Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 5 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

FROM KNOWLEDGE ORGANIZATION TO KNOWLEDGE REPRESENTATION

Fausto Giunchiglia, Biswanath Dutta,
Vincenzo Maltese

June 2013

Technical Report # DISI-13-027

From Knowledge Organization to Knowledge Representation

Fausto Giunchiglia[1], Biswanath Dutta[2], Vincenzo Maltese[1]

[1] DISI – University of Trento, Trento, Italy
[2] DRTC - Indian Statistical Institute, Bangalore, India

Abstract. So far, within the Library and Information Science (LIS) community, Knowledge Organization (KO) has developed its own very successful solutions to document search, allowing for the classification, indexing and search of millions of books. However, current KO solutions are limited in expressivity as they only support queries by document properties, e.g., by title, author and subject. In parallel, within the Artificial Intelligence and Semantic Web communities, Knowledge Representation (KR), has developed very powerful end expressive techniques which, via the use of ontologies, support queries by any entity property (e.g., the properties of the entities described in a document). However, KR has not scaled yet to the level of KO, mainly because of the lack of a precise and scalable entity specification methodology. In this paper we present DERA, a new methodology, inspired by the faceted approach, as introduced in KO, that retains all the advantages of KR and compensates for the limitations of KO. DERA guarantees at the same time quality, extensibility, scalability and effectiveness in search.

1 Introduction

So far, within the LIS community, KO has dealt with and developed its own very successful solutions, in terms of methodologies, systems and tools, for the classification, indexing and search of documents in libraries and digital archives. Documents are indexed and searched by their properties such as title, author and subject (the latter codifying what a document is about). Controlled vocabularies are employed in order to standardize the subject terminology, thus ensuring high precision in search. Recall is increased by expanding terms in queries with synonyms and more specific terms taken from the controlled vocabulary. Historically, this approach has scaled as it allows for the classification, indexing and search of millions of books, though at very high costs of training and maintenance (Library of Congress 2007). Several methodologies have been developed for the construction and maintenance, often centralized, of controlled vocabularies. Among them, the *faceted approach* (Ranganathan 1967) is known to have great benefits in terms of quality and scalability of the developed resources (Broughton 2006) (Broughton 2008). The above techniques are very effective for what concerns

searches exploiting document properties. A typical example of supported query is the following:

Give me documents with *author* "Nash, David" and *subject* "wood sculpture"

However, KO is limited in expressivity as it fails in situations when users do not know such properties directly, but they rather know, for instance, properties of the author or of any other entity the document is about, and want to search accordingly. For example, users may formulate the search need above as follows:

Give me documents about wood sculptures written by an artist born in Wales

The need for such kind of more expressive queries is proved by the fact that database and KR communities have spent decades in developing highly expressive query languages. It is enough to think to SQL within database management systems (Ramakrishnan and Gehrke 2000) for the first and SPARQL to query RDF (Prud'hommeaux and Seaborne 2006) for the second. Their usefulness (as well as limitations) is proved by plenty of studies. Questions like the ones suggested by us, meaning by queries requiring the same level of expressiveness, are in everyday use in many applications.

Addressing the query above in KO would require breaking it down into smaller search tasks and rely on scattered resources, such as catalogues and authority lists, to get all the relevant information which is necessary to reformulate it in terms of document properties only. This is actually one of the reasons making search by final users hard. It is a fact that search is often performed with the mediation of experts. In particular, for the query above it is necessary to identify the name of that artist born in Wales who wrote about wood sculptures. Supporting such situations requires appropriate sources of knowledge, the formalization of subjects, and a more expressive representation and query language.

In this respect, document search in KR is more expressive than in KO, as the former has developed very powerful end expressive techniques which, via the use of ontologies, support queries by any entity property (e.g., the properties of the entities described in a document). In fact, KR is concerned with the development of ontologies describing the relevant entities of a domain in terms of their basic properties, which enables an effective communication and information exchange, as well as automated reasoning (Berners-Lee et al. 2001) (Bouquet et al. 2004). Examples of entities include persons, places, organizations, and events. Taken from a KR perspective, documents are just one particular case of entity with its own properties (with title, author and subject being very important ones) and document search is a special case of reasoning. However, from a pragmatic point of view KR, so far, has failed as it currently lacks of appropriate entity specification methodologies which allow scaling as much as in KO.

In this paper we present DERA, a new *faceted KR approach* for the development of ontologies able to describe and reason about relevant entities of a domain, including documents. Domains include conventional fields of study (e.g., physics, mathematics), applications of pure disciplines (e.g., engineering, agriculture), any aggregate of such fields

(e.g., physical sciences, social sciences), or can even capture knowledge about our everyday lives (e.g., music, movie, sport, recipes, tourism). For instance, in the music domain, entities may include songs, singers and producers. DERA is faceted as the methodology engaged for the construction and maintenance of domain ontologies is inspired by the principles and canons of the *faceted approach* as originated in KO. This makes DERA capable of dealing with large-scale dynamic ever growing knowledge. DERA accounts for *entity classes* (E), *relations* (R) and *attributes* (A) of the relevant entities in the domain (D) and models them as *semantic facets*, i.e., facets where the semantics of the terms and the relations between them are made *explicit* (thus making each facet a formal ontology). The use of the fundamental categories E/R/A allows for a straightforward formalization of facets into Description Logics (DL) (Baader et al. 2002). This allows supporting the automation of complex tasks, such as highly expressive document search exploiting entity properties, via the usage of standard reasoning tools.

The remaining of the paper is organized as follows. Section 2 provides a motivation for our work showing the usefulness of moving from a purely KO to a KR approach to document search. Section 3 shows how descriptive ontologies - i.e., ontologies built at the purpose of describing and reason about real world entities - enable highly expressive document search exploiting entity properties. Section 4 explains how descriptive ontologies can be naturally formalized into DL ontologies, thus enabling complex forms of automated reasoning. Section 5 presents DERA as an innovative approach that inherits the benefits of both KO (in terms of methodologies for the development of scalable ontologies) and KR (in terms of expressiveness and effectiveness of search). Section 6 explains the steps followed in the DERA methodology for the construction of scalable descriptive ontologies. Section 7 describes related work. Finally, Section 8 concludes the paper by summarizing the work done and outlying the next steps.

