UNIVERSITY
OF TRENTO

DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

FACETED LIGHTWEIGHT ONTOLOGIES

Fausto Giunchiglia, Biswanath Dutta and Vincenzo Maltese

April 2009

Technical Report # DISI-09-022

Also: appears in: "Conceptual Modeling: Foundations and Applications",
Alex Borgida, Vinay Chaudhri, Paolo Giorgini, Eric Yu (Eds.)
LNCS 5600 Springer

Faceted lightweight ontologies

Fausto Giunchiglia (*), Biswanath Dutta (**), Vincenzo Maltese (*)

(*) Dipartimento di Ingegneria e Scienza dell'Informazione (DISI)
Università di Trento, Trento, Italy
(**) Documentation Research and Training Centre (DRTC), Indian Statistical Institute (ISI),
8th Mile Mysore Road, Bangalore- 560059, India

Abstract. We concentrate on the use of *ontologies* for the categorization of objects, e.g., photos, books, web pages. *Lightweight ontologies* are ontologies with a tree structure where each node is associated a natural language label. *Faceted lightweight ontologies* are lightweight ontologies where the labels of nodes are organized according to certain predefined patterns which capture different aspects of meaning, i.e., *facets*. We introduce facets based on the Analytico-Synthetic approach, a well established methodology from Library Science which has been successfully used for decades for the classification of books. Faceted lightweight ontologies have a well defined structure and, as such, they are easier to create, to share among users, and they also provide more organized input to semantics based applications, such as semantic search and navigation.

Keywords: Ontologies**,** Lightweight ontologies, facets, classifications, formal classifications.

1 Introduction

Ontologies are being used in different communities, for different purposes and with different modalities. There are various kinds of ontologies, according to the degree of formality, complexity of the graph structure, and expressivity of the language used to describe them [1]. Ontologies have two main applications. They can be used to *describe* objects or they can be used to *categorize* objects. In this paper, we concentrate on the second use, namely, we are interested in the problem of classifying, e.g., photos, Web pages, books.

Lightweight ontologies are ontologies with a tree structure where each node is associated a natural language label. We sometimes speak of *formal lightweight ontologies* meaning ontologies which can be obtained from lightweight ontologies by translating natural language labels into Description Logics (DL) [12] formulas which capture their meaning ([2] provides an example of how such translation can be done). In formal lightweight ontologies, node formulas stand in the subsumption relation, namely a formula in a node is always more general than the formula in the node below [1, 31]. In the following we talk of lightweight ontologies meaning sometimes formal lightweight ontologies. The context always makes clear what we mean.

Lightweight ontologies allow for automated document classification [1, 16], query answering [1, 21] and also for solving the semantic heterogeneity problem among multiple ontologies [15, 18, 19, 20]. They are definitely a very powerful tool which can be exploited towards the automation of reasoning in data and knowledge management. Still, the adoption of (lightweight) ontologies, so far, has not been as widespread as one would have expected when the work on the Semantic Web started. Among the problems which have been identified are the lack of interest or the difficulties on the user side in building such ontologies [4, 5], but also the fact that ontologies developed for one purpose can hardly being reused for other purposes, or by other users [5].

The goal of this paper is to introduce *faceted lightweight ontologies* as a very promising solution to the problem highlighted above. Faceted lightweight ontologies are defined in terms of *facets*. Recently, facets have been adopted with great success for the design of interfaces to web sites. See, for instance the survey by La Barre [23] and in particular the work done in Flamenco[1] (see for instance [24]), but see also [7,8,9] as an application to knowledge management which is somewhat related, in spirit, to our work. We construct facets following the approach which was first devised by Ranganathan at the beginning of the last century [22] [2] and, in particular, the POPSI Methodology, originally introduced in [26].

Taking the terminology of Library Science, facets are "*aspects of meaning*". They formalize, for any given domain (e.g., medicine, sports, music, science), the main characteristics of that domain and, in particular, the entities or objects which belong to that domain (e.g. in medicine, the body parts), the properties of objects (e.g., in medicine, the various kinds of disease) and the actions which can be taken (e.g., in medicine, surgery or medication). More precisely, a facet is a hierarchy of homogeneous group of terms (nodes), where each term in the hierarchy denotes a primitive atomic concept. Thus we have hierarchies of entities, properties, actions, and so on. We call *background knowledge* [17,14], a *faceted representation scheme*, namely a set of facets that represent the system a-priori knowledge about the domains of interest (see also [13] for an early attempt of defining a faceted representation schema not based on Ranganathan's theory). A faceted representation scheme allows for post-coordination, namely, for constructing complex labels (in Library Science terminology, also called *subjects*) by combining terms from facets at both indexing, classification and searching time. Faceted lightweight ontologies are lightweight ontologies where node labels (formulas) contain only atomic concepts which correspond to primitive concepts taken from the background knowledge.

The rest of the paper is organized as follows. Section 2 introduces and formally defines (classification) lightweight ontologies. Section 3 introduces facets. Section 4 introduces faceted subjects and, then, the notion of faceted lightweight ontology. Finally, Section 5 shows, via an example, how a faceted subject can be constructed according to the POPSI subject indexing system. Section 6 concludes the paper.

[1] http://flamenco.berkeley.edu

[2] This theory is widely recognized as a fundamental methodology that guides in the organization of the knowledge in a given domain (see for instance [30]) in terms of basic subjects and relations between them.

