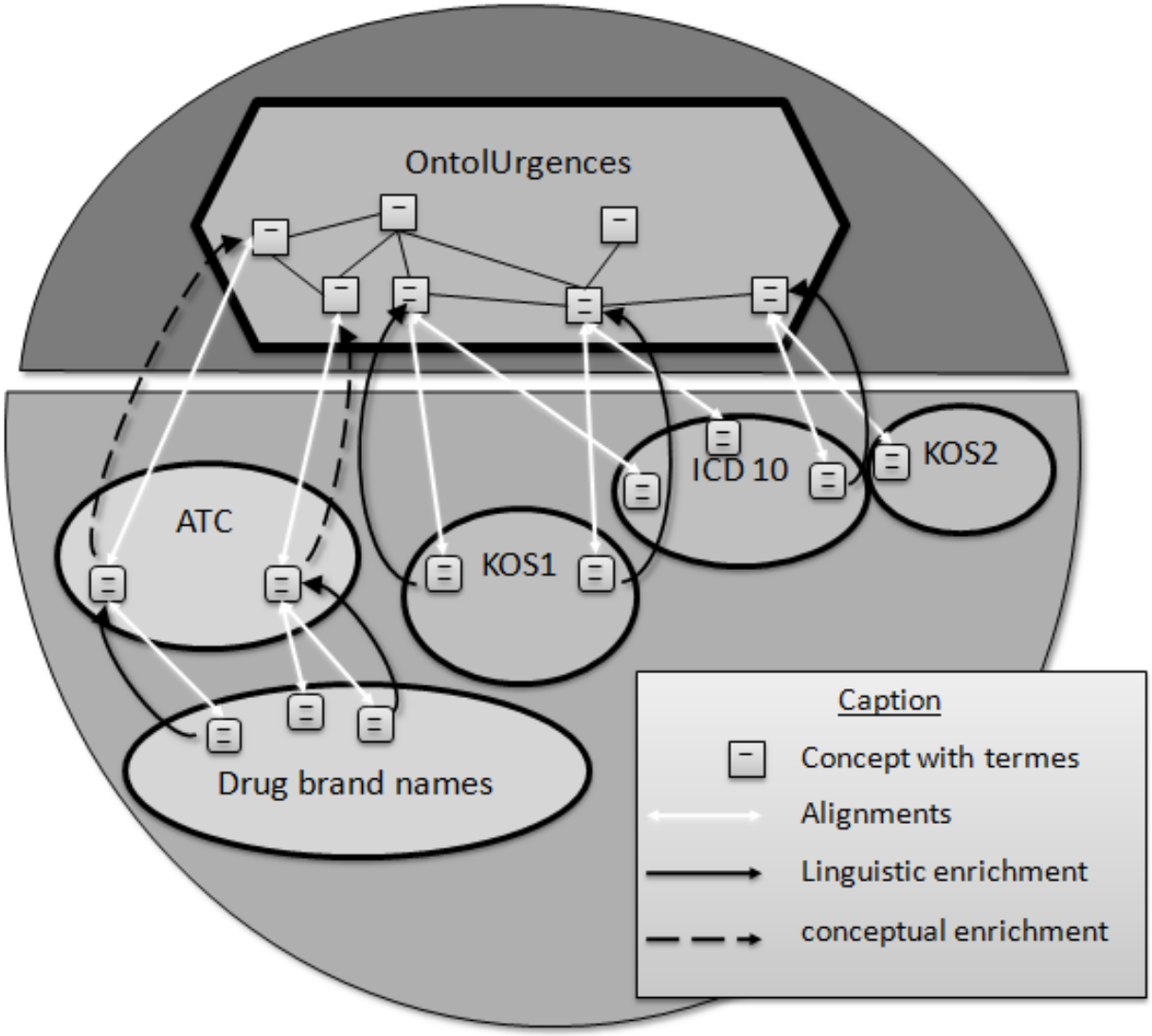

Fig. 5. Enhancement process of OntolUrgences through the alignment with KOS. Once the alignment between OntolUrgences and the KOS is performed, we enhance OntolUrgences mapped concepts with the terminological information coming from the KOS concepts. The enhanced ontology resulting from this operation is the so called TOR.

SKOS export. During the export phase, the TOR, now optimized for annotation and indexation, is made available in the SKOS format. Once the concepts of the ontology enhanced with lexical forms from the SOC, the representation model of the TOR is converted to SKOS. This operation of conversion is performed with the model transformation method described at the end of Section 4.2.

