UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

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

LIGHTWEIGHT ONTOLOGIES

Fausto Giunchiglia and Ilya Zaihrayeu

October 2007

Technical Report DIT-07-071

Lightweight Ontologies

Fausto Giunchiglia and Ilya Zaihrayeu
University of Trento, Italy
{fausto, ilya}@dit.unitn.it

SYNONYMS

Controlled vocabularies; Taxonomies; Thesauri; Business catalogues; Faceted classifications; Web directories; Topic hierarchies; User classifications

DEFINITION

Ontologies are *explicit specifications of conceptualizations* [8]. They are often thought of as directed graphs whose nodes represent *concepts* and whose edges represent *relations* between concepts. The notion of concept is understood as defined in Knowledge Representation, i.e., as a set of objects or individuals [2]. This set is called the *concept extension* or the *concept interpretation*. Concepts are often *lexically defined*, i.e., they have natural language names which are used to describe the concept extensions (e.g., concept `mother` denotes the set of all female parents). Therefore, when ontologies are visualized, their nodes are often shown with corresponding natural language concept names. The backbone structure of the ontology graph is a taxonomy in which the relations are "is-a", whereas the remaining structure of the graph supplies auxiliary information about the modeled domain and may include relations like "part-of", "located-in", "is-parent-of", and many others [9].

In their simplest version, one can think of *lightweight ontologies* as ontologies consisting of backbone taxonomies only. However, we generalize the "is-a" relationship to concept subsumption still matching the basic properties of backbone taxonomies: namely, in a lightweight ontology, the extension of the concept of a child node is a subset of the extension of the concept of the parent node. We formally define the notion of lightweight ontology as:

A *(formal) lightweight ontology* is a triple $O = \langle N, E, C\rangle$, where N is a finite set of nodes, E is a set of edges on N, such that $\langle N, E\rangle$ is a rooted tree, and C is a finite set of *concepts* expressed in a *formal language* F, such that for any node $n_i \in N$, there is one and only one concept $c_i \in C$, and, if n_i is the parent node for n_j, then $c_j \sqsubseteq c_i$.

The formal language F, used to encode concepts in C, belongs to the family of Description Logic (DL) languages [2] and it may differ in its expressive power and reasoning capability. However, the less expressive one with still useful reasoning capability has shown to be the *propositional DL language*, i.e., a DL language without roles (see [7, 4, 6] for examples of practical applications of formal lightweight ontologies based on the propositional DL language).

Taxonomies (e.g., NCBI [16]), thesauri (e.g., GLIN [17]), business catalogues (e.g., UNSPSC [18]), faceted classifications (e.g., Flamenco [1]), web directories (e.g., Yahoo![1]), and user classifications are examples of informal prototypes of formal lightweight ontologies. Hereinafter, we will refer to them as *(informal) lightweight ontologies*. Note that lightweight ontologies are much easer to be understood and constructed by ordinary users. In fact, as shown in [21], formal lightweight ontologies can be automatically constructed from user classifications as a by-product of normal computer use, whereas designing a full-fledged ontology (expressed, for example, in OWL-DL [14]) is a difficult and error-prone task even for experienced users [19].

HISTORICAL BACKGROUND

The notion of ontology was borrowed from philosophy and adopted in Computer Science under multiple definitions, where, probably, the first credible and the most commonly quoted one is "an explicit specification of a conceptualization" [8]. As ontology research evolved, new definitions and examples of what can be considered

---

[1] See http://www.yahoo.com.

