UNIVERSITÀ DI PISA

Master's Degree Course in
Humanities Computing

Modeling entity types in faceted lightweight ontologies

Candidate:
Raffaele Guarasci

Supervisor:
Prof. Alessandro Lenci

Co-supervisor 1:
Prof. Fausto Giunchiglia

Co-supervisor 2:
Prof. Maria Simi

Department of Philology, Literature and Linguistics

Academic Year 2012-2013

Table of Contents

Chapter 1

State of the art

Introduction

Since its creation, the Web has reached massive dimensions: so being able to categorize and classify this huge amount of information is really important.

A classification scheme, simply named classification, is a rooted tree made up of a set of documents: a natural language label is assigned to each node of this tree.

Aristotle was the first philosopher who, in the 4th century, “invented” a method of classification. Even nowadays classifications are used as a useful method for representing the various kinds of human knowledge, especially hierarchical classifications are the most pervasively approaches used.

One of the major advantages of this kind of classification is that natural language used to describe their contents is easily understood by human users. However, at the same time, this is also one of their main disadvantages as these same labels are ambiguous and very hard to be processed by software. Because of this hindrance, establishing classifications in the Semantic Web infrastructure is a very difficult task.

To compensate for this deal, web content should also be expressed in a language that can be unambiguously understood, interpreted and used by software to find, share and relate information more easily[1]. This solution is, in fact, the underlying idea of the Semantic Web, which is based on the idea of ontology.

An *ontology* defines a taxonomy of classes of objects and relations among them; unlike classifications, ontologies should be written in unambiguous and full machine-readable formal languages.

This work provides an overview of the various types of ontologies and their applications for knowledge representation. In particular, the work is focused on some approaches based on faceted lightweight ontologies to manage diversity in knowledge and, in the last chapter; results of an ontological modeling work on a specific domains are shown.

The thesis is structured as follows. Chapter 1 introduces the state-of the-art notions of classifications and ontologies for representing knowledge. From Section 1.1 to 1.3, we discuss the different kinds of ontologies, we introduce the classification scheme, lightweight ontology, and ontology. After that, a comparison between classification schemes and ontologies is provided. Section 1.4 discusses lightweight ontologies, their applications, and the problems involved in their applications. Section 1.5 discusses background knowledge for the ontologies. In Section 1.6 we present the faceted lightweight ontology as a solution to the problems of the lightweight ontology applications.

[1] T. Berners-Lee, J. Hendler, and O. Lassila. *The semantic web*. Scientific American, (284(5)):34–43, May 2001.

Chapter 2 focuses on the problem of knowledge representation. Section 2.1 discuss different types of knowledge bases. Section 2.2 introduce the fundamental notion of diversity in knowledge. Section 2.3 introduce an approach focused on concepts of *domain* and *context*. Section 2.4. describes the faceted approach since its origins. The rest of the chapter describes a methodology proposed for the construction of entity-centric data model and the creation of a flexible diversity-aware knowledge base.

Chapter 3 describes the work done on modeling some concepts in an ontology, starting from the approach explained in Chapter 2. Sections 3.1 analyzes the work done on the *Mind Product* type. Sections 3.2 and 3.3 describe the type *Information Object* and its sub-types *Image File*. Section 3.4 introduces first steps on definition of the *Event* type. Section 3.5 concludes the thesis by summarizing the work done and outlying the open issues.

1.1. Ontologies

An *ontology* may be broadly defined as an explicit formal description, or model, of the concepts of a domain. Ontologies are structures that formally define the nature and the structure of any organized system; they make explicit entities, concepts, objects, process and relations, highlighting hierarchical relations among them, and they can represent it in a formal language easily understandable by machine.

The concept of *ontology* comes from Aristotelian *theory of categories* and the notion of *metaphysics*, which firstly studies the essence of living beings (living beings as being), and secondly, the basic characteristics of reality as a whole (the being or principal entity upon which other entities depend). The original purpose was to provide a categorization of all existing things in the world.

In Aristotle times ontology was considered a branch of philosophy, that aims at explaining existence in a systematic manner, pertaining to the types and structures of objects, properties, events, processes and relations related to each part of reality: ontologies are nowadays adopted in several other fields, as Libraries and Information Science, or Artificial Intelligence too.

Since the last 20 years ontologies have become more important in the field of Artificial Intelligence, especially regarding Knowledge Engineering and Knowledge Representation, in so far that “Artificial Intelligence deals with reasoning about models of the world. Therefore, it is not strange that the term ontology was adopted to describe what can be computationally represented of the world in a program"[2].

The studies related to Artificial Intelligence that attempt to formalize knowledge representation languages led to the adoption of Description Logic (DL). Description Logics[3] is used in Artificial Intelligence for formal reasoning on the concepts of an application domain (known as terminological knowledge), and it is of particular importance in order to provide a logical formalism for ontologies in the Semantic Web.

Many definitions of ontologies have been provided, but the most quoted is the one that proposed by Gruber regarding to Artificial Intelligence and Knowledge Representation studies: "An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence[4]".

[2] Studer, R. et al. (1998). *Knowledge engineering: principles and methods*. Pennsylvania: School of Information Sciences and Technology (IST). Pennsylvania State University.

[3] Description logics is a family of formal knowledge representation languages, It is more expressive than propositional logic but has more efficient decision problems than first-order predicate logic. Description Logics are a subset of first-order logic, and this, with respect to propositional logic, which deals with simple declarative sentences, provides the tools to express propositions about objects, properties that objects may have in common, and the relationships between objects.

[4] Gruber, T. R. (1993). *A translation approach to portable ontology specifications*. Knowledge Acquisition, 5 (2), 199–220. Many other definitions are been proposed in recent years, some concerning the philosophical aspects, others that regard linguistic ones. In literature have been proposed many other definitions of ontology: "*An ontology defines the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary.*" (Neches, et al.); *"An ontology provides the means for describing explicitly the conceptualization behind the knowledge represented in a knowledge base."* (Bernaras, et al.); *"The subject of ontology is the study of the categories of things that exist or may exist in some domain."* (J. Sowa); "*an ontology is a catalogue of the types of things that*

The keywords on which this definition is focused are broadly explained below.

Conceptualization refers to an abstract model of the world in terms of basic cognitive units called concepts. Concepts represent the intension (i.e. the set of properties thanks to which a concept diverge from another), and summarize the extension (i.e. the set of objects having such properties).

Since concepts basically denote classes of objects, as an example the medicine domain can be modeled in terms of doctors, patients, body parts, diseases, symptoms and treatments used to cure or prevent diseases.

Explicit specification means that the abstract model is made explicit by providing names and definitions for concepts. In other words the name and the definition of each concept provide a specification of its meaning in relation with other concepts. The specification can be termed *formal* when the language used has formal syntax and formal semantics, as with a logic-based language; natural languages cannot be used for this purpose, because of their ambiguity. The conceptualization is *shared*, that means it captures knowledge when it is common to a people community. An ontology provides a common formal terminology and a grasp of a given domain of interest, and thus it allows for automation (logical inference), supports reuse and favor interoperability between applications and people. An ontology is called a *knowledge base* when it is made up of instances of the classes (the individuals).

are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D". (J.Sowa).

Although the common core is represented by a vocabulary of terms and the corresponding specification of their meaning, there are several kinds of ontologies, according to the specificity of the information described, and to the degree of formality and expressivity of the language used to build them[5]. An ontology can range from informal representations, like a user classification (e.g. the structure of folders in a file system) or a web directories (e.g. DMOZ[6], Yahoo![7] and Google[8]), to progressively more formal representations. These representations include enumerative classification schemes (e.g. the Dewey Decimal Classification[9] and the Library of Congress Classification[10]), *thesauri* (e.g. AGROVOC[11], NALT[12], AOD[13], and HBS[14]), faceted classification schemes (e.g., the Colon Classification[15]) , and, ultimately, formal ontologies expressed into a logic formal languages and represented using formal specifications such as DL or OWL[16].

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

[6] http://dmoz.org/;

[7] http://dir.yahoo.com/;

[8] http://directory.google.com/

[9] Dewey M. (1876), A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a library, OCLC 78870163

[10] See http://www.loc.gov

[11] See http://aims.fao.org/website/AGROVOC-Thesaurus/sub

[12] See http://agclass.nal.usda.gov/

[13] See http://etoh.niaaa.nih.gov/aodvol1/aodthome.htm

[14] See http://hul.harvard.edu/ois/ldi/

[15] S. R. Ranganathan, *The Colon Classification*, Rutgers Series on Systems for the Intellectual Organization of Information, S. Artandi (etd.), IV, Graduate School of Library Science, Rutgers University, New Brunswick, NJ, 1965.

[16] http://www.w3.org/TR/owl-features/

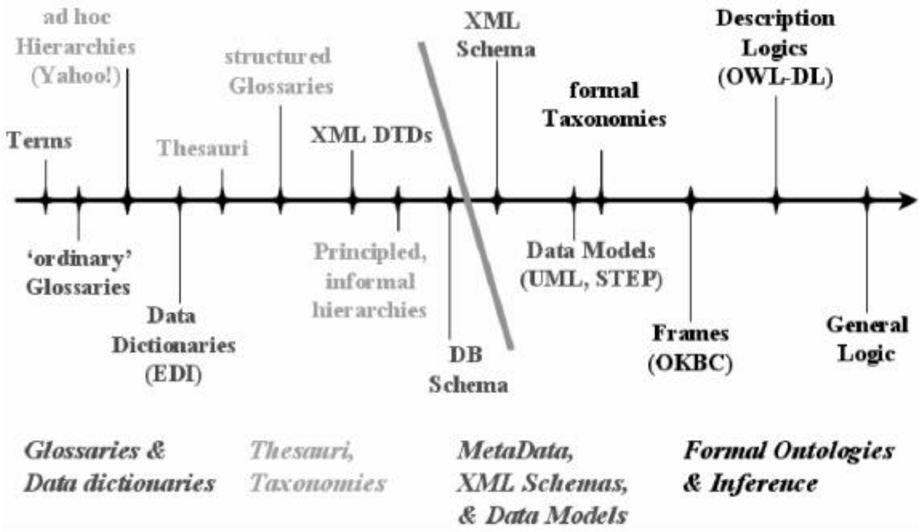


Figure 1.1 Kinds of ontologies, taken from Uschold and Gruninger(2004).

1.2. Types of ontologies

As described before, 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. The difference in the level of formality and expressivity is typically a function of the intended purpose. Since there are so many different kinds of ontology, it will be useful to share them into a functional division: the most known macro-division is based on the structure of information that ontology describes.

- **Top-level ontology** (or **upper ontology** or **foundation ontology**) describes very general concepts that are the same across all knowledge domains. The aim of a top-level ontology is to be totally domain-independent, in order to support very broad semantic interoperability;

- **Domain ontology** (or **domain-specific ontology**) describes and models a specific domain, narrowing concepts introduced in top-level ontology, which represents part of the world;

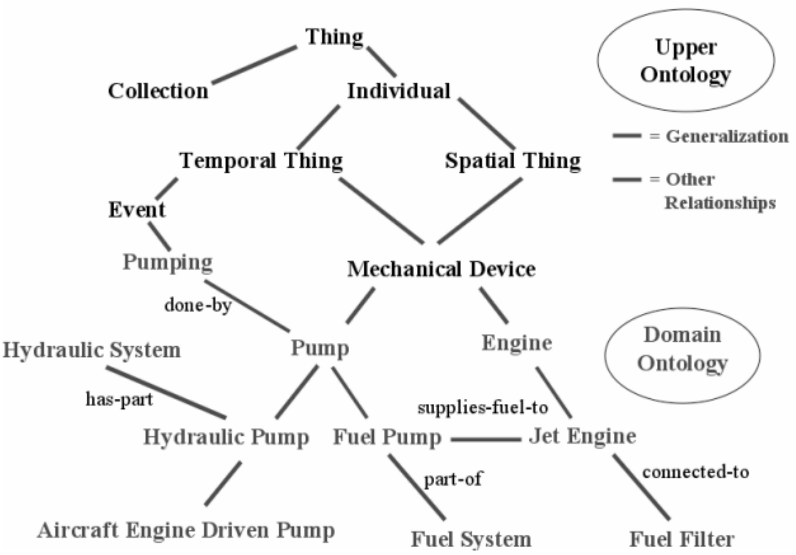


Figure 1.2 An example ontology, taken from Uschold and Gruninger(2004).

According to the terminology proposed in Giunchiglia and Zaihrayeu (2008)[17], ontologies can be mainly distinguish between:

1. **Descriptive ontologies**, which are mainly used to describe objects;
2. **Classification ontologies**, prevalently used for categorizing objects;

[17] Giunchiglia, F., Zaihrayeu, I. (2008). *Lightweight ontologies*. Encyclopedia of Database Systems.

Notice that to these different types of ontologies correspond different types of semantics, respectively called *real world semantics* and *classification semantics*.

1.2.1. Descriptive ontologies

In this kind of ontologies, concepts represent real world entities, and such entities can be connected via relations of the proper kind. Descriptive ontologies aim to specify the terms used in their original meaning, according to the nature and the structure of the domain they model[18]. These ontologies are in *real world semantics*[19], so the terms at nodes represent either individuals or classes of real world objects.

Two typical relations are used to build the trees/taxonomies, which provide the backbone to these ontologies: the *is-a* (Genus-species) and *part-of* (Whole-part) relations. Another important relation is *instance-of*, which indicates the relationship between classes and individuals represented in the schema.

The example reported in figure 1.3 (taken from Maltese and Farazi, 2011[20]) shows a scheme that describes some organizations and their location. White nodes represent classes while the black ones represent individuals. The first label at the nodes represents the preferred term and additional synonymous terms are in some

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

[19] Giunchiglia, F., Marchese, M., Zaihrayeu, I. (2007). *Encoding Classifications into Lightweight Ontologies*. Journal of Data Semantics, 8, 57-81.

[20] Maltese, V., Farazi, F. (2011). Towards the Integration of Knowledge Organization Systems with the Linked Data Cloud, UDC seminar.

cases provided in semicolon. Arrows represent relations and their direction indicates the direction of the relation. For instance, the term *country* (defined as: the territory occupied by a nation) denotes all the real world countries, while the term *Italy* indicates Italy as a country.

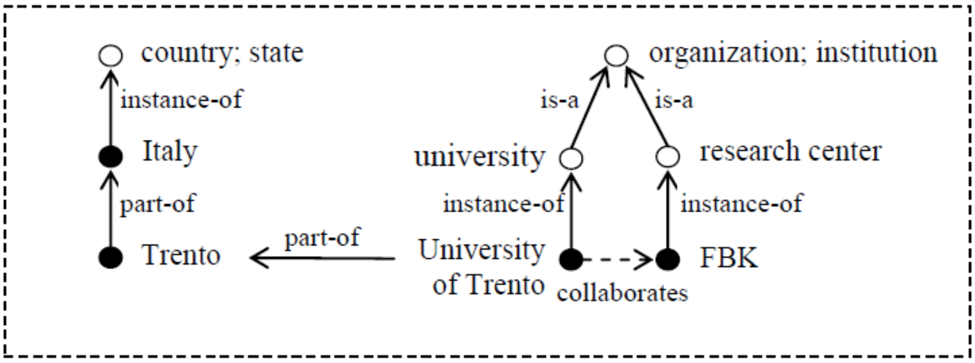


Figure 1.3 an example of descriptive ontology (taken from Malese, Farazi 2011)

Under this semantics, there is an *is-a* relation between the class named *organization* and subclass *university*, an *instance-of* relation between the class *country* and the individual *Italy* an *part-of* relation between the two classes *Trento* and *Italy*.

These kinds of schemes represent what it is known about the domain and they can be used to reason about it, since they provide knowledge about classes, attributes and relations.

For automating tasks, it is possible to translate these schemes into formal (descriptive) ontologies. By using Description Logics[21] terminology, *classes* can be translated into *concepts,* the *is-a* relation can be rendered into *logical subsumption*. The *is-a* relation constitutes the basic backbone of the hierarchical structure based on subsumption of a domain.

As emphasized by several works[22], taking the properties of the relations into account is also important, especially the transitivity of the relations[23].

Subsumption itself is assumed transitive, as well as the generic *part-of* relation. Nevertheless, if several kinds of part-of occur, this relation might lose the transitive property, especially when the various kinds are combined together[24].

In the previous example, the *part-of* relation between *Italy* and *Trento* is considered an *administrative part-of* relation, while the one between *Trento* and *University of Trento* can be characterized as a *topological part-of* or even just a generic associative relation. In fact, in this case, just the building that host the university as institution is located in Trento, rather than the institution as such.

[21] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. F. (2002). *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press.

[22] Varzi, A. C. (2006). *A note on the transitivity of parthood*. Applied Ontology, 1, 141-146.

[23] Cfr. Horrocks, I., Sattler, U. (1999). *A description logic with transitive and inverse roles and role hierarchies*. Journal of Logic and Computation, 9(3), 385-410.

[24] The typical example is the handle that is part of the door that is part of the house that after a chain of other *part-of* relations ends to be part of the universe.

1.2.2. Classification ontologies

Classification ontologies (in classification semantics) are built to describe, classify and search documents[25]. In *classification semantics* the terms labeled at nodes always represent classes of documents, while the instances are the documents themselves.

The example in figure 1.4., taken from Maltese and Farazi, (2011), shows a thesaurus made up with the purpose of classifying documents by country and organization. The labels at the nodes indicate the preferred term, optionally followed by synonymous terms in a separated semicolon, and arrows represent relations. If a documents is assigned to a node, it is marked by the letter *d* followed by an index.

In these schemes, hierarchical relations are represented by NT/BT (narrower term/broader term) relations, in which the direction of each arrow leaves the narrower term to reach the broader one. These relations are used to facilitate the indexing and searching tasks. In particular, following NT relations allows identifying progressively more specific concepts, since the extension (i.e. the set of documents about the concept) decreases. Following the inverse direction, in other words the BT relation, enables to identify progressively more general concepts, thus increasing the extension. Note that there are also RT (related term) relations, that are associative relations.

