Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

DERA: A FACETED KNOWLEDGE ORGANIZATION FRAMEWORK

Fausto Giunchiglia, Biswanath Dutta

March 2011

Technical Report # DISI-11-457

Submitted to the International Conference on Theory and Practice of Digital Libraries 2011 (TPDL'2011)

DERA: A Faceted Knowledge Organization Framework

Fausto Giunchiglia and Biswanath Dutta

Department of Information Engineering and Computer Science, University of Trento, Via Sommarive, 14 I-38123, Povo, Trento, Italy
{fausto, bisu}@disi.unitn.it

Abstract. The availability of a priori knowledge, also called background knowledge, is fundamental for the functioning of semantics based systems. In this paper we introduce a faceted knowledge organization framework called *DERA* (for *Domain, Entity, Relation, Attribute*) and describe its implementation inside a system, called UK (for Universal Knowledge) which is extensible and scalable and which allows for fully automated reasoning via a direct encoding into Description Logics (DL). Extendibility and scalability is obtained by allowing the definition of any number of domains, where a domain is taken to be "*an area of knowledge or field of study that we are interested in or that we are communicating about*". In turn, a domain is organized into a number of facets where a facet is taken to be "*a hierarchy of homogeneous terms describing an aspect of the knowledge being codified, where each term denotes a primitive atomic concept*". Domains, facets, terms can be added at any time, and the different applications can use any subset of them. The direct encoding of DERA into DL is obtained by allowing only three types of facets (i.e., Entity, Relation, Attribute) which can be directly translated into DL concepts, roles, attributes, or into instances whose properties are encoded using the terms occurring in the facets themselves. The current implementation of UK contains around 377 Domains, out of which 115 are in priority for development, more than 150,000 terms (encoding concepts, relations and attributes), around 10,000,000 instances and more than 93,000,000 axioms codified using the terms codified in the DERA facets.

Keywords: Knowledge organization framework, background knowledge, domain ontology

1 Introduction

The availability of a priori knowledge, also called background knowledge (BK) is fundamental for the functioning of semantics based systems. Many approaches have been developed for using the existing knowledge sources as BK ranging from lexical knowledge (e.g., WordNet[1]) to domain specific knowledge sources (e.g., UMLS[2],

[1] http://wordnet.princeton.edu/
[2] http://www.nlm.nih.gov/research/umls/

AGROVOC[3], NALT[4] etc.) [1, 2, 4]. Some attempts have also been made of exploiting the semantic web as background knowledge [5]. All of these approaches agree on one point, i.e., the usefulness of the high quality and high quantity domain specific knowledge [1, 34, 35, 36]. This paper follows this line of thought and proposes a faceted knowledge organization (KO) framework (*a classificatory structure for developing knowledge sources*), a methodology and its implementation inside a system called UK (for Universal Knowledge). The proposed framework, called *DERA* (where *DERA* stands for *Domain, Entity, Relation, Attribute*) is extensible and scalable to extensively large, virtually unbound quantities of knowledge, and is based on the following ideas:

1. Knowledge should be organized in domains (*where a domain is an area of knowledge or field of study that we are interested in or that we are communicating about*);
2. Each domain should be organized into a number of facets (*where a facet is a hierarchy of homogeneous terms describing an aspect of the knowledge being codified, where each term in the hierarchy denotes a primitive atomic concept* [3]);
3. UK, its domains, its facets should be designed following the Analytico-synthetic approach, a well established methodology from the Library Science which has been successfully used for several decades for the classification of books [8].

Domains, facets, terms can be added at any time (thus making the system *extendable*) and the different semantic based applications can use any subset of them (thus making the system highly *modular*). *Scalability* comes from the possibility to use any domain independently of the number and size of the domains. Furthermore, DERA allows for fully automated reasoning via a direct encoding in Description Logics (DL) [6]. A DERA domain can in fact be taken to specify a domain of interpretation in DL; this allows DERA to inherit all the "usual" properties and features of DL, e.g., *soundness*, *decidability*, and *decision procedures*. Our target is > 98% accuracy. In order to achieve the desired high quality, the DERA domains have been built manually, or when built (semi) automatically, a lot of human validation was enforced. Manual work is very well known to be expensive, to take a lot of time and to be error prone. The current implementation of UK has been developed via a large use of manpower, at different levels of skills and competence. The key idea towards the full development of UK is to use crowdsourcing integrated with a certification pipeline based on ideas already exploited on ESP games [30]. However, this work is not described here as being still preliminary and also because of the lack of space. The current specification of UK contains around 377 Domains, out of which 115 are in priority for development, more than 150,000 terms (encoding concepts, relations and attributes), around 10,000,000 instances and more than 93,000,000 axioms codified using the terms codified in the DERA facets.

This paper is organized as follows. In Section 2 we introduce DERA, its characteristics, and its components. In Sections 3, 4 and 5 we describe the DERA elementary components, namely entities, relations and attributes respectively. In Section 6 we show how DERA can be directly encoded in DL. In Section 7 we

[3] http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[4] http://agclass.nal.usda.gov/agt/agt.shtml

describe about the current status of UK. In Section 8 we discuss the related work. Finally, in Section 9 we provide some conclusive remarks.

2 DERA

DERA is a faceted knowledge organization framework. It allows for the organization of knowledge into a number of facets by defining any number of domains. The framework is independent of any particular domain. The DERA framework is characterized by a set of features that, as far as we know, are not present in any of the previous knowledge organization frameworks and that allow us to deal with the problems highlighted in the introduction.

