Use of shared lexical resources for efficient ontological engineering

Antonio Jimeno-Yepes[1], Ernesto Jiménez-Ruiz[2], Rafael Berlanga[2], and Dietrich Rebholz-Schuhmann[1]

[1] European Bioinformatics Institute,
Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
{yepes,rebholz}@ebi.ac.uk
[2] Dept. of Computer Systems and Languages,
Universitat Jaume I, Spain,
{ejimenez,berlanga}@uji.es

Abstract This paper is intended to approach one of the main problems in ontology engineering: *the lack of a shared terminology.* Nowadays there exists several biomedical ontologies describing overlapping domains, but there is not a clear correspondence between the concepts that are supposed to be *equivalent* or just *similar*. These resources are quite precious but their integration and further development are expensive. Terminological or lexical resources may support the ontological development in several stages of the lifecycle of the ontology including ontology integration and the labeling of concepts. In this paper we investigate the use of lexical resources during the ontology lifecycle using the example of the Health-e-Child (HeC) project. We claim that the proper creation and use of a shared lexicon is a cornerstone for the successful application of the Semantic Web technology within life sciences.

1 Introduction

Large domain ontologies are emerging from collaborative efforts in the life sciences, being its main aim to achieve the interoperability among the different research resources by assuming a common conceptualization. These resources mainly consist of both domain ontologies and terminological resources (e.g. thesauri), which allow researchers to process, store and share the ever increasing knowledge derived from their experiments. So far, these two kinds of resources have usually lived apart, being its later integration a very hard task. However, some exceptions exist where the lexicon is integrated with a semantic network (e.g. the Unified Medical Language System[3]). In this paper, we show that both cases have serious drawbacks.

Instead, we propose a loose coupling between the domain ontologies and a unique lexicon. Along this paper we show that the use and maintenance of such a shared lexicon will enable both a better integration of domain ontologies with

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

existing lexical resources and the proper evolution of the lexicon according to these ontologies. We claim that the use of a shared lexicon will ease some of the problems present during the development of ontologies and the interoperability of the ontologies.

In this paper we assume that domain ontologies and lexicons have different purposes, and therefore they cannot be treated with the same techniques nor simply merged into a common resource. A *lexicon* consists of a compendium of words enriched with information of its usage [1], being concern with the linguistic properties of words. We may encounter as well the term *terminology*, which is usually referred as a *specialized lexicon* [2]. Instead, a domain ontology is an explicit specialization of a conceptualization [3]. Domain ontologies have much more specific purposes than lexicons, as their intended consumers are computer applications rather than humans. Thus, domain ontologies do not need to care about variants and syntactic categories of the terms they use. In addition, the specific purpose of the ontologies motivates the development of different ontologies that still can label the concepts based on a shared lexicon.

Regarding the semantically equivalent groups defined in a lexicon (e.g. thesaurus entries, synsets, etc.), they also present significant differences with respect to the concepts of domain ontologies. These semantic groups do not offer a clear cut on their meaning as in an ontology where the concepts present disjoint interpretations. Rather lexicons present fuzzy frontiers to allow the slightly different interpretations humans can express with them.

In Figure 1 we have ordered the existing formalisms (denoted by boxes) according to their semantic expressiveness. Existing biomedical resources are placed to their closer formalism. Genuine lexical resources are placed closer to the left of the diagram like the Biolexicon[4], that contains terminology from several resources with some linguistic relevant information. We find as well the UMLS Specialist lexicon that has been used within several NLP and text mining applications. Closer to the limit between a lexicon and an ontology we find several resources that include links between lexical entries (e.g. UNIPROT). More complex resources lie in between the definition of ontology and lexicon like the NCI Metathesaurus, MeSH, ICD, the UMLS Metathesaurus and the OBO ontologies that account for more complex representations similar to semantic networks. Finally, at the end of the spectrum we find more formal ontologies such as FMA or Galen, which expresses stronger semantics. Unfortunately, these formal ontologies usually lack of lexical entries. As mentioned before, the aim of this paper is to approach the problem of making these resources interoperable.

The selected examples and use cases presented in this paper come from the the application domain of the EC FP6 Health-e-Child (HeC) project [7] that aims to develop an integrated health care platform for European paediatrics and decision support tools to access personalized health information. HeC project is mainly focused in paediatric heart diseases, inflammatory diseases (e.g. Juvenile Idiopathic Arthritis) and brain tumours.

The paper is organized as follows. Section 2 presents the ontology lifecycle steps and motivates the relevance of having and using a shared lexicon/thesaurus.