[25] Notice that, in this work, the term *document* is used in the sense of what archivists generically call documents, that is anything that can be archived, whether in a physical or digital format

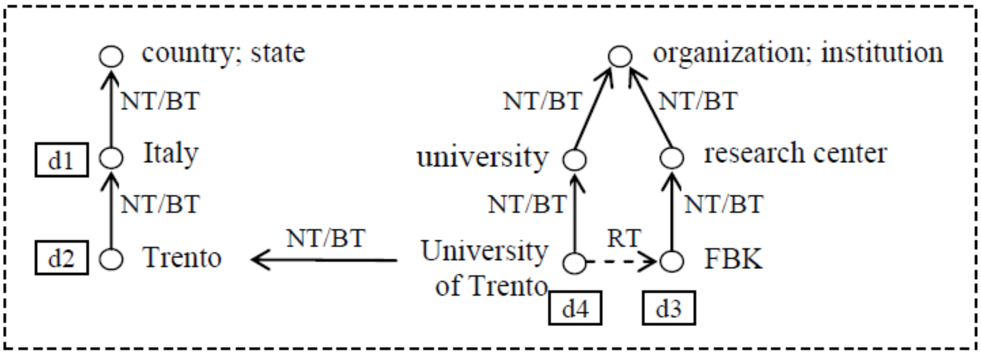


Figure 1.4 Example of thesaurus for classifying documents by country and organization

In the example, the term *country* indicates all the documents about countries. Following this semantics, NT/BT relations represent *subset/superset* relations, since NT and BT are one the inverse of the other. For instance, if the node *Italy* is linked to *country* through a BT relation, the semantics of the node *Italy* is the set of documents about Italy as a country.

To make explicit the intended semantics and to automate tasks it is possible to provide a formal representation of the schema using Description Logics terminology. As seen before, converting the schema into the corresponding formal (classification) ontology, classes becomes concepts, documents correspond to individuals in the domain of interpretation, and transitive NT/BT relations are translated into logical subsumption.

1.3. From descriptive to classification ontologies

It should now be clear how the difference between these two kinds of ontologies is reflected in two totally different semantics.

Despite of the different semantics it is possible to integrate a classification ontology with a descriptive one and vice versa, after a preliminary conversion of both ontologies, in order to have the same semantics.

If the aim is to classify, both schemes can be converted into classification ontologies; conversely, if the goal is to describe a domain, it is better to transform both ontologies into a descriptive one type.

According to the approach proposed by Ranganathan[26], it is possible to convert a descriptive ontology into the corresponding classification one, following these steps:

- converting instances into classes;
- converting *instance-of*, *is-a* and transitive *part-of* into NT/BT relations;
- converting other relations into RT relations;

This translation process implies an obvious loss of information. Real world classes and instances collapse into document classes, while *instance-of*, *is-a* and transitive *part-of* relations become undifferentiated hierarchical relations, and all

[26] This approach is in line with Ranganathan approach. Indeed he says that hierarchies are constructed on the basis of *genus-species* (*is-a* and *instance-of*) and *whole-part* (*part-of*) relations. See Ranganathan, S. R. (1967). *Prolegomena to library classification*. Asia Publishing House.

the other ones become associative relations. Anyway, the opposite conversion is possible too:

- each class has to be mapped to either a real world class or instance;
- each transitive NT/BT relation has to be converted to an *instance-of*, *is-a* or transitive *part-of*;
- each RT relation has to be codified into an appropriate real world associative relation;

Descriptive ontologies ensure maximum reusability. In fact, this is very useful for those applications that need to reason on a domain, and it requires a minimum effort to (automatically) convert them into classification ontologies when needed.

Conversely, if a scheme is made as a classification ontology, a significant human effort will be necessary to reconstruct its real world version. Then, to serve these different applications, both descriptive and classification ontologies are needed.

1.4. Lightweight ontologies

Classifications have been traditionally used as indexing and browsing structures for books and other bibliographic material in libraries. As described before, classifications are tree-like hierarchical structures, in which the content is described by attaching to each node a natural language label, while the links between nodes implicitly represent subset relations. For instance, when a node labeled *milk* is put under *cow*, this typically means that it contains documents about milk produced by cows, and that this set of documents is a subset of the documents about cows. Thus, in relation to their target application, a different interpretation of nodes and links of classifications[27] is possible. A classification can be defined as follows:

A classification is a rooted tree C = <N, E, L> where N is a finite set of nodes, E is a set of edges on N, and L is a finite set of labels expressed in natural language, such that for any node $n_i \in N$ these is one and only one label $l_i \in L$[*28*].

The example in figure 1.5, taken from Giunchiglia et al., (2012)[29], represents two very simple classifications: in these, white nodes represent categories, while the black ones exemplify annotated documents. Solid arrows between nodes represent sub-category relations, while dashed arrows means that a document is categorized into a certain category. Attached to nodes there are corresponding labels too.

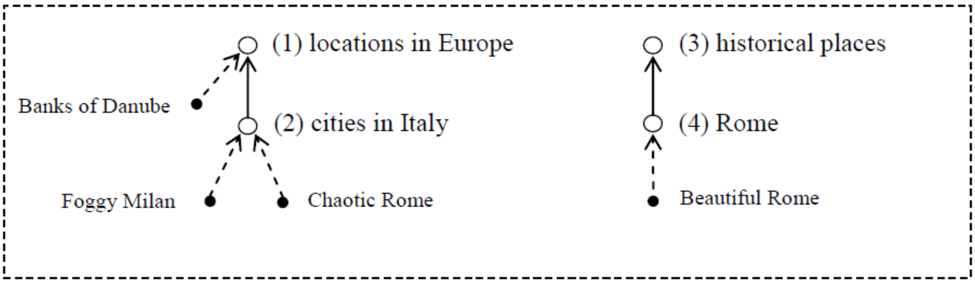


Figure 1.5 two example of classifications

[27] Giunchiglia, F., Marchese, M., Zaihrayeu, I. (2007). *Encoding Classifications into Lightweight Ontologies*. Journal of Data Semantics, 8, 57-81.

[28] *Ibid.*

[29] Giunchiglia, F., Maltese, V., Dutta, B. (2012). *Domains and context: first steps towards managing diversity in knowledge*. Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web.

Although classifications have no explicit formal semantics for edges, it is possible to represent both ontologies and classifications in the form of a graph. However, ontologies and classifications remain quite different in their uses, purpose, language, applications, and in other aspects.

- **Users:** A typical user of classifications is a human (e.g., a classifier in a library classification), whereas ontologies are primarily used by machines and, as such, are the key enablers of the Semantic Web.
- **Purpose:** Classifications are primarily used for the organization of (large) document collections into categories and subcategories, so that these documents can be easily accessed by humans. By contrast, ontologies are used for modeling a particular domain so that the resulting model represents a shared view of a group of individuals.
- **Language:** To describe nodes' categories, classifications use natural language that is well understood by humans but has an ambiguous nature. By contrast, ontologies are codified in a formal language, which is unambiguously interpreted by machines.
- **Nodes:** In an ontology, nodes normally represent atomic concepts (e.g., car, wine). In a classification, a label can represent a rather complex concept (e.g., "Open Source and Linux in Education") or an individual (e.g., "Napoleon Bonaparte").

- **Edges:** In an ontology graph, edges have well-defined semantics and they usually encode *sub-class-of*, *part-of* and other relations that hold between the two concepts connected by an edge. In a classification, an edge implicitly represents either a *specification* relation (like *is-a* relation) or as a *part-of* relation.

Classifications and ontologies are quite different and both have their pros and cons with respect to each other.

Classifications turn out to be very effective in manual tasks, but the automation of these processes need a modification into formal classification ontologies. For this purpose,[30] a series of techniques to formalize the meaning of labels and links in a classification, have been developed recently. This conversion procedure associates to each node in the classification a formula in a formal language, representing the meaning of the node in terms of classification semantics. The result is a *lightweight ontology*, a concept which links the gap between classifications and ontologies.

A lightweight ontology is defined as:

A (formal) lightweight ontology is a triple O = <N, E, C> where N is a finite set of nodes, E is a set of edges on N, such that <N, E> is a rooted tree, and C is a finite set of concepts expressed in a formal language F, such that for any node $n_i \in$

[30] See Giunchiglia et al., 2009, and Autayeu, et al., (2010) *Recommendations for Better Quality Ontology Matching Evaluations*. 2nd AISB Workshop on Matching and Meaning.

N, there is one and only one concept $c_i \in C$, and, if n_i is the par-ent node for nj, then $cj \sqsubseteq ci$.[31]

The formal language F used to encode concepts in C belongs to the family of DL languages and it may differ in its expressive power and reasoning capabilities[32]. The set of concepts C are taken from some form of *background knowledge*, for instance from WordNet. In fact, WordNet synsets, grouping words having the same meaning, is similar to concepts; hypernym and meronym relations between synsets can be considered as subsumption between concepts and the semantics is similar to the classification semantics.

The conversion of a classification into a lightweight ontology can be performed in two steps:

1. For all the labels in the classification compute the *concept at label*;
2. For all the nodes in the classification compute the *concepts at node*;

[31] F. Giunchiglia, B. Dutta, and V. Maltese. Faceted lightweight ontologies. In *Conceptual Modeling: Foundations and Applications: Essays in Honor of John Mylopoulos*, pages 36–51, Berlin, Heidelberg, 2009. Springer-Verlag.

[32] Autayeu, et al. (2010) shows that the expressive power necessary is very low, in fact node labels tend to be noun phrases and it is therefore sufficient to describe them in terms of conjunctions of atomic concepts representing intersections of sets of documents. Furthermore, in a recent experiment Giunchiglia et al, (2009) the labels of the classifications considered turn out to have a simple translation into propositional DL with a few *local* disjunctions and no negations.

In the first step, the nodes are labelled in isolation. Using NLP techniques, tuned for short noun phrases, such as proposed in Zaihrayeu et al. (2007)[33], their meaning is determined by constructing a corresponding formula, called the *concept at label*. Anyway, since the label alone does not provide enough clues for the disambiguation, it is necessary to keep all possible senses of the words. For instance, the concept at label of node 2 in figure 1.5 is (city#1 ⊔ city#2 ⊔ city#3) ⊓ Italy#1, where:

city#1: city, metropolis, urban center	- a large and densely populated urban area; may include several independent administrative districts; "Ancient Troy was a great city"
city#2: city	- an incorporated administrative district established by state charter; "the city raised the tax rate"
city#3: city, metropolis	- people living in a large densely populated municipality; "the city voted for Republicans in 1994"
Italy#1: Italy, Italian Republic,	- a republic in southern Europe on the Italian Peninsula; was the core of the

[33] Zaihrayeu, I., Sun, L., Giunchiglia, F., Pan, W., Ju, Q., Chi, M., Huang, X. (2007). *From web directories to ontologies: Natural language processing challenges*. International Semantic Web Conference (ISWC).

Italia	Roman Republic and the Roman Empire between the 4th century BC and the 5th century AD

In the second step, each formula is completed by taking the relative position of each node in the classification. This is done by taking the conjunction (⊓) of all the formulas along the path, from the root to the node and by filtering out the senses which are not compatible each other, i.e. not related by relations in WordNet. This formula is called the *concept at node*. For instance, to determine the concept at node for node 2 in figure 1.5 we need to consider that for the words *location* and *Europe* the following meanings are provided in WordNet:

location#1: location	- a point or extent in space
location#2: placement, location, locating, position, positioning, emplacement	- the act of putting something in a certain place
location#3: localization, localisation, location, locating, fix	- a determination of the place where something is; "he got a good fix on the target"
location#4: location	- a workplace away from a studio at which some or all of a

	movie may be made; "they shot the film on location in Nevada"

Europe#1: Europe	- the 2nd smallest continent (actually a vast peninsula of Eurasia); the British use `Europe' to refer to all of the continent except the British Isles
Europe#2: European Union, EU, European Community, EC, European Economic Community, EEC, Common Market, Europe	- an international organization of European countries formed after World War II to reduce trade barriers and increase cooperation among its members; "he took Britain into Europe"
Europe#3: Europe	- the nations of the European continent collectively; "the Marshall Plan helped Europe recover from World War II"

Table 1 Wordnet synsets Location and Europe

It is important to note that in WordNet only the first and second meaning of *city* are related (through a chain of hypernym relations) to the first meaning of *location*, and that the first meaning of *Europe* is related (through part-meronym) to the only sense available for *Italy*, while all the other senses are unrelated. After that, the sense filtering the concept at node of node 2 is computed as (location#1 ⊓ Europe#1) ⊓ ((city#1 ⊔ city#2) ⊓ Italy#1). The lightweight ontologies generated from the classifications in figure 1.5. are provided in following figure 1.6.

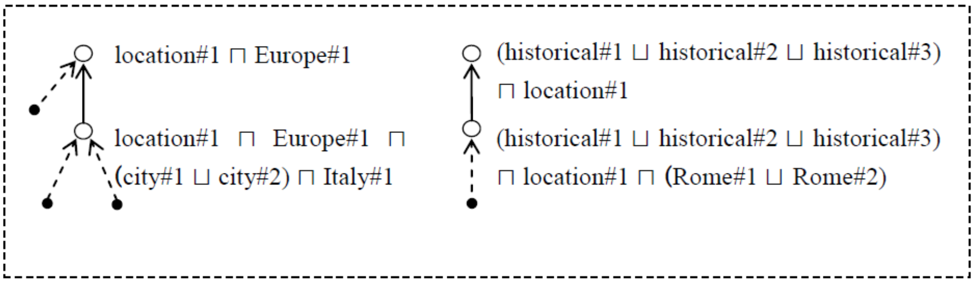

location#1 ⊓ Europe#1
location#1 ⊓ Europe#1 ⊓ (city#1 ⊔ city#2) ⊓ Italy#1
(historical#1 ⊔ historical#2 ⊔ historical#3) ⊓ location#1
(historical#1 ⊔ historical#2 ⊔ historical#3) ⊓ location#1 ⊓ (Rome#1 ⊔ Rome#2)

Figure 1.6 Lightweight ontologies generated from simple classifications

The level of accuracy in the translation process mostly depends on the accuracy of the NLP techniques used for the translation of the node labels into formal formulas.

As described in Giunchiglia and Zaihrayeu (2008)[34], lightweight ontologies can be used in many applications including document classification, semantic search, and matching of classifications, for instance for data integration. In all these applications, classifications are preliminary translated into lightweight ontologies:

[34] Giunchiglia, F., Zaihrayeu, I. (2008). *Lightweight ontologies*. Encyclopedia of Database Systems.

- **Document classification[35].** Document classification consists in assigning a document to one or more nodes in the classification based on the subject of a document, i.e. what the document is about. The basic idea is that each document is labelled with a formula in the formal language and is automatically classified by reasoning about subsumption on the nodes of the lightweight ontology. Note that this approach does not require the creation of a training dataset, which would normally be required in machine learning approaches.
- **Semantic search[36].** Semantic search, applied to classifications, is the problem of finding those documents in the classification, which correspond to a natural language query given in input. In brief, this problem can be solved by determining the concept corresponding to the query and by identifying, as answer to the query, those documents whose concept is more specific or equivalent to the concept of the query.
- **Semantic matching.** As a preliminary step towards integration and data coordination (interoperability in the broader sense) of heterogeneous repositories, semantic matching among classifications consists in identifying semantic relations among the nodes in the two schemas.

[35] This approach is proposed in Giunchiglia, F., Zaihrayeu, I., Kharkevich U. (2007). *Formalizing the get-specific document classification algorithm*. European Conference on Research and Advanced Technology for Digital Libraries.

[36] See Giunchiglia, F., Kharkevich, U., Zaihrayeu, I. (2009). *Concept search*. European Semantic Web Conference (ESWC).

1.5. Background Knowledge

The user of a lightweight ontology might be interested in the domain(s) his ontology belongs to. The knowledge base initially is built with the concepts imported from WordNet[37]. The domain specific knowledge a user is interested in for his/her own ontology, represented by a subset of a knowledge base, is called *background knowledge (BK).*

Background Knowledge can be modified by users and it is organized into two distinct parts: a *language-independent* and a *language-dependent* part[38].

In the language-independent part, knowledge is organized as a set of domains, each one is grouped into a set of facets, and each facet is made up of a hierarchy of a set of homogeneous concepts. Instances of concepts are named entities and are grouped into a set of entity types. Each concept can belong to a (possibly empty) set of domains. An entity type can correspond to a concept and a set of entities can be connected to a concept (i.e., its instances).

By contrast, in the language-dependent part, knowledge is organized as a list of words in a given language grouped into synsets. There are two kinds of synsets: concept synsets and entity synsets. Each concept synset is linked to a concept, but each concept may not have a synset representation in a human language. Similarly, each entity synset is connected to an entity, but an entity may not have a synset representation in a human language.

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

[38] F. Giunchiglia, B. Dutta, and V. Maltese. Faceted lightweight ontologies. In *Conceptual Modeling: Foundations and Applications: Essays in Honor of John Mylopoulos*, pages 36–51, Berlin, Heidelberg, 2009. Springer-Verlag.

In the Ontology part of the figure described above, location, country and city represent concepts. All the concepts in the ontology part are shown as circles and all the concepts in the Domain part are shown as dashed circles.

Links between the objects within a part are made of solid straight arrows and links across the parts are shown as dashed curved arrows. In the Entity part, *Italy* and *Trento* represent entities, and *Italy* is an instance of the concept *country*, *Trento* is an instance of the concept *city*, and the relation *part-of* connects the entities *Italy* and *Trento*.