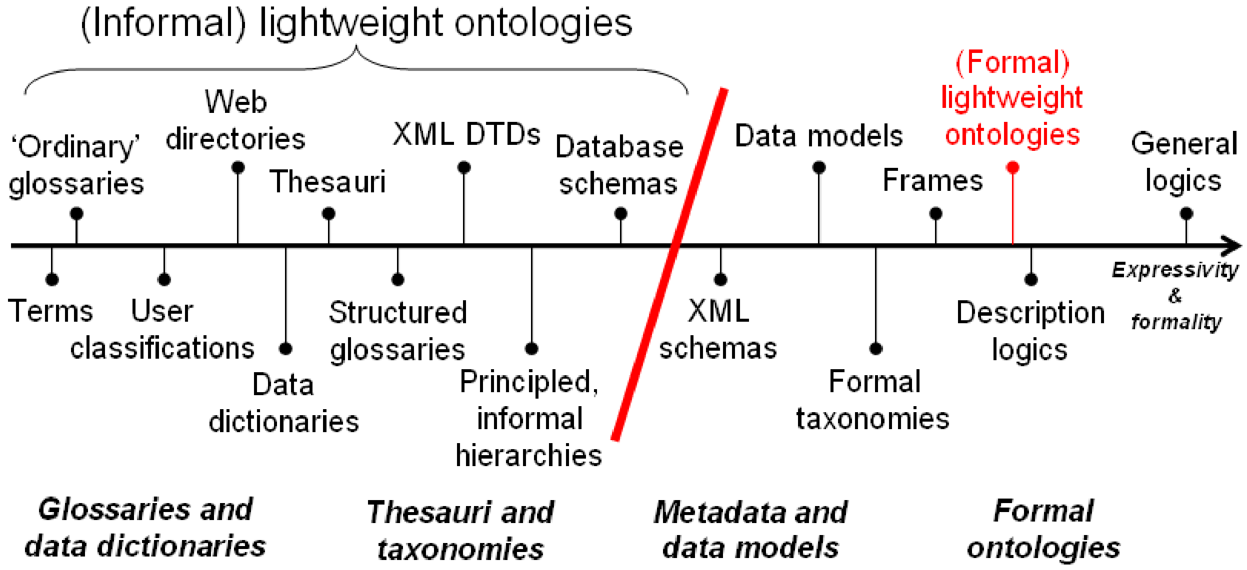

Figure 1: Kinds of ontologies. Adopted from [20].

to be an ontology started to appear. It is typical to characterize ontologies based on the degree of formality and expressivity of the language used to describe them. This characteristic form a continuum of ontology kinds (see Figure 1), starting from terms and web directories, and continuing to rigorously formalized logical theories [20]. However, most specifications agree on that an ontology should be defined in a formal language, which, in practice, usually means a logic-based language suitable for automating reasoning.

For a long time lightweight ontologies were not formally defined but were referred to by examples. For instance, they were referred to as terms, as controlled vocabularies, as thesauri, and as web directories like Yahoo! [20]. As it can be observed, informal lightweight ontologies largely cover the spectrum of informal ontology kinds shown in Figure 1. In some other approaches, lightweight ontologies are formal ontologies which use a computationally inexpensive logic language (e.g., see [4, 11, 13]). In practice, these ontologies often encode a hierarchy of classes which can be (automatically or semi-automatically) derived from web directories like Yahoo! (as in [4, 13]) or from more strictly defined but still informal structures such as thesauri and taxonomies (as in [11]).

To the best of our knowledge, the first attempt to formally define the notion of lightweight ontology as a kind of formal ontology was first made in [3]. In [3], the definition was restricted to formal web directories, whereas in the present document it has been generalized to provide a formal model for a larger spectrum of informal lightweight ontologies, such as controlled vocabularies, taxonomies, thesauri, business catalogues, faceted classifications, web directories, and user classifications.

SCIENTIFIC FUNDAMENTALS

In this section we discuss the main kinds of lightweight ontologies and their properties and we propose how formal lightweight ontologies can be generated from their informal prototypes.

Kinds of lightweight ontologies

Lightweight ontologies fall into two main kinds based on their usage: *descriptive* and *classification* lightweight ontologies. Descriptive lightweight ontologies are primarily used for defining the meaning of terms as well as the nature and structure of a domain [10], whereas classification lightweight ontologies are primarily used for describing, classifying, and accessing (large) collections of documents or, more generally, data items [10, 4]. Due to this difference, formal classification lightweight ontologies have a different domain of interpretation for their concepts. Namely, the extension of a concept in a formal classification lightweight ontology is the *set of documents about* the objects or individuals referred to by the (lexically defined) concept. For example, the extension of concept `mother` is the set of documents about female parents.