2 Motivation

With the purpose of providing effective mechanisms to make information timely available, several methodologies, systems and tools have been developed in KO for the classification, indexing and search of documents. In particular, documents are typically classified by subject and indexed by document properties such as *title*, *author* as well as *subject*. Indexing by title and author are straightforward as they are directly taken from the document. Indexing by subject is far more complicated as it requires an analysis of the document content and the application of precise principles and rules to construct corresponding subject strings as combinations of terms taken from a controlled vocabulary. Search is performed manually by using a card catalogue or electronically by issuing queries through Online Public Access Catalogue (OPAC) systems that provide access to classifications and indexes. In particular, OPAC systems allow identifying those entries matching a user query in input and return a corresponding set of relevant documents in output. Supported queries include conditions about single document properties, i.e., title, author, subject, or combinations of them. Typical examples of queries supported in KO are:

1. Give me documents with *title* “Il lago di Garda”
2. Give me documents with *subject* “Cromford Mill”
3. Give me documents with *subject* “Michelangelo”
4. Give me documents with *author* “Nash, David” and *subject* “wood sculpture”
5. Give me documents with *author* “Clinton, Bill” and *title* contains “autobiography”

In order to ensure a higher recall, OPAC systems sometimes support *semantic search* (Giunchiglia et al. 2009a), namely a search where terms in the subject are disambiguated and expanded with synonyms and more specific terms taken from the controlled vocabulary. For instance, the term *sculpture* could be expanded by adding the more specific term *statue*. Though, in practice a few OPAC systems really offer such functionality (Casson et al. 2009).

However, searching for documents by their properties is not always good enough. It fact, it requires users to know such properties in advance. Conversely, users may rather know, for instance, some of the properties of the author or of any other entity the document is about, and want to search accordingly. In this respect, document search in KR is more effective than in KO, as the former supports queries by any entity property. Typical examples of queries which are supported by KR and cannot be supported by KO are:

1. Give me documents about any lake with depth greater than 100 written by Italians
2. Give me documents about a factory in England established by Richard Arkwright during industrial revolution
3. Give me documents about any artist born in Italy between 1450 and 1550
4. Give me documents about wood sculptures written by an artist born in Wales
5. Give me autobiographies written by any president of the United States

Even if the queries in the second list above correspond, one by one, to the queries given in the first list, KO would fail in the above situation. In fact, though it is true that it is already possible to answer the queries in the second list in KO by looking into authority lists, catalogues and similar resources, this is not yet systematic in KO as it would still require breaking them down into smaller search tasks and rely on scattered resources to get all the relevant information which is necessary to reformulate the queries above in terms of document properties only. This is one of the reasons making search by final users hard. It is a fact that search is often performed with the mediation of experts. For instance, answering the third query above would require identifying the names of those Italian artists born between the given time interval.

In addition, a significant obstacle towards this to happen in KO is constituted by the fact that entries in the indexes codifying subjects are given as informal natural language strings. For instance, in the following subject strings:

(1) *Buonarroti, Michelangelo*
(2) *sculpture - Renaissance*

it is not explicitly specified that Michelangelo stands for the Italian artist, that sculpture is a term denoting a form of art, and that Renaissance denotes an historical period. The disambiguation of the terms occurring in the subjects is in fact possible if and only if for all the terms in the subjects there is a *unique entry as preferred term* in the controlled vocabulary, which is typically enforced for common nouns, but not always (given their potentially huge number) for proper nouns. Whenever this is done, for instance in thesauri, it is actually only in terms of underspecified hierarchical relations, for instance by placing *Buonarroti Michelangelo* as narrower term of *Italian artist*. This is still a limited and informal specification as it does not enable complex reasoning tasks based on rich entity descriptions. In fact, it only says that documents about *Buonarroti Michelangelo* are documents about *Italian artists.* Moreover, specifying only the name may cause trouble in search (e.g., drop in precision in case of homonymy or in recall in case an equivalent name is provided by the user). It is therefore necessary to make the meaning of subjects, in all their parts, explicit and unambiguous. Among other things, the lack of formality in the subjects makes their construction, maintenance and exploitation for search extremely difficult and costly. In fact, experts are needed *during construction* to select the appropriate terms from a controlled vocabulary and arrange them in the right citation order, *during maintenance* for instance to update terms that become obsolete, as well as *during search* to assist unskilled users who are not familiar with the domain terminology and the way terms need to be combined following the syntax and rules of the indexing language (Library of Congress 2007). Moreover, subjects and vocabularies alone do not say anything explicitly about Michelangelo in terms of his properties, e.g., his date and place of birth or his works, again in a way that is directly exploitable by reasoning tools. For instance, answering the third query above would require specifying in the subject, through appropriate unique identifiers pointing to an external knowledge resource, that Buonarroti Michelangelo refers to the artist born in Italy in the 1475.

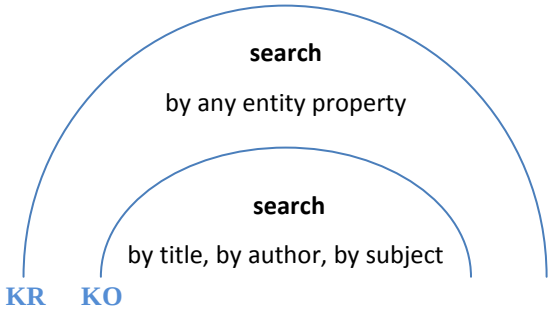

Fig. 1. From search by document properties to search by any entity property

As exemplified in Fig. 1 search by entity properties (typical of KR) actually includes search by document properties (typical of KO). However, while KO mainly relies on controlled vocabularies and indexes, KR employs supplemental knowledge resources (i.e.,

ontologies) providing an *explicit* description of the attributes of entities such as people (e.g., their date of birth), facilities and organizations (e.g., their date of establishment), events (e.g., when they happened) as well as relations between them (e.g., the fact that a certain person was born in a certain country). KR provides a more expressive representation and query language, able to codify and automatically query such knowledge. LIS seems to recognize the need for such resources. We can mention for instance the RDA[1], FRBR[2] and FRAD[3] initiatives as well as the recent OCLC work aiming to align BIBFRAME and Schema.org models (Godby 2013). However, KR already offers techniques for the representation and automatic exploitation of such resources.