We take a domain to be an *area of knowledge or field of study that we are interested in or that we are communicating about*. In other words, a domain is an organized field of knowledge that deals with specific kinds of subjects (*in this context we define a subject to be any piece of non-discursive information that summarises what a book or document (any body of information) is about* [7]). Domains provide a bird's eye view of the whole field of knowledge. They also offer a comprehensive context within which one can have large scale search [9]. In addition, domains are the way to deal with the well-known homographic disambiguation problem [10]. In DERA, domains can be conventional fields of study (e.g., *library science, mathematics, physics*), applications of the pure disciplines (e.g., *engineering, agriculture*), any aggregate of such fields (e.g., *physical sciences, social sciences*), and they may also capture knowledge about our day-to-day lives, which we call the *Internet domains* (e.g., *music, movie, sport, space, time, recipes, tourism*).

When we classify the subject of a document, the description may essentially need the combination of a number of its properties [11]. For example, in classifying the subject of a document, "*microscopic diagnosis of bacterial viruses on cells in India*", we may have to include terms for its constituent's body and its parts, for behavior, for processes, for action carried out on the body, for agents, for interaction with other objects, and so on. The combination of all these terms would allow us to exhaustively pinpoint the subject of this individual document. Each element of a subject provides an independent aspect of possible interest to an enquirer and these separately listed *aspects* are known as ''*facets*'' [8, 11]. Note that, by facet we mean *a hierarchy of homogeneous terms describing an aspect of the knowledge being codified, where each term in the hierarchy denotes a primitive atomic concept*.

Facets are derived following the methodology and principles [8, 12] of facet analysis, a well established technique introduced by Ranganathan [8] for building classificatory structures from atomic concepts which are analyzed into facets and arranged by the application of the system syntax [13]. Two typical relations, namely *is_a* (genus/ species) and *part_of* (whole/part), are used as the main means for structuring the hierarchies within a facet. Detailed examples of facets are provided in the next sections.

Any DERA domain consists of three *elementary components* namely *entity*, *relation*, and *attribute* and can be expressed as follows:

$$D = \langle E, R, A \rangle$$

Where each component, itself often called facet, contains a set of facets of a specific kind as described below.

- *E = Entity* – an elementary component consisting of facets built of *classes* and their instances, having either perceptual correlates or only conceptual existence within a domain in context. For example, in the *Space* domain, *natural elevations*, such as, *mountain, hill, seamount* etc. are entity classes, while the *Himalaya, Monte Bondone, Loihi seamount* etc. are entities. An example of "E" facet is provided in Fig. 1 in Section 3.
- *R = Relation* – an elementary component consisting of facets built of classes representing the *relation* between entities. For example, in the *Space* domain, *north, south, near, adjacent, in front,* etc. are spatial relations between entities. An example of "R" facet is provided in Fig. 2 in Section 4.
- *A = Attribute* – an elementary component consisting of facets built of classes denoting the *qualitative/ quantitative* or *descriptive properties* of entities. For example, in the *Space* domain, *altitude* (of a hill), *length* (of a river), *surface area* (of a lake), etc. are qualitative/ quantitative properties, while the *kinds of rocks* (of a mountain), *architectural style* (of a monument) are descriptive attributes. Two examples of "A" facets are provided in Fig. 3, 4 in Section 5.

3 Entities

An entity is something that has a distinct, separate existence, though it needs not be a material existence. According to Bhattacharyya [7], entity is "*an elementary category that includes manifestations having perceptual correlates or only conceptual existence, ...*". We define an entity as "*an elementary component that consists of classes (categories) and their instances, having either perceptual correlates or only conceptual existence in a domain in context"*. An entity can be therefore expressed as the pair:

$$E = <\{e\}, \{\mathcal{E}\}>^5$$

where,

- e = *Entity class* – consists of the core classes within a domain;
- $\mathcal{E}$ = *Entity* – consists of the real world (named) entities which are instances of the entity classes "e".

An entity Class (e) is the main means to denote what an object is. Every entity class is uniquely defined via its extension, i.e., the set of entities to which it refers. For example, in the *Space* domain, the extension of the class *mountain* is the set of real world mountains. An entity class represents the essence of the domain under consideration. It consists of the classes that represent the core idea of a domain, and does not contain the classes exposing the properties (e.g., quantitative, qualitative, etc.) of entities. To exemplify, *house, hut, school, hill, mountain* are core classes in the *Space* domain, while classes like, *latitude, longitude, altitude, architectural style, kind of rocks* are not. Similarly, *comedy, wacky comedy, horror, drama, spoof, vampire, monster, demon* are the core classes in context to a domain *Movie*.

[5] Notationally, by "{c}", we mean the set of objects c.

Within each entity class "e", the core classes are organized as facets. Fig.1(a) shows the facet *body of water* belonging to the *entity class* in the *Space* domain. The facet *body of water* is further divided into its sub-facets *stagnant body of water* and *flowing body of water*. We also see that the sub-facet *flowing body of water* is further divided into its sub-facets *natural flowing body of water* and *artificial flowing body of water*. Each of these facets further subsumes the classes like, *Stream, River, Brook, Canal, Aqueduct*, and so forth as shown in Fig. 1(a).