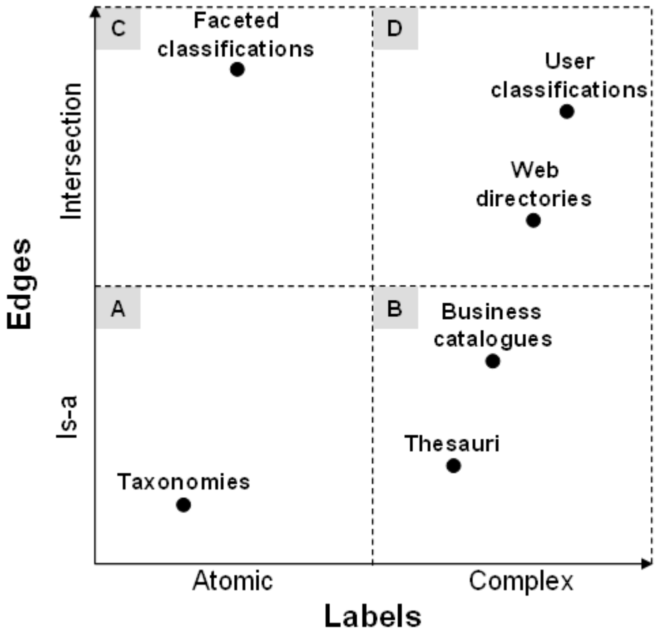

Figure 2: Kinds of labels and edges in lightweight ontologies.

Any descriptive lightweight ontology can be used as a classification lightweight ontology, but not vice versa. In fact, the two kinds differ in some principal properties which we summarize in Table 1 and discuss herein as follows. Classification lightweight ontologies are usually more complex than descriptive ones and the complexity is defined along two dimensions: *label complexity* (atomic vs. complex labels) and *edge complexity* ("is-a" vs. "intersection" edges). Below we discuss four categories of complexity (from the most simple to the most complex one) and we illustrate them in Figure 2.

Category A: atomic labels and "is-a" edges. Ontology labels in this category usually represent single atomic concepts (e.g., "theft", "cellular organisms") and relations between labels are usually "is-a" relations (e.g., "pens" is a child of "writing materials"). Typical examples of this category are (biological) taxonomies such as NCBI [16]. Ontologies in this category are mainly descriptive.

Category B: complex labels and "is-a" edges. Ontology labels in this category can be compound noun phrases which represent complex concepts (e.g., "Open Source and Linux in Education", "Pressure groups representation or participation services") and relations between labels are usually "is-a" relations. Typical examples of this category are thesauri such as GLIN [17] and business catalogues such as UNSPSC [18]. A higher complexity of labels (w.r.t. category A) in these domains is conditioned by the need of richer descriptions of indexing terms in thesauri and of e-commerce items in business catalogues. Most ontologies in this category are descriptive but some can be classification as well. For instance, the business catalogue UNSPSC can be used as a descriptive ontology or as a classification ontology in which e-commerce items are classified. Note that even if the labels can be complex, they are still mapped to atomic concepts in formal descriptive lightweight ontologies. In classification lightweight ontologies, complex labels represent a dimension of *power of classification* as one label can describe one complex concept that identifies a (very) specific set of documents. Moreover, complex labels can be mapped into complex concepts in formal classification ontologies, which allows for higher modularity in concept definitions. For instance, concept `baby_pictures` can be defined as the intersection of two concepts, `baby` and `picture`, whereas the interpretation of the former concept is the set of documents about babies (including pictures of babies as a kind of documents) and the interpretation of the latter concept is the set of pictures (including baby pictures). Note that in formal descriptive lightweight ontologies, the extension of concept `baby_pictures` (the set of baby pictures in the world) cannot be expressed as a function of the extension of concept `baby` (the set of babies in the world) and the extension of concept `picture` (the set of pictures in the world).