2 Lightweight classification ontologies

Ontologies have been used for centuries in different communities, for different purposes and with different modalities. The concept originated more than two thousand years ago from philosophy and more specifically from Aristotle's theory of categories[3]. The original purpose was to provide a categorization of all existing things in the world. Ontologies have been lately adopted in several other fields, such as Library and Information Science (LIS), Artificial Intelligence (AI), and more recently in Computer Science (CS), as the main means for describing how classes of objects are correlated, or for categorizing what archivists generically call documents.

Many definitions of ontologies have been provided. According to the most quoted, an ontology is "*an explicit specification of a conceptualization*" [10]. Their main purpose is to favour interoperability by providing a common terminology and understanding of a given domain of interest, which in turn allows for the assignment of clear meanings to information items. There are however different kinds of ontologies, or, in other words, several more specific concepts, according to the degree of formality and expressivity of the language used to describe them (see [2] for a discussion). They range from informal representations like user classifications (e.g. the structure of folders in a file system) and web directories (e.g. DMOZ, Yahoo! and Google[4]), to progressively more formal representations like enumerative classification schemes (e.g. the Dewey Decimal Classification[5] (DDC) and the Library of Congress Classification[6] (LCC)), Knowledge Organization Systems (KOS) such as thesauri (e.g. AGROVOC, NALT, AOD, and HBS) and faceted classification schemes (e.g., the Colon Classification), and, ultimately, formal ontologies which are expressed into a logic formal language and represented using formal specifications such as DL or OWL.

For the purpose of this work, however, following the terminology provided in [1], the core distinction is between

1. ontologies which are mainly used to *describe* objects, also called *descriptive ontologies*, and
2. ontologies which are mainly used to *categorize* objects, also called *classification ontologies*.

This distinction is reflected into the underlying semantics taken as reference, namely the *real world semantics* and the *classification semantics* described below. Based on this distinction then we further refine the notion of classification ontology into the notion of *classification lightweight ontology*, which is actually the core notion needed in this paper. Let us analyze these notions in detail.

[3] http://plato.stanford.edu/entries/aristotle-categories/

[4] http://dmoz.org/; http://dir.yahoo.com/; http://directory.google.com/

[5] http://www.oclc.org/dewey/

[6] http://www.loc.gov

2.1 Descriptive ontologies

In descriptive ontologies, *concepts represent real world entities, e.g., the extension of the concept animal is the set of real world animals*, which can be connected via relations of the proper kind. The purpose of descriptive ontologies is to specify the terms used in their original meaning, according to the nature and the structure of the domain they model [11]. Two typical relations are used to construct the trees (also called the taxonomies) which provide the backbone to these ontologies and they are *is-a* (Genus-species) and *part-of* (Whole-part) relations. In Fig. 1 (a), (b) we provide two examples of descriptive ontologies, based on these two relations, where each node represents a concept and each arrow represents a relation between them. The direction of the arrows represents the direction of the relations. Mixed situations are also possible.

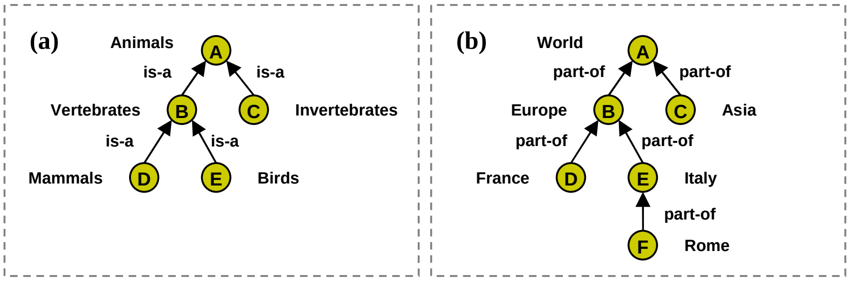

Fig. 1. (a) An is-a ontology; (b) A part-of ontology.

It is worth noticing that, when translating these ontologies in DL, the *is-a* relation is translated into subsumption ($\sqsubseteq$) or, more precisely, it is assumed to imply subsumption, while this is not the case for *part-of*. Therefore *is-a* constitutes the basic backbone of the subsumption based hierarchical structure of a domain.

2.2 Classification ontologies

Ontologies in classification semantics are built with the goal of indexing documents. As a consequence, *the extension of each concept (label of a node) is the set of documents about the entities or individual objects described by the label of the concept* [1,2]. For example, the extension of the concept animal is “the set of documents about animals” of any kind. This has three main consequences.

The first is that the semantic relation holding between nodes which are one above the other is *always* the *subset* relation. In other words the set of documents which can be classified in a node is always a subset of the documents which can be classified in the node above (and this motivates some techniques for minimizing the number of nodes where a document is classified, for instance the *get-specific* principle, see [16] for a formalization of this principle and its use in automatic classification). Fig. 2 (a), (b) provide the classification semantics version of the two ontologies reported in Fig. 1 (a), (b). As it can be noticed from Fig. 2, the standard relations of descriptive ontologies are translated into relations among sets. Thus, *is-a*, but also to *part-of*, when transitive, and *instance-of*, are translated into *subset*, while the others correspond to *overlap* (⊓). Fig. 2 (b) provides a case where the *part-of* relations of Fig. 1 (b) are translated into *subset* in classification semantics. One example where this is not possible is the chain of relations: handle *part-of* door *part-of* school *part-of* school system.