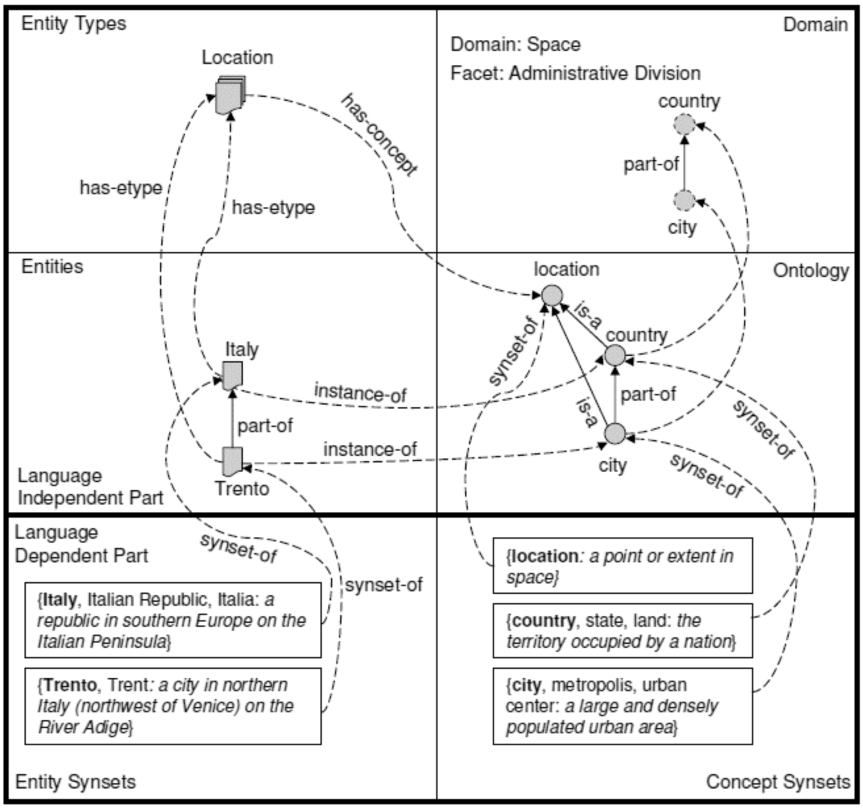

Entity Types
Location
has-etype
has-etype
has-concept
Domain
Domain: Space
Facet: Administrative Division
country
part-of
city
Entities
location
Ontology
is-a
country
synset-of
Italy
instance-of
part-of
part-of
is-a
synset-of
instance-of
city
synset-of
Language Independent Part
Trento
Language Dependent Part
synset-of
synset-of
{Italy, Italian Republic, Italia: a republic in southern Europe on the Italian Peninsula}
{Trento, Trent: a city in northern Italy (northwest of Venice) on the River Adige}
{location: a point or extent in space}
{country, state, land: the territory occupied by a nation}
{city, metropolis, urban center: a large and densely populated urban area}
Entity Synsets
Concept Synsets

Figure 1.7 Organization of background knowledge

1.6. Faceted Lightweight Ontologies

A *faceted lightweight ontology*[39] is a lightweight ontology in which the terms located in each node label, and their concepts, are available in the Background Knowledge, which is organized as a set of facets. Formally, it is defined as a quintuple $FLO = \langle LN, LE, LT, LC^{FL}, BK^{F} \rangle$, where LN is a finite (possibly empty) set of nodes, LE is a set of edges representing relations between nodes to form a rooted tree $\langle LN, LE \rangle$, LT is a set of terms, LC^{FL} is a finite set of concepts encoded in a formal language FL, such that for each term $lt_i \in LT$ there is one and only one concept $lc_i \in LC^{FL}$ and BK^{F} is background knowledge organized as a set of facets F such that $LT \in BK^{F}$ and $LC^{FL} \in BK^{F}$.

From this definition, it is easy to see that a faceted lightweight ontology is set up by a background knowledge and a lightweight ontology, where background knowledge plays the major role. Considering the figure 1.8, the term *fish* occurs in a node label in the hierarchy of a lightweight ontology, and this term represents an *aquatic vertebrate* if the background knowledge, attached to the lightweight ontology, is in the *animal* domain. On the other hand, when the background knowledge is in the *food* domain, the same term represents *the flesh of fish used as food*.

Therefore, by replacing the existing background knowledge with a new one selected from a different domain, we enable the same lightweight ontology to be

[39] Giunchiglia, F., Dutta, B., Maltese, V. (2009). *Faceted Lightweight Ontologies*. In: Conceptual Modeling: Foundations and Applications, A. Borgida, V. Chaudhri, P. Giorgini, Eric Yu (Eds.) LNCS 5600 Springer.

reused for another purpose. figure 1.8 shows that the use of a faceted lightweight ontology can be used for several purposes. For sake of simplicity, only the semantics of the lightweight ontology terms in different domains is provided, instead of the faceted background knowledge hierarchies.

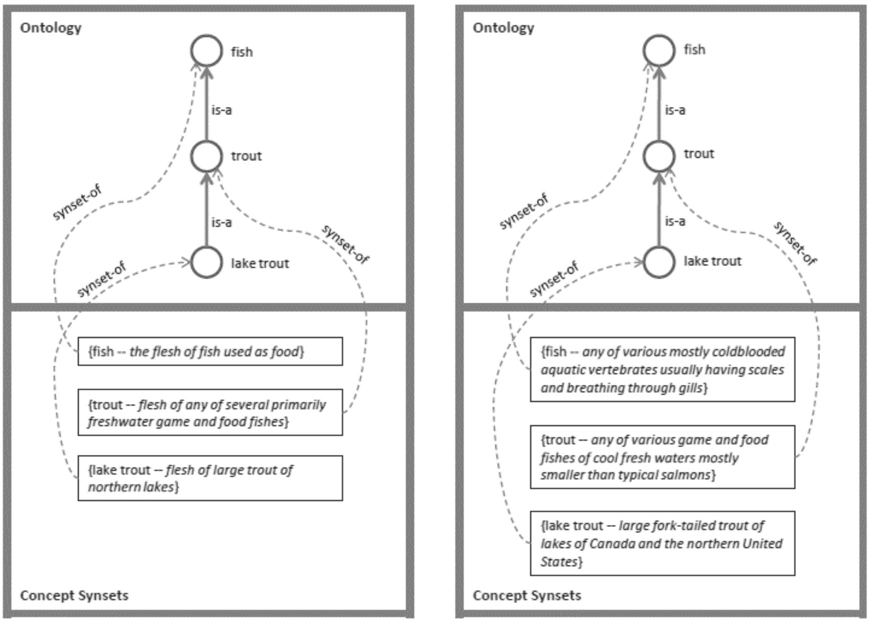

Ontology
fish
is-a
trout
is-a
lake trout
synset-of
synset-of
synset-of
{fish -- *the flesh of fish used as food*}
{trout -- *flesh of any of several primarily freshwater game and food fishes*}
{lake trout -- *flesh of large trout of northern lakes*}
Concept Synsets
Ontology
fish
is-a
trout
is-a
lake trout
synset-of
synset-of
synset-of
{fish -- *any of various mostly coldblooded aquatic vertebrates usually having scales and breathing through gills*}
{trout -- *any of various game and food fishes of cool fresh waters mostly smaller than typical salmons*}
{lake trout -- *large fork-tailed trout of lakes of Canada and the northern United States*}
Concept Synsets

Figure 1.8 An example of a faceted lightweight ontology: food domain and animal domain.

1.7. Conclusions

In this chapter we have provided a brief description of the classification schemes and ontologies, and have presented a comparison between them.

We have described lightweight ontologies, their applications, and the problems involved in their applications. We have proposed faceted lightweight ontologies as a solution to overcome these limitations.

Chapter 2

Approaches to knowledge representation

Constructing and modeling knowledge bases was the aim of many works in the last thirty years. As described before, an unavoidable and intrinsic property of the world that these bases would like to represent is diversity. To make diversity detectable and exploitable, favors interoperability and allows people to understand as well as a machine it is essential to make the meaning of the words explicit in a certain context (i.e. their semantics), so the information becomes unambiguous. Aiming at this goal, the preliminary step is the creation of a *diversity-aware knowledge base*: in order to fashion a knowledge base, and for representing, constructing and maintaining it, developing appropriate methodologies is necessary. A knowledge base can be seen as a collection of facts encoding knowledge of the real world that can be used to automate tasks. To be useful, a knowledge base should be very large, virtually unbound and able to capture the diversity of the world and, at the same time, to reduce the complexity of reasoning at run-time. At this purpose, as proposed by many studies, the notions of *domain* (as originated from library science) and *context* (as originated from Artificial Intelligence) have been indicated as essential for diversity-aware knowledge bases.

Domains have two important properties. They are the main tools in capturing diversity, in terms of language, knowledge and personal experience. For instance, according to the personal perception and purpose, the space domain may or may not include buildings and man-made structures; the food domain may or may not include dogs according to the local customs. Moreover, domains allow scaling up, so it is possible to add new knowledge to them at any time as needed.

Determining the context allows on the one hand a better disambiguation of the terms used, because it makes explicit some of the assumptions left implicit, and on the other hand allows the reduction of complexity of reasoning at run-time, since it selects from the domains the language and knowledge that are strictly necessary to solve the problem. It is important to note that diversity was also formalized in terms of diversity dimensions, i.e. the dimensions by which knowledge is framed. In library science topic, space and time are known to be the three fundamental diversity dimensions.

2.1. Knowledge bases types

A crucial point is that only a few existing knowledge bases can be considered diversity-aware[40]. Analyzing existing knowledge bases, they can be divided into two main broad categories: (a) automatically built and (b) hand-crafted knowledge bases.

a) Automatically built knowledge base

For automatic extraction of knowledge from freetext, tools like *KnowItAll*[41] and *TextRunner*[42] are the most known among the projects. However, these techniques typically achieve very low accuracy. For this reason, projects like *DBPedia*[43], *YAGO*[44] and BabelNet[45], which extract information from semi-structured knowledge sources (mainly Wikipedia infoboxes and categories), obtain more accurate results. While these systems generally lack in explicit quality control systems and semantics, BabelNet provides knowledge-based

[40] DENDRAL is widely considered the first expert system ever created embedding a knowledge base with domain specific knowledge (organic chemistry). See B. G. Buchanan, J. Lederberg, *The Heuristic DENDRAL program for explaining empirical data*, Stanford University, technical report (1971).

[41] O. Etzioni, M. J. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates, *Web-scale information extraction in KnowItAll*, WWW conference (2004).

[42] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, *Open information extraction from the web*, IJCAI conference (2007).

[43] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, C. Cyganiak, Z. Ives, *DBpedia: A Nucleus for a Web of Open Data*, 6th International Semantic Web Conference ISWC (2007).

[44] F. M. Suchanek, G. Kasneci, G. Weikum, *YAGO: A Large Ontology from Wikipedia and WordNet*, Journal of Web Semantics (2011).

[45] R. Navigli and S. Ponzetto. *BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network*. Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250.

algorithms to guarantee the high quality and in YAGO there is an explicit quality control mechanism mainly based on a unique entity reference system for which there cannot be two entities with the same name[46].

a) Manually built knowledge base

Among hand-crafted resources it is worth mentioning CYC[47] that is a general-purpose common sense knowledge base containing around 2.2 million assertions and more than 250,000 terms about the real world. Its open source version OpenCYC[48] contains 306,000 assertions and 47,000 terms. The content of CYC is organized according to the generality principle[49], and is split into three levels of abstraction, from broader and abstract knowledge (the upper ontology) and widely used knowledge (the middle ontology) to domain specific knowledge (the lower ontology).

SUMO (Suggested Upper Merged Ontology)[50] is a free formal ontology similar to CYC, made up of about 1,000 terms and 4,000 definitional statements. A SUMO extension, called MILO[51] (MId-Level Ontology) also exists, which covers individual domains, comprising overall 21,000 terms mapped with WordNet and 73,000 axioms. Both SUMO and MILO are therefore quite small.

[46] In YAGO there is a precise knowledge representation model based on RDFS

[47] C. Matuszek, J. Cabral, M. Witbrock, J. DeOliveira, *An introduction to the syntax and content of Cyc*, AAAI Spring Symposium (2006).

[48] http://www.opencyc.org/

[49] J. McCarthy, Generality in artificial intelligence, Communications of ACM 30 (1987), 1030–1035.

[50] A. Pease, G. Sutcliffe, N. Siegel, S. Trac, *Large theory reasoning with SUMO at CASC*, AI Communications, 23 2-3 (2010) 137–144.

[51]

2.1.1. Wordnet as knowledge base

CYC, SUMO and their extensions are built without targeting any particular range of reasoning tasks, and in DBPedia and in YAGO there is not an explicit notion of domain, but in both the entities include what is further differentiated into entities, classes, qualities and values. Everything is codified in terms of generic facts between entities (triples of the form source-relation-target). In CYC there is a notion of domain, but it is used only to partition knowledge into easier to manage components. Moreover, in CYC too there is a generic notion of entity.

Even if not specifically developed for supporting reasoning tasks, WordNet - as demonstrated by the thousands of citations - is the most widely used linguistic resource nowadays, because it is manually constructed and it exhibits a significant quality and size. For this reason, it is also frequently adapted for semantic applications. Anyway, even if it is not tailored for any particular domain, it is often considered too fine grained to be really useful in practice[52].

Wordnets has different multilingual extensions; there are at least two main approach to build a multilingual wordnet. The model proposed in EuroWordNet[53] project is based on the construction of different language specific wordnets independently from each other, trying in a second phase to find correspondences

[52] R. Mihalcea, D. I. Moldovan, *Automatic generation of a coarse grained WordNet*, NAACL Workshop on WordNet and Other Lexical Resources (2001).

[53] P. Vossen, *Categories and classifications in EuroWordNet*, Proceedings of the First International Conference on Language Resources and Evaluation. Granada, 399-407, 1998; *Special Issue on EuroWordNet*, Computer and the humanities, 2-3, 73-251, 1998.

between them. Another approach is the the model adopted in MultiWordNet[54], which consists of building language specific wordnets keeping as much as possible of the semantic relations available in the English WordNet, by building the new synsets in correspondence with the Wordnet synsets, whenever possible, and importing semantic relations from the corresponding English synset. Anyway, in digital library communities it is possible to find other valuable resources, especially domain specific knowledge encoded in informal or semi-formal knowledge organization systems such as subject headings and thesauri[55].

Hand-crafted resources are surely more accurate but very difficult to construct and maintain, to alleviate this problem are born some recent projects like *Freebase*[56] that follow a collaborative approach that relies on volunteers to fill the knowledge base. This approach is the main drawback of Freebase, which does not guarantee consistency in the use of the terminology, leaving its users *free* to independently define their axioms without effective mechanisms for quality control.

[54] Pianta et al., *MultiWordNet: developing an aligned multilingual database*, Proceedings of the First International Conference on Global WordNet, Mysore, India, 2002.

[55] For instance,about agriculture we can mention AGROVOC20 and NALT21; about medicine the most widely known is UMLS. In general, their main drawback is the lack of an explicit semantics.

[56] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, *Freebase: a collaboratively created graph database for structuring human knowledge*, ACM SIGMOD international conference on Management of data (2008), 1247-1250.

2.2. Diversity in knowledge

As seen in previous chapter, the main problem in knowledge representation and semantic tasks is the lack of background knowledge, defined as the a-priory knowledge necessary to make semantics effective. In fact, as several studies have have demonstated[57] that it is impossible to achieve accurate enough results without high quality and contextually relevant background knowledge, which is not easy to obtain. In fact, the background knowledge should be theoretically very large and virtually unbound, in order to provide all the possible meanings of the words and how they are related to each other. At the same time, the background knowledge should be context sensitive, capturing the diversity of the world, which is exhibit in the inherent ambiguity of language.

In fact, it is possible to refer to the same real world object in different words in different communities and in different languages. Otherwise, the same word may denote different notions in different domains; for instance, *bug*, which depicts an insect in entomology and a failure in a computer program in computer science. Many other elements contribute in characterizing the meaning of a word, as space, time, needs, culture, opinions and personal experience also.

[57] See: F. Giunchiglia, P. Shvaiko, M. Yatskevich, *Discovering missing background knowledge in ontology matching*, European Conference on Artificial Intelligence ECAI (2006), 382–386; B. Lauser, G. Johannsen, C. Caracciolo, J. Keizer, W. R.van Hage, P. Mayr, *Comparing human and automatic thesaurus mapping approaches in the agricultural domain*, International Conference on Dublin Core and Metadata Applications (2008); P. Shvaiko, J. Euzenat. *Ten Challenges for Ontology Matching*, 7th Int. Conference on Ontologies, Databases, and Applications of Semantics, ODBASE, (2008); Z. Aleksovski, W. ten Kate, F. van Harmelen, *Using multiple ontologies as background knowledge in ontology matching*, ESWC workshop on collective semantics (2008); B. Magnini, M. Speranza, C. Girardi, *A semantic-based approach to interoperability of classification hierarchies: Evaluation of linguistic techniques*, COLING (2004).

Diversity/ambiguity is an intrinsic property, that aims at minimizing the effort and maximizing the gain[58].

According to Giunchiglia et al. (2012)[59], diversity can be considered emerging at least along three main dimensions:

- ***Diversity in natural language***: terms may denote classes (common nouns), entities (proper nouns), properties, qualities and other modifiers (adjectives and adverbs); different terms can be used to denote the same notion (synonymy); the same term may denote different things (polysemy). For instance, e.g. the term *bank* in the first classification of figure 2.1 may mean a sloping land or a financial institution. At the entity level, *Rome* the capital of Italy is also known as the *Eternal City*; there might be different places in the world (and in general different entities) called *Rome*;
- ***Diversity in formal language***: when disambiguated, each term corresponds to a concept written in some formal language. Different classifications, according to their specific scope and purpose, may use different formal languages.
- ***Diversity in knowledge***: in this level, the relations between concepts are recognized. The amount of knowledge, in terms of axioms, necessary for a

[58] F. Giunchiglia, *Managing Diversity in Knowledge*, Invited Talk at the European Conference on Artificial Intelligence ECAI, Lecture Notes in Artificial Intelligence 2006.

[59] F. Giunchglia, V. Maltese, B. Dutta, *Domains and context: First steos towards managing diversity in knowledge*, Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web (2012).

certain task is also a function of the local goals, culture, opinions and personal experience. For instance, while dogs are mainly perceived as pets, they are regularly served as food in China (culture); while someone may consider beautiful the city of Rome in Italy, somebody else may consider it too chaotic (opinion); somebody may consider climate change an urgent problem to be solved, while somebody else may even negate its existence (school of thought).

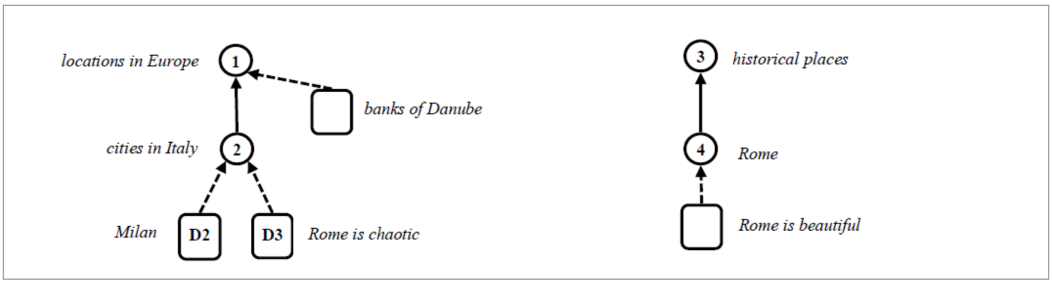


Figure 2.1 two examples of classification

The intrinsic ambiguity of natural language is a critical issue: for that identifying resources that provide the background knowledge relevant for the disambiguation[60] it is fundamental. However, most of the time the meaning of the words and the context of use is left implicit. For this reason, implicit knowledge[61] is relevant and necessary in understanding and disambiguation process. It is also

[60] P. Shvaiko, J. Euzenat. *Ten Challenges for Ontology Matching*, 7th Int. Conference on Ontologies, Databases, and Applications of Semantics, ODBASE, (2008); Z. Aleksovski et al. *Using multiple ontologies as background knowledge in ontology matching*, ESWC workshop on collective semantics (2008).