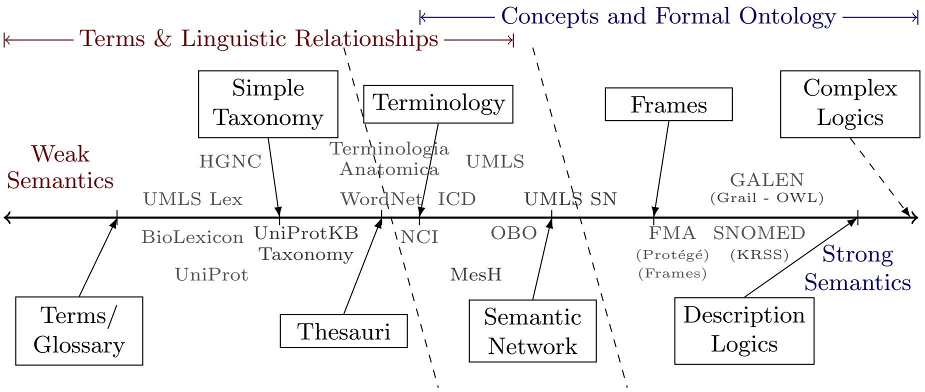

Figure 1. Adapted Ontology Spectrum based on [5, 6, 2]

A discussion about current efforts, limitations and desired requirements for a shared lexicon are presented in Section 3, moreover the main lexicon engineering techniques are introduced. Section 4 comments the main experiences carried out within the HeC domain. Finally some conclusions are given in Section 5.

2 The Role of Lexicons in the Ontology Lifecycle

Lexical forms present in available resources can be used for labeling ontological concepts. The reuse of these labels in different ontologies in combination with a proper definition of the ontological concepts may enable better integration of ontologies. This section is intended to show the main problems that experts, knowledge engineers and ontology engineers find in the different stages of the lifecycle of the ontology development and how the use of a shared lexicon could ease these problems.

In this paper we adopt the METHONTOLOGY methodology [8] to illustrate how a shared lexicon can help the development of an ontology and vice versa. METHONTOLOGY proposes several steps for the lifecycle of an ontology: Requirements Specification, Knowledge Acquisition, Conceptualization, Integration with top ontologies, Implementation, Evaluation and Evolution/Maintenance. As Figure 2 shows, the shared lexicon interacts with almost all the development phases. Moreover, external resources like domain protocols, domain ontologies and research articles will also play an important role as sources of knowledge. In the following subsections we describe in detail the role of the lexicon at each development phase.

2.1 Requirements Specification

Within the objectives of the HeC project, an ontology to describe a kind of arthritis called JIA (Juvenile Idiopathic Arthritis) requires to be created. This

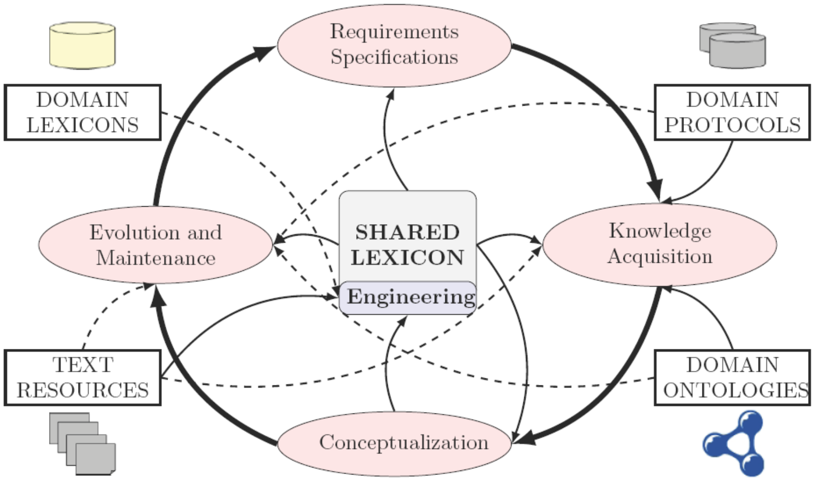

Figure 2. The Lexicon within the Ontology Life Cycle. *Solid arrows represent an essential role, whereas dashed arrows mean auxiliary role.*

ontology is intended to represent the involved knowledge in JIA by means of different levels of granularity: molecular (e.g. genomic and proteomic data), cellular (e.g. results of blood tests), tissue (e.g. synovial fluid tests), organ (e.g. affected joints), body (e.g. damage index, rheumatology examinations, treatments), population (e.g. epidemiological studies). The purpose of this multilevel representation is to give a complete characterization of the different JIA subtypes in order to provide a rich ontological layer to the HeC System. This semantic layer will be applied in *Query Enhancement* over the patient data, and in the *Decision Support Systems*. JIA is a rare kind of Arthritis and there is not yet a consensus about its classification nor even its name [9]. So far, three classification schemes have been proposed, namely: ACR (American College of Rheumatology), which uses *Juvenile Rheumatoid Arthritis (JRA)* as preferred name and proposes three disease subtypes, EULAR (European League Against Rheumatism), which opts for *Juvenile Chronic Arthritis (JCA)* and proposes six disease subtypes, and finally ILAR (International League of Associations for Rheumatology) which prefers JIA and proposes eight subtypes. In this stage, a classification criterion should be chosen and the initial set of terms for describing the disease and subtypes must be defined. Clearly, the use of a lexicon would make easier the selection of terms (synonyms) for labeling the desired concepts.