The second consequence is that, in the DL translation of classification ontologies, the subset relation is translated into subsumption between the formulas of nodes which are one above the other. It is important to observe that the DL translation of the same ontology, if taken with real world semantics or with classification semantics, leads to a different DL theory (compare again Fig. 1(b) and Fig. 2(b))).

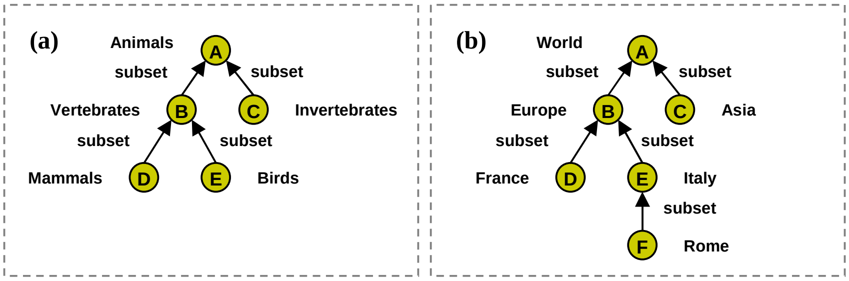

Fig. 2. Two ontologies in classification semantics.

Notice that the labels in both ontologies in Fig. 2 are such that each of them represents a proper subset of the label of the node above. Thus, for instance, *vertebrates* represents a proper subset of *animals*. However, and this is the third consequence, in classification ontologies, the situation above can be generalized to consider labels which denote sets which are not in the subset relation, but, rather, in the *overlap* relation. As a matter of fact, this is what happens in most classification ontologies [2]. Consider for instance the classification ontology in Fig. 3 (a). The intuition is that node B should contain all documents which are about "*the research on Java*". In other words, the meaning of a node (so-called "*concept at a node*" in [1,14]) can be constructed by taking the DL conjunction of (semantically, the intersection of the sets denoted by) the concepts of all the labels in the path from the root to the node itself. The application of this rule to the example in Fig. 3 (a) leads to the ontology in Fig. 3(b). As it can be noticed, the concept associated to a node is in the subsumption relation with any node above and this is obtained by applying the conjunction operator over the path. The numbers after each label are used to denote the concept which is obtained by disambiguating it (each word may correspond to more than one concept, e.g. Java can be an island, a programming language or a kind of coffee beans). It is easy to notice how the situation in Fig. 3 (a) collapses into the situations in Fig. 2 (a) once we return to the subset relation: all the conjunctions become redundant due to the fact that if A ⊑ B, then A ⊓ B is equivalent to A.

2.3 Lightweight classification ontologies

All the theory on classification of Library Science and, as a consequence, the theory of facets, as originally devised by Ranganathan and later refined in the POPSI methodology, is based on classification semantics. And it is correctly so, as these methodologies were invented in order to classify books and position them in shelves. In the following of this paper we also concentrate on classification ontologies and classification semantics. The motivation is quite similar to that in Library Science. It is a fact that, e.g., on line catalogs, file systems, web directories and library classifications are used for classifying objects and can be translated, exactly or with a certain degree of approximation, into classification ontologies.

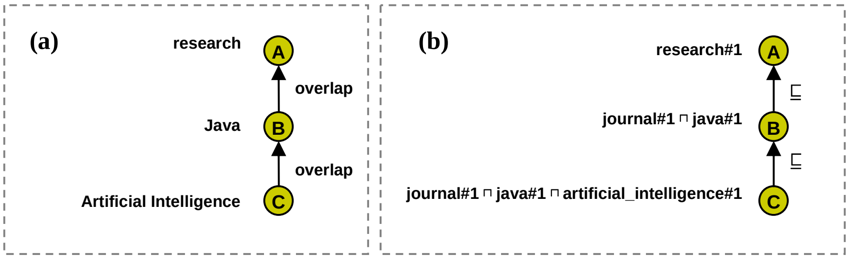

Fig. 3. (a) A classification ontology with no *subset-of* relation between labels, (b) the corresponding formal ontology.

As a matter of fact, in all applications from Library Science and also in our reference applications, the classification ontologies which are needed are quite simple and consists of trees, possibly multi-rooted, where most of the nodes in the father-child relation do *not* have labels whose denotation stands in the subset relation. Each node label can therefore be translated into a logic formula (typically built as a combination of conjunctions and disjunctions of atomic concepts) representing the meaning of the node taking into account its context, namely the path from the root to the node [3]. This leads to the definition of classification lightweight ontology, as originally defined in [25] (the word "classification" did not appear in the original definition):

*A **lightweight classification ontology** O is a rooted tree <N,E,L^F> where:*

a) N is a finite set of nodes;

b) E is a set of edges on N;

c) L^F is a finite set of labels expressed in a Propositional DL language such that for any node $n_i \in N$, there is one and only one label $l_i^F \in L^F$;

d) $l_{i+1}^F \sqsubseteq l_i^F$ with n_i being the parent of n_{i+1}.