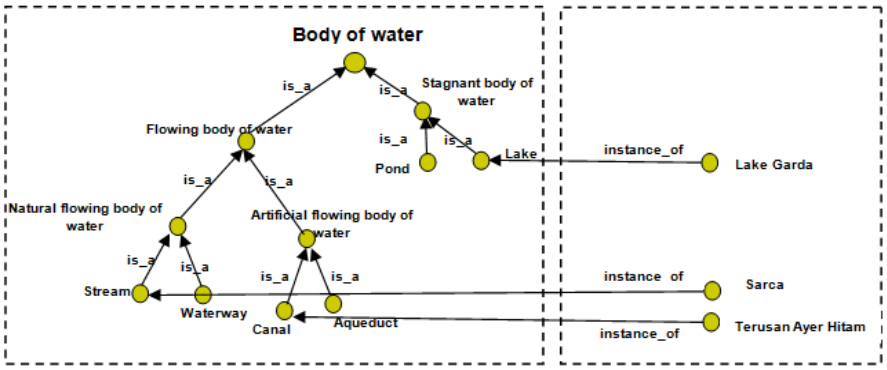

Fig. 1: 1(a). A fragment of the *body of water* facet

Fig. 1(b). Entities in instance_of relation with their entity classes

By the entities (℮), we mean real world named entities. The idea of using entities as modelling constructs to represent instances of things is widely held. Coad and Yourdon [14] for instance argue that an entity is "*an abstraction of something in the problem domain*". Similarly, Chen [15] argues that, "*an entity is a 'thing' which can be distinctly identified*". In DERA, entities are linked with the entity classes by the *instance_of* relation. For instance, *Lake Garda instance_of Lake*; while the linkages between entities are established by *part_of* relation (not shown in the figure). For instance, *West Bengal part_of India*, *India part_of Asia*. Fig. 1(b) presents the entities against their entity classes. For instance, we have *Sarca instance_of Stream*, *Terusan Ayer Hitam instance_of Canal*.

4 Relations

This elementary *component consists of facets built of relations inside a domain"*. Relations play an important role for effective knowledge discovery. Consider for instance the following queries:

- Retrieve all the secondary schools *within* 500 meters of the Dante railway station in Trento.
- Find all the highways of the Trentino province *adjacent* to marine areas.

within and *adjacent* are two relations of the *Space* domain which describe the spatial relation between two entities. Some other important examples of relations (in context of other domains) are: *friend, father, mother,* etc. describing *social relations* between two persons; *born_in*, *lives_in*, etc. describing relations between a person and a

location; *painter* describing a relation between a painting and a person. The elementary component relation is defined:

$$R = <\{r\}>$$

where

- r = *Relation* - consists of the classes representing the relations between entities.

A relation is a mutual property (one or more) of a thing in the real world [16]. More precisely, a relation is a link between two entities. According to Stockdale and Possin [17], a relation can be between oneself and the environment or between two or more objects outside of oneself. Each relation builds a semantic relation between two entities. Relations are also structured into facets. For instance, *spatial relation* is a relation facet within the Space domain. The *spatial relation* facet can have any number of sub-facets, for example, *Direction*, *Internal spatial relation*, *External spatial relation*, *Position in relation to border or frontier*, *Longitudinal spatial relation*, *Sideways spatial relation*, *Relative level* and so forth (for a detailed view of these facets see [12]). Fig. 2 (right side) shows two such sub-facets *External spatial relation* and *Internal spatial relation*. Fig. 2 also demonstrates how a relation can be used. For example, by using a relation *near*, we express the knowledge that *Lake Caldonazzo* is *near Lake Garda*.

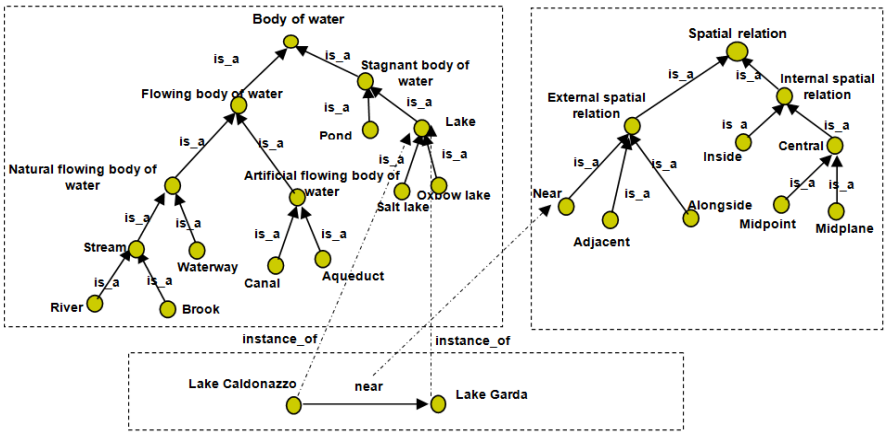

Fig. 2. An extension of Fig. 1 with an additional relation facet

Note that, in some cases, classes belonging to the entity class (e) facet of a domain can be reused as relations. For example, the domain *Agent* is designed as a common-purpose domain[6] and some of the facets belonging to the entity class of this domain are *biological agent* (e.g., *bacteria, virus*), *profession* (e.g., *actor, teacher*) and so forth. The facet *profession* can be partially reused as relation facet within a *Movie* domain. This is because the classes (e.g., *actor, actress, director*) belonging to the facet *profession* are basically the roles (*actions and activities assigned to or required*

[6] Common-purpose domains are domains can be used for common purposes and can be reused fully or partially in the context of any other domains. For example, the entity class facet of a general-purpose domain *Material* can be reused in context of other domains like, *Numismatics*, *Sculpture*, etc.

or expected of a person or group) played by the agents in the *Movie* domain (here *role* is used with the meaning defined in [23]).