2.2 Knowledge Acquisition

The knowledge acquisition in HeC is based on a set of medical protocols (in [10] several techniques to automatically extract the main concepts from HeC acquisition protocols are proposed) and the correspondent specifications of the mentioned classification criterion. Each subtype of JIA is characterized by affecting different set and number of joints, the occurrence of some symptoms like fever or rash, the laboratory tests that are analysed, the different treatments that are applied, etc. The development of the ontology from scratch would imply the conceptualization of the different joints of the body, the classification of the drugs for the treatments, the characterization of the different laboratory tests, etc. Nevertheless this knowledge is already well known by the community (unlike JIA) and it is assumed to be already defined in the available biomedical ontologies. As far as we know, the NCI thesaurus[4], the GALEN ontology[5], and the OBO ontologies[6] contains information that is relevant to JIA such as descriptions of diseases, drugs, laboratory tests, cells, human anatomy, etc.

The reuse of knowledge represented in ontologies (see [11] for a survey) could be interesting due to the following reasons: (a) developers save time through reusing existing ontologies rather than writing their own; (b) the used knowledge is commonly accepted by the community and used in similar applications; (c) developers are not always experts in all the areas covered by a concrete disease (i.e. drug classification). However, in practice important drawbacks arise when merging ontologies. In this case, *Ontology Matching*[7] should be performed, that is, to discover the correspondences between entities of the different ontologies. This task is rather hard [12] since in most cases there is not a common nomenclature for the entity names. String matching techniques could provide an approximate results in some cases like *"NCI:Juvenile_Rheumatoid_ Arthritis"* and *"Galen:JuvenileArthritis"*. However in other examples like *"DiseaseOntology:Chronic Childhood Arthritis"*[8] *additional knowledge* should be provided in other to establish the matching between concepts.

Additionally, *Semantic Compatibility* should be also taken into account. Once the lexical correspondence between concepts has been established, the ontologies (or ontology modules) can be merged. At this point new challenges about the semantic compatibility between the ontology axioms (e.g. unsatisfiability when merging) arise, but they are outside the scope of this paper.

Currently there exist several efforts in the creation of large biomedical ontologies. However, it seems they are evolving in a rather independent way. We can understand that the conceptualization and formalization evolve with respect to the specific requirements of a specific application, but the used nomenclature should be shared. For example, the use of the concept *Chronic Childhood Arthritis*[9] could vary between different domain ontologies, but the used term (JIA, JCA or JRA) should refer to the same entity in the domain. The reuse of terms for labels from a shared lexicon (e.g. UMLS) will relax considerably the required

[4] NCI thesaurus: `ftp://ftp1.nci.nih.gov/pub/cacore/EVS/NCI_Thesaurus`

[5] GALEN Ontology: `http://www.co-ode.org/galen`

[6] Open Biomedical Ontologies: `http://www.obofoundry.org/`

[7] Ontology Matching Initiative: `http://www.ontologymatching.org/`

[8] *Disease Ontology* is an example of ontology enriched with synonyms coming from shared thesauri like UMLS or ICD. However, we propose that ontologies should only maintain a link to the correspondent thesaurus, see Section 3.4

[9] *Chronic Childhood Arthritis* is the preferred concept name in UMLS

matching tasks between ontologies. Ontology concepts could use any preferred nomenclature (no spaces, use of hyphens, acronyms, short expressions, etc.) but they will be *annotated* with a unique concept interpretation, that is, they will point to an entry in the shared lexicon.

As commented earlier, *knowledge acquisition* can require merging different sources and ontologies. METHONTOLOGY already proposes the creation of a glossary to enrich ontologies with synonyms and definitions in order to facilitate the integration with other resources. Undoubtedly, this proposal should be kept, but we should go further by making such a glossary available to all community.

2.3 Conceptualization

As commented above, the granularity of the ontology will be connected to the purposes of the application, in this sense, the same entry in a lexicon could have different interpretations within different ontologies. This characteristic is related to the *localized semantics* proposed in [13], in which the concept *context* is defined as local models representing a partial or concrete view of the domain. For our purposes the concepts and theory treated in [13] are rather complex but the general idea of the local use of a shared *concept* is important. For example, following the mentioned classification criteria, the concept *Chronic Childhood Arthritis* may have the interpretations given in axioms 1 to 3.

$$\begin{aligned}
\mathsf{ACR : JRA} &\equiv \mathsf{SystemicJRA} \sqcup \mathsf{PolyArticularJRA} \sqcup \mathsf{PauciarticularJRA} \quad (1)\\
\mathsf{EULAR : JCA} &\equiv \mathsf{SystemicJCA} \sqcup \mathsf{PolyArticularJCA} \sqcup \mathsf{PauciarticularJCA} \sqcup \\
&\quad \mathsf{Juvenile\_Psoriatic\_Arth.} \sqcup \mathsf{Juvenile\_Ankylosing\_Spondylitys} \quad (2)\\
\mathsf{ILAR : JIA} &\equiv \mathsf{SystemicJIA} \sqcup \mathsf{PolyArticularJIA} \sqcup \mathsf{OligoarticularJIA} \sqcup \\
&\quad \mathsf{Psoriatic\_Arthritis} \sqcup \mathsf{Enthesisrelated\_Arthritis} \quad (3)
\end{aligned}$$

Such interpretations may belong to three different JIA ontologies used, probably, for different application purposes. If in some moment they are required to be integrated in only one ontology (perhaps a consensus is achieved and it is established a unique classification criteria) the matching between labels (terms) would be easier if a shared lexicon was used to annotate the concepts (i.e. *Chronic Childhood Arthritis* $\equiv$ *JIA* $\equiv$ *JCA* $\equiv$ *JRA*). The semantic integration, as commented in Section 2.2, will depend on the compatibility of the used axioms within the conceptualization and formalization of the merged JIA ontologies.