3 Facets

According to the Analytico-Synthetic approach [14], facets are defined following two steps:

1. examine the field (domain) to identify relevant terms. They can be gained by consulting domain experts and all sorts of information sources over the domain. This process starts in the so called "*idea plane*", the language independent con-

ceptual level, where primitive concepts are identified. Each identified concept, in turn, is expressed in the "*verbal plane*" in a given language, for example in English, trying to articulate the idea *coextensively*, namely identifying a term which exactly and unambiguously expresses the concept;

2. group the identified terms (also called *isolate ideas*) according to their common properties or characteristics, and order them (in hierarchies) in a meaningful sequence. The set of homogeneous terms form a *facet*. For example, *Nose, Larynx, Trachea, Bronchi, Lung, Pleural sac, Mediasinum* form a facet called *Respiratory system* (these entities are in the *part-of* relation with *Respiratory system*). Now the terms *Outer nose* and *Nasal*, which are again *part-of* Nose, can form a facet called *Nose* which will be treated as *sub-facet* of the facet *Respiratory system*.

These two steps construct a *faceted representation scheme* and correspond to what in our previous work we call the definition and construction of the so called *background knowledge* [17, 21], namely the a-priori knowledge which must exist in order to make semantics effective. Notice that the grouping of terms of step 2 have real world semantics, namely, they are descriptive ontologies which are formed using *part-of*, *is-a* and *instance-of* . Facets have the following two key properties:

1. They are organized as a set of independent *domains* which are completely modular and can be developed independently.
2. For each domain, facets are grouped into specific elementary *categories*. Originally, Ranganathan defined five fundamental categories: *Personality*, *Matter*, *Energy*, *Space* and *Time (PMEST)*. Later on, Bhattacharyya proposed a refinement which consists of four main categories, called DEPA: *Discipline* (D) (what we now call a domain), *Entity* (E), *Property* (P) and *Action* (A), plus another special category, called Modifier (m).

In our approach we organize facets according to the DEPA categories. Let us describe them in some detail:

- **Discipline** (or **domain**): it includes conventional fields of study (e.g., *Library Science*, *Mathematics* and *Physics*), applications of the traditional pure disciplines (e.g., *Engineering* and *Agriculture*), any aggregates of such fields (e.g., *Physical Sciences* and *Social sciences*), or also, in more modern terms, fields like *music*, *sports*, *computer science*, and so on.
- **Entity**: the elementary category Entity is manifested in perceptual correlates or in conceptual existence. It is distinct from their properties and actions performed by them or on them. Basically the concepts represent the core idea of a domain treated as under this elementary category. For example, "*Teachers*", "*Students*", *"Courses"* are the core concepts to a domain *"Education"*.
- **Property**: it includes concepts denoting quantitative or qualitative attributes. For example, *Quality*, *Quantity*, *Measure*, *Weight*, *Taste*, etc;
- **Action**: it includes concepts denoting the notion of *"doing"*. It includes "*processes*" and "*steps*" of doing. An action can manifest as "*Self-action*" or "*Ex-*

ternal action". A self-action is an action done by some agent (explicit or implicit) on or in itself. For example, *Imagination*, *Interaction*, *Reaction*, *Reasoning*, *Thinking*, etc. An external action is an action done by some agent (explicit or implicit) on a concept of any of the elementary categories described above. For example, *Organization*, *Cooperation*, *Classification*, *Cataloguing*, *Calculation*, *Design*, etc.

- **Modifier**: it includes concepts used or intended to be used to qualify other concepts. With the help of a modifier, the extension of a concept is decreased and the intension is increased without disturbing its conceptual wholeness. For example, "Mining in India", here India modifies Mining. By implication, any concept from the elementary categories above or combination of two or more concepts may serve as the basis of deriving a modifier. There are many kinds of modifiers, in particular we can distinguish *common modifiers* (e.g., *space-modifier*, *time-modifier*, *environment-modifier*, *form-modifier*, *language modifier*) and *special modifiers* (e.g., *Infectious*, *Bacterial*, *Fungus*, etc. modify the concept "*Diseases*" in the *Medicine* domain). Common modifiers are common to all disciplines used to modify manifestations of more than one elementary category, occurring singly or in combination. Special modifiers modify manifestations of one and only one elementary category. However, following the principle of reusability (described below), some modifiers can be shared by a set (but not all) domains (for instance chemical substances are used both in Chemistry and in Agriculture, possibly under different categories).

The basic rule for formulating subject headings is *Discipline* (base) first, followed by *Entity* (core), which is followed by *Property* and/or *Action*. Property and/or Action may be further followed by Property and/or Action as the case may be, followed by *Common modifiers*. The *species/types* and/or *modifiers* and/or *parts* and/or *constituents* for each of the *elementary categories* follow immediately the manifestation to which they are respectively *species/types* or *modifiers* or *parts* or *constituents*. In Fig. 4 we provide an example of facets grouped in the DEPA categories in the Medicine domain. Notice that, even if this is not the case in the example above, in each category we can potentially have more than one facet.

Facets possess some essential properties as listed below:

- **Hospitability**: they are easily extensible. New terms representing new knowledge can be accommodated without difficulty in the hierarchical structure. Terms in the hierarchies are clearly defined, mutually exclusive and collectively exhaustive.
- **Compactness**: facet based systems need less space with respect to the other hierarchical knowledge organization systems to classify the universe of knowledge. There is no explosion of the possible combinations as the basic elements (facets) are taken in isolation.
- **Flexibility**: hierarchical knowledge organization systems are mostly rigid in their structure, whereas facet based systems are flexible in nature.
- **Reusability**: a facet based ontology developed for a particular domain could be partially usable into another related domain.