[61] F. Giunchiglia, *Contextual reasoning*, Epistemologica - Special Issue on I Linguaggi e le Macchine, 16 (1993), 345–364.

important to note that the amount of implicit knowledge is potentially infinite; therefore it is quite impossible to completely determine them, a considerable portion of knowledge remains in the human minds[62].

2.3. Domain and context

A recent approach63 proposed to take into account this diversity and exploit it to make explicit the local semantics, i.e. the meaning of words in a certain context, such that information becomes unambiguous to both humans and machines. Towards this goal, a preliminary step is the creation of a diversity-aware knowledge base, which requires appropriate methodologies for its representation, construction and maintenance.

This approach is centered on the fundamental notions of domain and context. As already described, domains capture diversity in terms of language, knowledge and personal experience and allow the addition of new knowledge at any time. Context allows the disambiguation of the terms used (i.e. by making explicit some of implicit assumptions) and it can reduce the complexity of reasoning, by selecting from the domains the language and the knowledge. In the

[62] L. Prusak, *Knowledge in Organizations*, Cap. 7: The tacit dimension by M. Polanyi, 1997.

[63]

approach proposed by Giunchiglia et al. (2012)64, this problem solution can be summarized into three subsequent steps:

1. Develop an extensible diversity-aware knowledge base explicitly codifying the differences in language, natural and formal, and knowledge in multiple *domains*;
2. Given the specific problem, build the corresponding *context* as a formal local theory by determining from the knowledge base the implicit assumptions which are relevant to understand it and building the corresponding context as a logical
3. Solve the problem in context.

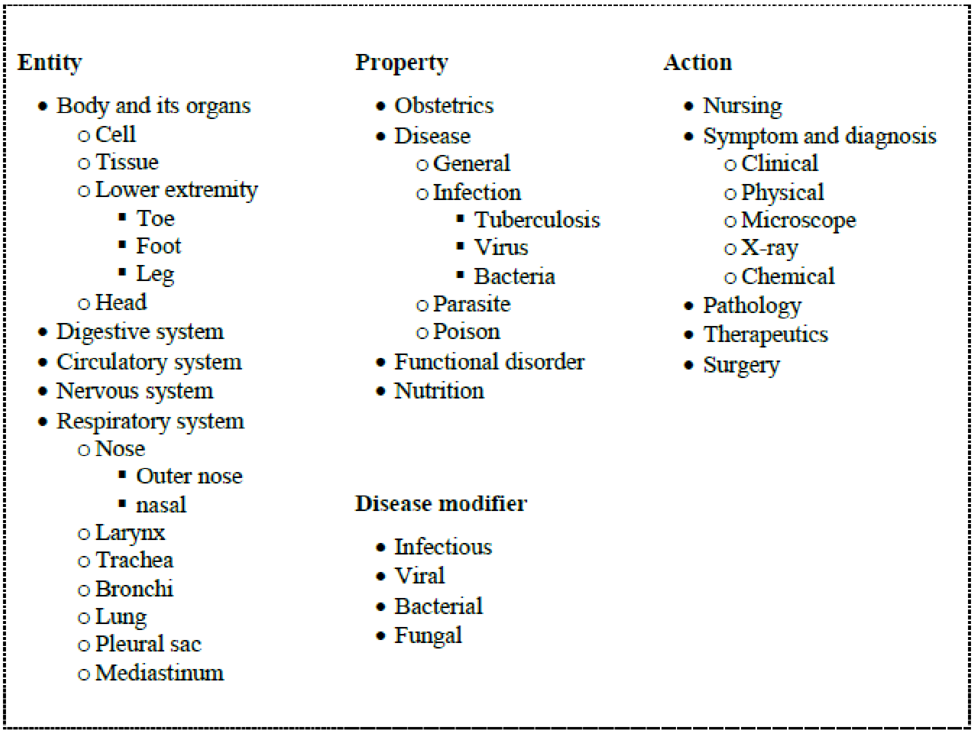
For this aim, the proposed method adapt the *faceted approach*, a library science methodology mainly used for the organization of knowledge in libraries[65]. The fundamental notion of the *faceted approach* is the concept of *domain* and its components, called *facets*, which allow both capturing diversity and an incremental growth of the knowledge base.

[64] F. Giunchglia, V. Maltese, B. Dutta, *Domains and context: First steos towards managing diversity in knowledge*, Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web (2012).

[65] S. R. Ranganathan, *Prolegomena to library classification*, Asia Publishing House (1967)

2.4. Faceted approach

The Indian librarian Ranganathan was the first one to propose the theory of *faceted analysis*, as the fundamental methodology that guides in the creation of a faceted classification for a domain[66]. The first faceted classification scheme, named *Colon Classification*, was developed in the late 1930's. There are five main fundamental categories proposed: *Personality*, *Matter*, *Energy*, *Space* and *Time*, plus facets of general applicability called *common isolates* or *modifiers* (e.g. Language and document Form). Figure 2.2 provides a small example for the medicine domain.



Entity

- Body and its organs
  - Cell
  - Tissue
  - Lower extremity
    - Toe
    - Foot
    - Leg
  - Head
- Digestive system
- Circulatory system
- Nervous system
- Respiratory system
  - Nose
    - Outer nose
    - nasal
  - Larynx
  - Trachea
  - Bronchi
  - Lung
  - Pleural sac
  - Mediastinum

Property

- Obstetrics
- Disease
  - General
  - Infection
    - Tuberculosis
    - Virus
    - Bacteria
  - Parasite
  - Poison
- Functional disorder
- Nutrition

Disease modifier

- Infectious
- Viral
- Bacterial
- Fungal

Action

- Nursing
- Symptom and diagnosis
  - Clinical
  - Physical
  - Microscope
  - X-ray
  - Chemical
- Pathology
- Therapeutics
- Surgery

Figure 2.2. An example of the medicine domain taken from the Colon Classification

[66] V. Broughton, *The need for a faceted classification as the basis of all methods of information retrieval*, Aslib Proceedings 58 1/2 (2006), 49-72.

According to the *analytico-synthetic* approach proposed by Ranganathan[67], facets for a given domain are defined following two steps:

- **Analysis**. In this step relevant terms of the domain are identified and gained by consulting domain experts and all sorts of information sources about the domain. This process starts in the so-called *idea plane*, the language independent conceptual level, where atomic 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;

- **Synthesis.** In this step the identified terms (also called *isolate ideas*) are grouped into *facets*, according to their common properties or characteristics, and they are ordered in hierarchies. The set of homogenous terms form a *facet*, for example, in the medicine domain described in figure 2.2, the facet called *Respiratory system* is made up of the terms *Nose, Larynx, Trachea, Bronchi, Lung, Pleural sac, Mediastinum, which*

[67] S. R. Ranganathan, *The Colon Classification, Rutgers Series on Systems for the Intellectual Organization of Information*, S. Artandi (etd.), IV, Graduate School of Library Science, Rutgers University, New Brunswick, NJ, 1965.

are entities in the *part-of* relation with *Respiratory system*. Moreover, the facet *Respiratory system* has a *sub-facet* called *Nose*, composed by terms *Outer nose* and *Nasal*.

These two steps bulid a *faceted representation scheme* and correspond to the so-called *background knowledge*, namely the a-priori knowledge, which must exist in order to make semantics effective. Notice that the grouped terms of step 2 are formed using *part-of* and *instance-of* relations, so they can be considered descriptive ontologies. Created facets are organized in a set of independent *domains* and, for each domain; they are grouped into specific elementary *categories*[68].

2.4.1. Facets properties

As described in Giunchiglia et al. (2009)[69], facets possess the essential properties listed below:

- **Hospitability**. They are easily extensible. It is possible to accommodate without difficulty in the hierarchical structure new terms representing new

[68] 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).

[69] Giunchiglia, F., Dutta, B., Maltese, V. (2009). *Faceted Lightweight Ontologies*. In: Conceptual Modeling: Foundations and Applications, A. Borgida, V. Chaudhri, P. Giorgini, Eric Yu (Eds.) LNCS 5600 Springer.

knowledge. Terms in the hierarchies are clearly defined, mutually exclusive and collectively exhaustive;

- **Compactness**. Facet-based systems need less space to classify the universe of knowledge with respect to the other hierarchical knowledge organization systems. There is no explosion of the possible combinations because of the isolation in which the basic elements (facets) are taken.
- **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).

Even if underestimated for years, facet analysis is at the basis of modern classification systems, such as the second edition of the Bliss Bibliographic

Classification (BC2)[70] and projects like FAKTS[71], a project which attempts to provide facets useful in online environments by reorganizing BC2 and UDC auxiliary tables. Moreover, it is more and more used in several other traditional classification systems for the definition of facets of general use as an add-on to the standard classification schemes. As shown by Broughton (2008)[72] facet classification is often used as a guideline for the generation of thesauri since it helps in the identification of terms and relationships between them.

Anyway, some researchers argue that faceted classifications are not a particular kind of library classification, but rather the only viable form enabling the locating and relating of information to be optimally predictable[73].

2.4.2. Managing diversity using faceted approach

The approach proposed by *facet classification* and its key notion of facets, that allow domain knowledge modeling by exploiting and making explicit the different aspects of knowledge within the domain, can be followed and adapted to define the data model underlying the creation of the diversity-aware knowledge base.

[70] http://www.blissclassification.org.uk/

[71] Broughton, V., Slavic, A. (2007). *Building a faceted classification for the humanities: principles and procedures*. J Doc 63(5), 727–754.

[72] Broughton, V. (2008). *A Faceted Classification as the Basis of a Faceted Terminology: Conversion of a Classified Structure to Thesaurus Format in the Bliss Bibliographic Classification*, 2nd Edition. Axiomathes Journal, Springer Online Issue, 18 (2), 193-210.

[73] Mills, J. (2004). *Faceted classification and logical division in information retrieval*, Library trends, 52 (3), 541-570.

Faced based systems proved their usefulness and effectiveness in organizing and searching documents in conventional library systems, however, as emphasized by many studies, the major drawback of these systems lies in their structure. All these systems only consider the syntactic form in which subjects in natural language (syntax) are described, so they fail to make explicit the way the meaning (semantics) of subjects (what the document is about) is built starting from the semantics of their constituents. Consequently, it is not possible to perform a direct translation of their elements into a formal language. They do not explicitly specify the relations constituting the facets, the taxonomical *is-a* and *instance-of* (genus/species) and mereological *part-of* (whole/part) relations between the classes, thus limiting their applicability. Therefore, making them explicit is a fundamental step towards automation and interoperability.

To overcome these limitations it was proposed to define facets as descriptive ontologies. Following this approach, a *domain* can be defined as composed by:

- *classes* of real world objects
- *entities*, that represent the instances of the classes;
- *relations* between entities and classes that provide structure to the domain. They include *is-a, instance-of, part-of* relations and other additional relations according to the scope of the ontology;
- qualitative, quantitative and descriptive *attributes*

Each domain, for instance the Space domain reported in figure 2.3, is organized in three levels:

- ***Formal language level***: it provides the terms used to denote the elements of the domain. Terms that denotes classes (e.g. *lake*, *river* and *city*), entities (e.g. *Garda lake*), name of a relation (e.g. *direction*) and attribute name (e.g. *depth*) or value (e.g. *deep*) are called *formal terms*, indicating their independence from language and that they have a precise meaning and role in (logical) semantics. These elements are arranged into facets using *is-a*, *part-of* and *value-of* relations.
- ***Knowledge level***: it codifies what is known about the entities in terms of attributes (e.g. *Garda lake* is *deep*), the relations between them (e.g. *Tiber* is part of *Rome*) and with corresponding classes (e.g. *Tiber* is an instance of *river*). The knowledge level is codified using the formal language described in the item above and is, therefore, also language independent;
- ***Natural language level***: it defines set of words (*natural language terms*) such that words with same meaning within each natural language are grouped together and mapped to the same formal term. This level can be instantiated to multiple languages.

This methodology is similar to WordNet and follows the same terminology, so words are disambiguated by providing their meaning, also called *sense*. It is possible to describe the meaning of each word labelling it with a natural language description. Synonymous words are grouped into a *synset*. For instance, since *stream* and *watercourse* have the same meaning in English, they are part of the same synset. Similarly, homonymous words belong to different synsets. In this data model, within a language each synset is associated to a set of words (the synonyms), a natural language description, a part of speech (noun, adjective, verb or adverb) and a corresponding formal term.

Notice that the knowledge level correspond to what in previous chapter was called *background knowledge*[74], i.e. the a-priori knowledge that must exist to make semantics effective. Each facet corresponds to a *lightweight ontology*, and plays a fundamental role in task automation. The natural language level provides instead an interface to humans and can be exploited for instance in Natural Language Processing (NLP).

[74] Giunchiglia, F., Shvaiko, P., Yatskevich, M. (2006). *Discovering missing background knowledge in ontology matching*. ECAI conference, 382–386.

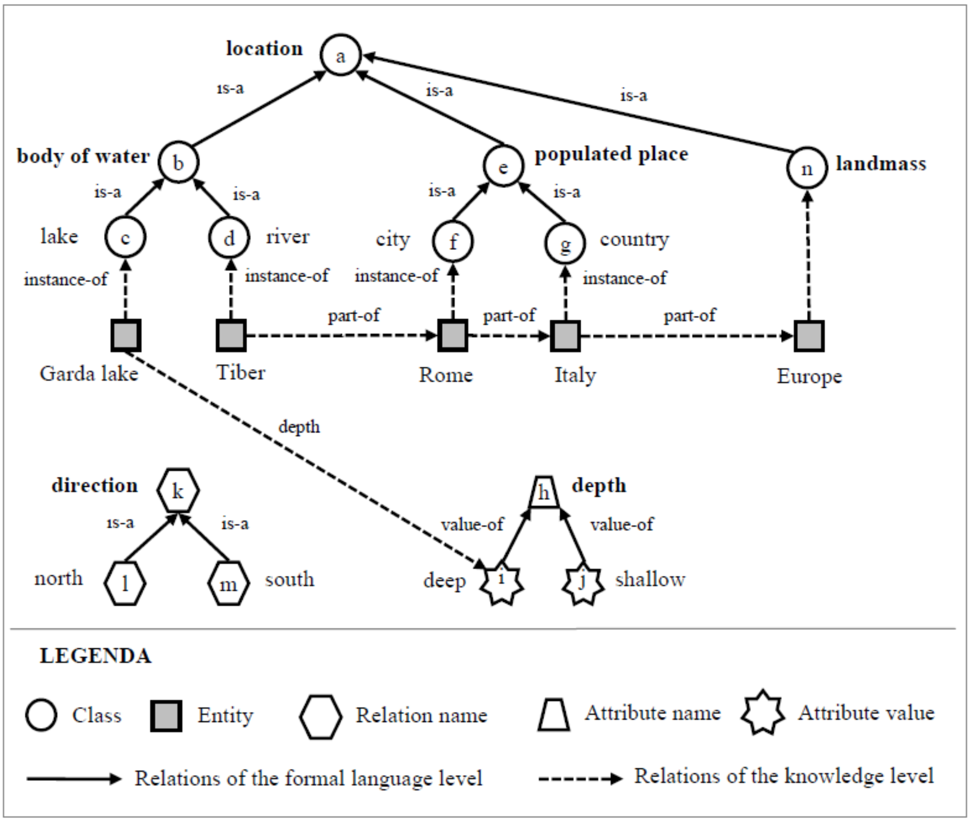


Figure 2.3. a small fragment of Space domain

As an example, Fig. 2.3 (taken from Maltese and Farazi 2011) provides a small fragment of the *Space* domain following the proposed data model: classes are represented with circles, entities with squares, relation names with hexagons, attribute names with trapezoids and attribute values with stars. Letters inside the nodes (capital letters for entities and small letters for classes, relations and attributes) denote formal terms, while corresponding natural language terms are provided as labels of the nodes (in the figure synonyms are not represented). Arrows denote relations between the elements: solid arrows represent those relations constituting the facets (*is-a*, *part-of* and *value-of* relations) and which are part of the formal language level; dashed arrows represent *instance-of*, *part-of* and the other relations (*depth* in this case) which are part of the knowledge level. Here the hierarchies rooted in *body of water*, *populated place* and *landmass* are facets of entity classes and are subdivisions of *location*, the one rooted in *direction* is a facet of relations and the one rooted in *depth* is a facet of attributes.

2.4.3. Methodology

Following these statements, the process to build a faceted ontology is organized in five subsequent phases:

- **Step 1: Identification of the terminology.** It consists in collecting and classifying the natural language terms. In general, in the faceted approach this is mainly done by interviewing domain experts and by reading available literature about the domain under examination including indexes, abstracts, glossaries, reference works. In this approach each natural language term is analyzed and disambiguated by reconstructing the corresponding sense, by grouping into synsets those with the same meaning, and by associating each synset to a formal term. Each formal term is then classified as a class, entity, relation or attribute (name or value). For instance, in the construction of *Space*, this step consists in the selection of the resources that allow identifying the natural language terms representing the geospatial classes, the entities, the relations, the attributes and their disambiguation into formal terms. Best resources for Space specific terminology are identified in *Thesaurus of Geographical Names* (TGN)[75] and *GeoNames*[76]. GeoNames was used as main source, TGN instead, being a thesaurus, was used for consultation to better disambiguate GeoNames classes and relations.

- **Step 2: Analysis.** The formal terms collected during the previous phase are analyzed per *genus et differentia*, i.e. in order to identify their commonalities and their differences. Analysis have as aim to identify as many characteristics as possible of the real world entities represented by each of the terms. Doing this, the result would be as fine grained as wanted in differentiating among them. For instance, for the term *river*, defined as *"a large natural stream of water (larger than a brook)"*, following characteristics can be identified: a body of water; a flowing body of water; no fixed boundary; confined within a bed and stream banks; larger than a brook.
- **Step 3: Synthesis.** In this step, formal terms are arranged into facets. This is done by referring to their lexicalization in a language, e.g. to the corresponding synsets, and according to the characteristics identified with the previous phase. Grouping the terms into arrays by a common characteristic progressively form the levels of the facet hierarchies. For

[75] TGN is a poly-hierarchical (i.e., multiple parents are allowed) structured vocabulary containing 688 classes and around 1.1 million place names.
http://www.getty.edu/research/conducting_research/vocabularies/tgn.

[76] GeoNames provides 8million place names in various languages amounting to 7 million unique places and corresponding attributes such as latitude, longitude, altitude and population.

instance, in the *Space* ontology, considering the list of characteristics selected with the analysis it is possible to create different categories. Based on analyzed characteristics *stream* and *river* can be grouped in the same category *flowing body of water,* and a further analysis suggests the creation of a more general facet, called *body of water*. Note that *river* is a natural stream, and therefore a special kind of stream. In particular, this means that all the properties of *stream* are inherited by *river* (but not the vice versa). This is reflected in the facet hierarchy by putting *river* under *stream*.