It is worth mentioning that the design requirements of an ontology may involve concepts with labels that are not present in most of the available lexicons. For example, not all the subtypes of JIA are properly described in UMLS. As commented, lexicons will help ontologies to use a common nomenclature, but ontologies will also help lexicons to evolve. In general, ontologies will require a finer granularity than the initially expressed by lexicons and will demand the necessity of new concepts given the specific requirements of the domain. Obviously, a new challenge arises, that is, how to maintain consensual and shared lexicons up-to-date with respect to the new specific ontologies and their evolutions. Additionally, *hypernym* relationships within the lexicon may be useful in order to check the coherence of the ontology conceptualization, that is, it may be helpful to consider desired subsumptions (e.g. $JIA \sqsubseteq Systemic\_Disease$) or even to avoid non desired ones (e.g. $JIA \sqsubseteq Non\_Systemic\_Disease$).

2.4 Evolution and Maintenance

The evolution and maintenance (addition of new concepts, the deletion of obsolete ones, the re-structuring of the already defined concepts, the addition of new facts, etc.) of an ontology may be produced due to different reasons: requirements changed, the domain has changed (e.g. new facts were discovered) or the point of view of the domain changed (e.g. use of a different classification criterion). The evolutions will imply to come back to previous steps in order to acquire new knowledge and to integrate this knowledge within the ontology. Again, the lexicon will play a key point providing the concepts necessities when possible or being updated with new ontology requirements in order to keep up-to-date for further ontology demands.

In biomedicine the change and extension of the domain evolves quickly. Publications represent an important source of *brand new* facts of domain knowledge. For example Medline[10] indexes more than 800,000 new journal papers per year containing the last research done in more than 700 topics. Text mining techniques try to identify within the text concepts and facts relating them. These techniques usually use domain lexicons in order to detect interesting entities within text. However several studies (e.g. [14]) have already shown that the link between the most relevant biomedical resources and the literature is not obvious. This is not only due to the complexity of the required matching algorithms but also due to the decouple of the ontology/lexicon development effort and the literature. In an important number of cases current lexical resources (in Section 3.1 some examples are given) do not provide useful synonyms to be detected within the text. In order to overcome these problems lexicons should better select the synonyms to characterize their concepts considering, at least, the lexical variants used in texts.

3 Towards a Thesaurus for Life Sciences

We have presented the relevance of the lexicon in the ontology lifecycle and how this lexicon could be updated accordingly. Basically, the lexicon will provide the necessary terminology required for the existing concepts. In case there is no entry (e.g. subtypes of JIA) for a concept, the current process may suggest the creation of this new entry.

The proper creation of new entries will require the selection of the appropriate terms (i.e. preferred name and synonyms). These terms may be provided by a

[10] Medline: http://medlineplus.gov/

community effort, where several domain experts study the appropriate set of terms, and/or using natural language processing (NLP) and text mining [15] to extract such terms from the literature [16]. A proposal for automatic term management (ATM) can be found in [17]. This approach identifies three modules. The first module is about automatic term recognition, which identifies lexical structures that can be mapped to domain concepts. The second module implies term structuring for identifying relevant relations or term associations, mainly by using classification and clustering techniques. The last module consists of an intelligent term manager that in addition of storing the terms accordingly it may provide links and definitions to existing resources. Existing resources can be reused either to train the classifiers or to use them in dictionary approaches to term recognition. In addition to this approach we can use approaches that collect existing structures from available terminological resources. For example, UMLS is the result of merging several medical resources and thesauri. In this case, similar issues to the alignment of ontologies have to be addressed. It is worth mentioning that even this approach requires ATM solutions for extending and maintaining the resulting *meta thesaurus*.

As a consequence, the existence of a common thesaurus can help to link concepts from existing resources at the same time that it ensures no duplicate entries for the same concepts. This thesaurus will collect the different terms in a common repository allowing ontologies to be linked accordingly. Thus, the final scenario consists of one thesaurus and many specific ontologies. These ontologies may be designed according to different criteria, for they are usually applied in different contexts. We find the best example in the OBO ontologies where several ontologies can overlap in some of their concepts.