3 Classification Ontologies and Descriptive Ontologies

Ontologies constitute high level descriptions of a domain, which can be used by intelligent applications to draw implicit consequences from explicitly represented knowledge (Baader et al. 2002). This is achieved through some form of automated reasoning. It has been observed that KO and KR, having different purposes, employ different kinds of ontologies (Giunchiglia et al. 2006) (Giunchiglia et al. 2009b). In fact, Giunchiglia et al. (2006) introduced the key distinction between classification ontologies and descriptive ontologies.

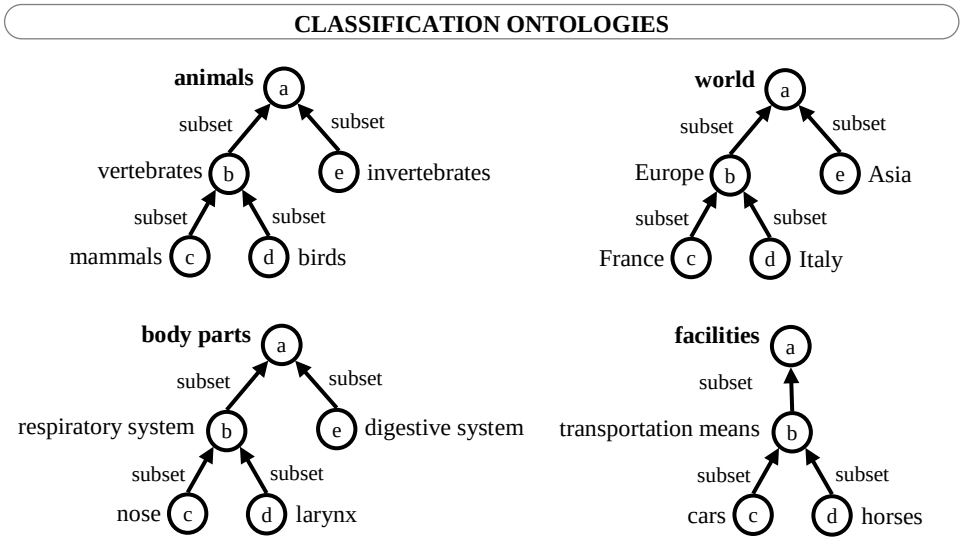

Fig. 2. Example of classification ontologies

[1] http://metadataregistry.org/
[2] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records
[3] http://www.ifla.org/publications/functional-requirements-for-authority-data

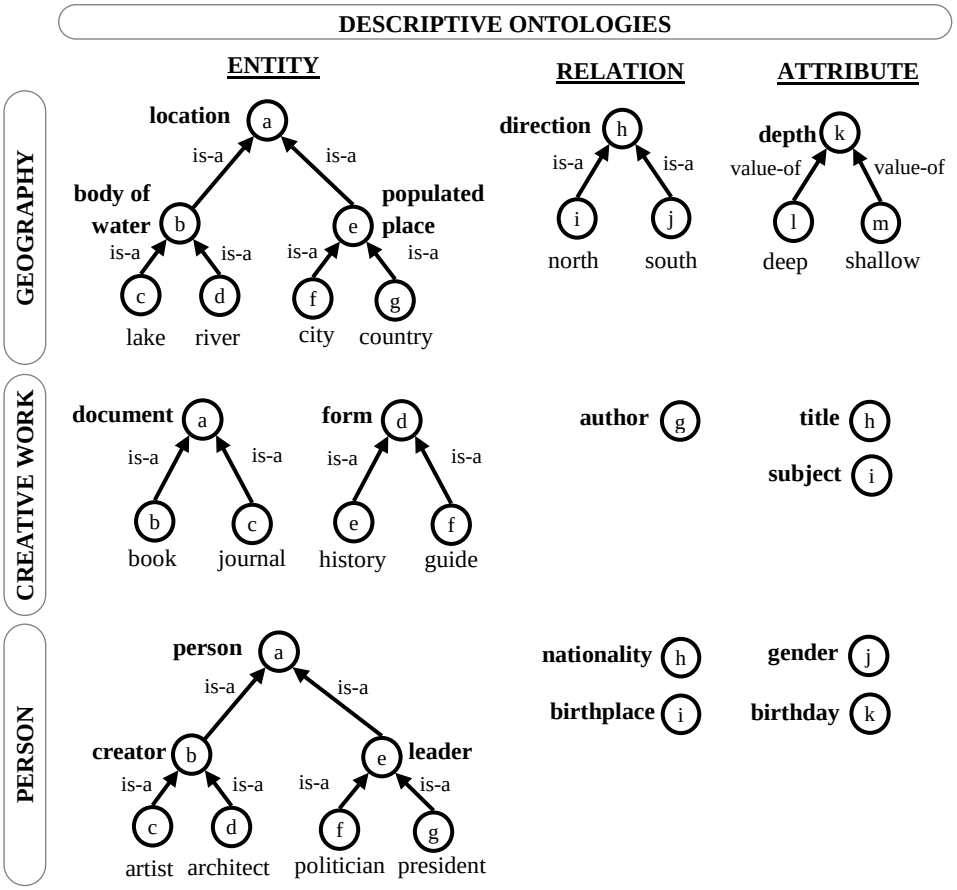

Fig. 3. Example of descriptive ontologies in different domains

KO employs ***classification ontologies***, i.e., ontologies mainly used to describe, classify and search for documents. In these ontologies, as the main focus of KO is on documents, terms denote sets of documents, hierarchical BT/NT relations between terms denote superset/subset relations, and the individuals are the documents themselves. All knowledge organization systems including classifications, thesauri, or subject indexes follow such semantics (Zaihrayeu et al. 2007). An example of such ontologies is given in Fig. 2. For instance, the term *horses* denotes documents about horses (animals), while the fact that it is placed under *transportation means* indicates that documents about horses are also documents about transportation means (at least in the context in which the classification is used). This is called *classification semantics* in (Giunchiglia et al. 2009b). The only simple form of reasoning carried out for document search in KO is based on the transitivity of the

hierarchical relations. In fact, this is what is needed to enable semantic search (Giunchiglia et al. 2009a). For instance, documents about horses can be returned when searching for documents about facilities, because *horses BT transportation means* and *transportation means BT facilities*.