5 Attributes

This elementary component consists of classes belonging to or that are characteristic of entities. Entities can be distinguished through attributes. Attributes are effective for Named Entity Recognition (NER) [18] and for efficient information retrieval [19]. For example, in the current version of UK there are 14 locations called *Rome* in *United States of America* (USA), one in *Italy* (the capital city of *Italy*) and one in *France*. Using the *latitude* and *longitude* we can easily distinguish them [12]. Attributes are primarily "*qualitative/ quantitative*" and *descriptive* in nature. As a consequence we define two kinds of attributes:

$$A = <\{\mathcal{A}\}, \{\wp\}>$$

where,

- $\mathcal{A}$ = *Datatype attribute* – consists of classes which qualify or quantify the properties of entities;
- $\wp$ = *Descriptive attribute* – consists of classes describing entities.

A datatype attribute ($\mathcal{A}$) includes the attributes that specify the quality or quantity of the entities within a domain. Consider for example, *deep lakes*; here, *deepness* is a datatype attribute that can be shared by all deep lakes. On the other hand we could also quantify the exact *depth* of the lake (e.g., 346 m). Similarly consider for instance, *red car*; here, *redness* is a datatype attribute that can be shared by all red cars.

For each of the datatype attributes (whenever applicable), DERA allows for storing the possible qualitative values in the knowledge-base along with their attribute names. This provides a controlled vocabulary for them. The attribute values are mostly *adjectives*, whereas in some cases they are *intransitive verbs*. For example, in the *Space* domain, some of the datatype attributes are, *latitude, longitude, height, length, width, depth, altitude, population, climate*, and so forth. The values encoded for the attribute *depth* are {deep, shallow}; similarly the values for *length* are {long, short}. In linking the attribute values with their corresponding attribute names, we use the relation *attribute* when the values are *adjectives* (see Fig. 3). We use the relation *attribute*, because for instance, *deep* is not a kind of *depth*, instead it is an attribute that qualifies the *depth*.

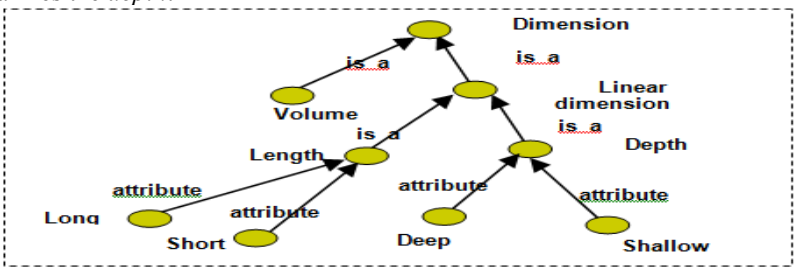

Fig. 3: A fragment of a datatype attribute facet.

A descriptive attribute ($\wp$) is a facet consisting of attributes that describe the entities under a domain in consideration. A descriptive attribute describes entities (as one would expect). For example, consider the fact that "India is a democratic country". This statement entails the knowledge that the *political system* of a country India is a *democracy*. In the *Space* domain, *political system* can be treated as a descriptive attribute, while *democracy* stands as a possible value. Here, *political system* is a descriptive attribute, primarily because of its descriptive behavior that characterizes the Indian political system. In analogy to datatype attributes, in case of descriptive attributes, DERA allows to store the possible values along with their descriptive attribute names. The values could be atomic or compound concepts. For example Fig. 4 shows an example of a descriptive attribute namely *architectural style* of a monument and the corresponding possible set of values.

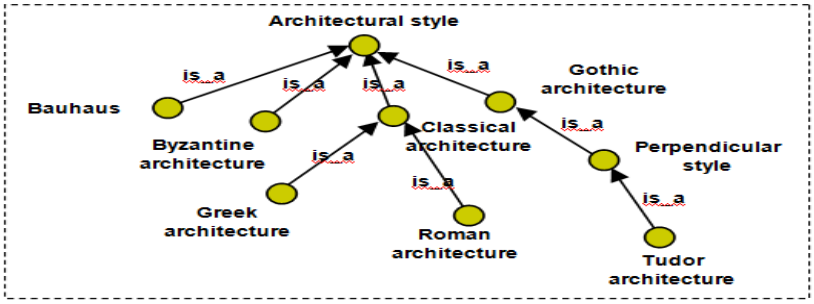

Fig. 4. A fragment of a descriptive attribute facet.

6 From DERA to Description Logics

DERA allows for the definition of any number of domains. In turn, any such domain can be formalized as a Description Logics (DL) theory. The DL formalization of a domain is a direct encoding from the DERA facets into DL formulas and is done by modeling the three components (i.e., Entity, Relation, Attribute) as DL concepts, roles, attributes or into instances whose properties are encoded using the terms occurring in the facets. In the following of this section we describe how in DL it is possible to define entity classes, entities and relations, and to build facets.

Entity classes are formalized as atomic concepts. Relations and attributes are formalized as DL roles. Entities are formalized as DL individuals.

$e_1,\ldots,e_m$	(entity classes)
$\mathcal{E}'_1,\ldots,\mathcal{E}'_n$	(entities)
$R_1,\ldots,R_s$	(relations)
$\mathcal{A}_1,\ldots\mathcal{A}_t$	(datatype attributes)
$\wp_1,\ldots,\wp_u$	(descriptive attributes)

where $e_m(i = 1,\ldots,m)$ are concepts for entity classes, $\mathcal{E}_n(j = 1,\ldots,n)$ are individuals for entities, $R_k(k = 1,\ldots,s)$ are roles for relations, $\mathcal{A}_x(x = 1,\ldots,t)$ are roles for datatype attributes, $\wp_u(y = 1,\ldots,u)$ are roles for descriptive attributes.