Category C: atomic labels and "intersection" edges. Ontology labels in this category usually represent

	Descriptive lightweight ontologies	**Classification lightweight ontologies**
	Informal	
Primary use	Defining the meaning of terms as well as the nature and structure of a domain	Describing, classifying, and accessing (large) collections of documents or, more generally, data items
Labels	Single nouns or simple noun phrases denoting atomic concepts as the most typical case	Often, compound noun phrases denoting complex concepts
Edge relations	"is-a" relation	"is-a", "part-of", or, more generally, "intersection" relation
Examples	Taxonomies (e.g., NCBI [16]), thesauri (e.g., GLIN [17])	Business catalogues (e.g., UNSPSC [18]), faceted classifications (e.g., Flamenco [1]), web directories (e.g., Yahoo!), user classifications
	Formal	
Concept extension	Set of objects or individuals belonging to the class denoted by the concept	Set of documents about the objects or individuals belonging to the class denoted by the concept
Node concepts	Atomic concepts	Atomic and complex concepts

Table 1: A classification of lightweight ontologies.

single atomic concepts and relations between labels are usually "intersection" relations, which means that the label of a parent node *specifies* the meaning of the label of its child node. For example, parent node "Italy" specifies the meaning of its child node "Vacation" to the meaning "Vacation in Italy". A typical example of this category is a faceted classification such as Flamenco [1], in which child nodes represent aspects or facets of their parent nodes along atomic orthogonal dimensions (e.g., time, space, function, material, etc). All ontologies in this category are classification ontologies, for which the "intersection" relation creates an additional dimension of power of classification by allowing it to describe a specific set of documents thought levels of categories in the ontology. Note that the interpretation domain of formal classification ontologies allows it to treat edges as the *intersection* of parent and child concepts and, therefore, compute concepts of nodes given their position in the ontology tree. For example, the intersection of root concept `italy` with its child concept `vacation` results into a concept whose extension is the set of documents about vacations in Italy, which is the actual meaning of the child node, given its position in the tree.

Category D: complex labels and "intersection" edges. Ontology labels in this category usually represent complex concepts and relations between labels are usually "intersection" relations. All ontologies in this category are classification ontologies for which the combination of complex labels and "intersection" edges creates the maximum classification power. Labels in this category can represent names of individuals. These labels are mapped to concepts whose extension is the set of documents about the individuals (e.g., the extension of concept `moscow` is the set of documents about the city Moscow). Typical examples of this category are web directories like Yahoo! (in which web pages are classified) and user classifications (in which email messages, favorites, and files are classified). Note that user classifications may have more complex labels and more "intersection" relations than web directories due to the fact that there are basically no rules and restrictions for user classifications which are commonly followed in web directories.

Note that the propositional DL is sufficient for representing formal descriptive lightweight ontologies in categories A and B, as the only thing that needs to be encoded is the subsumption hierarchy of atomic classes. It is also sufficient for representing formal classification lightweight ontologies in categories A and C, as the only relations that need to be represented are subsumption and intersection of atomic classes. The propositional DL is capable of capturing the semantics of labels in classification lightweight ontologies in categories B and D to a significant extent, while approximating the meaning of labels in some cases, as it is discussed in the following section.

From informal to formal lightweight ontologies

Formal descriptive lightweight ontologies can be generated from informal ones by transforming their organizational structure into a rooted tree (where necessary) and by converting their term labels, expressed in natural language, into concepts in the formal language F. The generation of formal classification lightweight ontologies requires an extra step, namely, node concepts are computed as the intersection of the concepts corresponding to the term labels on the path to the root node [4]. Note that the rooted tree structure of formal lightweight ontologies allows it to capture the backbone organization of many their informal prototypes. In fact, taxonomies, business catalogues, faceted classifications, web directories, and user classifications use rooted trees to organize their categories; in the simplest case, the hierarchy of thesaurus terms, built based on the Broader Term relation, is a rooted tree; and, a controlled vocabulary can be seen as one level rooted tree, where the root node represents the Top concept and its child nodes are the controlled vocabulary terms. Below we discuss the principal steps of the process of conversion of term labels into concepts in F for classification lightweight ontologies. The conversion of labels of descriptive lightweight ontologies follows the same principles even if it is an easier case due to the relative simplicity of their labels. In the following we assume that F is the Propositional DL language.