- **Step 4: Standardization.** For each formal term in a facet, a standard (or preferred) term should be selected among the natural language terms associated to the corresponding synset. In the faceted approach this is usually done by identifying the term which is most commonly used in the domain and which minimizes the ambiguity. This is similar to the WordNet approach where words are ranked in the synset, and the first word is the preferred one. For instance, for the synsets created with the words from GeoNames, original terms were changed based on standard vocabularies. For instance, *mountain range* (geology terminology) was substituted with *mountains* (more general) or *submarine hill* (oceanography terminology) was changed in *hill* (that includes undersea entities).

- **Step 5: Ordering.** Formal terms in each array are ordered following many possible criteria, e.g., by chronological order, by spatial order, by increasing and decreasing quantity (for instance by size), by increasing complexity, by canonical order, by literary warrant and by alphabetical order. The criteria should be based upon the purpose, scope and subject of the ontology; in fact, it is not always possible to establish an order, especially when the classes do not share any characteristic (e.g. body of water and landform). In these cases it is used the *canonical order*, i.e., the order traditionally followed in library science.

2.5. The Entitypedia project to knowledge representation

Entitypedia[77] is a framework "developed to build a diversity-aware knowledge base with an initial set of domains and extensible according to the local scope, purpose, language and personal experience". Entitypedia was developed following the approach and data model presented in previous sections. Following the domain-centric data model, Entitypedia is centered on concepts of domain and context.

Entitypedia is totally modular, and allows therefore plugging an arbitrary number of domains; classes (concepts), entities (their instances) and their relations and attributes are clearly defined, while different vocabularies in different

[77] http://entitypedia.org/

languages (initially English and Italian), that are nearly distinguished from the formal language used in task automation, are provided.

2.5.1. DERA and faceted approach

A new methodology called DERA has been proposed[78] for the construction of the domain knowledge. The DERA framework is entity oriented and its aim is to develop domains to be used for automation. Its development make a point on the real world entity representations in mind, including inter-alia locations, people, organizations, songs, movies, which are relevant to a given domain. Entitypedia is based on and adapted from the faceted approach[79], which represents an effective methodology for domain construction and maintenance. Decades of research in library science proves that the use of the principles at the basis of the faceted approach allows the creation of better quality domain ontologies (in terms of robustness, extensibility, reusability, compactness and flexibility) and make them easier to maintain[80]. By using the DERA methodology Entitypedia has been incrementally filled up with domain knowledge starting with space[81]. By taking

[78] F. Giunchiglia, B. Dutta, V. Maltese, *From knowledge organization to knowledge representation*, DISI Technical Report (2013).

[79] Ranganathan, S. R. (1967). *Prolegomena to library classification*, Asia Publishing House.

[80] V. Broughton, *The need for a faceted classification as the basis of all methods of information retrieval*, Aslib Proceedings 58 1/2 (2006), 49-72; V. Broughton, Building a Faceted Classification for the Humanities: Principles and Procedures, Journal of Documentation (2007); L. Spiteri, A Simplified Model for Facet Analysis, Journal of Information and Library Science 23 (1998), 1-30.

[81] Giunchiglia, F., Maltese, V., Dutta, B. (2012). *Domains and context: first steps towards managing diversity in knowledge*. Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web. DOI: 10.1016/j.websem.2011.11.007

GeoNames, WordNet and MultiWordNet as main sources, the work on space led to the creation of GeoWordNet[82] a very large open source geo-spatial ontology containing overall more than 1000 classes, 7 million entities, 70 different kinds of relations and 35 kinds of attributes. Region, administrative division, populated place, facility, abandoned facility, land, landform, body of water, agricultural land and wetland, are some of the facets emerging from classes. However, to enabling diversity-aware applications more suitable for a web application is fundamental to identify those domains, which are more likely to play a role in everyday life and in particular on the Web.

2.5.2. Building Entitypedia

To incrementally populate Entitypedia is following described the general strategy.

Step I: bootstrapping the knowledge base.

Entitypedia was initially made up with general terminology imported from WordNet 2.1 and the Italian section of MultiWordNet. This essentially provided the requirement for bootstrapping the natural language level.

[82] F. Giunchiglia, V. Maltese, F. Farazi, B. Dutta, *GeoWordNet: a resource for geo-spatial applications*, Extended Semantic Web Conference ESWC (2010); B. Dutta, F. Giunchiglia, V. Maltese, A facet-based methodology for geo-spatial modelling, GEOS (2011).

Words, synsets and lexical relations between them have been imported from WordNet and MultiWordNet to the natural language part of the knowledge base, instantiated for the English and Italian language, respectively[83]. It is worthy to note that WordNet instances/entities were not imported since they are not a significant number and no attributes are provided for them and because a huge quantities of entities and corresponding metadata were imported from other resources. Note that the official number of entities in WordNet is 7671[84], while 683 of them are common nouns instead. The identification of the wrong ones occurred by manually verifying those with no uppercased lemma, hereafter they were converted into noun synsets, while the other 6988 were considered still entities.

Figures are provided in Table 2.1. Excluding the 6988 entities and corresponding relations, WordNet was completely imported. MultiWordNet instead was only partially imported. In particular, 92.47% of the words, 94.28% of the senses and 94.30% of the synsets were imported, while The 318 Italian lexical and semantic relations provided were not imported.

[83] These two languages were selected because of the importance that the English and Italian languages have respectively in the context of the Living Knowledge (http://livingknowledge-project.eu) and the Live Memories (http://www.livememories.org) projects we are involved in.

[84] Miller, G. A., Hristea, F. (2006). *WordNet Nouns: classes and instances*. Computational Linguistics, 32(1), 1 – 3.

Object	Quantity
Natural language part	
English synsets	110,609
English words	147,252
Italian synsets	33,356
Italian words	45,156
Formal language part	
Classes, qualities and values	110,609
Entities	~9.5 millions
Domains	2 (*Spale* and *Time*)
Classes, qualities and values in the domains	>1000 (*Space*) >200 (*Time*)
Knowledge part	
Is-a and *part-of* relations	204,481
Instance-of relations	~9.5 millions

Table 2.1. statistics about the current size of the knowledge base Entitypedia

For each synset in the two languages, a language-independent concept was created at formal language level. If it is possible to express the same notion in the two languages, then corresponding synsets are linked to the same concept. Because of the partial coverage of the language in MultiWordNet and the well-known problem of gaps in languages (i.e. given a lexical unit in a language, it is not always possible to identify an equivalent lexical unit in another language) not all concepts have a corresponding synset in Italian. *Hypernym* (is-a) and transitive *part meronym* (part-of) relations were elected as semantic hierarchical relations (corresponding to subsumption under classification semantics). All the other relations were defined as generic associative relations.

Step II: building the Space and Time domains.

First domains used to start populating the knowledge base are *Space* and *Time* domains. To construct the *Space* domain was used a semi-automatic approach. Domain specific terms were extracted mainly from GeoNames[85] and TGN[86], WordNet and some scientific literature about geography and its related areas. These terms were analyzed, organized into facets and mapped with the concepts created in the previous phase. The *analysis* aim to enlist the characteristics of division the use of which is necessary to form the facets. In other words, these features were used to form the different levels of abstraction of the conceptual categories. The concepts were analyzed using the topological, geometric or geographical characteristics of corresponding entities. The main principle followed in this determination phase is exhaustiveness. Exhaustiveness may allow the formation a huge number of very fine grained groups of concepts. On the other hand, the purpose of *synthesis* is to arrange the concepts into facets by

[85] http://www.geonames.org/

[86] http://www.getty.edu/research/tools/vocabularies/index.html

characteristics. At each level of the hierarchy - each of them representing a different level of abstraction - similar concepts are grouped by a common characteristic. Concepts sharing the same characteristic form what in jargon is known as an *array* of homogeneous and mutually disjoint concepts.

For instance the concepts for *river* and *lake* Have as the primary characteristic that both are bodies of water. Since they share the same characteristic and are disjoint, both of them are categorized in the same array under *body of water*. Their characteristics were enlisted as follows:

- *river* is a flowing body of water; has no fixed boundary; is confined within a bed and stream banks; is larger than a brook
- *lake* is a stagnant body of water and has fixed geographical boundary.

Because of this sort of detailed list of concept characteristics, it is possible not only to distinguish them but also to identify the more general categories. In the complete facet[87], under the root concept *body of water,* there are two broad categories identified, i.e. *stagnant body of water* and *flowing body of water*. Now for instance, it is necessary to include the new concept *pond*, which characteristics are a stagnant body of water and smaller than a lake: the facet can be easily extend by adding it under *stagnant body of water*. This shows that the facets at the array level are exhaustive enough to accommodate new concepts.

[87] The complete facet is provided in B. Dutta, F. Giunchiglia, V. Maltese, *A facet-based methodology* for geo-spatial modelling, GEOS (2011).

This process led to the creation of a set of facets containing overall more than 1000 concepts, but at this time the facets are not explicitly provided. Conversely, the concepts and relations constituting them were rather merged with WordNet. Similarly to *Space*, the *Time* domain was built by using WordNet and Wikipedia[88] as main sources and arranging identified concepts by common characteristics. For instance, *holidays* are grouped *by religion*; *Christian holydays* include *Easter* and *Christmas*; *Islamic holidays* include *Ramadan and Muharram*.

The fact that in this approach, inside the facets generated following this methodology, the distinction between classes, entities, qualities and values is made explicit, unlike the Analytic-Synthetic approach, is worthy of note. The *is-a*, *instance-of*, *part-of* and *value-of* relations between the entities are explicit too. In other words, the facets produced by the Analytic-Synthetic approach correspond to *classification ontologies*, i.e. ontologies built in order classify documents. Conversely, this approach produces *descriptive ontologies*, i.e. ontologies built to describe a domain.

Step III: populate the knowledge base with entities.

In this step 7 million entities from GeoNames were automatically imported at knowledge level in the knowledge base[89]. A significant part of this data were

[88] http://www.wikipedia.org/

[89] Around 600,000 additional locations as well as 700,000 persons and 150,000 organizations are currently been imported from YAGO

released as an open source geo-spatial ontology, *GeoWordNet*[90]. Notice that it is possible to use GeoWordNet, distributed in WordNet format, instead of WordNet as background knowledge.

Step IV, next steps: building the Internet domains.

As already seen, the Entitypedia long term goal is not to build the world knowledge, but to identify those domains more suitable to be used in the Web[91], for enabling diversity-aware applications for it. A prioritized list of around 350 domains was identified. On the very top of this list we find domains such as *Space*, *Time*, *food*, *sports*, *tourism*, *music* and *movie*[92], which were called *Internet domains* or also *everyday domains*.

[90] F. Giunchiglia, V. Maltese, F. Farazi, B. Dutta, *GeoWordNet: a resource for geo-spatial applications*, Extended Semantic Web Conference ESWC (2010). http://geowordnet.semanticmatching.org/

[91] In the context of the Living Knowledge EU project, this has been identified as strategic towards enabling diversity-aware applications for the Web. http://livingknowledge-project.eu/

[92] One of the other domain developed is the *political science* domain (see D. P. Madalli, A.R.D. Prasad, Analytico synthetic approach for handling knowledge diversity in media content analysis, UDC seminar (2011). Another domain under development is the *food* domain.

2.5.3. Entitypedia vs other knowledge bases

Entitypedia settles between the two approaches described before. Its modeling has required import of knowledge from existing resources, such as GeoNames and YAGO, but also a significant amount of manual work to provide the data high quality. Moreover, experts in library science, following a precise methodology and guiding principles, manually build domain knowledge.

By comparing it with respect to pre-existing systems, Entitypedia has at least the following distinctive features, summarized in table 2.2.

- There is a clear split between natural language, formal language and knowledge
- There is an explicit definition of domain as a way to codify knowledge which is local to a community thus reflecting their specific purpose, needs, competences, beliefs and personal experience
- There is an explicit distinction between classes, entities, qualities and values
- It is totally modular, able to be continuously extended with knowledge about new domains and new vocabularies
- Domain knowledge is created following a precise methodology and principles inspired by well-established library science methodologies and practices

- Domain knowledge is used to construct the context formalized (given the specific tasks we want to serve) as a propositional DL theory and therefore the complexity of reasoning is limited to propositional reasoning
- It does not only consist of a data repository, but it comes with a framework to support a precise set of basic semantic tasks including natural language understanding, automatic classification, semantic matching and semantic search by encoding knowledge in the most appropriate semantics according to the task at hand[93].

Entitypedia provides the proof of the applicability of *faceted approach* to knowledge representation. Thanks to this methodology, which has as a focus the fundamental notions of *domains* and *facets*, it is possible to create a knowledge base completely modular and extensible adding a virtually endless number of domains and facets with corresponding classes, entities and vocabularies. The usefulness of Entitypedia, in particular in the *Space* domain, was proved in real scenarios. However, one of the main drawback of Entitypedia approach is its requirement of a significant amount of manual work for providing the high quality of the knowledge.

[93] V. Maltese, F. Farazi, *Towards the Integration of Knowledge Organization Systems with the Linked Data Cloud*, UDC seminar (2011).

Knowledge base	#entities	#facts	Domains	Distinction Concepts instances	Distinction Natural language Formal language	Manually built
YAGO	2.5 M	20 M	No	No	No	No
CYC	250k	2.2 M	Yes	No	No	Yes
OpenCYC	47k	306k	Yes	No	No	Yes
SUMO	1k	4k	No	Yes	Yes	Yes
MILO	21k	74k	Yes	Yes	Yes	Yes
DBPedia	3.5 M	500M	No	No	No	No
Freebase	22 M	?	Yes	Yes	No	Yes
Entitypedia	10 M	80 M	Yes	Yes	Yes	Yes

Table 2.2 Comparison of existing knowledge bases in terms of support to diversity

Chapter 3

Modeling Entity Types: cases of study

In the context of Entitypedia project described in previous chapter, the University of Trento is developing an entity-centric knowledge representation schema. The lattice under development has a structure similar to the lattice proposed by schema.org94. Schema.org is an attempt to organize web knowledge in a pragmatic way without specific formal criteria. It provides 293 types into its structure, but it cannot be considered a reliable resource. In fact, besides the lack of any formal criteria, the descriptions of types are confused and incomplete and only few types introduce valid attributes. Thus schema.org should be considered only as a good starting point to get an overview about general structure and types, but not as a valuable resource.

[94] http://schema.org/

- **Thing**
  - **CreativeWork**
    - Article
    - Blog
    - Book
    - ItemList
    - MediaObject
      - ImageObject
      - VideoObject
    - Movie
    - MusicPlaylist
      - MusicAlbum
    - MusicRecording
    - Recipe
    - Review
    - TVEpisode
    - TVSeason
    - TVSeries
    - WebPage
  - **Event**
  - **Organization**
    - EducationalOrganization
    - **LocalBusiness**
      - FoodEstablishment
    - PerformingGroup
      - MusicGroup
  - **Person**
  - **Place**
    - CivicStructure
    - **LocalBusiness**
  - **Product**
  - **Intangible**
    - Offer
      - AggregateOffer
    - Rating
      - AggregateRating
    - StructuredValue
      - ContactPoint
        - PostalAddress
      - GeoCoordinates
      - NutritionInformation

Table 2 Schema.org types

In this context, the purpose of the work described in following sections is to define and model some types, starting from the structure proposed in schema.org lattice. Entity types (and their sub-types) analyzed are:

- **CreativeWork**: in schema.org creative work includes 43 types and in this work it corresponds to the notion of *mind product* and *information object* (or *computer file*), defined as one of the possible manifestation of *mind product*.
- **Event**, with its 25 types, is pretty minimal in schema.org. These types are very broad (e.g. BusinessEvent, MusicEvent, SocialEvent) and it can be organized in many sub-trees.

The analysis proposed follows these key steps: it starts from the definition of the concept, and then follows the step of validation on standards. Notice that not only the codified standards are considered, but also standards that are the most used *de facto*, especially in web context. After these preliminary steps, it is possible to model the types, using a minimal set of attributes needed to describe them and validated by standards.

3.1. Mind Product

The first problem is to understand what a MIND PRODUCT is, which are its attributes and its possible manifestations. In a broad sense, a MIND PRODUCT can be defined as any product of a mental act, whose main feature are to act as model from which multiple copies can be generated. The proposed approach is to start with a concept closer to mind product idea, the borrowed legal notion of *creative work*:

> *"A creative work is a manifestation of creative effort such as artwork, literature, music, paintings, and software. Creative works have in common a degree of arbitrariness, such that it is improbable that two people would independently create the same work. Creative works are part of property rights. The term is frequently used in the context of copyright law.*[95]*"*

[95] http://en.wikipedia.org/wiki/Creative_work

Based on this definition it is possible to extract some key features suitable for featuring main aspects of a MIND PRODUCT:

- MIND PRODUCT is a manifestation of a mental process. For instance, *literary work* is considered a mind product because it is produced by a mental act of creation and by the process of putting something in a written form. Note that MIND PRODUCT is not the mental process itself, but the result of this process.
- as an original product of someone's intellect, mind products can be able to receive a copyright, that provides exclusive rights to specific kinds of creations of mind (like literary, musical and artistic works[96]).
- MIND PRODUCTS can have different types of “manifestation”, physical (e.g. books, paintings) or virtual (e.g. software).

Based on these features, a first definition of **MIND PRODUCT** can be “any product of human intellect, resulting from a mental act of creation, copyrightable and realizable in different copies or reproductions and in different forms”.

Note that each MIND PRODUCT is not necessarily realizable in several copies, e.g. pieces of art such as paintings and sculptures are created in a single original. Copies created following the original they are just reproductions.

Here it is important to mention the difference between “instance” and “manifestation”, in a terminological view. In modeling, “instance” indicates an

[96] Cfr. http://en.wikipedia.org/wiki/Intellectual_property

individual that belongs to a class, e.g. “Gulliver's Travels” is an instance of a mind product, which is a novel, a specific kind of literary work. Instead, with the term “manifestation” we indicate the manner in which a certain MIND PRODUCT is made concrete. Therefore, manifestations are not instances of a MIND PRODUCT, but they are instances of its concretization, whether physical or not.

3.1.1. Copyright, Intellectual Property and Mind Product

Actually the crucial point is to define the kinds of MIND PRODUCTS and how to distinguish between MIND PRODUCTS and their concretizations. As seen before, as a creative product, MIND PRODUCTS can be eligible for copyright protection. Indeed copyright laws provide the most authoritative starting point to define mind product features and types; moreover, they can help to distinguish a mind product from its possible manifestations.

a) Intellectual Property and Mind Product definition

Concerning mind product definition, it is important to note that copyright law protects the form of expression of ideas, not the abstract ideas themselves. For instance, a musical composition like *Chopin's Piano Sonata No. 2* is not protected by copyright as a process of creating a new piece of music. Copyright laws protect Chopin’s creativity in the sense of the choice and arrangement of musical notes, sounds and musical form. This is important to understand MIND PRODUCT as a "product" of an idea, not as the mental act itself.