An *Interpretation* $\mathcal{I}$ of a DERA domain consists of an Interpretation Function *I* and a non empty set D (the Domain of Interpretation) of entities, namely,

$$\mathcal{I} = \langle D, I \rangle$$

D contains the set of entities ($\mathcal{E}^I$) which provide the extensions of concepts, relations, datatype attributes and descriptive attributes e^I, R^I, $\mathcal{A}^I$, and $\mathcal{C}^I$ respectively. Thus, for instance, $Lake^I \in e^I$ is a concept with name *Lake*, while $Lake\ Garda^I \in \mathcal{E}^I$ is an individual for a concept *Lake*. Similarly, we interpret a relation R as a binary relation $R^I \subseteq D \times D$, a datatype attribute $\mathcal{A}$ as a binary relation $\mathcal{A}^I \subseteq D \times D$ and a descriptive attribute $\mathcal{C}$ as a binary relation $\mathcal{C}^I \subseteq D \times D$. To sum up, we have therefore:

$$e^I \subseteq D, \qquad \mathcal{E}^I \in D, \qquad R^I \subseteq D \times D, \qquad \mathcal{A}^I \subseteq D \times D, \qquad \mathcal{C}^I \subseteq D \times D$$

We formulate the DERA facets as subsumption axioms, namely as axioms of the form Ai $\sqsubseteq$ Aj, where Ai, Aj can be entity classes, relations, datatype attributes and descriptive attributes. For instance, the left nodes of Fig. 1(a), right side and the lower right nodes of Fig. 2, the left node of Fig. 3, and left node of Fig. 4 are axiomatized as follows:

FlowingBodyOfWater $\sqsubseteq$ *BodyOfWater*
NaturalFlowingBodyOfWater $\sqsubseteq$ *FlowingBodyOfWater*
Stream $\sqsubseteq$ *NaturalFlowingBodyOfWater*
InternalSpatialRelation $\sqsubseteq$ *SpatialRelation*
Central $\sqsubseteq$ *InternalSpatialRelation*
Midplane $\sqsubseteq$ *Central*
Volume $\sqsubseteq$ *Dimension*
Bauhaus $\sqsubseteq$ *ArchitecturalStyle*

Notice that, following the standard for Analytico-synthetic approach (for related work, see in [3]) as defined originally in Library Science, there is no need to use disjointness or negations, thus leading to the use of a rather inexpressive version of DL (with individuals).

7 UK - the Universal Knowledge

For the last four years, while refining the DERA methodology, we have used it to develop what has now become an ever growing, large scale, knowledge organization system, that we call UK. The first step in the implementation of UK was to build the first universal domain i.e., *everything*. This domain was built by uploading WordNet 2.1. We started with WordNet because of its size and quality. We uploaded 117,597 synsets, 354,057 relations, 147,252 terms and 207019 senses from WordNet. We also uploaded 33,156 synsets, 45,156 terms and 59,656 synsets from the Italian MultiWordNet[7].

After implementing what constitute the first version of the universal domain, called *everything*, the next step was to build a second domain, namely *Space*. Our goal was to create large-scale semantically enriched geo-spatial knowledge-base. Unfortunately

[7] http://multiwordnet.fbk.eu/english/home.php

WordNet has quite limited coverage in geo-spatial information and lacks of latitude and longitude coordinates [20]. Therefore, it was essential to look elsewhere as we wanted an adequate amount of geo-spatial information. We evaluated several geo-spatial related information resources that include Wikipedia[8], DBPedia[9], GEMET[10] and the ADL gazetteer[11], but they are limited either in locations, classes, relations or metadata. GeoNames[12] and TGN[13], instead, both met our requirements. As a result we developed GeoWordNet, a semantic resource (now available as open source[14]), which is the outcome of the full integration of GeoNames, with TGN and WordNet and the Italian part of MultiWordNet (see in [21] for details).

At this early stage we had nearly 7 million locations from all over the world. But we wanted to test extendibility of the UK. We achieved this thanks to the SGC project in collaboration with the Autonomous Province of Trento (PAT) in Italy. In this project a dataset of 20,162 locations of the province was analyzed and integrated with the GeoWordNet. We also automatically generated an Italian and English gloss for each entity imported from PAT. The inclusion of PAT data into our knowledge-base provided some evidence that the UK is flexible and extendable. In fact limited to the area we considered, we moved from 2,000 to around 18,000 locations and at the same time we had to add only a few entity classes, relations and attributes. After the *Space* domain we concentrated on the second most significant domain i.e., *Time*. In its current implementation, the Time domain consists of 157 entity classes, 3 relations and 53 attributes.

As a next step we imported 600,000 locations from YAGO[15]. In addition we also imported 719,512 persons and 153,764 organizations (Table 1 provides detailed statistics about the current size of UK). The uploading of these general-purpose entities (e.g., *person, organization, video, song*, etc.) allowed us to create the basis for the development of a large number of domains. To exemplify, *person* entities are linked to domains like, *Medicine, Literature, Movie, Music, Painting, Sculpture* and so forth.