5.2. Enhancement of the TOR through the analysis of noun phrases

To improve the terminological completeness of OntolUrgences TOR, a complementary semi-automatic enhancement procedure was introduced. This procedure incorporates the principles of the *bottom-up* methodology used by the designers of domain ontologies. It includes the following steps: 1) we first analyze with NLP tools the content of the documents produced by the operating health professionals (*i.e.* the EHRs), in order to extract (this time by mobilizing statistical methods) the noun phrases likely to be among the most structuring of the considered field of knowledge, that is the terms that are specific and essential to the field; 2) Once these terms are identified,

health professionals (emergency physicians): A) perform a filtering operation to retain only the terms actually belonging to the medical field and likely to be clinically relevant during the process of IR in EHRs and B) validate the relevance of the identified synonymous terms; 3) these terms are then: A) added as synonyms (`skos:altLabel` tag) when they meet medical concepts already present in OntoLUrgences TOR, or B) converted into new concepts, when they refer to notions that do not yet have a conceptual representation in the TOR (in that specific case these terms correspond to the so-called *candidate-terms* of the *bottom-up* methodology (Mazuel and Charlet, 2009)). This conceptual conversion step requires to produce a formal definition of the concept being considered, which means firstly positioning the concept in the existing ontological hierarchy (and therefore assign to the new concept a « father concept », that is to say a concept which subsumes its meaning and of which it inherits the defining features).

Imagine that the above procedure selects the phrase « adenocarcinoma of the bile ducts », which does not exist in the TOR (nor as « preferred term » `skos:prefLabel`, nor as synonymous term `skos:altLabel`). The phrase is added as a synonym `skos:altLabel` if it matches an existing concept, it is converted into a new concept otherwise. In the specific case of this example, a concept [Bladder adenocarcinoma], which `skos:prefLabel` is « Adenocarcinoma of bladder », already exists: the phrase « adenocarcinoma of the bile duct » is then added as `skos:altLabel` the concept. Now, the detection of this phrase in an EHR will enable its annotation with the concept of [Bladder adenocarcinoma].

6. Conceptual enhancement of drugs

6.1. ATC

The Anatomical Therapeutic Chemical (ATC[11]) is a classification system used for the classification of drugs (Martindale et al., 2005). The ATC is divided in 14 groups each one refined in five levels (anatomical, therapeutic, therapeutic/pharmacological, chemical/therapeutic/pharmacological and chemical substance). ATC is a classification and not an ontology. Indeed, one drug can have more than one code (depending on the administration route or localization). Also, two concepts in the hierarchy may have the same label but not the same code (e.g. the concept "Antimycotis for systemic use" with the code "J02" has the same label than its child with code "J02A").

The ATC classification is widely accepted and used by health actors. For this reason, we made the choice to reuse this resource in our work for the description of drugs and their chemical substances. The Agence Française de Sécurité Sanitaire des Produits de Santé (AFSSAPS[12]), which is responsible for the introduction of new drugs in France, provides the names of authorized drugs with their associated ATC codes[13]. Thess links will be used in the linguistic enhancement process of the ATC concepts and will be describe later.

Rather than redefine the concepts of chemical molecules and drugs in OntoLUrgences, we preferred to integrate the ATC. To maintain a "formally correct" ontology and to ease the integration of ATC, we only included the 14 concepts that form the first level of the classification. We then developed a conceptual enhancement process that consists of integrating the rest of the ATC concepts during the development of the TOR.

[11] See `http://www.whocc.no/atc_ddd_index/`
[12] In english: French Agency for Sanitary Safety of Health Products.
[13] Available at the adress: `http://bit.ly/qg2755`

6.2. Conceptual enhancement process

As we use SOC to enhance the set of terms available in the TOR, we use the resources made available by the AFSSAPS French agency to enhance concepts from the ATC. This enhancement makes possible the detection of drug brand names in documents.

The procedure consists in adding brand names to the ATC concepts before including them into the ontology. For instance, we add as synonyms to the ATC concept "pseudoephedrine, combinations", identified by the code "R01BA52", the labels of drug names authorized in France using this chemical substance from which "ACTIFED", "ANADVIL" or "VICKS". This concept is then integrated into the TOR as a more specific (*narrower*) concept than the concept with the code "R01B".