The generation of a common thesaurus requires the resolution of several issues like an agreement concerning the meaning of the entries in the lexicon. As we have seen, JIA already presents a difficult conceptualization even among domain experts. The outcome of the research in the field may require not only to create new concepts but also to split existing ones. This will imply the necessity of maintaining the ontology up-to-date, since some of the links have become obsolete (see Section 3.4). Additionally, another way of solving the problem would consist of the generation of several versions.

Although current approaches represent an important initiative for the construction of a shared lexicon they still lack some important requirements to allow a straight forward interoperability with ontologies and text resources. Next section presents the main limitations of current efforts and proposes some requirements to be followed in order to get the intended lexicon.

3.1 Limitations of current reference lexicons

The UMLS Metathesaurus (UMLS-Meta) represents the best effort for the creation of a *reference thesaurus*. However it has several drawbacks, most of them because of its complexity, since in some cases the UMLS-Meta is closer to an ontology than to our intended thesaurus/lexicon. The UMLS-Meta contains concepts from more than 100 terminologies, classifications, and thesauri, for example: MesH, SNOMED CT or ICD. This makes UMLS-Meta a really rich source of knowledge, but also a source of ambiguity, redundancy and meaningless entries. In the literature we can find some efforts [18, 19] to normalize the UMLS-Meta by filtering redundancy and solving a basic level of the ambiguity [11]. However some ambiguity cases are rather hard to solve. This is the case of the term *Prostate Cancer* which has associated two UMLS-Meta entries: $C0600139$ and $C0376358$. Both concepts refer to the Neoplastic Processes, *Carcinoma of prostate* and *Malignant tumor of prostate*, respectively. These Neoplastic Processes have a close relationship, indeed the former is represented as a child of the later within the NCI and UMLS-Meta taxonomies.

After filtering redundant cases and solving some of the trivial ambiguity problems the UMLS-Meta still contains a huge number of concept labels that surely will not have a correspondence neither in ontology labels nor texts. Next we present some representative cases (extracted from a portion of UMLS-Meta related to the JIA domain) that our intended thesaurus should avoid:

Descriptive names Some synonyms are closer to a text definition than to a concept name. For example, *Therapeutic or Preventive* concepts $C0199105$: *"Anaesthesia for open procedure on knee joint Procedure"*, and $C0580168$: *"Amputation of finger through distal interphalangeal joint"*. Nevertheless, not all concepts can be described with a few words. Indeed, such complex concepts should be described in formal ontologies by combining somehow smaller units of meaning of the lexicon, e.g. concept $C0580168$ can be formally described as $Amputation \sqcap \exists involve.Finger \sqcap \exists through.InterphalangealJoint$, where the semantics for each of its elements is defined in a formal ontology. Additionally, each of the concept constituents can be linked to entries of the lexicon.

Parametrization in the label The *Clinical Drug* $C1614077$ has the preferred name *"Etanercept 50 mg/mL subcutaneous solution"*. This term indicates not only the drug name but also the dosage for this pharmaceutical product. Therefore the lexicon should contain only the generic name, and then the formal ontologies should represent *"Etanercept 50 mg/mL subcutaneous solution"* as either a subclass of *"Etanercept"* or just as an instance.

Complex nomenclature Chemicals formulae as in concept $C0255404$: *"N-methyltropan-3-yl 2-(4-bromophenyl)propionate"* are useful as a definition of the concept, but an ontology concept or a thesaurus term should not use this nomenclature. Moreover, entities detected in text rarely will match with this term.

Inappropriate syntax Concept $C0366794$ with Semantic Type *Clinical Attribute* has the preferred name string *"Hemoglobin C/Hemoglobin.total:Mass Fraction:Point in time:Whole blood"*. Obviously this string is encoding some data perhaps only understandable in the source vocabulary.

As commented above, UMLS-Meta is more complex than a simple thesaurus or a glossary of terms and it does not only contain synonymy relations but also

[11] Filtering UMLS and solving Ambiguity: `http://skr.nlm.nih.gov/papers/`

inclusion relations like hyponymy and hypernymy (i.e. is-a or subsumption relations in ontologies) and part-whole relationships like meronymy and holonymy (i.e. has-part, part-of). This makes UMLS-Meta really hard to evolve and maintain properly, and the inclusion of new vocabularies may introduce unexpected classifications of the concepts (i.e. cycles). The desired thesauri should contain a clearer hierarchy with or only *hypernymy* or only *meronymy*. The granularity of the lexicon hierarchy could vary from a top level ontology classification (e.g. UMLS Semantic Network) to fine granularity hierarchies like the OBO classifications or the UMLS-Meta hypernymy hierarchy itself. More complex classification of the concepts should be delegated to the ontology conceptualization process.