Table 1. Detailed statistics about the current size of UK

Object	**Number**
Concepts	110,609
Relations	204,481
Axioms	93,000,000
Entities	9,500,000

However, it is worthwhile noting that since the knowledge in WordNet is organized as per the linguistic structure, it was not useful for us to use it in its original

[8] http://www.wikipedia.org/
[9] http://dbpedia.org/
[10] http://www.eionet.europa.eu/gemet/
[11] http://www.alexandria.ucsb.edu/gazetteer/
[12] http://www.geonames.org/
[13] http://www.getty.edu/research/tools/vocabularies/index.html
[14] http://geowordnet.semanticmatching.org/
[15] http://www.mpi-inf.mpg.de/yago-naga/yago/

form. We had therefore to organize the UK as a set of facets and domains (initially only the universal domain *everything*). This work is leading to a profound restructuring of the original WordNet structure. In fact, while being good from linguistic point of view, WordNet presents many problems from knowledge organization point of view. Table 2 presents one such example, which shows how the notion of domain based faceted Knowledge Organization system has led us in restructuring WordNet. The left column of Table 2 shows a toy example of the *Climatography* facet (*the description or study of climate*) consisting of the concepts *Weather*, *Atmospheric pressure*, *Humidity*, *Cloud*, *Rainfall*, *Snow*, has been developed in the context of *Physical Geography* domain, while the right column shows the position of those concepts in the sub-tress of original WordNet.

Table 2: Comparison of classifications following the notions of facet based knowledge organization system and linguistic approach

UK	WordNet sub-trees

By Climatography
 	
• Weather
• Atmospheric pressure
• Humidity
  o Cloud
  o Rainfall
  o Snow
	entity > physical entity > process > phenomenon > natural phenomenon > physical phenomenon > atmospheric phenomenon > **weather**
	
	 entity > physical entity > process > phenomenon > natural phenomenon > physical phenomenon > pressure > gas pressure > **atmospheric pressure**
	
	 entity > abstract entity > abstraction > state > condition > wetness > **humidity**
	
	 entity > physical entity > process > phenomenon > natural phenomenon > physical phenomenon > atmospheric phenomenon > **cloud**
	
	 entity > physical entity > process > phenomenon > natural phenomenon > physical phenomenon > atmospheric phenomenon > weather > precipitation > **rainfall**
	
	 entity > physical entity > process > phenomenon > natural phenomenon > physical phenomenon > atmospheric phenomenon > weather > precipitation > **snow**

As part of the definition of the domains we have carried out a very thorough research and identified a set of 377 domains and sub-domains as reported below. We have defined these domains after a careful analysis[16] of a query log of 20,000,000 queries from America Online (AOL). The entire set of domains was divided into three groups namely,

[16] This work has been carried out by A.R.D. Prasad, D. P. Madalli and their research team at DRTC, ISI, Bangalore, India as part of the LivingKnowledge project. A detailed report describing this work is being wrtitten.

- *primary domains (17 in total),* namely, domains at the first-level of the hierarchy of domains, e.g., *Health*, *Computer*, *Arts*;
- *sub-domains (350 in total),* namely, domains beneath the primary domains, e.g., *Artificial Intelligence*, *Social Networking*, *Fine arts*;
- *common-purpose domains(10 in total),* namely, domains can be used for common purposes and can be reused fully or partially in the context of any other domains. For example, the *entity class* facet of a general-purpose domain *Material* can be reused in context of other domains like, *Numismatics*, *Sculpture*, etc.

Out of the 377 domains we enlisted 115 top-priority domains, for example, e.g., *Space, Time, Food, Recipe, Hotel, Sports, Tourism, Medicine, Agents, Social relations, Software, Hardware, Social networking, Artificial Intelligence (AI), Sculpture, Drawing, Plastic arts, Music, Real estate, Political system*, *Transportation*. Some of them have either already implemented or under implementation. To exemplify, the current implementation of the domain Movie consists of 196 entity classes, 21 relations and 30 attributes, while the current implementation of the domain Sports consists of 263 entity classes, 23 relations and 29 attributes.

8 Related Work

We split the related work in two parts.

Knowledge Organization Frameworks. In traditional libraries, fully faceted classification systems like the Colon Classification (CC), the Bibliographic Classification[17] (BC) and partially faceted classification systems like the Universal Decimal Classification[18] (UDC) are very popular as Knowledge Organization Systems (KOS). They have been used for several decades as knowledge organization (KO) tools in libraries for classifying and shelving the library documents. DERA uses the Analytico-synthetic approach and as such, it is a direct evolution of Ranganathan's Colon Classification [22] which is where we focus our comparison in the following of this section.

Ranganathan (1933), in his colon classification defined five fundamental categories in which to arrange facets: Personality [P], Matter [M], Energy [E], Space [S] and Time [T], plus an additional category to characterize the domain, called Basic Subject [BS]. The Classification Research Group (1960s), in its Bibliographical Classification System [27], further refined the Ranganathan's fundamental categories into thirteen categories: Thing/entity, Kind, Part, Property, Material, Process, Operation, Patient, Product, By-product, Agent, Space, and Time. Similarly, Bhattacharyya (1975) in describing his subject indexing technique called POPSI [7] proposed five categories: Domain [D], Entity [E], Property [P], Action [A] and Modifier [m]. Modifiers include those facets which can be used across the domains such as Space, Time, Form and Language.