This process is repeated for each concept of the ATC classification and results in the addition of 5,496 concepts to the TOR. Using data from the AFSSAPS, we have added over 17,000 drug brand names to ATC concepts. This enhancement is done automatically through a JAVA program that runs transformation rules based on SPARQL over RDF data. This method is described in the literature (Polleres et al., 2007; Morbidoni et al., 2007).

7. Validation processes

7.1. Why using validation procedures?

After one year of work, it appeared that the implementation of control procedures was necessary to maintain the quality of OntolUrgences TOR, and that these procedures had to be replayed regularly. Indeed, *a*) many stakeholders, physicians as well as modelers, are working together on the ontology, and despite all our efforts, we have not always been able to correctly apply the guidelines for the maintenance of quality and the homogeneity of the TOR. In addition, *b*) many instructions are binding and a person may apply them one day and forget them another.

These instructions mainly concern the terms related to concepts and the annotations related to the relevance of concepts for indexing documents (boolean annotation <terminologicalConcept> - *cf.* 3.6).

In a first step, these procedures do not address the structure of the ontology. The main reason is that at this level of development of the TOR and given the skills of the team, the problems we encountered were first terminological problems. But it is clear that problems of structuration, also present, call for future treatments (*cf.* 8). Our procedures are based on patterns, or anti-patterns when managing mistakes to be avoided. This work falls under the current research area concerned with the control of the quality of ontologies, as can be read on more structural points in (Roussey et al., 2010) or (Rector et al., 2004).

7.2. Which meta-model?

The quality control procedures were designed to ensure that the TOR meets the criteria of a specific meta-model. As far as this part is concerned, the meta-model can be expressed by the list of following rules:

- Each concept must carry an annotation <*terminologicalConcept*> in Boolean format;
- Each terminological concept (*cf.* previous rule) must have one, and only one, `skos:prefLabel` in French.
- Each terminological concept must have zero or one Skos:prefLabel in another language. The other relevant languages are: English, for the communication, and Latin, widely represented in the etymology of medical concepts;

- Due to the IR algorithms functioning, two different concepts must not have the same `skos:prefLabel` or the same `skos:altLabel` (string of identical characters);
- The `skos:hiddenLabel` proposed by the SKOS norm is used to store the part of the concept identifier that appears in the arborescence (the frag-URI) according to its language. Additionally, skos:hiddenLabel bears a meaning.

7.3. The procedures

The procedures are implemented by uploading the ontology to a SESAME store, and via SPARQL requests. Consequently, the quality criteria are verified in the *triplestore*:

- *Each terminological class must have a prefLabel* (see figure 6).
- *Each class must have one, and only one, prefLabel in the same language.*
- *Each prefLabel must be associated with a language.*
- *Each altLabel must be associated with a language.*
- *Each class must carry a hiddenLabel.*
- *Each class must carry only one hiddenLabel in the same language.*
- *Two classes must have the same prefLabel for the same language.*
- *Two classes must not have the same altLabel for the same language.*
- *Two classes must not have the same hiddenLabel in the same language.*
- *Two classes must not have one identical prefLabel and altLabel in the same language.*
- *Tracking of multiple parent classes.* The fact that one concept has two parent concepts is not a problem in OWL. However, this multiple parenting can be symptomatic of flawed modeling. In our methodology, the ontology is first designed based on a differential approach. The double heritage appears with the implementation of the defined concepts. Two parents can be allocated to one concept as an intermediary solution before the enhancement of the modeling. However, this double heritage, either intended or not, must be tracked and listed.
- *Additional requests.* Most of the additional requests have two primary optimization purposes – (a) to standardize the frag-URI in relation to lowercase or uppercase letters, and (b) to standardize the labels and remove, to the possible extend, the characters (such as brackets or parentheses) that could hinder proper matching.

```
select distinct ?conceptOnto
where{
       ?conceptOnto rdf:type <http://www.w3.org/2002/07/owl#Class>.
       OPTIONAL{?conceptOnto <http://cwi.nl/~troncy/DOE#prefLabel> ?prefLabel.}
       filter( !bound(?prefLabel))
              filter (!isblank(?conceptOnto))
}
```

Fig. 6. SPARQL Request verifying that each terminological class has one, and only one, prefLabel.

8. Conclusion and perspectives

Lerudi is a project that applies a specific methodology to the requirements of the medical emergency environment. The goal of this project is to build a TOR capable of retrieving information efficiently. Trough the complete description of the building process and the TOR validation in a large team, we have shown that:

1. concepts and terms must be precisely articulated within such a resource;
2. the developed meta-modeling architecture must allow the modeling of all necessary KOS and other knowledge structures;
3. standardized procedures based on this architecture may be implemented to enable the modeling.