- **Clear, but rigorous, structure**: the faceted approach aims at the identification of the logical relations between concepts and concepts groups. Sibling concepts must share a common characteristic.
- **The methodology**: a strong methodology for the analysis and categorization of concepts along with the existence of reliable rules for synthesis is provided.
- **Homogeneity**: a facet represents a homogeneous group of concepts, according to the specified common characteristic(s).

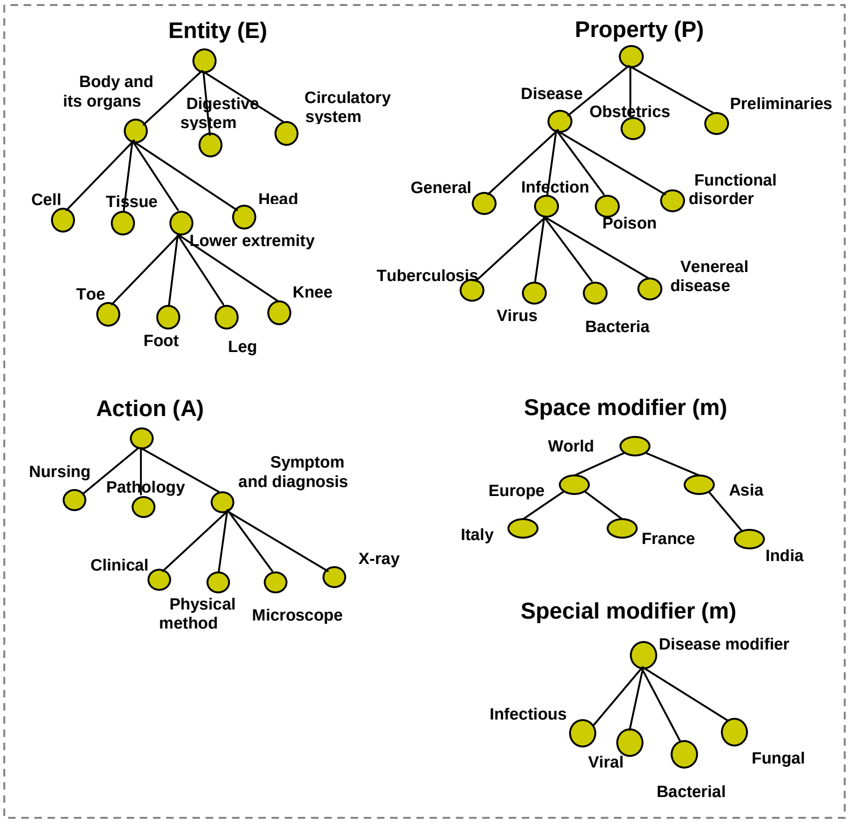

Fig. 4. The set of facets for the *Medicine* domain.

4 Faceted Lightweight Ontologies

Once the background knowledge is constructed, the next step is to see how to use facets in order to index or classify documents (in our case, inside lightweight ontologies). As from above, for us this corresponds to associating to each document and node in a classification a DL formula [1,2]. This step happens inside what Ranganathan called the “*notational plane*”. Here an unambiguous notation is used to synthetically attach meaning and provide order to the managed objects, typically books on the shelves. Following G. Bhattacharyya [26], the key idea is to associate to a node or document a *subject*, namely "*a piece of non-discursive information that summarises indicatively what a book or document (any body of information) is about*". A subject, in our terms, is the label and corresponding concept associated to a document or a node in a lightweight ontology. Since in lightweight ontologies we use classification semantics, a document will be classified in any node whose subject is more general than the subject of the document [1,16].

We define subjects in terms of facets. The key intuitions are three:

1. We associate to each term in the subject a label and corresponding concept taken from a faceted classification scheme (in POPSI the concept is given by the preferred term and its context);
2. For each term in a facet, the context is constructed by associating to it all the terms from the root to the term itself, thus disambiguating the intended concept. Notice that this means that, in the step from the background knowledge to the subject concept, we need to translate from real world semantics (used in the background knowledge) to classification semantics (used in lightweight ontologies).
3. Each subject contains terms (concepts) from *potentially all* the DEPA categories, thus allowing for the complete disambiguation of the subject. However, the user is supposed to provide, explicitly or implicitly, at least the *discipline* and the main *entity*.

In POPSI, in order to construct the context, each *leading heading* (also called *lead term* or *term-of approach*) is followed by the *context heading*, namely the set of auxiliary terms which preserves the context (in terms of the *discipline* and the path from the root of the facet to the term). For instance, the context of the term *Cell* is:

Cell (lead term)
Medicine, Body and its organs > Cell (context heading)

In the above example, "," separates the *isolate ideas* (i.e. the concepts) belonging to the different fundamental categories as shown in section 3, while ">" identifies the increasing intension and decreasing extension of isolate ideas within a facet. Notice that, from Fig. 4 above, Medicine is the name of the domain while the second part is the complete path in the entity facet. Consider, furthermore, the subject "*Microscopic diagnosis of bacterial viruses on cells in India*". Its terms are completely contextualized in POPSI as follows (the sequence of concrete steps necessary to identify them is described in the next section):

(Domain):	Medicine,
(Entity):	Body and its organs > Cell,
(Property):	Disease > Infection > Virus,
(Modifier of P.)	Bacterial,
(Action):	Symptom and diagnosis > Microscope,
(Space modifier):	Asia > India

The main advantage of the faceted approach is that it makes explicit the logical relations among the concepts and concept groups and removes the limitations of traditional hierarchies. It allows for viewing a complex entity from a variety of perspectives or from different angles. For example, a *cow* can be described as an animal, as a pet, as a food item, as a commodity, as a God for a particular community, and so on, depending on the domain. Therefore, each time, by providing the context, the faceted approach allows for the representation of different concepts.