WordNet [15] senses of adjectives and common nouns in the label become atomic concepts in F. The extension of a common noun concept is the set of documents about objects of the class, denoted by the noun; and, the extension of an adjective concept is the set of documents about objects, which possess the qualities, denoted by the adjective. Proper names become atomic concepts as well, whose extension is the set of documents about the individual referenced by the proper name. Notationally, we represent adjective and common noun atomic concepts using the following syntax: *lemma-pos-sn*, where *lemma* is the lemma of the word, *pos* is its part of speech, and *sn* is the sense number in WordNet [15]. We use $_{NNP}$ to tag proper name atomic concepts.

Syntactic relations between words in the label are translated into logical connectives of F in order to build complex concepts from atomic concepts. For example, a set of adjectives followed by a noun group is translated into the logical conjunction ($\sqcap$) of the concepts corresponding to the adjectives and to the nouns; prepositions like "of" and "in" are translated into the conjunction; coordinating conjunctions "and" and "or" are translated into the logical disjunction ($\sqcup$). The final formula for the label is built following these rules and taking into account how words are coordinated in the label.

Let us consider a relatively complex label: "*Bank and personal details of George Bush*". Its correct translation to F produces the following concept:

$$(\texttt{bank-noun-1} \sqcup \texttt{personal-adj-1}) \sqcap \texttt{detail-noun-1} \sqcap \texttt{george\_bush}_{NNP}$$

The extension of the concept above is the intersection of three sets of documents: (i) documents about the President George W. Bush, (ii) documents containing isolated facts about something (i.e., details), and (iii) the union of documents about bank institutions and documents concerning a particular person or his/her private life. Note that the extension comprises (all and only) documents one would classify under the above given label.

Despite its seeming simplicity, the translation process is subject to various mistakes originating from incorrect natural language processing (NLP)[2]. For instance, due to a mistake in part-of-speech tagging, the word *personal* might be recognized as a noun, which has only one sense in WordNet defined as "*a short newspaper article about a particular person or group*"; due to a mistake in word sense disambiguation, the sense of the word *bank* might be identified as "*sloping land (especially the slope beside a body of water)*"; due to a mistake in named entity locating, the proper name *George Bush* might not be recognized and might then be considered as two distinct nouns, where the noun *bush* means "*a low woody perennial plant usually having several major branches*"; finally, due to a mistake in (syntax) parsing, the input label might be erroneously translated into:

$$\texttt{bank-noun-1} \sqcup \texttt{personal-adj-1} \sqcap \texttt{detail-noun-1} \sqcap \texttt{george\_bush}_{NP}$$

a concept, whose extension is the union of documents about bank institutions and documents discussing personal details of the President George W. Bush.

Note that the propositional DL can capture the semantics of complex labels to a significant extent only when these labels are built from noun phrases possibly connected through coordinating conjunctions such as "and" and "or" (e.g., "Big city life and civil protection"). It approximates the meaning of some other kinds of labels, e.g.,

---

[2] How the NLP problems, described in this paragraph, can be solved is beyond the scope of this document. Interested readers are referred to [21] for a first account.

labels with prepositions. For instance, labels “life in war” and “life after war” will collapse into the same formula, which approximates the meaning of both labels.

KEY APPLICATIONS

Formal (descriptive and classification) lightweight ontologies can be used in various domains, such as document classification (e.g., see [7]), semantic search (e.g., see [4]), and data integration (e.g., see [6]). In the following subsections we briefly discuss these three application domains.