KR employs ***descriptive ontologies***, i.e., ontologies built at the purpose of describing and reason about real world entities. In these ontologies, terms denote sets of real world entities, hierarchical *is-a* relations provide the backbone structure to these ontologies and indicate a subset relation, while the individuals include any real world entity. For instance, the relation *horse is-a animal* indicates that horses are a subset of all animals. This is called *real world semantics in* (Giunchiglia et al. 2009b). Descriptive ontologies provide knowledge about entities in terms of classes, attributes and relations. For instance, they may specify that animals are affected by certain kinds of diseases and that certain cures are needed to defeat them. An example of complex reasoning is searching for cures to a certain disease affecting a given animal. In essence, the purpose of KR is much broader than KO. In fact, taken from a KR perspective, documents are just one particular case of entity with its own properties (with title, author and subject being very important ones) and document search is a special case of reasoning.

An example of descriptive ontologies covering the geography, creative work and document domains is given in Fig. 3. In the picture, each node denotes a different *entity class*, *relation* or *attribute*. Relevant entities in the geography domain are locations and more specific entities, such as rivers and lakes; relevant entities in the person domain are people; *documents* are modeled as those entities which are target of the creative work domain, with *title*, *author* and *subject* being their properties. In particular, while title and subject are attributes, author is represented as a relation between a document and a person.

Descriptive ontologies are populated with entities and the value of their properties in corresponding domains. For instance, in Fig. 4 the geography domain includes the entities *Garda Lake* (as instance of lake) and *Italy* (as instance of country), the creative work domain includes the entity *Book#1* (as instance of book, which in turn is more specific than document) having corresponding title, author and subjects. Notice how the subject string *Garda Lake - history - guide* is represented as three different subject attributes.

In KR, document search is a standard reasoning task over descriptive ontologies. For instance, answering the query

Give me documents about any lake with depth greater than 100 written by Italians

over the descriptive ontologies in Fig. 3 and corresponding entities in Fig. 4 amounts to identifying all those entities which (a) are instances of the entity class *document* and (b) with "subject" set to entities that are instances of the entity class *lake* having "depth" greater than *100* and (c) with "author" set to entities having "nationality" equal to *Italy*. This would return *Book#1*, because (a) it is an instance of the entity class *book* which is more specific than *document*, (b) it has *Garda Lake* as subject which is an instance of *lake* and has a *depth* of 346 m which is greater than *100* and (c) its author is *Solitro Giuseppe* who has *nationality* set to *Italy*.

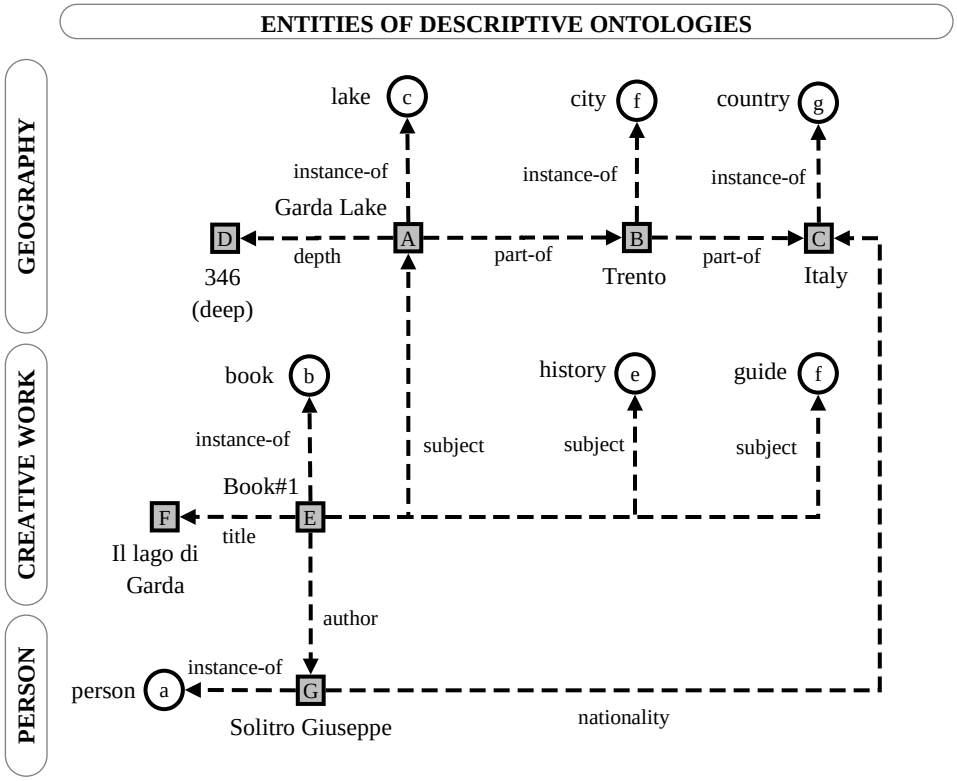

Fig. 4. Example of entities and their properties populating the descriptive ontologies given in Fig. 3

4 From Descriptive Ontologies to Description Logics

Descriptive ontologies have a straightforward formalization into DL ontologies. With the formalization (summarized in Table 1), DL concepts denote either sets of entities or sets of attribute values. DL roles denote either relations or attributes. In other words, a DL *interpretation* $\mathcal{I} = \langle \Delta, I \rangle$ consists of the *domain of interpretation* $\Delta = F \cup G$ where:

- F is a set of individuals denoting real world *entities*
- G is a set of *attribute values*

and of an interpretation function I where:

$$E_i^I \subseteq F \qquad R_j^I \subseteq F \times F \qquad A_k^I \subseteq F \times G \qquad v_r^I \in G \qquad (1)$$

that is, each entity class E_i corresponds to a DL concept whose interpretation is a subset of the entities in F; each relation R_j corresponds to a DL role whose interpretation is a binary relation between entities in F; each attribute A_k corresponds to a DL role whose interpretation is a binary relation between entities in F and attribute values in G, restricted by the interpretation of the concepts denoting corresponding attribute values v_r (connected through *value-of* relations); *is-a* relations correspond to subsumption ($\sqsubseteq$) between concepts or between roles; *part-of* relations and associative relations correspond to DL roles. And where:

$$e_p^I \in F \qquad r_q^I \in F \times F \qquad a_s^I \in F \times G \tag{2}$$

that is, instances e_p of entity classes (connected through *instance-of* relations) correspond to entities in F; instances r_q of relations are elements of the Cartesian product F x F; instances a_s of attributes are elements of the Cartesian product F x G.