Notice that in a strict legal sense, copyright legislation is part of the wider body of law known as intellectual property.

Broadly speaking, the Convention Establishing the World Intellectual Property Organization (WIPO)[97] gives the following list of subject matter protected by intellectual property rights:

- literary, artistic and scientific works;
- performances of performing artists, phonograms, and broadcasts;
- inventions in all fields of human endeavor;
- scientific discoveries;
- industrial designs;
- trademarks, service marks, and commercial names and designations;

The term intellectual property refers broadly to the creations of the human mind. Note that in a strict legal sense intellectual property is usually divided into two branches, namely industrial property, which protects inventions, industrial design and trademarks and copyright, which protects "literary and artistic works",

[97] *Ibidem.*

understood to include every original work, irrespective of its literary or artistic merit[98].

The main criterion to distinguish between industrial property and copyright is based on the difference between inventions and literary and artistic works:

- Inventions may be defined as new solutions to technical problems. These new solutions are ideas, not necessarily represented in a physical embodiment. The protection for inventions gives a monopoly right to exploit an idea
- Copyright law protects only the form of expression of ideas, not the ideas themselves. So copyright law protects unauthorized use of the expressions of ideas.

Notice that more than one type of protection may be employed to the same work. For example, the particular design of a bottle may qualify for copyright protection as a sculpture, or for trademark protection based on its shape, or the trade dress appearance of the bottle as a whole may be protectable. Titles and character names from books or movies may also be protectable as trademarks while the works from which they are drawn may qualify for copyright protection as a whole.

[98] Cfr. Berne Convention for the Protection of Literary and Artistic Works (article 2) http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html#P85_10661

b) Intellectual Property and Mind Product types

Both industrial property and copyright laws can be a useful source to have a preliminary idea of potential MIND PRODUCT classes, their attributes and their behavior. Copyright laws can be used to define possible MIND PRODUCTs types, while industrial property laws can specify how these MIND PRODUCTs are protected. Trademarks and patents seek to protect the idea behind a product, a service or an invention; by comparison, copyright seeks to protect the manifestation of these ideas in the form of a creative work.

According to "Paris Convention for the protection of Industrial property"[99], industrial property takes a range of forms, the main types (*primary rights*) of which are outlined in following table. Note that in some jurisdiction there are also more specialized varieties of *sui generis* exclusive rights, such as circuit design rights or industrial design rights.

Patents	A patent grants an inventor exclusive rights to make, use, sell, and import an invention for a limited period of time, in exchange for the public disclosure of the invention.
Geographical indications	A geographical indication is a sign used on goods that have a specific geographical origin and possess qualities or a reputation that are due to that place of origin.
Trademarks	A trademark is a recognizable sign, design or expression which identifies products or services of a particular source from those

[99] http://www.wipo.int/treaties/en/ip/paris/trtdocs_wo020.html#P71_4054

	of others
Trade dress	Trade dress is a legal term of art that generally refers to characteristics of the visual appearance of a product or its packaging (or even the design of a building) that signify the source of the product to consumers
Trade name	A commercial or trade name is the name or designation that identifies an enterprise
Trade secrets	A trade secret is a formula, practice, process, design, instrument, pattern, not generally known, by which a business can obtain an economic advantage over competitors or customers.

Table 1 Industrial property types

The “Berne Convention for the Protection of Literary and Artistic Works”[100], an international agreement governing copyright, provides a set of works that can be protected by copyright rights. This provides an overview on main categories of creative works.

Literary works	Literary works include every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression, such as books, pamphlets and other writings; lectures, addresses, sermons and other works of the same nature
Musical works	musical compositions with or without words, dramatic-

[100] http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html#P82_10336

	musical works and choreographic works
Cinematographic works	cinematographic works to which are assimilated works expressed by a process analogous to cinematography;
Artistic works	works of drawing, painting, architecture, sculpture, engraving and lithography
Photographic works	photographic works to which are assimilated works expressed by a process analogous to photography
Design works Applied arts	illustrations, maps, plans, sketches and three-dimensional works relative to geography, topography, architecture or science.

Table 2 Protected works according to Berne Convention for the Protection of Literary and Artistic Works

In later years new forms of expression have been added to the list, not included in the original text. In 1996 the “World Intellectual Property Organization Copyright Treaty” (WIPO)[101] was adopted to address the issues raised by information technology and the Internet, which were not addressed by the Berne Convention. In according to Trade-Related Aspects of Intellectual Property Rights (TRIPS)[102] agreement, software, computer-implemented inventions, whether in source or object code, and multimedia productions can be protected as literary works under the Berne Convention[103].

[101] http://www.wipo.int/treaties/en/ip/wct/trtdocs_wo033.html#P56_5626

[102] http://www.wto.org/english/docs_e/legal_e/27-trips_01_e.htm

[103] Cfr. TRIPS, Article 10: http://www.wto.org/english/docs_e/legal_e/27-trips_04_e.htm

Software	Computer programs, whether in source or object code, shall be protected as literary works under the Berne Convention
Multimedia productions	Combination of sound, text and images in a digital format, accessible by a computer program

Table 3 Protected works in according to WIPO/TRIPS

Notice that in some countries there is a *sui generis* right concerning databases[104]. Database right is considered comparable to, but distinct from copyright that exists to recognize the investment that is made in compiling a database, even when this does not involve the creative aspect that is reflected by copyright.

3.1.2. Mind Product sources

Concerning MIND PRODUCT categories, it should be considered that there are other available resources: the web-based hierarchy of creative works provided by schema.org; *Wikisaurus*, a thesaurus built from *Wikipedia Wiktionary* and Wordnet lexical database.

[104] https://en.wikipedia.org/wiki/Database_right

a) Schema.org

As described before, Schema.org[105], as an attempt to organize web knowledge in a pragmatic way without formal criteria, is not considered a good resource. It is regarded only to evaluate main concepts coverage. Schema.org defines a creative work in a broad and confused way: “The most generic kind of creative work, including books, movies, photographs, software programs, etc.”; it provides a hierarchy to describe many types of creative work.

- Article
  - BlogPosting
  - NewsArticle
  - ScholarlyArticle
  - TechArticle
- Blog
- Book
- Code
- Comment
- DataCatalog
- Dataset
- Diet
- ExercisePlan
- ItemList
- Map
- MediaObject
  - AudioObject
  - DataDownload
  - ImageObject
  - MusicVideoObject
  - VideoObject
- Movie
- MusicPlaylist
- MusicRecording
- Painting
- Photograph
- Recipe
- Review
- Sculpture
- SoftwareApplication
- TVEpisode
- TVSeason
- TVSeries
- WebPage
- WebPageElement

Table 4 schema.org creative work full hierarchy[106]

[105] http://schema.org/

[106] http://schema.org/docs/full.html

This hierarchy seems to be quite detailed for web-oriented types (e.g. MediaObject has specific subtypes), but for other categories it is incomplete (e.g. book has no subtypes) or confused (e.g. TVEpisode, TVSeason and TVSeries could be grouped in a single type).

b) Wikipedia Wikisaurus

Wikisaurus is a *Wiktionary* subproject and a wiki namespace aiming at creating an electronic thesaurus, a dictionary of synonyms, antonyms and further semantically related terms such as hyponyms, hypernyms, meronyms and holonyms.

Wiktionary defines *creative work* as following: “a tangible manifestation of creative effort, such as literature, music, paintings, and software[107]”, Wikisaurus provides a list of hyponyms that defines possible creative work categories:

- work of art
- artwork
- art piece
- painting
- photograph
- photo
- dance
- performance
- show
- work of fiction
- fanfic
- prose
- midquel
- software
- electronic game
- transcreation
- pot boiler
- pot-boiler

[107] Cfr. http://en.wiktionary.org/wiki/creative_work

• motion picture • short • theatermusic	• poetry • sequel • prequel	• potboiler

Tabella 5 wikisaurus hyponyms of creative work

Wikisaurus appears as a work in progress resource: there is no a hierarchical criterion, all types are on the same level and representation turns out to be disordered. Furthermore, there are problems with synonymous terms, for instance both pot-boiler and pot boiler refer the same concept, they are just graphic variations, but they are represented as two different types.

c) Wordnet

Although Wordnet is based on lexical criteria and it has an exclusively linguistic purpose, it can be considered a good resource. Wordnet can be useful as a terminological source, to evaluate and extend terms coverage and to get an overview of the terms organization in a taxonomic form. In a different way it also provides a list of creative work types. Synsets that refer more to a broad concept of a "creative work" are following:

- **Creation**: something that has been brought into existence by someone

- **Abstraction**: a general concept formed by extracting common features from specific examples.

As said before, Wordnet approach is exclusively linguistic; there is not any distinction between abstract creative works and manifestations. For instance, browsing *creation* hyponyms, there are some synsets that refer to some aspects of MIND PRODUCT (e .g. definitions like “the products of human creativity”) and other synsets related to artifact features (e.g. “a visual or tangible rendering of someone or something”).

Creation	something that has been brought into existence by someone
=> art, fine_art	the products of human creativity; works of art collectively
=> Representation	a visual or tangible rendering of someone or something
=> document	anything serving as a representation of a person's thinking by means of symbolic marks
=> picture, image	a visual representation of an object or scene or person produced on a surface

=> piece (Art)	an artistic or literary composition
=> musical_composition	a musical work that has been created

=> Abstraction (Factotum)	a general concept formed by extracting common features from specific examples
=> communication	
=> movie, film	a form of entertainment that enacts a story by a sequence of images giving the illusion of continuous movement
=> signal, signaling, sign	any communication that encodes a message
=> file, data file	a set of related records (either written or electronic) kept together
=> written communication,	communication by means of written symbols
=>document, written document	writing that provides information

Table 6 wordnet creative works taxonomy

3.1.3. Mind Product types and their manifestations

Starting from categories of copyrightable products, it can be assumed a preliminary list of potential MIND PRODUCT types, related to their protection mechanism

Mind product type	Protection types
Literary works	copyright
Musical works	copyright
Photographic works	copyright
Cinematographic works	copyright
Artistic works	copyright
Design works	copyright + industrial property
Products	industrial property
Inventions	industrial property

This list, however, is not intended to be exhaustive. However, it suggests some guidelines to define macro-types of MIND PRODUCTS. Moreover, these categories are so broad as to be easy to extend and extremely adaptable, for instance "literary works" includes Dante's Divina Commedia, priest homilies, academic articles and forms of expression heterogeneous between them, that only share the written form, etc. The weak point of these macro-types is due to the total lack of attributes specification.

MIND PRODUCT main aspect is due to the fact that they act as models from which multiple occurrences or copies can be generated. As mentioned before, intellectual property laws are also helpful to make a clear distinction between abstract model and its manifestations.

As seen before, Intellectual property protects form of expression of ideas, which do not require to be represented in a physical embodiment. Intellectual property is related to items of information or knowledge expressed in a creative work, not in unlimited number of derivative copies[108].

This is a crucial point to distinguish between MIND PRODUCT and manifestations: MIND PRODUCT is completely independent from any manifestation. Manifestations are just multiple copies generated following the MIND PRODUCT abstract model. E.g., a car model is a MIND PRODUCT, while a car built from that project is its physical manifestation, reproducible in many copies.

If the manifestation is a physical artifact, copies are limited in number, because they come at a cost; however, virtual objects can be reproduced in a virtually infinite number of copies, since they cost only memory size. All copies, in any form of manifestation, have features aligned to the model. Model is a flexible template where attributes denoting the distinctive features of the copies are subject to fixed or variable constraints. This is done by fixing a set of possible values or range restriction. For instance, a car can be distributed in different but restricted range of colors and with different optional features (to be chosen among

[108] Cfr. http://www.wipo.int/freepublications/en/intproperty/909/wipo_pub_909.html#works

a limited set of options); a book can be distributed in different formats, e.g. as a physical paper volume or as e-book.

It is important to take into account that the distinction between model and its copies is not at all considered in literature.

A MIND PRODUCT can generate two types of manifestations, physical or virtual. Physical manifestations are *artifact* and virtual manifestations are *information objects*.

- An **artifact** represents any physical object that makes concrete a MIND PRODUCT. As a tangible object, it have physical properties, like weight, dimension or shape.
- An **information object** (described in the following section) is a virtual copy in a digital format. An information object is a computer file that has typical digital objects properties like size, file name or URI.

At the present, there are only these two types of manifestations, but this could be a limitation of current technologies. So cannot be excluded other future possible forms of MIND PRODUCT manifestations. In the vast majority of cases, the same MIND PRODUCT can have both physical and virtual manifestations. E.g. *literary work* can be materialized as *book* (artifact) that has some physical properties like weight, paperback, number of pages, or as *digital document* (information object) with typical computer file properties like size, format, compatibility. Similarly, *photographic work* can be made concrete as artifact *photo* or as information object *image file*.

Summing up a MIND PRODUCT is a copyrightable entity totally independent of any concretization and it provides the abstract model from which copies are created. Manifestations are distinguished by their form, physical or virtual, and by properties that characterize them.

3.1.4. Kinds of mind products

Based on macro-types of MIND PRODUCTs seen before and on relationship between model and its copies, here are proposed some kinds of MIND PRODUCTs.

1. **"Literary, scientific and artistic works", such as books, songs, movies, industrial design works**: the work can be reproduced in multiple copies and in multiple formats. There is 1 abstract model and n physical or virtual copies generated following the model.
2. **Information objects like computer programs**: they are directly created as information objects.
3. **Piece of art such as paintings and sculptures**: they are directly created as artifacts or information object (computer art) and when copies are created following the original they are just reproductions.

Notice that all of them can be modified, transformed or adapted into a new MIND PRODUCT. This is especially true, but not limited to, for reproductions. For instance, a book can be commented or translated in a different language. A painting can be modified by adding some distinctive features. When this happens, the modifications can in turn be protected by copyright. They are called derivative works. However, to be copyrightable, a derivative work must be different enough from the original to be regarded as a "new work" or must contain a substantial amount of new features.

3.2. Information Object

In a general sense, a COMPUTER FILE (a.k.a. INFORMATION OBJECT) is a block of arbitrary information or resource for storing information, which is available to a computer program and is usually based on some kind of durable storage. A file is durable in the sense that it remains available for programs to use after the current program has finished. Computer files can be considered as the modern counterpart of paper documents, which traditionally were kept in offices' and libraries' files, which are the source of the term.

The following sections provide the state of the art in computer file metadata with focus on metadata standards specifications. Based on the state of the art, it is proposed a set of attributes to define an entity type able to describe INFORMATION OBJECTS.

3.2.1. Standard for Computer Files

Here an overview on the most used standards to describe a computer file, from the Dublin Core metadata terms, a well-known standard to describe many kinds of resources, to the sets of standards and properties used by major operating systems to manage computer files.

3.2.1.1. Dublin Core Metadata Initiative

The Dublin Core[109] metadata terms are a set of vocabulary terms, which can be used for multiple purposes, from simple resource description to interoperability combining metadata vocabularies of different metadata standards. The terms can be used to describe a full range of web resources (video, images, web pages, etc.), physical resources, such as books, and objects like artworks.

The Dublin Core standard includes two levels, Simple and Qualified. Simple Dublin Core, also known as Dublin Core Metadata Element Set[110] (version 1.1) comprises 15 elements; Qualified Dublin Core[111] includes additional elements and a list of qualifiers, or element refinements, that refine the semantics of the elements in ways that may be useful in resource discovery. Following tables provide the complete list of Dublin Core elements and a list of most common Element Refinement Dublin Core terms. For each term a simple description is provided.

[109] http://dublincore.org/

[110] http://dublincore.org/documents/dces/

[111] http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/#description

Name	Description
DC.Contributor	An entity responsible for making contributions to the resource.
DC.Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
DC.Creator	An entity primarily responsible for making the resource (person, organization, service).
DC.Date	A point or period of time associated with an event in the lifecycle of the resource.
DC.Description	An account of the resource (e.g. abstract, table of contents, summary).
DC.Format	The file format, size and duration.
DC.Identifier	An unambiguous reference to the resource within a given context (e.g. URL, URI, ISBN).
DC.Language	A language of the resource.
DC.Publisher	An entity responsible for making the resource available.
DC.Relation	A related resource preferentially identified by means of a string conforming to a formal identification system.
DC.Rights	Information about various property rights associated with the resource, including intellectual property rights.
DC.Source	A related resource from which the described resource is derived. The described resource may be derived from the related resource in whole or in part.
DC.Subject	The topic of the resource typically represented using keywords, key phrases, or classification codes.

DC.Title	A name given to the resource that is typically the name by which the resource is formally known.
DC.Type	The nature or genre of the resource.

Table 7 Dublin Core Metadata Element Set Version 1.1.

DC Element	Element Refinement	Description
Title	Alternative	Substitute to the title (e.g. abbreviations or translations).
Description	Table Of Contents	A list of subunits of the content of the resource.
	Abstract	A summary of the content of the resource.
Date	Created	Date of creation of the resource.
	Valid	Date (often a range) of validity of the resource.
	Available	Date (often a range) that the resource will become or did become available.
	Issued	Date of formal issuance (e.g., publication) of the resource.
	Modified	Date on which the resource was changed.
Format	Extent	The size or duration of the resource.
	Medium	The material or physical carrier of the resource.
Relation	Is Version Of	The described resource is a version, edition, or adaptation of the referenced

		resource
	Has Version	The described resource has a version, edition, or adaptation, namely, the referenced resource.
	Is Replaced By	The described file is supplanted by the referenced file.
	Replaces	The described resource supplants the referenced resource.
	Is Required By	The described file is required by the referenced resource, either physically or logically.
	Requires	The described resource requires the referenced resource to support its function, delivery, or coherence of content.
	Is Part Of	The described resource is a physical or logical part of the referenced resource.
	Has Part	The described resource includes the referenced resource either physically or logically.
	Is Referenced By	The described resource is referenced or otherwise pointed to by the referenced resource.
	References	The described resource references, cites, or otherwise points to the referenced resource.
	Is Format Of	The described resource is the same of

		the referenced resource presented in another format.
	Has Format	The referenced resource is the same of the described resource presented in another format.
Coverage	Spatial	Spatial characteristics of the intellectual content of the resource.
	Temporal	Temporal characteristics of the intellectual content of the resource.