Document classification

Document classification is the problem of assigning a document to one or more categories based on the document contents. In the context of classification lightweight ontologies, document classification refers to assigning a document to: (i) controlled vocabulary terms; (ii) categories in taxonomies, business catalogues, faceted classifications, web directories, or user classifications. The approach reported in [7] presents fully automatic classification of documents into web directories based on the *get-specific* document classification algorithm. The underlying idea is that a web directory is converted into a formal lightweight ontology, that a document is assigned a concept, and that the document classification problem is then reduced to reasoning about subsumption on the formal lightweight ontology. Note that this classification approach does not require the creation of a training dataset which would normally be required in machine learning approaches [7].

Semantic search

In the context of lightweight ontologies, semantic search is the problem of finding categories and/or documents (when applicable) classified in categories of (informal) lightweight ontologies, such that the found objects *semantically* correspond to a provided natural language query. Loosely speaking, semantic correspondence of an object to a query means that the meaning associated with the object is more specific or equivalent to the meaning given to the query under common sense interpretation. For instance, document about *Ethiopian villages* semantically corresponds to a query about *African settlements.* The approach reported in [4] formalizes the above informal description and introduces a semantic search algorithm for lightweight classification ontologies populated with documents. The underlying idea is that the user query is converted to a concept in the way presented earlier in this document and that the answer to the query is computed as the set of documents whose concepts are more specific or equivalent to the concept of the query. In order to reduce the computation complexity, the query is first run on the structure of the corresponding formal lightweight ontology in order to identify the scope of relevant nodes and then it is run on the documents populated in some of the nodes from the scope.

Data integration

Data integration is the process of combining data residing at different sources and providing the user with a unified view of these data [12]. Often, a data source can be represented in the form of a rooted tree, whose nodes are assigned natural language labels, and, in this case, data integration can be facilitated by discovering semantic relations which exist between nodes of the source trees [6]. A semantic relation between two nodes can be more/less general, equivalent, or disjoint. In the domain of lightweight ontologies, semantic relations can be found between elements of controlled vocabularies, taxonomies, thesauri, business catalogues, faceted classifications, web directories, and user classifications. Found relations can then be used for enabling integration or inter-operation of web directories, for merging business catalogues, and so on.

FUTURE DIRECTIONS

There are two major problems related to formal lightweight ontologies which drive future directions of research in this area. The problems are:

Natural language processing. Since formal lightweight ontologies are supposed to be often generated from their informal prototypes, the quality of these ontologies strongly depends of the correctness and completeness of NLP procedures involved in the conversion process. Note that NLP for informal lightweight ontologies is a potentially new domain in the NLP research due to the particular characteristics of term labels (e.g., they are usually short noun phrases with little context) [21];

- **Lack of background knowledge.** Reasoning on formal lightweight ontologies, which is used, for example, in document classification, in semantic search, and in data integration as discussed above, strongly depends on the set of axioms which must be known a priori [4]. These axioms are extracted from a knowledge base such as WordNet [15]. It was shown that lack of background knowledge is the main source of a relatively low recall in reasoning-based tasks on formal lightweight ontologies [5].

EXPERIMENTAL RESULTS

First evaluation studies of document classification show that re-classification of 1217 HTML pages into a part of the DMoz hierarchy[3], that has 157 nodes distributed in a tree of depth 6, allows it to reach 41% in the micro-averaged F1 measure [7]. Document concepts in the conducted experiments were built by computing the conjunction of the formulas corresponding to the first 10 most frequent words appearing in the documents (excluding stop words).

The performance of S-Match, a tool that, among other things, facilitates the integration of lightweight ontologies, reaches about 50% and higher in F-measure on some data sets [6].

DATA SETS

Some informal lightweight ontologies are available for download in the form of data files. These include the DMoz web directory[4], business catalogues UNSPSC[5] and eCl@ss[6], NCBI taxonomy[7], and many others.

CROSS REFERENCE

I(a) – DATABASE FUNDAMENTALS: Data models (including semantic data models)
VIII(i) – DATA INTEGRATION: Metadata management

RECOMMENDED READING

Between 3 and 15 citations to important literature, e.g., in journals, conference proceedings, and websites.