Knowledge in (1) corresponds to what in DL is called the *intentional knowledge (TBox)*, i.e., a set of general statements about what is known in terms of concepts, denoting sets of individuals, and concept properties; such statements constitute the basic terminology and theory of the domain (e.g., persons have a date of birth). Knowledge in (2) corresponds to what in DL is called the *extensional knowledge (ABox)*, i.e., a set of assertions about specific individuals and the actual value of their properties (e.g., the date of birth of Michelangelo Buonarroti is 6th March 1475).

	Descriptive ontology element	**DL formalization**	
$E_1, \ldots, E_p$	entity classes	Concepts	**TBox**
$R_1,\ldots, R_q$	relations between classes	Roles	
$A_1,\ldots, A_s$	Attributes	Roles	
value-of	hierarchical relation	role restrictions	
is-a	hierarchical relation	subsumption ($\sqsubseteq$)	
part-of	hierarchical relation	Roles	
any other relation	associative relations	Roles	
$e_1,\ldots, e_n$	entities instances	individuals in F (entities)	**ABox**
$v_1,\ldots, v_r$	attribute values	individuals in G (values)	
$r_1,\ldots, r_m$	relations between entities	role assertions	
$a_1,\ldots, a_t$	attributes of entities	role assertions	
instance-of	hierarchical relation	concept assertions	

Table 1. Formalization of a descriptive ontology into DL

For instance, the descriptive ontology given in Fig. 3 for the geography domain and corresponding entities in Fig. 4 can be formalized into the TBox and ABox below:

TBox	**ABox**
location ⊑ ∀direction.location ⊓ ∀depth.{deep,shallow}	lake(Garda-lake)
body-of-water ⊑ location	city(Trento)
populated-place ⊑ location	country(Italy)
lake ⊑ body-of-water	depth(Garda-lake, deep)
river ⊑ body-of-water	part-of(Garda-lake, Trento)
city ⊑ populated-place	part-of(Trento, Italy)
country ⊑ populated-place	
north ⊑ direction	
south ⊑ direction	

5 The DERA approach

DERA provides a concrete answer to the need for a suitable approach and methodology for the development of *descriptive ontologies* which allow scaling to the production of ever growing knowledge, and their exploitation for a highly expressive document search. This in turn allows us to build, on demand, on the basis of the query the necessary DL theory as described in Section 4.

DERA is a new *faceted KR approach* for the development of descriptive ontologies and their exploitation for automated reasoning. DERA is faceted as it takes inspiration from category-based systems and in particular from the faceted approach introduced by Ranganathan (1967) and later simplified by Bhattacharyya (1975), thus aiming at the same quality and scalability benefits. However, it clearly differs from them as the original approach aims at the development of classification ontologies.

DERA is entity-centric rather than document-centric. We take an ***entity*** to be *any object so important to be denoted with a name*. They include concrete real world entities such as locations, persons, organizations and events, as well as documents, any creative work, piece of art, and also fictional objects, such as comics' characters. One immediate consequence of adopting a KR approach is that DERA is a system of *semantic categories*, namely categories supporting the specification of the terminology of a domain for the *representation* (rather than the organization) of the relevant *entities* (rather than only documents) by their basic *properties* (thus, not only the subject).

We adopt and extend the notion of ***domain*** as originally given in LIS. In DERA, a domain is *any area of knowledge or field of study that we are interested in or that we are communicating about that deals with specific kinds of entities*. They include conventional fields of study (e.g., physics, mathematics), applications of pure disciplines (e.g., engineering, agriculture), any aggregate of such fields (e.g., physical sciences, social sciences), or can even capture knowledge about our everyday lives (e.g., music, movie, sport, recipes, tourism). Domains provide a bird eye view of the whole field of knowledge,

offer a comprehensive context within which classification and search can be supported (Mills 2004), and words disambiguated (Ciaramita and Altun 2006). Domains have two fundamental properties (Giunchiglia et al. 2012a). They are the main means by which diversity is captured, in terms of language, knowledge and personal experience. For instance, according to local customs the *food* domain may or may not include bugs and dogs. In addition, domains allow scaling as they account for the evolution of knowledge. For instance, in evolving the *transportation* domain we may extend *ground transportation means* with *electrical cars*.

Within each domain, entities are described in terms of basic properties and in particular of their *entity classes*, *relations* and *attributes* which therefore become the ***fundamental categories*** of our categorization system. Under each fundamental category, terms are arranged into facets, each of them covering a different aspect of the domain. More precisely, we define a ***facet*** to be *a hierarchy of homogeneous terms describing an aspect of the domain, where each term in the hierarchy denotes a different atomic concept* (Giunchiglia et al. 2009b). Facets are further subdivided into sub-facets. Facets (and their subdivisions) are mutually disjoint.

A **DERA domain** is a triple **D = <E, R, A>** where:

- ***E*** *(for Entity)* is a set of facets grouping terms denoting *entity classes*, whose instances (the entities) have either perceptual or conceptual existence. Terms in these hierarchies are explicitly connected by *is-a* or *part-of* relation.
- ***R*** *(for Relation)* is a set of facets grouping terms denoting relations between entities. Terms in these hierarchies are connected by *is-a* relation.
- ***A*** *(for Attribute)* is a set of facets grouping terms denoting *qualitative/quantitative* or *descriptive* attributes of the entities. We differentiate between attribute names and attribute values such that each attribute name is associated corresponding values. Attribute names are connected by *is-a* relation, while attribute values are connected to corresponding attribute names by *value-of* relations.

The mapping of E/R/A above to DL should be obvious. *is-a*, *part-of* and *value-of* relations form the backbone of facets, are assumed to be transitive and asymmetric, and hence are said to be *hierarchical*. Other relations, whenever defined, not having such properties are said to be *associative* and connect terms in different facets. All together facets constitute the TBox of a descriptive ontology.

For instance, within the geography domain relevant entities are *locations* (the main E facet) that may include *inter-alia* land formations (e.g., continents, islands), bodies of water (e.g., seas, streams), geological formations (e.g., mountains, valleys), administrative divisions (e.g., wards and provinces) and populated places (e.g., cities, villages). Each of them generates a different sub-facet of entity classes. Spatial relations between them may include near, adjacent, in front. They generate sub-facets of relations. Entities may be described in terms of their length (e.g., of a river, with values long and short) or depth (e.g., of a lake, with values deep and shallow). They generate sub-facets of attributes. See the example in Fig. 5.