Finally, the integration of the KOS in the same format and the RDF transformation service (capable of operating pre-treatments) allow to generate a termino-ontological resource with a lexicalization able to carry out the annotation, inference and indexation actions of the patients files. This project demonstrates the possibility to accommodate multiple KOS and to provide an efficient resource based on different request and transformation treatments.

References

Aitken, J., Webber, B., and Bard, J. (2003). Part-of relations in anatomy ontologies: a proposal for rdfs and owl formalisations. In *Pacific Symposium on Biocomputing 2004: Hawaii, USA, 6-10 January 2004*, page 166. World Scientific Pub Co Inc.

Bachimont, B., Isaac, A., and Troncy, R. (2002). Semantic Commitment for Designing Ontologies: A Proposal. In Gomez-Pérez, A. and Benjamins, V., editors, *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02)*, volume (2473) of *Lecture Notes in Artificial Intelligence*, pages 114–121, Sigüenza, Espagne. Springer Verlag.

BS8723 (2008). Structured vocabularies for information retrieval, part 4: Interoperability between vocabularies,.

Charlet, J., Bachimont, B., Bouaud, J., and Zweigenbaum, P. (1994). Ontologie et réutilisabilité : expérience et discussion. In *Actes des 5 ^{es} Journées Acquisition des Connaissances*, pages C1–C14, Strasbourg, France.

Charlet, J., Bachimont, B., Bouaud, J., and Zweigenbaum, P. (1996). Ontologie et réutilisabilité : expérience et discussion. In Aussenac-Gilles, N., Laublet, P., and Reynaud, C., editors, *Acquisition et ingénierie des connaissances : tendances actuelles*, chapter 4, pages 69–87. Cepaduès-éditions.

Charlet, J., Bachimont, B., Mazuel, L., Dhombres, F., Jaulent, M.-C., and Bouaud, J. (2009a). Ontomenelas : motivation et retour d'expérience sur l'élaboration d'une ontologie noyau de la médecine. In *3^{e} Journées Francophones sur les Ontologies*, Poitier, France. ACM.

Charlet, J., Baneyx, A., Steichen, O., Alecu, I., Daniel, C., Bousquet, C., and Jaulent, M.-C. (2009b). Utiliser et construire des ontologies en médecine : Le primat de la terminologie. *Techniques et Sciences Informatiques*, 28(2).

Clarke, S. (2008). Iso 2788+ iso 5964+ much energy= iso 25964. *Bulletin of the American Society for Information Science and Technology*, 35(1):31–33.

Cormont, S., Vandenbussche, P.-Y., Buemi, A., Delahousse, J., Charlet, J., and Lepage, E. (2011). Implementation of a platform dedicated to the biomedical analysis terminologies management. In *American Medical Informatics Association (AMIA) Annual Symposium*.

Després, S. and Crampe, M., editors (2010). *Actes des 21 ^{es} Journées Ingénierie des Connaissances*, Nîmes, France. Presse des Mines.

Dhombres, F., Jouannic, J., Jaulent, M., and Charlet, J. (2010). Choix méthodologiques pour la construction d'une ontologie de domaine en médecine prénatale. In Després and Crampe (2010).

Dhombres, F., Vandenbussche, P., Rath, A., Hanauer, M., Olry, A., Urbero, B., Choquet, R., and Charlet, J. (2011). Projet OrphaOnto - première étape de l'ontologisation des bases de connaissances d'Orphanet. In *Acte s des 22e Journées Francophones d'Ingénierie des Connaissances*, pages 1–12, Chambéry, France. [accepté].

Gayet, P., Charlet, J., Josseran, L., and Miroux, P. (2010). Représentation de la médecine d'urgence dans le corpus des abstracts du congrès urgence. In *Actes du congrès URGENCES 2010*. Poster.

Giroud, M. (2009). L'accès au dossier médical personnel par le médecin régulateur du samu. In *Actes du congrès URGENCES 2009*. Ce texte s'appuie sur les travaux du Groupe animé par l'ASIP santé (ex GIP-DMP) : M. Baudot, M. Bloch, E. Clout, N. Janin, J.-M. Picard, C. Attali, M. Thonnet, H. Vu Than P. Liot, R. Picard, J. Charlet, I. Colombet, M.-C. Jaulent, S. J. Darmoni, M. Joubert, M. Fieschi, P. Lesteven, Y. Lannehoa, F. Braun, M. Giroud, P. Menthonnex.