3.2 Limitations of current reference ontologies

The OBO ontologies present a huge community effort in the development of ontologies, but we still miss the use of a common lexicon/thesaurus to normalize the used nomenclature. Moreover, the OBO ontologies, like the UMLS-Meta , are in the middle of what we expect from an ontology and from a lexicon. The underlying logic of the OBO ontologies is not too complex, being in most cases limited to simple taxonomies (e.g. *Disease Ontology*). The Gene Ontology has also assertions but in the most of cases they refer to concept metadata. The use of more complex logic would give more expressive power to express complex concepts that can not be described only by a name an a set of subsumptions. Moreover this kind of ontologies will provide a framework to classify facts or concepts of the world according to a concept definition without making an explicit specification of the subsumption. This would make easier the introduction of new concepts within the hierarchy by only giving the definition of the concept. For example, from the simple set of axioms 4 to 6 we would infer that $JIA \sqsubseteq Systemic\_Disease$, but without defining explicitly this axiom.

$$\mathsf{Disease} \sqcap \exists \mathsf{affects.Whole\_Body} \sqsubseteq \mathsf{Systemic\_Disease} \tag{4}$$

$$\mathsf{JIA} \sqsubseteq \mathsf{Arthritis} \sqcap \exists \mathsf{affects.Whole\_Body} \tag{5}$$

$$\mathsf{Arthritis} \sqsubseteq \mathsf{Disease} \tag{6}$$

Nevertheless, complex logics have also drawbacks in the sense of computability and therefore a good balance between efficiency and expressivity should be achieved. Apart from the expressivity issues, the OBO ontologies also present some lexical problems. On one hand some of the concepts names used in OBO ontologies, like some UMLS-Meta entries, are closer to definitions than to a concept name (e.g. GO:0007180 *"transforming growth factor beta ligand binding to type II receptor"* (biological_process) or GO:0016456 *"X chromosome located dosage compensation complex, transcription activating"* (cellular_component)). Like in UMLS, these concept names are of little help when performing for example text mining tasks. On the other hand, these ontologies are overloaded with too much metadata (i.e. synonyms, definitions, references) making hard their management. For example, the *Human Disease Ontology* contains 14772 classes (and 1 property), 18593 subsumption relationships and 442168 entity annotations (i.e. synonyms, references to entries of other thesaurus, mainly UMLS, ICD, SNOMED and MESH), therefore an average of almost 30 annotations per class. The case of the *Gene Ontology* is similar, containing more than 450000 entity annotation axioms for less than 30000 classes, and 150000 assertion axioms being used as a annotation values.

More formal ontologies like the Foundational Model of Anatomy (FMA is available as Protégé Frames) and Galen (available in Grail and OWL) seem to be projects being developed independently with respect to the UMLS-Meta and OBO foundry efforts. Galen contains some information about synonymy but, as far as we know, it does not provide an explicit connection with a public lexicon, indeed they present the problem of label (i.e. term) selection[12] to better describe the concepts without ambiguity. As commented previously some concepts are hard to describe and the selection of a proper label for them is not a straightforward task. The lexicon should provide a consensual term for the label and the corresponding definition. As known, natural language could be rather ambiguous when describing complex and similar concepts, for this reason the ontology should provide a logic based and non-ambiguous description of the desired concept. On the other hand, the development of FMA[13] represents a quite interesting initiative since FMA uses *Terminologia Anatomica (TA)* [20] as an official source of anatomical terms. In this way they are making a clear distinction between terms and concepts, and which role they have within the terminology and the ontology respectively. Perhaps *TA* it does not represents the desired lexicon since is not general enough but it seems to overcome some of the proposed limitations.

3.3 Necessity of a Lexicon for properties

Finally, the use of the proper properties in the ontology lifecycle will have also an essential role. On the one hand a *lexicon of properties* will be really helpful in the correct connection between concepts, on the other hand the identification of which properties are mainly used to relate concepts will help text mining techniques to discover interesting knowledge from texts. In the literature we can find some efforts in this line, mainly the ones proposing *Ontology Design Patterns* (e.g. [21]) in which the set of allowed properties and expression are established in advance. Additionally the UMLS Semantic Network[14] also provides a specification of 96 properties (86 of them with an associated inverse). These efforts represent good initiatives but they should be integrated within a standard and shared lexicon, that is, the desired lexicon not only should store information about terms but also the proper description of the properties that will relate terms.

[12] Problems of Labels: http://www.opengalen.org/themodel/labels.html

[13] About FMA: http://sig.biostr.washington.edu/projects/fm/FME/aboutFME.html

[14] UMLS Semantic Network: http://semanticnetwork.nlm.nih.gov/

3.4 Thesauri-Ontology Linkage

As commented previously, we intend to have per each domain one shared thesaurus and several ontologies using subsets of the thesaurus concepts for different application purposes. Each ontology concept will be annotated (i.e. entity annotation axioms in OWL[22]) with the corresponding term identifier of the thesaurus. Optionally information from the lexicon can be integrated in the ontology to make faster its processing.