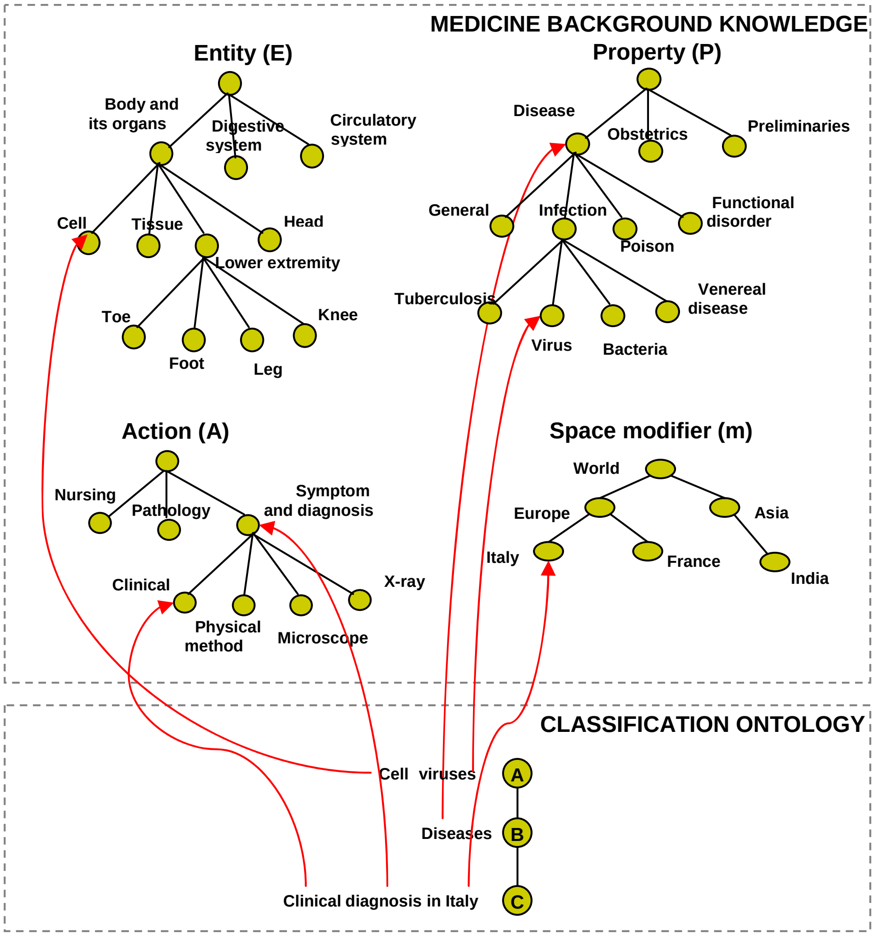

Fig. 5. A faceted lightweight ontology.

Based on the notion of subject, we can now define *a faceted lightweight (classification) ontology* as follows:

A ***faceted lightweight (classification) ontology*** *is a lightweight ontology where each term and corresponding concept occurring in its node labels must correspond to a term and corresponding concept in the background knowledge, modeled as a faceted classification scheme.*

Fig. 5 provides an example of how this can be done. Notice that in faceted lightweight ontologies there might be nodes, as in Fig. 5, whose labels contain terms from multiple DEPA categories, while in other cases we will have one node per DEPA category. The more terms and corresponding DEPA categories there will be, the more specific the lightweight classification ontology will be.

5 Subject indexing

How do we use faceted classifications schemes, in practice? As already mentioned in the previous section, documents will be classified under those nodes whose subject is more general than theirs. But, the real challenge is that in most cases the subject specification is only partial. To this extent, POPSI provides a methodology for providing the missing *contextual information*. The solution lies mainly in the appropriate representation of the extension and intension of the thought content (subject matter) of the indexed documents. Let us now discuss the steps involved in POPSI in deriving the subject strings starting from the titles associated to documents to index along with an example. Let us consider the example of a subject given in the previous section:

“*Microscopic diagnosis of bacterial viruses on cells in India*”.

The analysis is organized in eight steps, as described below:

<u>Step 1</u> (**Analysis of the subject indicative expression):** it concerns the analysis of the subject indicative expression pertaining to the source of information. It may be the title of a book, article etc. For the example above, we derive the following terms:

D = Medicine (implicit in the above title)
E = Cells (explicit)
P = Viruses (explicit)
m of P = Bacterial (explicit)
A = Microscopic diagnosis (explicit)
m = India (explicit) (Space modifier)

In our approach this step is performed analogously. Notice that implicit categories must be provided manually by the user or computed automatically by the system.