Joubert, M., Merabti, T., Vandenbussche, P.-Y., Abdoune, H., Dahamna, B., Fieschi, M., and Darmoni, S. (2011). Modeling and integrating terminologies into a french multi-terminology server. In *Poster presented at MedInfo*.

Martindale, Sweetman, S., and Martindale, W. (2005). *The complete drug reference*. Pharmaceutical Press (2002). Martindale: The complete drug reference, thirty-fifth edition, London.

Mazuel, L. and Charlet, J. (2009). Alignement entre des ontologies de domaine et la Snomed: trois études de cas. In Gandon, F., editor, *Actes des 20 ^{es} Journées Ingénierie des Connaissances*, pages 133–144, Hammamet, Tunisie.

Mazuel, L. and Charlet, J. (2010). Alignment between domain ontologies and snomed: three case studies. In Safran, C., Marin, H. F., and Reti, S. R., editors, *MEDINFO 2010 - Proceedings of the 13 th World Congress on Medical and Health Informatics - Partnerships for effective e-Health solutions*, volume 160, Cape Town, South Africa. IOS Press. Poster.

Mazuel, L. and Sabouret, N. (2007). Degré de relation sémantique dans une ontologie pour la commande en langue naturelle. In Trichet, F., editor, *Actes des 18 ^{es} Journées Ingénierie des Connaissances*, pages 73–85, Grenoble, France. Cépaduès. ISBN 978.2.85428.790.5.

Merabti, T., Massari, P., Joubert, M., Sadou, E., Lecroq, T., Abdoune, H., Rodrigues, J., and Darmoni, S. J. (2010). An automated approach to map a french terminology to UMLS. *Studies in Health Technology and Informatics*, 160(Pt 2):1040–1044.

Miles, A. (2006). Skos: requirements for standardization. In *DC-2006: Proceedings of the International Conference on Dublin Core and Metadata Aplications*, pages 55–64.

Mohammed, S. A. and Sahroni, M. N. (2010). Information quality as success determinant for health information systems. In *Proceedings of of the 2010 Regional Conference on Knowledge Integration (ICT 2010)*, pages 674–679.

Morbidoni, C., Polleres, A., Tummarello, G., and Le Phuoc, D. (November 2007). Semantic web pipes. Technical report, DERI.

Polleres, A., Scharffe, F., and Schindlauer, R. (2007). Sparql++ for mapping between rdf vocabularies. *Lecture Notes in Computer Science*, 4803:878.

Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H., and Wroe, C. (2004). Owl pizzas: Practical experience of teaching owl-dl: Common errors & common patterns. In *In Proc. of EKAW 2004*, pages 63–81. Springer.

Reymonet, A., Thomas, J., and Aussenac-Gilles, N. (2007). Modélisation de ressources termino-ontologiques en owl. In Trichet, F., editor, *Journées Francophones d'Ingénierie des Connaissances (IC)*, pages 169–180, http://www.cepadues.com/. Cépaduès Editions.

Rodrigues, J.-M., Trombert-Paviot, B., Rector, A., Baud, R., Clavel, L., Abrial, V., Idir, H., and Very, J.-M. (1999). GALEN, il existe quelque chose après les mots : leur signification et au delà le savoir médical. *Innovation Stratégique en Information de Santé*, (2–3):48–62.

Rosenbloom, S. T., Miller, R. A., and Johnson, K. B. (2006). Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. *J Am Med Inform Assoc*, 13(3):277–88.

Roussey, C., Scharffe, F., Corcho, O., and Zamazal, O. (2010). Une méthode de débogage d'ontologies OWL basées sur la détection d'anti-patrons. In Després and Crampe (2010), pages 43–54.

van Heijst, G., Schreiber, A. T., and Wielinga, B. J. (1997). Using explicit ontologies in KBS development. *International Journal of Human-Computer Studies*, 45(2/3):183–292.

Vandenbussche, P.-Y. and Charlet, J. (2009). Méta-modèle général de description de ressources terminologiques et ontologiques. In *Ingénierie de la Connaissance (IC)*.

Zweigenbaum, P., Bachimont, B., Bouaud, J., Charlet, J., and Boisvieux, J.-F. (1995). Issues in the structuring and acquisition of an ontology for medical language understanding. *Methods of Information in Medicine*, 34(1/2).