We share with these systems the key notion that facets allow modeling domain specific knowledge by exploiting and making explicit the different aspects of

[17] http://www.blissclassification.org.uk/
[18] http://www.udcc.org/about.htm

knowledge within a domain. The previous facet based systems proved their usefulness and effectiveness in organizing and searching conventional library documents [24]. However their major drawback was into their structure. All these systems fail in making explicit the way the meaning (semantics) of subjects (what the document is about) is built starting from the semantics of their constituents. In fact, they only consider the syntactic form by which subjects are described in natural language (syntax). Consequently, they do not allow for a direct translation of their elements - terms and arcs in the facets - into a formal language, e.g. in form of DL axioms. They do not explicitly specify the taxonomical is-a (genus/ species) relation and mereological part-of (whole/ part) relation between the classes. This makes reasoning in these systems very hard to automate, which is instead the main advantage and goal of DERA.

Large Scale Knowledge Sources. In the last 3-4 decades, there have been many attempts of constructing large scale knowledge resources. Primarily they can be categorized into two: handcrafted and automatically extracted [25]. In case of handcrafted knowledge sources, two highly ambitious live projects can be named, such as, WordNet and Cyc[19]. WordNet is one of the most often referred sources. It is widely used as background knowledge to support numerous language processing tasks. It is a lexical source organized the words in English language according to their meaning and terms are related by the terminological relations such as synonymy, hypernymy/hyponymy, meronymy/holonymy, antonymy, and derivational. On the other hand, Cyc is a general purpose commonsense knowledge source has been developed by mapping or integrating several number of ontologies including SENSUS, FIPS 10-4, pharmaceutical thesauri, large portions of WordNet, MeSH/Snomed/UMLS and CIA World Factbook [26].

Examples of automatically generated knowledge sources are YAGO, DBPedia and Freebase[20]. YAGO is built by combining the Wikipedia categories with WordNet. It includes the relations extracted from Wikipedia infoboxes. YAGO consists of 10,000,000 entities (include concepts, relations and individuals) and more than 80,000,000 facts [29]. DBPedia is a large scale repository of assertions extracted from Wikipedia. It covers entity types such as geographic information, people, companies, films, music, genes, drugs, books and scientific publication and consists of 2,600,000 entities including 198,000 persons, 328,000 places, 101,000 musical works, 34,000 films and 20,000 companies [28]. Freebase, a social knowledge source consists of concepts and axioms automatically extracted from Wikipedia and merged them with other resources (e.g., Baseball Almanac, Chickipedia, MusicBrainz, the Notable Names Database). It consists of approximately 20,000,000 entities.

The first fundamental differences between UK and this work is that UK is organized by domains and each domain by a set of facets. The main advantage of UK is that, since each facet encodes a homogeneous group of terms, it can be grown up over time without restructuring the entire UK. Furthermore, since each facet describes an aspect of a domain, hence, it can be reused across many domains. Finally, UK is developed according to a precise methodology which follows the Analytico-synthetic

[19] http://cyc.com/cyc/technology/whatiscyc
[20] http://www.freebase.com/

approach and it maintains a very high level of consistency (see [12] for a detailed explanation of how the DERA methodology is applied to the Space domain).

9 Conclusion

In this paper we have introduced DERA, a new Faceted Knowledge Organization Framework which allows for the use of domains and facets and as a consequence, to deal with problems such as extendibility, modularity and scalability and of the automation of reasoning using DL. We have also shown how the use of DERA has allowed us to develop a large scale knowledge base that we call UK, which contains millions of entities, thousands of terms and tens of millions of axioms. A lot of the UK has been developed manually or at least with a thorough manual quality control. This allows for a level of correctness and data quality which as far as we know, is quite unique.

Acknowledgement. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231126 "LivingKnowledge: LivingKnowledge - Facts, Opinions and Bias in Time". Thanks to Vincenzo Maltese and Feroz Farazi for their continuous and intense work on the data sets. We also want to thank Ilya Zaihrayeu and Marco Marasca for their contribution to the definition of the data structures of our implementation of DERA and Abdelhakim Freihat for importing the Italian MultiWordNet.

References

1. Aleksovski, Z., Klein, M., ten Katen, W. and Harmelen, F. van: Matching Unstructured Vocabularies using a Background Ontology. In: S. Staab and V. Svatek, editors, Proc. of EKAW, LNAI. Springer-Verlag (2006).
2. Giunchiglia, F., Marchese, M., and Zaihrayeu, I.: Encoding Classifications into Lightweight Ontologies. In: Proc. of the 3rd European Semantic Web Conference (ESWC 2006), Budva pp. 80--94 (2006).
3. Guinchiglia, F., Dutta, B., Maltese, E.: Faceted Lightweight Ontologies. In: Conceptual Modeling: Foundations and Applications. A. Borgida, V. Chaudhri, P. Giorgini and Eric Yu (Eds.), LNCS, Vol. 5600, pp. 36--51, Springer-Verlag, Heidelberg (2009)
4. Stuckenschmidt, H., Harmelen, F. van, Serafini, L., Bouquet, P. and Giunchiglia, F.: Using C-OWL for the Alignment and Merging of Medical Ontologies. In Proc. of the First Int. WS. on Formal Biomedical K. R. (KRMed) (2004).
5. Sabou, M., d'Aquin, M. and Motta, E.: Using the Semantic Web as Background Knowledge for Ontology Mapping. In Proc. of the Int. Workshop on Ontology Matching (OM-2006).
6. Baader, F., Calvanese, D., McGuinness, D, Nardi, D., Patel-Schneider, P. F.: The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press (2002)