The desired thesaurus will require a unique entry identifier, the link to the words representing the terms including the preferred term and the synonyms. The different words can be kept in a common table referenced by the entries in the thesaurus that allows ambiguity analysis, i.e. how many entries are related to the same word. Metadata added to the thesaurus entries will ease the search for existing entries and solve ambiguous cases. The link of each entry to a *Semantic Category* (e.g. disease, gene, drug, organ, etc.) has been shown helpful for disambiguation purposes in many fields. Additionally, Semantic Categories can belong a top ontology similar to the UMLS Semantic Network. In this case, the finer the granularity of this semantic network is, the more precise will be the searches. However, if such a network is too intricate, the resulting lexicon will be hard to maintain for coherence. Thus, lexicon entries and their relationships must be just focused on the definition, origin and purposes of their entries according to the community requirements. Table 1 shows an example of entry for the desired lexicon:

Identifier:	SWAT4LS0000001
Preferred Name:	Chronic Childhood Arthritis
Pref. Ontology Label:	*Juvenile_idiopathic_arthritis*
Synonyms:	*Juvenile idiopathic arthritis, Juvenile rheumatoid arthritis, Juvenile arthritis, JIA, JRA, JCA, . . .*
Semantic Category:	*Disease*
Hypernymy:	*Rheumatoid Arthritis*
Definition:	*Rheumatoid arthritis of children occurring in three . . .*
Status:	*Up-to-date*

Table 1. Example of Lexicon Entry *(Source: UMLS Metathesaurus)*

The evolution of the ontology may imply changes over the thesaurus like the addition of new entries, the deprecation of obsolete entries or the split of entries in several ones. Obviously the evolution of the thesaurus will also affect the referencing ontologies. For this reason, the lexicon should release stable versions periodically if important changes were made. Moreover each entry of the thesaurus should also have metadata about the status of the entry, indicating if the entry is being reviewed (new entries), is obsolete (pointing to which entry or entries should be used instead), or just if the entry is up-to-date. Referencing ontologies should periodically check if the referenced version of thesaurus is the last one and if the used lexical entries suffered any change or become obsolete.

The ontology and the lexicon are going to be quite interconnected during all the ontology lifecycle stages, therefore the used ontology editor should allow the connection to the lexicon in order to search for lexicon terms and to annotate ontology concepts with the proper lexicon entry. The UMLS Tab[15] for the ontology editor Protégé[16] was a good initiative trying to integrate UMLS-Meta within the ontology lifecycle. On the other hand, the OBO ontology editor[17] also allows the linking (i.e. cross references) of defined concepts to synonyms coming from other resources.

4 Experiences within the Health-e-Child Project

The Health-e-Child project has provide us an excellent real application domain for our experiments. The *Biomedical Knowledge Representation Workpackage* is intended to give an ontology-based representation of the HeC domains (e.g. JIA disease) and to link that knowledge with external resources (e.g. text resources, thesauri, biomedical databases, etc.) We have mainly focused our efforts on the linkage to external knowledge. For this end we have worked on three main issues: text mining, annotation of medical protocols and ontology reuse. Currently, the development of the HeC ontologies is still an ongoing task.

The work presented in [19] analyses different techniques to annotate textual resources with UMLS-Meta terms, and it compares the results with an annotated corpus. Concerning the recall results, we found that some lexical variants are not covered by UMLS-Meta, that is, it lacks the desired entry or it does not provided the proper synonym to identify the concept. Concerning precision results, ambiguous entries within UMLS-Meta and partial annotations usually lead to errors. As commented in Section 3.1, UMLS-Meta represents the main effort on building a medical reference thesaurus, however it still needs to be further polished, refined and extended.

Within this project, another interesting task is to extract information contained in medical protocols (e.g. patient data forms) [10]. For this purpose, we regard these medical protocols as a set of input controls (input fields in patient data forms), where each control has an associated text label (e.g. *Date of Diagnosis, Bone Erosion Evaluation (BEE)*). UMLS-Meta based annotations[19] were used to assign a set of UMLS-Meta terms to each form control. Afterwards, a set of logical representations are associated to each form control (e.g. $BEE \sqsubseteq \exists hasUMLS.C0587240 \sqcap \exists hasUMLS.C1261322$). Moreover, this logic representations have been integrated within a classification purpose ontology (see [10] for a more comprehensive explanation) which aims to classify controls into categories (e.g. *Medical Procedure*, *Measurement*, etc.). Again, incomplete and wrong annotations due to ambiguous entries were the main problems. Unlike text mining, where wrong annotations not necessarily have an important consequence, wrong and incomplete annotations may imply a wrong characterization

[15] UMLS tab: http://protegewiki.stanford.edu/index.php/UMLS_Tab

[16] Protégé: http://protege.stanford.edu/

[17] OBO-Edit: `http://oboedit.org/`

of the medical protocols. Hence, a richer and cleaner controlled vocabulary will be necessary in order to improve the quality of semantic annotations.

Finally, regarding our ontology reuse, our main experience stems from building modules from Galen and NCI ontologies [23]. Modules allow us to extract the desired portion of knowledge from a target ontology, given a set of concepts of interest (e.g. *Juvenile Arthritis, Joint*). However, the definition of the *exact* concept labels of interest became a really hard and ontology dependent task since no common terminology were used in NCI and Galen. Moreover the integration of the extracted modules was again a cornerstone since not only a different conceptualization were found but also different concepts names representing the same reality (e.g. *"NCI:Juvenile_Rheumatoid_ Arthritis"* and *"Galen:JuvenileArthritis"*).