<u>Step 2</u> (Formalization of the Subject Proposition): in this stage the formalization of the sequence of the terms appearing in the subject derived by Step 1 (Analysis) is done. According to the principles of sequence, the components are sequenced in the following way:

> Medicine (D), Cells (E), Viruses (P), Bacterial (m of P), Microscopic diagnosis (A), India (m)

In our approach this step is not needed.

Step 3 (**Standardization of the Subject Proposition):** It consists in the identification of the standard terms, when synonyms of the same term are available, denoting the atomic concepts present in the subject proposition. For our example, this step is not applicable. So, the subject proposition remains the same:

> Medicine, Cells, Viruses, Bacterial, Microscopic diagnosis, India

In our approach this step is performed analogously. This information is codified in the background knowledge.

Step 4 (**Modulation of the Subject Proposition):** It consists of augmenting the standardized subject proposition by interpolating and extrapolating, as the case may be, the successive super-ordinates of each concept by using standard terms with indication of their synonyms. In practice, it corresponds to the identification of corresponding contextual terms, namely the correct disambiguation of each concept used, providing the right amount of hierarchically related concepts:

> Medicine, Body and its organs > Cell, Disease > Infection > Virus, Bacterial, Symptom and diagnosis > Microscope, Asia > India

In our approach this step is performed analogously: we extract from the background knowledge, the concept of each natural language term occurring in the subject.

Step 5 (**Preparation of the Entry for Organizing the Classification):** This step consists of preparing the main entries in the so called associative index in alphabetical arrangement. This is done by assigning a systematic set of numbers as given in [26] to indicate the categories and positions of the subject propositions. In our example:

> Medicine 8 Body and its organs 8.3 Cell 8.2 Disease 8.2.4 Infection 8.2.4.4 Virus 8.2.4.4.6 Bacterial 8.2.1 Symptom and diagnosis 8.2.1.4 Microscope 4 Asia 4.4 India

In our approach this step is not needed.

Step 6 (**Decision about the Terms-of Approach):** It consists of deciding the terms-of approach, namely the lead terms, for generating associative classifications, and of controlling synonyms. For controlling synonyms, each standard term is to be referred to from each of its synonyms. For example (this is not part of our running example),

> Chemical treatment (Medicine)
> *see*
> Chemotherapy

In our approach this step is not needed.

__Step 7__ **(Preparation of the Entries for Associative Classification):** It consists of preparing entries under each term-of-approach by cyclic permutation. For example (all other entries can be treated similarly):

Body and its organ

Medicine, Body and its organs > Cell, Disease > Infection > Virus, Bacterial, Symptom and diagnosis > Microscope, Asia > India

Cell

Medicine, Body and its organs > Cell, Disease > Infection > Virus, Bacterial, Symptom and diagnosis > Microscope, Asia > India

In our approach this step is not needed.

__Step 8__: **Alphabetical Arrangement of Entries**
It consists of arranging all the entries including the reference entries in alphabetical sequence according to a set of standardized rules ignoring the signs and punctuation marks.

Asia
Medicine, Body and its organs > Cell, Disease > Infection > Virus, Bacterial,
Symptom and diagnosis > Microscope, Asia > India
Bacterial
Medicine … India
…
…
Virus
Medicine … India

In our approach this corresponds to indexing or classifying inside a faceted lightweight ontology using the concepts of the nodes and the documents.

6 Conclusion

In this paper, we have introduced the notion of faceted lightweight ontology as a lightweight ontology whose terms are extracted from a background knowledge organized in terms of facets. Using facets allows us to have much more control on the language and concepts used to build ontologies and also on their organization, which in general will exploit the structure and terms of the four basic DEPA categories.

This work has been done as part of the FP7 Living Knowledge FET IP European Project.