7. Bhattacharyya, G.: POPSI: its fundamentals and procedure based on a general theory of subject indexing languages. Library Science with a Slant to Documentation, 16 (1), pp. 1--34 (1979)
8. Ranganathan, S. R.: Prolegomena to library classification. Asia Publishing House, London (1967)
9. Mills, J.: Faceted classification and logical division in information retrieval. Library Trends, 52(3), pp. 541--570 (2004)
10. Ciaramita, M. and Altun, Y.: Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In: Conference on Empirical Methods in Natural Language Processing (EMNLP, Sydney, Australia), pp. 594--602 (2006)
11. Vickery, B.: Faceted classification for the Web. Axiomathes, 18, 145--160 (2008)
12. Dutta, B., Giunchiglia, F., Maltese, V.: A facet-based methodology for geo-spatial modelling. In: Claramunt, C., Levashkin, S. and Bertolotto, M. (Eds.) GeoS-2011, LNCS, vol. 6631, pp. 133--150, Springer-Verlag, Berlin Heidelberg (2011)
13. Broughton, V.: Building a Faceted Classification for the Humanities: Principles and Procesures. Journal of Documentation (2007)
14. Coad, P. and Yourdon, E.: Object-oriented analysis. In: 2nd ed. Yourdon Press Computing Series. Yourdon Press, Upper Saddle River, N.J. (1991)
15. Chen, P. P.: The entity-relationship model: toward a unified view of data. In: ACM Trans. Database Syst., 1(1), pp. 9-36 (1976)
16. Wand, Y., Story, V. C., and Weber, R.: An ontological analysis of the relationship construct in conceptual modeling. In: ACM Transactions on Database Systems, 24(4), pp. 494--528 (1999)
17. Stockdale, C. and Possin, C.: Spatial Relations and Learning. (2001). http://www.newhorizons.org/spneeds/inclusion/teaching/stockdale.html
18. Y. Kalfoglou, M. Schorlemmer. Ontology mapping: the state of the art. Knowledge Engineer Review, 18(1), pp. 1--31 (2003).
19. Jones, C. B., Adbelmoty, A.I., G. Fu: Maintaining Ontologies for Geographical Information Retrieval on the Web. In Proc. of On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, LNCS (2003)
20. Buscardi, D., Rosso, P.: Geo-wordnet: Automatic Georeferencing of wordnet. In: Proc. of the 5th Int. Conference on Language Resources and Evaluation (LREC) (2008).
21. Giunchiglia, F., Maltese, V., Farazi, F., and Dutta, B.: GeoWordNet: a Resource for Geo-Spatial Applications. In: Aroyo et al (Eds.), LNCS, vol. 6088, pp. 121--136. Springer-Verlag, Berlin Heidelberg (2010)
22. Ranganathan, S. R.: Colon Classification. SRELS, Bangalore (1987)
23. Guarino, N. Concepts, Attributes and Arbitrary Relations: some Linguistic and Ontological Criteria for Structuring Knowledge Bases. (1992). http://www.loa-cnr.it/Files/DKE92.pdf
24. Broughton, V. The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings, 58(1/2), pp. 49--72 (2006).
25. Nastase, V., Strube, M., Borschinger, B., Zirn, C. and Elghafari, A. WikiNet: a very large scale multi-lingual concept network. In: Proc. of the 7th International Conference on Language Resources and Evaluation, La Valetta, Malta, pp. 17--23 (2010). http://www.lrec-conf.org/proceedings/lrec2010/pdf/
26. Reed, S. L. and Lenat, D. B.: Mapping ontologies into Cyc. In: AAAI 2002 Conference Workshop on Ontologies for the Semantic Web. Edmonton, Canada (2002).
27. Mills, J. and Broughton, V.: Bliss Bibliographic Classification. 2nd ed. London: Butterworth and Bowker-Saus (1977).
28. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellmann, S.: DBPedia- a crystallization point for the web of data. In: Web Semantics: Sci. Serv. Agents on the World Wide Web (2009).

29. Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-Kelham, E., de Melo, G. and Weikum, G.: YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages. In: Proc. of the 20th international conference companion on World Wide Web (WWW 2011) (2011).
30. Von Ahn, L.: Games With A Purpose. IEEE Computer Magazine. pp 96—98 (2006).
31. Guarino, N. and Welty, C.: An Overview of OntoClean. In Steffen Staab and Rudi Studer, eds., The Handbook on Ontologies, Berlin:Springer-Verlag, pp. 151-172 (2004).
32. Oltramari, A., Gangemi, A., Guarino, N. and Masolo, C.: Restructuring Wordnet's Top-level: the OntoClean Approach. pp. 17--26 (2002).
33. Hage, W. R. van, Katrenko, S. and Schreiber, G.: A Method to Combine Linguistic Ontology-Mapping Techniques. In: Proc. Of ISWC. pp. 732-744 (2005).
34. Lauser, B., Johannsen, G., Caracciolo, C., Keizer, J., Hage, W. R. van and Mayr, P.: Comparing Human and Automatic Thesaurus Mapping Approaches in the Agricultural Domain. In: Int'l Conf. on Dublin Core and Metadata Applications (2008).
35. Shvaiko, P. and Euzenat. J.: Ten Challenges for Ontology Matching. In: 7th Int. Conference on Ontologies, Databases, and Applications of Semantics, ODBASE, (2008).
36. Aleksovski, Z., Kate, W. ten and Harmelen, F. van: Using Multiple Ontologies as Background Knowledge in Ontology Matching. In: ESWC workshop on collective semantics (2008).