5 Conclusions

In this paper we have presented a still opened issue: the necessity of use and maintenance of a lexicon for ontology engineering, specially for the Life Sciences. We have also emphasized the main limitations and problems of current approaches, which should be better coordinated, integrated and reused. The gap between knowledge representation languages and domain expert skills is another important issue to be addressed. In this way, very expressive languages like OWL are being used for representing simple taxonomies. Instead, defining more complex biomedical concepts requires good skills in Description Logics, which are difficult to understand by domain experts.

Future work will be focused on applying the ideas of this paper to the development of the HeC domain ontologies. We also aim at creating a light-weight thesaurus following the guidelines of this paper so that it provides all the necessary lexical information required by the HeC ontologies and their applications. Moreover, we will study how to filter and enrich existing lexical resources in order to create this new thesaurus.

Acknowledgments

The authors wish to thank the EU project Health-e-Child (IST 2004-027749) for providing us the application domain. This work has been partially funded by the Spanish National Research Program (contract number TIN2005-09098-C05-04). Ernesto Jimenez-Ruiz was supported by the PhD Fellowship Program of the *Generalitat Valenciana*. Antonio Jimeno-Yepes was supported by funding from the EC STREP project *BOOTStrep* (FP6-028099, http://www.bootstrep.org).

References

[1] Hirst, G.: Ontology and the lexicon. In: Handbook on Ontologies in Information Systems, Springer (2004) 209–230

[2] Bodenreider, O.: Lexical, terminological and ontological resources for biological text mining. In: Text mining for biology and biomedicine. Artech House. (2006)

[3] Gruber, T.R.: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In Guarino, N., Poli, R., eds.: Formal Ontology in Conceptual Analysis and Knowledge Representation. (1993)

[4] Pezik, P., Jimeno, A., Lee, V., Rebholz-Schuhmann, D.: Static dictionary features for term polysemy identification. Building and evaluating resources for biomedical text mining, LREC Workshop (2008)

[5] Bechhofer, S.: Ontology language standardisation efforts. OntoWeb. Technical Report. http://www.ontoweb.org/About/Deliverables/d4.0.pdf (2002)

[6] McGuinness, D.L.: Ontologies come of age. In: Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. (2003)

[7] Freund, J., et al.: Health-e-child: An integrated biomedical platform for grid-based pediatrics. In: Proc of Health-Grid 2006, Valancia, Spain (2006) 259–270

[8] Fernandez, M., Gomez-Perez, A., Juristo, N.: Methontology: from ontological art towards ontological engineering. In: Proceedings of the AAAI. (1997)

[9] Duffy, C.M., et al.: Nomenclature and classification in chronic childhood arthritis: Time for a change? Arthritis and Rheumatism **52**(2) (2005) 382–385

[10] Berlanga, R., Jimenez-Ruiz, E., et al.: Medical data integration and the semantic annotation of medical protocols. In: The 21th IEEE International Symposium on Computer-Based Medical Systems (CBMS). (2008)

[11] Pinto, H.S., Martins, J.P.: Reusing ontologies. In: AAAI 2000 Spring Symposium on Bringing Knowledge to Business Processes, AAAI Press (2000) 77–84

[12] Shvaiko, P., Euzenat, J.: Ten challenges for ontology matching. In: Proceedings of ODBASE. (2008)

[13] Bouquet, P., Giunchiglia, F., Harmelen, F., Serafini, L., Stuckenschmidt, H.: C-OWL: Contextualizing ontologies. In: Proc. of ISWC. LNCS 2870 (2003)

[14] Beisswanger, E., Poprat, M., Hahn, U.: Lexical Properties of OBO Ontology Class Names and Synonyms. In: 3rd International Symposium on Semantic Mining in Biomedicine. (2008)

[15] Spasić, I., Schober, D., Sansone, S., Rebholz-Schuhmann, D., Kell, D., Paton, N.: Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. BMC Bioinformatics **9**(5) (2008) S5

[16] Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the c-value/nc-value method. International Journal on Digital Libraries (2000)

[17] Ananiadou, S., Nenadic, G.: Automatic terminology management in biomedicine. Text mining for biology and biomedicine. Artech House (2006) 67–97

[18] Aronson, A.R.: Mapping text to the umls metathesaurus. Technical report: http://skr.nlm.nih.gov/papers/index.shtml (2001)

[19] Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., Rebholz-Schuhmann, D.: Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinformatics **9**(Suppl 3) (2008) S3

[20] Rosse, C.: Terminologia anatomica: Considered from the perspective of next-generation knowledge sources. Clinical Anatomy **14**(2) (2001) 120–133

[21] Egana, M., Antezana, E., Kuiper, M., Stevens, R.: Ontology design patterns for bio-ontologies: a case study on the cell cycle ontology. BMC Bioinformatics (2008)

[22] Cuenca-Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: OWL 2: The next step for OWL. Journal of Web Semantics (2008) To Appear.

[23] Jimenez-Ruiz, E., et al.: Safe and economic re-use of ontologies: A logic-based methodology and tool support. In: European Semantic Web Conference. (2008)