Table 8 List of most common Element Refinement Dublin Core terms

3.2.1.2. File System/Operating System Standard

Every file system/operating system uses different properties to identify and manage computer files. Following is provided a list of the most common file properties used by major operating systems.

a) Microsoft Windows

In Microsoft Windows systems, by default, a dialog box displays basic properties for every file, including such things as file name, file size, and the file creation, last access, and last modification date[112]. In addition to these basic properties, the Windows operating system also tracks a number of extended file

[112] In NTFS file systems general informations about a file (Creationtime, LastModificationTime, LastChangeTime, and LastAccessTime) are stored in $STANDARD_INFORMATION attribute type.

properties[113], typically hidden. The following table provides a list of extended file properties used.

Index	Property	Index	Property	Index	Property	Index	Property
0	Name	8	Owner	16	Artist	24	Camera Model
1	Size	9	Author	17	Album Title	25	Date Picture Taken
2	Type	10	Title	18	Year	26	Dimensions
3	Date Modified	11	Subject	19	Track Number	30	Company
4	Date Created	12	Category	20	Genre	31	Description
5	Date Accessed	13	Pages	21	Duration	32	File Version
6	Attributes	14	Comments	22	Bit Rate	33	Product Name
7	Status	15	Copyright	23	Protected	34	Product Version

[113] http://technet.microsoft.com/en-us/library/ee176615.aspx

Tabella 9 Windows Extended File Properties (Notice that index 26, 27 and 28 are not used)

Notice that in Windows Operating System there is a difference between file properties and file attributes. Although one can say the file size and the file date/time are file attributes (i.e., any properties associated with a file other than the file contents), with the more narrow definition and popular usage, the file attributes are collection of flags, which describes various aspects of the file[114]. Microsoft Windows provides a command, `attrib`[115], useful to display, set, or remove file attributes.

Name	Description
Read-Only	This attribute is useful to make a file write-protected by software. It's not possible to delete read-only file under normal circumstance. (e.g. certain system files are kept as read-only by default)
Hidden	It makes the file invisible in certain applications' file list display. Since many file applications has the feature to ignore the Hidden attribute bit, the hidden file is not always invisible.
System	It is the least rigorously defined in its usage. Applications

[114] http://www.xxcopy.com/xxcopy06.htm

[115] http://technet.microsoft.com/en-us/library/bb490868

	treat the System attribute similarly to the Hidden attributes for directory listing.
Archive	It determines whether a file requires a backup (archiving). The Archive attribute is set whenever an existing file is either overwritten or modified by the file system. A new file is usually created with the Archive attribute set.

Tabella 10 Microsoft Windows file attributes

Windows NTFS volumes allow to set security permissions on files, which grant or deny access to the files[116]. These permissions can be viewed using Explorer Properties dialog box.

Permission	Meaning for Files
Read	Permits viewing or accessing of the file's contents. Read is the only permission needed to run scripts
Write	Permits writing to a file
Read & Execute	Permits viewing and accessing of the file's contents as well as executing of the file
Modify	Permits reading and writing of the file; allows deletion

[116] To a complete and specific definition of all file and folders permissions and advanced and special permissions see: http://technet.microsoft.com/en-us/library/bb727008.aspx

	of the file
Full Control	Permits reading, writing, changing and deleting of the file

Tabella 11 Windows files basic permissions

b) Unix-based Systems: Linux and Mac OS

In linux-based operating system and Apple Mac OS, the core properties used to describe a file are similar to Windows $STANDARD_INFORMATION: Name, Creationtime, LastModificationTime, LastChangeTime, and LastAccessTime. These main properties are displayed in a dialog box in a visual interface or using various shell command like `ls`[117]. Both Linux-based Operating Systems and Apple Mac OS share Unix methods to manage file attributes and permissions. Unix file permissions are divided into three groups: for the file owner, for the group owner, and for everyone else. Each group can have up to three attributes for reading (r), writing (w) and executing (x). The references (or classes) are used to distinguish the users to whom the permissions apply. They are represented by one or more of the following letters.

Reference	Class	Description

[117] http://unixhelp.ed.ac.uk/CGI/man-cgi?ls; http://developer.apple.com/library/mac/documentation/Darwin/Reference/ManPages/man1/ls.1.html

u	User	the owner of the file
g	Group	users who are members of the file's group
o	Others	users who are not the owner of the file or members of the group
a	All	all three of the above

In Unix-based System permissions given to users, groups and/or the other class to access files are called *Modes*. The modes indicate which permissions are to be granted or taken away from the specified classes. There are three basic modes which correspond to the basic permissions [118](Note that Modes can be changed with `chmod`[119] command).

Mode	Name	Description
r	Read	read a file or list a directory's contents
w	Write	write to a file or directory
x	Execute	execute a file or recourse a directory tree

Tabella 12 Basic Unix Modes

Although file permissions are clearly specified, files attributes and properties do not have any unique specification. Unix-based systems do not have

[118] For a complete list of Unix modes and file permissions: http://en.wikipedia.org/wiki/Modes_(Unix).

[119] http://www.gnu.org/software/coreutils/manual/html_node/chmod-invocation.html

a specific files properties list, except general properties like name, path or date/time.

Following tables provide a list of most common properties in Unix-based System. In the former there are the most common properties available to users, in the latter some properties normally hidden to users and visible only using external tools or shell command.

Property/Attribute	Description
File Type	The 'real file type' according to Unix.
Access	Read, Write, Execute bits divided up for User, Group, and Other.
User Flags	Extended flags that override ordinary file permissions.
System Flags	Extended flags that override all other file permissions.

Table 13 Basic Unix file properties displayed in a property dialog box

Property/Attribute	Description
Device	The ID of the physical device the file resides on.
Device Type	This will be '0' unless the target is a 'device file' in /dev.
Size	The size of the file in bytes.

Mode	The file's mode displayed in octal. Note this includes more than ordinary permissions.
Blocks	Implies the full storage requirements of the file.
Links	The number of (hard) links to a file. A physical Unix file can be referenced by any number of file system paths.
Owner	This is the 'user' - the account that created the file. Only the root account can change the ownership of a file.
System/User Flags	These are the extended flags that override ordinary file permissions.
Group	The group the file belongs to.

Table 14 Some of Unix Extended File Attributes (normally hidden from users)

3.2.2. Modeling Information object Entity type

After an analysis of state of the art standards, we propose a list of essential attributes to describe computer files. Note that there are several changes from types that already exist in schema.org:

- Webpage is not a separate subtype. Because the only distinguishing property of Webpage was to have links to other files, now this is a property common to all files, whether they are stored online and offline.

- Focus on category temporal, that includes date/time about files and new attribute Modification Time, because it is one of the most common properties used by Operating System to identify a file
- DeviceModel and DeviceID Attributes have been moved from Computer File to Image File subtypes, since they are used only for images or media.

Description	A **computer file** (a.k.a. Information object) is a block of arbitrary information or resource for storing information, which is available to a computer program and is usually based on some kind of durable storage. A file is durable in the sense that it remains available for programs to use after the current program has finished. Computer files can be considered as the modern counterpart of paper documents which traditionally were kept in offices' and libraries' files, which are the source of the term. However, differently from such artifacts, computer files have the possibility to be reproduced at almost zero cost (only the storage).
Standards	Dublin Core, Windows Properties, Unix Properties
Subtypes	Image File, Video File, Audio File

Category: general

Name	Reference	Description

Creator	DC.creator Windows.Author Unix.Owner	Creator of the file
File Name	DC.Title Windows.Name	The name of the file (without the format)
URL	DC.Identifier Windows.path Unix.path	The URL pointing to the physical location where the file is stored
Format	DC.Format Windows.type Unix.type	The format of the file, denoting a particular way to encode information
Size	DC.Format.Extent Windows.Size Unix.Size	Measures the actual amount of disk space consumed by the file (in bytes)
Tag	DC.Subject	Keywords or terms associated with or assigned to a file
Mind Product	DC.relation	The mind product on the basis of which the file has been created.
Link	DC.Relation	A linked computer file

Category: creation		
Name	**Reference**	**Description**
Source	DC.Source	A software or device from which the file is generated, e.g. "smartphone"

Category: temporal		
Name	**Reference**	**Description**
Creation Time	DC.Date.Created Windows.Date.Created Unix.Date.Created	The time at which the file was created
Modification Time	DC.Date.Modified Windows.Date.Modified Unix.Date.Modified	Date on which the file was changed

3.3.Information object Subtype: Image file

Main INFORMATION OBJECT subtypes are IMAGE FILE, VIDEO FILE and AUDIO FILE (not considered in schema.org lattice). The following section proposes an analysis of image file subtype, based on most used standards for digital images.

To describe the attributes related to an image file it must be considered the two families of digital images: raster (or bitmap) images and vector images.

3.3.1. Raster Image

A raster image, or bitmap, is a dot matrix data structure representing a generally rectangular grid of pixels. To describe a raster image it is possible define at least three aspects for an image description: photoparametrical aspect, semantic aspect and low-level pixel aspect.

Following table provides some of the most used standards for raster images.

Standard	Description
EXIF[120]	It is a specification for the image file format used by digital camera. The specification defines a set of attributes which covers a broad domain of conditions under which the image was taken
DIG35[121]	It defines a standard set of attributes that improve semantic

[120] Exchangeable Image Format www.exif.org

[121] DIG35 – Digital Imaging Group http://xml.coverpages.org/FU-Berlin-DIG35-v10-Sept00.pdf

	interoperability between devices, services and software
IPTC[122]	It was developed to facilitate media exchange between news agencies. The set of attributes was defined with the aim of handling the full semantics related to the image content.

Table 15 Raster Image Standards

3.3.2. Vector Image

A vector image is represented using geometrical primitives (points, lines, curves, shapes or polygons), which are based on mathematical expressions.

The most used standard for vector images is SVG[123] (Scalable Vector Graphics). SVG is an XML-based open standard developed by the World Wide Web Consortium (W3C). SVG Specifications provide a set of attributes to describe structure of an image (Regular attributes) and some attributes for styling properties (Presentation attributes).

Attribute	Description
x (coordinate)	The x-axis coordinate of one corner of the rectangular region into which an embedded svg element is placed.
y (coordinate)	The x-axis coordinate of one corner of the rectangular

[122] IPTC Photo Metadata - International Press Telecommunications Council Photo Metadata http://www.iptc.org/std/photometadata/2008/specification/IPTC-PhotoMetadata-2008_2.pdf

[123] SVG - Scalable Vector Graphics W3C Specifications: http://www.w3.org/TR/SVG11/

	region into which an embedded svg element is placed.
Width	the intrinsic width of the svg document fragment.
Height	the intrinsic height of the Svg document fragment
Viewport	the position and size of the viewport that corresponds to this svg element.
currentScale	It indicates the current scale factor relative to the initial view.
currentTranslate	It indicates the translation factor.
pixelUnitToMillimeterX	Size of a pixel along the x-axis of the viewport.
pixelUnitToMillimeterY	Size of a pixel along the y-axis of the viewport.

Table 16 some of the most used SVG regular attributes

In addition to these attributes, SVG provides a set of properties to describe many other aspects of vector images, like shape, color model, filler, etc.

3.3.3. Modeling Image File entity type

After an analysis of more common digital images standards, following is proposed a list of attributes shared by all types of images, both raster and vector. Changes from previous version are:

- New standard for vector image introduced: SVG.
- Device Model and Device ID attributes are grouped into category Creation

Description	An **image file** is a two-dimensional digital representation of something or somebody, image, picture, figure, photograph, icon, painting, exposure, illustration.
Standards	Raster: EXIF, DIG35, IPTC
	Vector: SVG

Category: general		
Name	**Reference**	**Description**
Height	SVG.Height EXIF.ImageHeight	The horizontal size of an image in pixels
Width	SVG.Height EXIF.ImageWidth	The vertical size of an image in pixels
Byte per pixel	EXIF.BitPerSample	The number of bits per single pixel used to represent the color of the pixel
Color model	SVG.color-profile EXIF.ColorSpace	The abstract mathematical model that describes the way colors are represented (e.g. RGB)

Category: creation		
Name	**Reference**	**Description**
Device Model	DC.Source Unix.DeviceMOdel EXIF.DeviceModel Windows.CameraModel	The model name of the device from which the photo has been generated, e.g. "iPhone 3G"
Device ID	Unix.Device EXIF.DeviceID	The unique identifier of the device from which the photo has been generated

3.4.Event

An EVENT is defined as "something that happens at a given place and time". So, key features needed to describe an event are a location, a start time and an (optional) end time.

The standard used to describe and classify events is EventsML-G2[124]. It is an exchange standard of the IPTC[125] (International Press Telecommunications Council), optimized to share events informations. EventsML-G2 is a member of the family of G2-Standards[126], a family of news exchange format standards which provides state-of-the-art metadata and XML technology to combine rich functionality, compactness and compatibility with the Semantic Web.

EventsML-G2 provides a set of attributes to describe in a general way the EVENT and many attributes to define the participants and the event organizer:

EML.StartTime	The starting date of the event
EML.EndTime	The ending date of the event
EML.DateConfirmation	The status of confirmation of start and and date (start and end date confirmed, start date confirmed, end date confirmed, start and end date approximative)
EML.AccessStatus	The current state of the event (e.g. expected, canceled, confirmed)
EML.Name/headline	The name of the event
EML.Web-url	The web-url or homepage related to the event

[124] http://www.iptc.org/site/News_Exchange_Formats/EventsML-G2/

[125] http://www.iptc.org/site/Home/

[126]http://iptc.cms.apa.at/cms/site/single.html?channel=CH0087&document=CMS1206527645546

EML.Location	The place where the event occurs (e.g., city, city quarter, building, set of locations, etc.)
EML.City	The city of the event
EML.Country	The country of the event
EML.Description/details	A description of the event
EML.Participant	The list of participants (persons or organizations) of the event
EML.Organizer	The person or organization taking care of the organization of the event
EML.ContactInfo	The person or organization that is the reference point for the event
EML.Language	The main spoken language

In addition to these features, EventML-G2 provides a set of subjects to describe different types of events. In following table IPTC Subject codes, used to classify events types are provided.

IPTC Subject codes	
Arts, culture and entertainment	Health
Crime, law and justice	Human interest
Disaster and accident	Weather
Economy, business and finance	Lifestyle and leisure
Education	Politics
Environmental issues	Religion and belief
Science and technology	Social issues
Sport	Unrest, conflict and war

3.4.1. Modeling Event Entity type

As shown before, EventsML standard provides the basis for the attributes of this type. Notice that some of the EML proposed fields pertain to specific subtypes, but not all of them are relevant for the basic definition of EVENT type (e.g. Registration, DateConfirmation). Nevertheless, attributes Status, Organizer and Contact were kept, because the vast majority of events have one of these. Notice that fields EML.StartDate and EML.EndDate are not reported, because they correspond to temporal attributes inherited from the parent node Entity.

Description	Something that happens at a given place and time.
Standards	Event-ML (EML)
Subtypes	

Name	Reference	Description
Participant	EML.Participant	The list of participants (persons or organizations) of the event
Location	EML.Location	The place where the event occurs (e.g., city, city quarter, building, set of locations, etc.)
Status	EML.AccessStatus	The current state of the event (e.g. expected, canceled, confirmed)
Organizer	EML.Organizer	The person or organization taking care of the organization of the event
Contact	EML.ContactInfo	The person or organization that is the reference point for the event

3.5.Summing up

The modeling work proposed here is based on the Entity Centric Representation approach described in previous chapters. The aim of the work is to show a better methodology to model entity types, starting from some simple top level concepts. MIND PRODUCT, INFORMATION OBJECT and EVENT are entity types chosen for the analysis, because they have two characteristics in common. They constitute central nodes (with many subtypes in their sub-trees) in schema.org taxonomy and, on the other hand, they are defined in a very approximate and confused way in schema.org structure.

The analysis starts from schema.org taxonomy, but it introduces formal criteria to build and describe entity types. Entity types are modeled starting from their attributes and, these attributes are extracted from existent standards or available resources about this specific entity type. In order to maximize reusability and interoperability, standard *de facto* are preferred to standard *de iure*. In other words, most widely used web standards, though less precise, (e.g. Dublin Core) are considered more relevant than “official” standards, which are more accurate but less used in practical purposes (e.g. ISO standards).

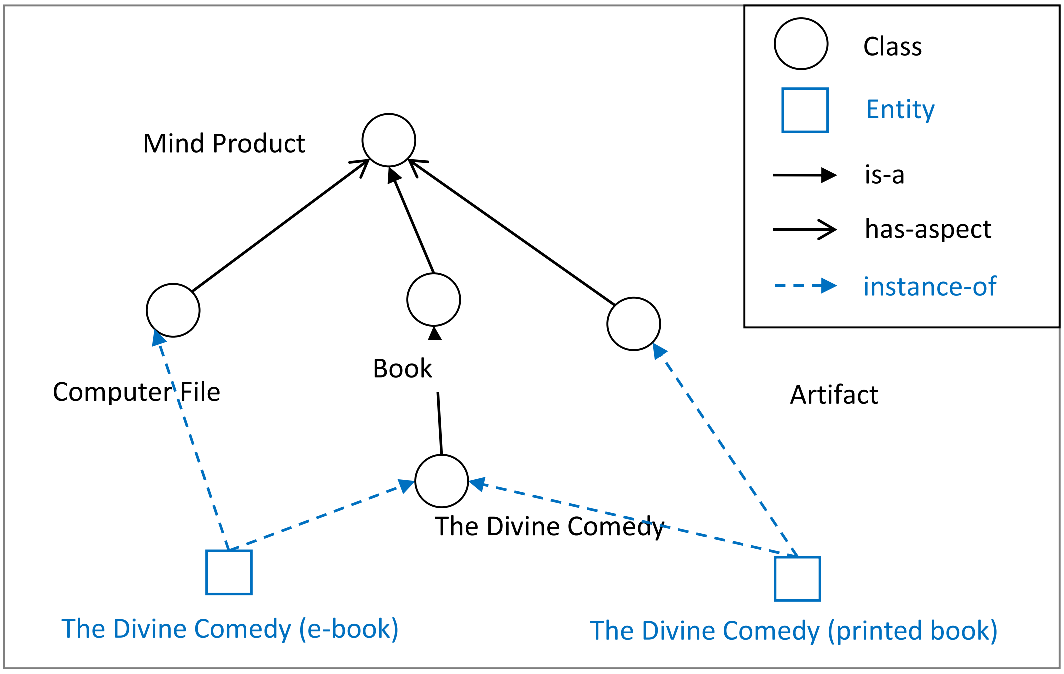
3.5.1. Mind Product problems

MIND PRODUCT is the most interesting case in this modeling work. This is due to many reasons. First, in modeling types as EVENT and INFORMATION OBJECT many official well-codified standards can be used. For instance, EventsML-G2 XML format is a widely used standard already known in literature to categorizing Events.