[1] Flamenco demos. See http://flamenco.berkeley.edu/demos.html. Retrieved on September 10, 2007.

[2] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.

[3] F. Giunchiglia, M. Marchese, and I. Zaihrayeu. Towards a theory of formal classification. In *Proceedings of the AAAI'05 International Workshop Contexts and Ontologies: Theory, Practice and Applications*, Pittsburgh, USA, July 2005. AAAI Press. ISBN 1-57735-237-8.

[4] F. Giunchiglia, M.Marchese, and I. Zaihrayeu. Encoding classifications into lightweight ontologies. In *JoDS VIII: Special Issue on Extended Papers from 2005 Conferences*, Winter 2006. A shorter version of the paper appeared in ESWC-2006.

[5] F. Giunchiglia, P. Shvaiko, and M. Yatskevich. Discovering missing background knowledge in ontology matching. In *ECAI*, pages 382–386, 2006.

[6] F. Giunchiglia, M. Yatskevich, and P. Shvaiko. Semantic matching: Algorithms and implementation. *Journal on Data Semantics*, IX, 2007.

[7] F. Giunchiglia, I. Zaihrayeu, and U. Kharkevich. Formalizing the get-specific document classification algorithm. In *11th European Conference on Research and Advanced Technology for Digital Libraries*, Budapest, Hungary, September 2007.

[8] T. R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, 1993.

[9] N. Guarino. Some ontological principles for designing upper level lexical resources. In *First International Conference on Lexical Resources and Evaluation*, volume 2830, Granada, Spain, May 1998.

[10] N. Guarino. Helping people (and machines) understanding each other: The role of formal ontology. In *CoopIS/DOA/ODBASE (1)*, page 599, 2004.

---

[3] See http://www.dmoz.org.

[4] See http://rdf.dmoz.org.

[5] See http://www.unspsc.org.

[6] See http://www.eclassdownload.com.

[7] See ftp://ftp.ncbi.nih.gov/pub/taxonomy.

[11] M. Hepp and J. de Bruijn. GenTax: A generic methodology for deriving OWL and RDF-S ontologies from hierarchical classifications, thesauri, and inconsistent taxonomies. In *4th European Semantic Web Conference*, LNCS 4519, pages 129–144, Innsbruck, Austria, June 3-7 2007. Springer.

[12] M. Lenzerini. Data integration: a theoretical perspective. In *PODS '02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246, New York, NY, USA, 2002. ACM Press.

[13] B. Magnini, L. Serafini, and M. Speranza. *AI*IA 2003: Advances in Artificial Intelligence*, volume 2829/2003 of *LNCS*, chapter Making Explicit the Hidden Semantics of Hierarchical Classifications, pages 436–448. Springer Berlin/Heidelberg, 2003.

[14] D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview. Technical report, World Wide Web Consortium (W3C), February 10 2004. See http://www.w3.org/TR/owl-features/.

[15] G. Miller. *WordNet: An electronic Lexical Database*. MIT Press, 1998.

[16] National Institutes of Health National Center for Biotechnology Information, National Library of Medicine. Taxonomy browser. Available at http://www.ncbi.nlm.nih.gov/Taxonomy/.

[17] Library of Congress. Thesaurus for the Global Legal Information Network (GLIN). Available at http://www.loc.gov/lexico/servlet/lexico/.

[18] United Nations Development Programme. United nations standard products and services code (UNSPSC). Available at http://www.unspsc.org/. Retrieved on March 15, 2007.

[19] A. L. Rector, N. Drummond, M. Horridge, J. Rogers, H. Knublauch, R. Stevens, H. Wang, and C. Wroe. OWL pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. In *EKAW*, pages 63–81, 2004.

[20] M. Uschold and M. Gruninger. Ontologies and semantics for seamless connectivity. *SIGMOD Rec.*, 33(4):58–64, 2004.

[21] 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)*. Springer, 2007.