References

1. F. Giunchiglia, M. Marchese, I. Zaihrayeu. Encoding Classifications into Lightweight Ontologies. Journal of Data Semantics 8, pp. 57-81, 2006. Short version in: Proceedings of the 3rd European Semantic Web Conference (ESWC), 2006,
2. F. Giunchiglia and I. Zaihrayeu. Lightweight ontologies. In S. LNCS, editor, Encyclopedia of Database Systems, 2008.
3. I. Zaihrayeu, L. Sun, F. Giunchiglia, W. Pan, Q. Ju, M. Chi, and X. Huang. From web directories to ontologies: Natural language processing challenges. In 6th International Semantic Web Conference (ISWC 2007), 2007.
4. Mai, J.-E.: Classification in Context: Relativity, Reality, and Representation. Knowledge Organization. 31(1), 39-48, 2004.
5. Duval, E., Hodgins, W., Sutton, S. & Weibel, S, L.: Metadata Principles and Practicalities. DLib Magazine, 8(4), 2002. http://www.dlib.org/dlib/april02/weibel/ 04weibel.html
6. Nicholson, D., Neill, S., Currier, S., Will, L. Gilchrist, A., Russell, R. and Day, M.: HILT: High Level Thesaurus Project – Final Report to RSLP & JISC. Centre for Digital Library Research, Glasgow, UK, 2001. http://hilt.cdlr.strath.ac.uk/Reports/Documents/HILTfinalreport.doc.
7. Yannis Tzitzikas, Nikos Armenatzoglou, Panagiotis Papadakos: FleXplorer: A Framework for Providing Faceted and Dynamic Taxonomy-Based Information Exploration. DEXA Workshops: 392-396, 2008.
8. Yannis Tzitzikas, Anastasia Analyti, Nicolas Spyratos, Panos Constantopoulos: An algebra for specifying valid compound terms in faceted taxonomies. Data Knowl. Eng. (DKE) 62(1):1-40, 2007.
9. Yannis Tzitzikas, Nicolas Spyratos, Panos Constantopoulos, Anastasia Analyti: Extended Faceted Ontologies. CAiSE:778-781, 2002.
10. T. R. Gruber. A translation approach to portable ontology specifications. Knowledge Aquisition, 5(2): pp.199–220, 1993.
11. N. Guarino. Helping people (and machines) understanding each other: The role of formal ontology. In CoopIS/DOA/ODBASE (1), p. 599, 2004.
12. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. F. Patel-Schneider. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, 2002.
13. D. Soergel. A Universal Source Thesaurus as a Classification Generator. Journal of the American Society for Information Science 23(5), pp. 299–305, 1972.
14. F. Giunchiglia, M. Yatskevich, P. Shvaiko, 2007. Semantic Matching: algorithms and implementation. Journal on Data Semantics, IX, 2007.
15. Fausto Giunchiglia, Fiona McNeill, Mikalai Yatskevich, Juan Pane, Paolo Besana, Pavel Shvaiko “Approximate Structure-Preserving Semantic Matching”, 7th International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2008), Monterrey, Mexico, Nov 2008.
16. Fausto Giunchiglia, Ilya Zaihrayeu, and Uladzimir Kharkevich. "Formalizing the get-specific document classification algorithm", in 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007), Budapest, Hungary, September 2007. LNCS Springer Verlag.

17. Giunchiglia F., Shvaiko P., Yatskevich M., "Discovering Missing Background Knowledge in Onology Matching". In: vol.141, Leipzig University, Germany:Brewka et al., 2006. p. 382-386. Proceedings: "17th European Conference on Artificial Intelligance - ECAI 2006", Riva del Garda, Italy, August 29th - September 1st, 2006
18. Giunchiglia F., Shvaiko P., Yatskevich M., "Semantic schema matching". In: On the move to meaningful internet systems 2005: COOPIS, DOA, and ODBASE: OTM Confederated International Conferences, vol. 1. Proceedings: "CoopIS, DOA, and ODBASE", Agia Napa, Cyprus, 2005 Editors: Meersman R., Tari Z., Berlin Heidelberg:Springer, LNCS, Vol. 3760/2005, p. 347-365, 2005.
19. Giunchiglia F., Yatskevich M., Giunchiglia E., "Efficient semantic matching". In: Proceedings of the 2nd European semantic web conference (ESWC'05). Editors: Gomez-perez A., Euzenat J., Heidelberg:Springer, 2005. Lecture Notes in Computer Science, Vol. 3532/2005, p. 272-289, Proceedings: "Second European Semantic Web Conference, ESWC", Heraklion, Crete, Greece, 29 May - 1 June 2005, Note: ISBN: 3-540-26124-9
20. Giunchiglia F., Yatskevich M., "Element Level Semantic Matching". Workshop on "Meaning Coordination and Negotiation”. ISWC04, Hiroshima, Japan, November 2004
21. Fausto Giunchiglia, Uladzimir Kharkevich, Ilya Zaihrayeu: "Concept Search: Semantics Enabled Syntactic Search". in Semantic Search 2008 workshop (SemSearch2008) at the 5th European Semantic Web Conference (ESWC2008), 2008.
22. S. R. Ranganathan. The Colon Classification. In S. Artandi, editor, Vol IV of the Rutgers Series on Systems for the Intellectual Organization of Information, 1965. New Brunswick, NJ: Graduate School of Library Science, Rutgers University.
23. K. La Barre. Adventures in faceted classification: A brave new world or a world of confusion? Knowledge organization and the global information society: proceedings 8th ISKO conference, London, 13-16 July 2004.
24. M. Hearst. Design Recommendations for Hierarchical Faceted Search Interfaces. In ACM SIGIR Workshop on Faceted Search, Seattle, WA, 2006.
25. F. Giunchiglia, V. Maltese, A. Autayeu. Computing minimal mappings. University of Trento, DISI Technical Report, 2008.
26. Bhattacharyya, G. "POPSI: its fundamentals and procedure based on a general theory of subject indexing languages", Library Science with a Slant to Documentation, Vol. 16 No. 1, March, pp. 1-34, 1979.
27. Ranganathan, S. R. Prolegomena to library classification. London: Asia Publishing House, 1967.
28. Ranganathan, S.R. Elements of library classification. Bombay: Asia Publishing House. pp. 3, 1960.
29. Aptagiri, D.V., Gopinath, M.A. and Prasad, A.R.D. A frame-based knowledge representation paradigm for automating POPSI, 1995.
30. V. Broughton. The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings, 58(1/2) pp. 49-72, 2006.
31. Zaihrayeu I., Marchese M., Giunchiglia F., "Encoding Classifications into Lightweight Ontologies". In: Proceedings of the 3rd European Semantic Web Conference (ESWC 2006). Lecture Notes in Computer Science, Vol. 4011, p. 80-94. Budva, Montenegro, June 11-14, 2006.