By contrast, MIND PRODUCT lacks of any univocal definition and in literature, an ontology that categorize its types does not exist.

Furthermore, a MIND PRODUCT has a particular status: it is a totally abstract concept which can assume different instantiations or materializations, both physical or virtual ones. A MIND PRODUCT can be considered as “an abstract product of the human intellect”, its physical instantiations are called *artifacts*, defined as “a man-made object taken as a whole”, and virtual instantiations are called *information objects*, defined “a file maintained in computer-readable form”.

A crucial point is to define various kinds of relations between a mind product and its instantiations and manifestations. For instance, as shown in the following figure, a combination of many kinds of relations can be used to represent these situations: *is-a*, *has-aspect* and *instance-of* relations.



In modeling work, MIND PRODUCT becomes the abstract concept of *creative work*. This abstract concept can assume a material or virtual form and these forms can be instantiated in different copies.

3.5.2. Possible applications

The goal of this modeling work is to show first steps part of a bigger work of creation of a new lattice under development in the University of Trento. This entity-centric lattice, based on faceted approach, will be mapped with schema.org ontology. New types have been developed from existing schema.org ones, because only few types provided by schema.org introduce new attributes. In this document the focus is only on schema.org types and attributes are ignored.

Starting from schema.org types it is possible to create and model corresponding concepts able to be placed in a new structure.

The analysis shows that schema.org concepts are not always well categorized and placed in the taxonomy, so the structure needs some form of cleaning and reorganization. For the concepts which are not in schema.org, it is possible to extend the lattice creating new concepts and validating them using standards and other available resources.

The following table indicates an example of this work of reorganization of some types, analyzed and validated using standards as reference, are mapped with (the symbol > indicates an *is-a* relation).

Schema.org types	New entity types
Thing > Event	Event Conference Session Speech Meeting Lecture Journey
Thing > Creative Work Thing > Creative Work > Media object Thing > Creative Work > Media object > Image object Thing > Creative Work > Media object > Video object Thing > Creative Work > Media object > Audio object	Mind Product Information Object Image File Video File Audio File

3.6.Conclusions

In this study, we have shown an overview of some approaches to knowledge representation, in particular approaches focused on the use of lightweight ontologies.

First, we have discussed different types of classifications and ontologies, their possible applications and their main problems: the inaccuracy of the natural language in processing tasks and the limits due to the lack of background knowledge and the resulting difficulty in catching the intrinsic knowledge diversity.

In particular we have explained how this lack of background knowledge represents one of the main problems for the success of current approaches to knowledge representation, so a huge virtually unbound knowledge base able to capture the diversity of the world, as well to reduce the complexity of reasoning, is a crucial point.

Then, to solve these limitations we have introduced the approach proposed by Giunchiglia et al. (2007)[127], that allows to formalize classifications into faceted lightweight ontologies. The faceted approach[128], a well-established methodology centered on key notions of domain and facet and used in library science for the

[127] Giunchiglia, F., Marchese, M., Zaihrayeu, I. (2007). *Encoding Classi-fications into Lightweight Ontologies*. Journal of Data Semantics, 8, 57-81.

[128] Ranganathan, S. R. (1967). Prolegomena to library classification. Asia Publishing House.

organization of knowledge in libraries, is an effective methodology that allows to catch the diversity of the world in language, knowledge and personal experience.

In this thesis we have introduced the project Entitypedia, currently under development in the University of Trento, in order to create this large scale diversity-aware knowledge base, having as central points the concepts of domain and context.

We have defined a domain as *any area of knowledge or field of study that we are interested in or that we are communicating about*. Domains are the main means by which diversity is captured, in terms of language, knowledge and personal experience. On the other hand, according to Giunchiglia et al. (2006)[129], the notion of *context* allows reducing the complexity, by selecting from the domains the language and the knowledge, which are strictly necessary to solve the problem.

The knowledge base built using this approach can be seen as a proof of the applicability of the faceted approach: it is completely modular since at any moment it allows plugging an arbitrary number of domains and facets with corresponding classes, entities, qualities and values as well as vocabularies in different languages and for different communities.

In the context of this overall approach, we have performed an experiment of modeling: starting from an preexistent taxonomic structure, we have tried to

[129] Giunchiglia, F. (2006). *Managing Diversity in Knowledge*. Invited Talk at the European Conference on Artificial Intelligence ECAI, Lecture Notes in Artificial Intelligence.

define, model and analyze some simple entity types. Schema.org provided the basic structure that allows creation and modeling of types; we analyzed the types *Event*, *Mind Product* and *Information Object*, on the basis of the equivalent nodes in schema.org.

Differently from schema.org methodology, in modeling process we have followed some formal criteria, where entity types have been defined starting from a minimal set of attributes needed for characterizing them into an efficient way.

To validate the effectiveness of these attributes and to make efficient entity types, we introduced a control that considers all the available resources, in particular existent codified standards and also standards that are more used *de facto* in web context. In some cases, as types EVENTS or IMAGE FILE, there are many official standards well-codified, as EventML-G2 XML format or EXIF standard format, but in other cases it has been necessary to combine existent standards with other types of resources.

For instance, the mind product case is particularly interesting because of two reasons: the lack of an univocal definition and the lack of any existent taxonomy not strongly related to a specific domain. Furthermore, MIND PRODUCTS have an unique property, not yet considered in literature: a mind product is a totally abstract concept which can assume different instantiations or materializations, both physical or virtual ones. The relations between a MIND PRODUCT and its manifestation and its subtypes is worthy of interest and may deserve future studies.

Of course, it is important to note that the performed analysis has been focused only on a little part of the ontology and has described only the first starting steps towards the building up and modeling of some abstract upper level entity types.

However, it should be an interesting task to perform a deeper analysis on particular types of relations between types and subtypes in the ontology and use this methodology for the construction and the analysis of specific domains of interests.

Bibliography

- Aleksovski Z. et al. *Using multiple ontologies as background knowledge in ontology matching*, ESWC workshop on collective semantics (2008).
- Aleksovski Z., Ten Kate W., van Harmelen F., *Using multiple ontologies as background knowledge in ontology matching*, ESWC workshop on collective semantics (2008);
- Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak C., Ives Z., *DBpedia: A Nucleus for a Web of Open Data*, 6th International Semantic Web Conference ISWC (2007).
- Autayeu, et al., (2010) *Recommendations for Better Quality Ontology Matching Evaluations*. 2nd AISB Workshop on Matching and Meaning.
- Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. F. (2002). *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press.
- Banko M., Cafarella M. J., Soderland S., Broadhead M., Etzioni O., *Open information extraction from the web*, IJCAI conference (2007).
- Berners-Lee T., Hendler J., Lassila O., *The semantic web*. Scientific American, (284(5)):34–43, May 2001.
- Bollacker K., Evans C., Paritosh P., Sturge T., Taylor J., *Freebase: a collaboratively created graph database for structuring human knowledge*, ACM SIGMOD international conference on Management of data (2008), 1247-1250.

- Broughton V., *The need for a faceted classification as the basis of all methods of information retrieval*, Aslib Proceedings 58 1/2 (2006), 49-72;

- Broughton, V. (2008), *A Faceted Classification as the Basis of a Faceted Terminology: Conversion of a Classified Structure to Thesaurus Format in the Bliss Bibliographic Classification*, 2nd Edition. Axiomathes Journal, Springer Online Issue, 18 (2), 193-210.

- Broughton, V., Slavic, A. (2007). *Building a faceted classification for the humanities: principles and procedures*. J Doc 63(5), 727–754.

- Buchanan B.G., Lederberg J., *The Heuristic DENDRAL program for explaining empirical data*, Stanford University, technical report (1971).

- Dewey M. (1876), *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a library*, OCLC 78870163

- Dutta B., Giunchiglia F., Maltese V., *A facet-based methodology for geo-spatial modelling*, GEOS (2011).

- Etzioni O., Cafarella M. J., Downey D., Kok S., Popescu A., Shaked T., Soderland S., Weld D. S., Yates A., *Web-scale information extraction in KnowItAll*, WWW conference (2004).

- Giunchglia F., Maltese V., Dutta B., *Domains and context: First steos towards managing diversity in knowledge*, Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web (2012).

- Giunchiglia F., *Contextual reasoning*, Epistemologica - Special Issue on I Linguaggi e le Macchine, 16 (1993), 345–364.

- Giunchiglia F., Dutta B., Maltese V., *Faceted lightweight ontologies*. In *Conceptual Modeling: Foundations and Applications: Essays in Honor of John Mylopoulos*, pages 36–51, Berlin, Heidelberg, 2009. Springer-Verlag.

- Giunchiglia F., Dutta B., Maltese V., *From knowledge organization to knowledge representation*, DISI Technical Report (2013).

- Giunchiglia F., Maltese V., Farazi F., Dutta B., *GeoWordNet: a resource for geo-spatial applications,* Extended Semantic Web Conference ESWC (2010);

- Giunchiglia F., *Managing Diversity in Knowledge*, Invited Talk at the European Conference on Artificial Intelligence ECAI, Lecture Notes in Artificial Intelligence 2006.

- Giunchiglia F., Shvaiko P., Yatskevich M., *Discovering missing background knowledge in ontology matching*, European Conference on Artificial Intelligence ECAI (2006), 382–386;

- Giunchiglia, F., Dutta, B., Maltese, V. (2009). *Faceted Lightweight Ontologies*. In: Conceptual Modeling: Foundations and Applications, A. Borgida, V. Chaudhri, P. Giorgini, Eric Yu (Eds.) LNCS 5600 Springer.

- Giunchiglia, F., Kharkevich, U., Zaihrayeu, I. (2009). *Concept search*. European Semantic Web Conference (ESWC).

- Giunchiglia, F., Maltese, V., Dutta, B. (2012). *Domains and context: first steps towards managing diversity in knowledge*. Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web.

- Giunchiglia, F., Marchese, M., Zaihrayeu, I. (2007). *Encoding Classifications into Lightweight Ontologies*. Journal of Data Semantics, 8, 57-81.

- Giunchiglia, F., Shvaiko, P., Yatskevich, M. (2006). *Discovering missing background knowledge in ontology matching*. ECAI conference, 382–386.

- Giunchiglia, F., Zaihrayeu, I. (2008). *Lightweight ontologies*. Encyclopedia of Database Systems.

- Giunchiglia, F., Zaihrayeu, I., Kharkevich U. (2007). *Formalizing the get-specific document classification algorithm*. European Conference on Research and Advanced Technology for Digital Libraries.

- Gruber, T. R. (1993). *A translation approach to portable ontology specifications*. Knowledge Acquisition, 5 (2), 199–220.

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

- Horrocks, I., Sattler, U. (1999). *A description logic with transitive and inverse roles and role hierarchies*. Journal of Logic and Computation, 9(3), 385-410.

- Lauser B., Johannsen G., Caracciolo C., Keizer J., Van Hage W. R., Mayr P., *Comparing human and automatic thesaurus mapping approaches in the agricultural domain*, International Conference on Dublin Core and Metadata Applications (2008).

- Madalli D. P., Prasad A.R.D., Analytico synthetic approach for handling knowledge diversity in media content analysis, UDC seminar (2011). Another domain under development is the *food* domain.

- Magnini B., Speranza M., Girardi C., *A semantic-based approach to interoperability of classification hierarchies: Evaluation of linguistic techniques*, COLING (2004).

- Maltese V., Farazi F., *Towards the Integration of Knowledge Organization Systems with the Linked Data Cloud*, UDC seminar (2011).

- Matuszek C., Cabral J., Witbrock M., DeOliveira J., *An introduction to the syntax and content of Cyc*, AAAI Spring Symposium (2006).

- McCarthy J., *Generality in artificial intelligence*, Communications of ACM 30 (1987), 1030–1035.

- Mihalcea R., Moldovan D. I., *Automatic generation of a coarse grained WordNet*, NAACL Workshop on WordNet and Other Lexical Resources (2001).
- Miller G., *WordNet: An electronic Lexical Database*. MIT Press, 1998.
- Miller, G. A., Hristea, F. (2006). *WordNet Nouns: classes and instances*. Computational Linguistics, 32(1), 1 – 3.
- Mills, J. (2004). *Faceted classification and logical division in information retrieval*, Library trends, 52 (3), 541-570.
- Navigli R., Ponzetto S., *BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network*. Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250.
- Pease A., Sutcliffe G., Siegel N., Trac S., *Large theory reasoning with SUMO at CASC*, AI Communications, 23 2-3 (2010) 137–144.
- Pianta et al., *MultiWordNet: developing an aligned multilingual database*, Proceedings of the First International Conference on Global WordNet, Mysore, India, 2002.
- Prusak L., *Knowledge in Organizations*, Cap. 7: The tacit dimension by M. Polanyi, 1997.
- Ranganathan S.R., *The Colon Classification*, Rutgers Series on Systems for the Intellectual Organization of Information, S. Artandi (etd.), IV, Graduate School of Library Science, Rutgers University, New Brunswick, NJ, 1965.
- Ranganathan, S. R. (1967). *Prolegomena to library classification*, Asia Publishing House.

- Shvaiko P., Euzenat J., *Ten Challenges for Ontology Matching*, 7th Int. Conference on Ontologies, Databases, and Applications of Semantics, ODBASE, (2008);
- Spiteri L., *A Simplified Model for Facet Analysis, Journal of Information and Library Science*, 23 (1998), 1-30.
- Studer, R. et al., *Knowledge engineering: principles and methods*. Pennsylvania: School of Information Sciences and Technology (IST). Pennsylvania State University (1998).
- Suchanek F.M., Kasneci G., Weikum G., *YAGO: A Large Ontology from Wikipedia and WordNet*, Journal of Web Semantics (2011).
- Uschold, M., Gruninger, M. (2004). *Ontologies and semantics for seamless connectivity*. SIGMOD Rec., 33(4), 58–64.
- Varzi, A. C. (2006). *A note on the transitivity of parthood*. Applied Ontology, 1, 141-146.
- Vossen P., *Categories and classifications in EuroWordNet*, Proceedings of the First International Conference on Language Resources and Evaluation. Granada, 399-407, 1998; *Special Issue on EuroWordNet*, Computer and the humanities, 2-3, 73-251, 1998.
- Zaihrayeu, I., Sun, L., Giunchiglia, F., Pan, W., Ju, Q., Chi, M., Huang, X. (2007). *From web directories to ontologies: Natural language processing challenges*. International Semantic Web Conference (ISWC).

Webliography

- A complete list of Unix modes and file permissions: http://en.wikipedia.org/wiki/Modes_(Unix).
- Berne Convention for the Protection of Literary and Artistic Works (article 2) http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html#P85_10661
- DIG35 – Digital Imaging Group: http://xml.coverpages.org/FU-Berlin-DIG35-v10-Sept00.pdf
- Exchangeable Image Format specifications and related resources: www.exif.org
- The National Agricultural Library's Agricultural Thesaurus: http://agclass.nal.usda.gov/
- Agricultural Information Management Standards : http://aims.fao.org/website/AGROVOC-Thesaurus/sub
- OS X Manual Page – Apple developer: http://developer.apple.com/library/mac/documentation/Darwin/Reference/ManPages/man1/ls.1.html
- Yahoo directories: http://dir.yahoo.com/;
- Google directories: http://directory.google.com/
- The Open Directory Project (ODP) main page: http://dmoz.org/;
- Dublin Core Metadata Initiative: http://dublincore.org/

  - Extended Dublin Core qualifiers description page: http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/#description
  - Dublin Core Element specifications: http://dublincore.org/documents/dces/
- Wikipedia pages related to Intellectual Property and Creative work
  - http://en.wikipedia.org/wiki/Creative_work
  - http://en.wikipedia.org/wiki/Intellectual_property
  - http://en.wiktionary.org/wiki/creative_work
  - https://en.wikipedia.org/wiki/Database_right
- The Entitypedia project official page: http://entitypedia.org/
- Alcohol and Other Drug Thesaurus: http://etoh.niaaa.nih.gov/aodvol1/aodthome.htm
- GeoWordNet project main page: http://geowordnet.semanticmatching.org/
- Harvard University Library: http://hul.harvard.edu/ois/ldi/
- EventsML-G2 Standard web resources:
  - http://iptc.cms.apa.at/cms/site/single.html?channel=CH0087&document=CMS1206527645546
  - http://www.iptc.org/site/Home/
  - http://www.iptc.org/site/News_Exchange_Formats/EventsML-G2/
- Living Knowledge project webpage: http://livingknowledge-project.eu/
- Schema.org main page: http://schema.org/
  - Schema.org full taxonomy: http://schema.org/docs/full.html
- Microsoft Windows file system specifications:

  - http://technet.microsoft.com/en-us/library/bb490868
  - http://technet.microsoft.com/en-us/library/ee176615.aspx
- Unix man pages:
  - http://unixhelp.ed.ac.uk/CGI/man-cgi?ls;
  - http://www.gnu.org/software/coreutils/manual/html_node/chmod-invocation.html
- Bliss Classification Association: http://www.blissclassification.org.uk/
- GeoNames geographical database: http://www.geonames.org/
- Getty vocabularies and the Getty Thesaurus of Geographic Names:
  - http://www.getty.edu/research/conducting_research/vocabularies/tgn.
  - http://www.getty.edu/research/tools/vocabularies/index.html
- Library of Congress home page: http://www.loc.gov
- OpenCyc Ontology project: http://www.opencyc.org/
- OWL Web Ontology Language specifications: http://www.w3.org/TR/owl-features/
- The main page of Wikipedia project: http://www.wikipedia.org/
- World Intellectual Property Organization web resources:
  - http://www.wipo.int/freepublications/en/intproperty/909/wipo_pub_909.html#works
  - http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html#P82_10336

    - http://www.wipo.int/treaties/en/ip/paris/trtdocs_wo020.html#P71_4054
    - http://www.wipo.int/treaties/en/ip/wct/trtdocs_wo033.html#P56_5626
- WTO Intellectual Proerty (TRIPS) resources:
    - http://www.wto.org/english/docs_e/legal_e/27-trips_01_e.htm
    - http://www.wto.org/english/docs_e/legal_e/27-trips_04_e.htm
- A complete list of file attributes: http://www.xxcopy.com/xxcopy06.htm
- IPTC Photo Metadata - International Press Telecommunications Council Photo Metadata: http://www.iptc.org/std/photometadata/2008/specification/IPTC-PhotoMetadata-2008_2.pdf
- SVG - Scalable Vector Graphics W3C Specifications: http://www.w3.org/TR/SVG11/
- List of all Microsoft file and folders advanced and special permissions: http://technet.microsoft.com/en-us/library/bb727008.aspx