ENTITY	RELATION	ATTRIBUTE
Location	**Direction**	**Name**
Landform	(is-a) East	**Latitude**
(is-a) Natural elevation	(is-a) North	**Longitude**
(is-a) Continental elevation	(is-a) South	**Altitude**
(is-a) Mountain	(is-a) West	**Area**
(is-a) Hill		**Population**
(is-a) Oceanic elevation	**Relative level**	
(is-a) Seamount	(is-a) Above	**Depth**
(is-a) Submarine hill	(is-a) Below	(value-of) deep
(is-a) Natural depression		(value-of) shallow
(is-a)Continental depression	**Containment**	
(is-a) Valley	(is-a) part-of	**Length**
(is-a) Trough		(value-of) long
(is-a) Oceanic depression		(value-of) short
(is-a) Oceanic valley		
(is-a) Oceanic trough		
Body of water		
(is-a) Flowing body of water		
(is-a) Stream, Watercourse		
(is-a) River		
(is-a) Brook		
(is-a) Still body of water		
(is-a) Lake		
(is-a) Pond		

Fig. 5. Exemplification of the geography domain in DERA

When facets are populated with specific entities of a domain, *instance-of* relations connect entities to their respective classes in E. Entities are described in terms of attributes (A) and relations (R), each of them being in turn a pair <*n*, *v*> where *n* is the attribute or relation name and *v* is its value consistent with what is defined in A for the attributes and R for the relations, respectively. Entities and their properties which populate the facets constitute the ABox of a descriptive ontology.

For instance, the *Garda Lake* (an entity) can be described as an instance of *lake* (entity class in the body of water sub-facet), located in *Italy* (part-of relation) with *depth* (attribute name) of 346 m (quantitative value) which can be considered *deep* (qualitative value).

6 Descriptive Ontologies in DERA

The methodology engaged in DERA follows a minimal set of guiding principles, extensively described in (Giunchiglia et al. 2012b), which are inspired by the canons and

principles described by Ranganathan in (Ranganathan 1967) and guides though the whole process of constructing and maintaining facets, each of them covering a different aspect of the domain. However, differently from the original approach, DERA aims at the development of facets as *descriptive ontologies* (rather than classification ontologies). The main steps in the methodology are as follows:

- ***Step 1: Identification of the atomic concepts***. Relevant terms of the domain in natural language (e.g., in English or Italian) are collected, examined and disambiguated into atomic concepts. Terms are collected primarily by interviewing domain experts and by reading available literature about that particular domain including *inter-alia* indexes, abstracts, glossaries, reference works. Analysis of query logs, when available, can be extremely valuable to determine user's interests. Collected terms are then examined and disambiguated into atomic concepts. Terms with same meaning (synonyms) are grouped together and are given a natural language description that makes explicit the intended meaning. This corresponds to what in the faceted approach is called the *verbal plane* and what in (Giunchiglia et al. 2006) (Giunchiglia et al. 2012a) is called the *natural language level*. Each group of terms denotes a different atomic concept and is subsequently classified alternatively as an *entity class (E)*, *relation (R)* or *attribute (A)*. This corresponds to what in the faceted approach is called the *idea plane* and what in (Giunchiglia et al. 2006) (Giunchiglia et al. 2012a) is called the *formal language level*. For instance, we can recognize that in the geography domain the terms *stream* and *watercourse* are synonyms whose meaning can be described as "*a natural body of running water flowing on or under the earth*" (natural language) and that the group denotes an entity class (one atomic concept at formal language level), that is:

  (E) watercourse, stream: a natural body of running water flowing on or under the earth

  This is different from the original faceted approach, not only in terms of categories, but also because in Ranganathan's approach synonyms and definitions are not explicitly given. Vocabulary control is instead considered by Battacharyya (1982).

- ***Step 2: Analysis***. The atomic concepts are analyzed per *genus et differentia*, namely in order to identify their commonalities and their differences. The main goal is to identify as many distinguishing properties - called *characteristics* - as possible of the real world objects represented by the concepts. This allows being as fine grained as wanted in differentiating among the concepts. For instance, we can recognize that in geography for the concept *river* we can identify the following characteristics:

  - a body of water
  - a flowing body of water
  - no fixed boundary
  - confined within a bed and stream banks
  - larger than a brook

This is similar to the faceted approach.

- ***Step 3: Synthesis***. Collected terms are arranged into facets such that at each level of the hierarchy - each of them representing a different level of abstraction - concepts are grouped by a common characteristic. Concepts sharing the same characteristic form an *array* of homogeneous concepts. Concepts in each array can be further organized into sub-groups (or sub-facets), thus generating a new level in the hierarchy. Child concepts are connected to their parent concept through an explicit is-a (*genus-species*) or part-of (*whole-part*) relation. For instance, we can recognize that under the *body of water* facet *stream is-a flowing body of water* and that, due to their commonalities, we could declare *river is-a stream* and *brook is-a stream* by placing them under the same array. Thus, we may progressively obtain the following facet:

  Body of water
  (is-a) Flowing body of water
  (is-a) Stream
  (is-a) Brook
  (is-a) River
  (is-a) Still body of water
  (is-a) Pond
  (is-a) Lake

  This is different from the original faceted approach, where genus-species and whole-part relations are left implicit. In fact, as it aims at the creation of classification ontologies, terms are arranged in facets by means of generic hierarchical relations. Among other things, explicit relations make maintenance more rigorous. For example, it facilitates the distinction between transitive and non-transitive relations (Maltese and Farazi 2011).

- ***Step 4: Standardization***. Each atomic concept can be potentially denoted with any of the terms in the group of synonyms. When the group contains more than one term, a standard (or preferred) term should be selected among the synonyms. 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[4] approach where terms are ranked within the synset and the first one is the preferred. For instance, in WordNet the term *stream* is preferred to *watercourse*:

  (E) stream, watercourse: a natural body of running water flowing on or under the earth

[4] http://wordnet.princeton.edu/

This is different from the original Ranganathan's approach, where only one term is kept in the classification scheme while the others are discarded and external resources are needed to identify synonyms and to get definitions whenever needed. Synonyms and definitions are instead typically provided in more recent faceted schemes.

- ***Step 5: Ordering***. Concepts in each array are ordered. There are several criteria devised by Ranganathan. They include by chronological order, by spatial order, by increasing and decreasing quantity, by increasing complexity, by canonical order (the order traditionally followed in LIS), by literary warrant and by alphabetical order. For instance, in the geography domain one may follow the canonical order.

  This is similar to the faceted approach. Ordering is not considered essential in KR, but it turns out to be very useful for maintenance purposes, for instance to check the level of coverage of a facet or to facilitate the identification of a suitable position for a new concept.

- ***Step 6: Formalization.*** The fundamental categories E/R/A are such that this allows for an obvious formalization of corresponding facets into DL ontologies.

  This step is implicitly performed in LIS. In fact, the formalization includes what in the faceted approach is called the *notational plane*, i.e., the level where an unambiguous notation is used to synthetically attach meaning and provide order to terms. However, the way in which this is done in DERA makes automation of non-trivial tasks, such as highly expressive document search by entity properties, possible. In fact, document search can be framed in DL as an *instance retrieval* problem (Baader et al. 2002).

7 Related work

In LIS several methodologies have been developed for the construction and maintenance of classification ontologies. In particular, in *category-based subject indexing systems* relevant terms of a domain are organized into a classification scheme of a few fundamental categories. As the ultimate purpose is the construction of document *subjects*, such systems are grounded on *syntactic categories*, namely categories playing a role in the syntax of the subject indexing language, i.e., the language used to construct the subject strings stored in subject indexes. Hierarchies under each fundamental category encode different aspects or *facets* of the domain knowledge. Approaches differ in the kind and number of categories. Kaiser (1911) proposed *Concrete*, *Process* and *Country*; Vickery (1960) adopted thirteen categories. Ranganathan (1967) postulated *Personality*, *Matter*, *Energy*, *Space* and *Time*. Bhattacharyya (1975) simplified the categories proposed by Ranganathan by proposing only *Discipline*, *Entity*, *Property* and *Action*. In these approaches, facets of general applicability are called *common isolates* or *modifiers* (e.g. *Language* and document *Form*).

Ranganathan was the first who proposed and formalized a theory of *facet analysis* which is widely recognized as a fundamental methodology that guides in the creation of high quality classification schemes, in terms of robustness, extensibility, reusability, compactness and flexibility (Broughton 2006) (Broughton 2008). In particular, Ranganathan's approach allows scaling as with domains it is possible to add new knowledge at any time as needed.

On the contrary, KR currently lacks of methodologies to the development of descriptive ontologies which allow scaling as much as in KO. In KR, existing approaches to ontology construction and maintenance focus on ontology evaluation (Guarino and Welty 2002), supporting tools (Corcho et al. 2004), general design criteria (Gruber 2003), or on the ontology building process itself (Fernandez-Lopez 1999). In particular, OntoClean (Guarino and Welty 2002) provides meta-properties that impose a set of constraints on the taxonomic structure of ontologies that turn out to be very useful during the building process, in evaluating and improving those (Welty et al. 2004). Welty and Jenkins (1999) proposed an ontology specifically for the description of documents and their subjects, but they neither address any methodological issue nor provide any explicit implementation. Since developing ontologies from scratch is an extremely time-consuming and error prone task, many approaches have attempted to reuse existing sources (Stuckenschmidt et al. 2004). They range from lexical (e.g., WordNet) to domain-specific resources (such as UMLS and AGROVOC). All these approaches underline the usefulness of domain-specific knowledge (Laursen et al. 2008).

8 Conclusions

We have shown that, despite the very successful solutions developed, existing KO approaches to document indexing and search, by employing *classification ontologies*, are limited in expressivity as they only support queries by document properties. In this respect KR is very powerful and potentially boundless as, by employing *descriptive ontologies*, it supports queries by any entity property. This motivates the usefulness to move from a purely KO to a KR approach to document search. Though, from a pragmatic point of view KR, so far, has failed as it lacks of appropriate methodologies which allow scaling as much as in KO.

In this paper we presented the new DERA faceted KR approach and a corresponding methodology, inspired by the faceted approach, for the development of high quality and scalable descriptive ontologies. It allows modeling relevant entities of the domain (including documents) and their properties and enables automated reasoning. In particular, it supports a highly expressive search of documents exploiting entity properties. By bridging between KO and KR, we compensate for the limitations and leverage on the respective strengths of these two approaches. In fact, we inherit quality and scalability properties of the faceted approach from KO as well as the expressiveness and effectiveness of search from KR. Because of the methodology followed, DERA domains are flexible, reusable, and allow scaling and coping with the diversity of the world and the evolution of

knowledge. Automated reasoning is made possible because the fundamental categories E/R/A are such that this allows for a straightforward formalization of corresponding facets into standard DL ontologies.

As future work, we plan to experiment DERA in vertical domains and to develop a collaborative platform for the construction and maintenance of domains. Up to this point, the methodology has already proved effective in experiments conducted in the *geography* domain, for instance for the encoding of the relevant knowledge (Dutta et al. 2011) (Giunchiglia et al. 2012b) and the search of maps in semantic geo-catalogues (Shvaiko et al. 2010) (Farazi et al. 2012). In particular, in (Dutta et al. 2011) and (Giunchiglia et al. 2012b) we describe the development of a faceted descriptive ontology using DERA for the geography domain, that we called Space, which includes more than 1000 concepts and around 7 million spatial entities; in (Shvaiko et al. 2010) and (Farazi et al. 2012) we describe how the usage of a faceted descriptive ontology in combination with standard AI tools results in a significant improvement in search. Furthermore, in the recent years our efforts have been directed to the development of a new system, that we called Universal Knowledge Core, and a collaborative platform for the employment of experts for the construction and maintenance of such ontologies. It is our plan to evaluate the costs of these activities even if our guess is that it will be comparable to the costs required for standard knowledge organization systems.

Acknowledgments. The research leading to these results has received funding from the CUbRIK Collaborative Project, partially funded by the European Commission's 7th Framework ICT Programme for Research and Technological Development under the Grant agreement no. 287704. We are grateful to Silvano Groff (Central Library of the municipality of Trento) and Claudio Gnoli (Science and Technology Library of the University of Pavia) for the interesting and fruitful discussions.

References

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

Battacharyya (1975). Battacharyya, G. (1975). POPSI: its fundamentals and procedure based on a general theory of Subject Indexing Languages. Library Science with a slant to doc. 16 (1), 1-34.

Battacharyya (1982). Battacharyya, G. (1982). Classaurus: its fundamentals, design and use. Universal classification: subject analysis and ordering systems. Proceedings of the 4th International Study Conference on Classification Research, 6th Annual Conference of Gesellshaft für Klassifikation, Vol. 1. Edited by I. Dahlberg. Frankfurt : Indeks Verlag (1982), 139-148.

Berners-Lee et al. (2001). Berners-Lee, T., Hendler, J., Lassila, O. (2001). *The semantic web*. Scientific American, 284 (5), 28-27.

Bouquet et al. (2004). Bouquet, P., Giunchiglia, F., Harmelen, F. van, Serafini, L. and Stuckenschmidt, H. (2004). *Contextualizing ontologies*. Journal of Web Semantics, 1(4), 325-343.

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

Broughton (2008). 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.

Casson et al. (2009). Casson, E., Fabbrizzi, N., Slavic A. (2009). *Subject search in Italian OPACs: An opportunity in waiting?* IFLA 2009 Satellite Meetings.

Ciaramita and Altun (2006). Ciaramita, M., Altun, Y. (2006). *Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger*. Conference on Empirical Methods in Natural Language Processing, 594-602.

Corcho et al. (2004). Corcho, O., Gomez-Perez, A., Gonzalez-Cabero, R., and Suarez-Figueroa, C. (2004). *ODEVAL: a tool for evaluating RDF(S), DAML + OIL, and OWL Concept Taxonomies*. First Conference on AI Applications and Innovations, 369–382.

Dutta et al. (2011). Dutta, B., Giunchiglia, F., Maltese, V. (2011). *A facet-based methodology for geo-spatial modelling*. In 4th International Conference on GeoSpatial Semantics (GEOS), vol. 6631, 133-150.

Farazi et al. (2012). Farazi, F., Maltese, V., Dutta, B., Ivanyukovich, A. and Rizzi, V. (2012). *A semantic geo-catalogue for a local administration*. Artificial Intelligence Review, 1-20.

Fernandez-Lopez (1999). Fernandez-Lopez, M. (1999). *Overview of methodologies for building ontologies*. Workshop on ontologies and problem-solving methods at IJCAI.

Giunchiglia et al. (2006). Giunchiglia, F., Marchese, M., Zaihrayeu, I. (2006). *Encoding Classifications into Lightweight Ontologies*. Journal of Data Semantics, 8, 57-81, 2006.

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

Giunchiglia et al. (2009b). Giunchiglia, F. Dutta, B. Maltese, V. (2009). *Faceted lightweight ontologies*. In "Conceptual Modeling: Foundations and Applications", LNCS Springer.

Giunchiglia et al. (2012a). Giunchiglia, F., Maltese, V., Dutta, B. (2012). *Domains and context: first steps towards managing diversity in knowledge*. Journal of Web Semantics, 12-13, 53-63.

Giunchiglia et al. (2012b). Giunchiglia, F., Dutta, B., Maltese, V., Farazi, F. (2012). *A facet-based methodology for the construction of a large-scale geospatial ontology*. Journal on Data Semantics, 1(1), 57-73

Godby (2013). Godby, C. J. (2013). The Relationship between BIBFRAME and the Schema.org 'Bib Extensions' Model: A Working Paper. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2013/2013-05.pdf.

Gruber (2003). Gruber, T. R. (2003). *Towards principles for the design of ontologies used for knowledge sharing*. In N. Guarino and R. Poll (Eds.), Formal ontology in conceptual analysis and knowledge representation, Padova, Italy.

Guarino and Welty (2002). Guarino, N., Welty, C. (2002). *Evaluating Ontological Decisions with OntoClean*. Communications of the ACM. 45(2), 61-65. New York: ACM Press.

Kaiser (1911). Kaiser, J. (1911). *Systematic indexing*. London: Isaac Pitman & Sons

Laursen et al. (2008). Lauser, B., Johannsen, G., Caracciolo, C., Keizer, J., Van Hage, W. R. Mayr, P. (2008). *Comparing Human and Automatic Thesaurus Mapping Approaches in the Agricultural Domain*. International Conference on Dublin Core and Metadata Applications.

Library of Congress (2007). *Library of Congress Subject Headings: Pre- vs. Post-Coordination and Related Issues*. Library of Congress - Cataloging Policy and Support Office - Technical Report.

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

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

Prud'hommeaux and Seaborne (2006). Prud'hommeaux, E., Seaborne, A. (2006). *SPARQL Query Language for RDF*. W3C Working Draft. http://www.w3.org/TR/2006/WD-rdf-sparql-query-20061004/

Stuckenschmidt et al. (2004). Stuckenschmidt, H., Van Harmelen, F., Serafini, L., Bouquet, P. Giunchiglia, F. (2004). *Using C-OWL for the Alignment and Merging of Medical Ontologies*. First International workshop on Formal Biomedical Knowledge Representation (KRMed).

Ramakrishnan and Gehrke (2000). Ramakrishnan, R., Gehrke, J. (2000). *Database Management Systems*. McGraw-Hill.

Ranganathan (1967). Ranganathan, S. (1967). *Prolegomena to library classification*. London: Asia Pub. House.

Shvaiko et al. (2010). Shvaiko, P. Ivanyukovich, A., Vaccari, L., Maltese, V., Farazi, F. (2010). *A semantic geo-catalogue implementation for a regional SDI*. In INSPIRE Conference, 2010.

Vickery (1960). Vickery, B. C. (1960). *Faceted classification: A guide to the construction and use of special schemes*. London: ASLIB.

Welty and Jenkins (1999). Welty, C., Jenkins, J. (1999). *Formal ontology for subjects*. Journal on Data and Knowledge Engineering, 32 (1), 155-181.

Welty et al. (2004). Welty, C., Mahindru, R., Chu-Carroll, J. (2004). *Evaluating ontology cleaning*. In D. McGuinness & G. Ferguson (eds), AAAI2004. San Jose, CA: AAAI/MIT Press.

Zaihrayeu et al. (2007). Zaihrayeu, I., Sun, l., Giunchiglia, F., Pan, W., Ju, Q., Chi, M., Huang, X. (2007). *From web directories to ontologies: Natural language processing challenges*. In 6th International Semantic Web Conference (ISWC), Springer.