Dipartimento di Informatica,
Sistemistica e Telematica

Istituto di Matematica Applicata e
Tecnologie Informatiche

Semantic and Visual Analysis of Metadata to Search and Select Heterogeneous Information Resources

Submitted by

Riccardo Albertoni

DIST, Università di Genova
Via All'Opera Pia 13, 16145 Genova, Italy
http://www.dist.unige.it

CNR-IMATI, Sezione di Genova
Via de Marini 6, 16149 Genova, Italy
http://www.ge.imati.cnr.it

Università degli Studi di Genova

Dipartimento di Informatica,
Sistemistica e Telematica

Consiglio Nazionale Delle Ricerche

Istituto di Matematica Applicata e
Tecnologie Informatiche

Sezione di Genova

Dottorato di Ricerca in Scienze e Tecnologie dell'Informazione
e della Comunicazione
Indirizzo Ingegneria Elettronica e Informatica
(XVIII ciclo)

Dissertation

Semantic and Visual Analysis of Metadata to Search and Select Heterogeneous Information Resources

by

Riccardo Albertoni

Supervisors

Dott.ssa ***Bianca Falcidieno*** *– IMATI-GE/CNR*

Dott.ssa ***Monica De Martino*** *– IMATI-GE/CNR*

March, 2007

Abstract

An increasing number of activities in several disciplinary and industrial fields such as the scientific research, the industrial design and the environmental management, rely on the production and employment of informative resources representing objects, information and knowledge. The vast availability of digitalized information resources (documents, images, maps, videos, 3D model) highlights the need for appropriate methods to effectively share and employ all these resources. In particular, tools to search and select information resources produced by third parties are required to successfully achieve our daily work activities.

Headway in this direction is made adopting the metadata, a description of the most relevant features characterizing the information resources. However, a plenty of features have to be considered to fully describe the information resources in sophisticated fields as those mentioned. This brings to a complexity of metadata and to a growing need for tools which face with this complexity.

The thesis aims at developing methods to analyze metadata easing the search and comparison of information resources. The goal is to select the resources which best fit the user's needs in specific activities. In particular, the thesis faces with the problem of metadata complexity and supports in the discovery of selection criteria which are unknown to the user. The metadata analysis consists of two approaches: visual and semantic analysis. The visual analysis involves the user as much as possible to let him discover the most suitable selection criteria. The semantic analysis supports in the browsing and selection of information resources taking into account the user's knowledge which is properly formalized.

Sommario

Un numero crescente di attività in settori disciplinari e industriali quali la ricerca scientifica, la progettazione industriale, la pianificazione territoriale e la gestione ambientale, si basano sulla produzione ed utilizzo di risorse informative (oggetti, informazioni e conoscenza) la cui controparte è digitalizzata. La grande disponibilità di risorse informative digitalizzate (documenti, immagini, mappe, filmati, modelli 3D) evidenzia l'importanza di definire strumenti adeguati per la loro condivisione ed il loro efficace utilizzo. In particolare, emerge la necessità di strumenti per la ricerca e selezione di risorse informative prodotte da terzi che sono determinanti per il successo di attività lavorative specifiche.

In questo senso i metadati, informazioni che descrivono le risorse, si stanno affermando come strumento per la gestione ed il reperimento di risorse informative digitali ed eterogenee. Tuttavia, specialmente in settori specialistici, numerose ed articolate caratteristiche devono essere incluse nel metadato per ottere una soddisfacente caratterizzazione delle risorse. Ciò porta ad una nuova complessità del metadato ed una crescente esigenza di strumenti appositi per la sua analisi.

L'obiettivo di questa tesi è studiare e sviluppare metodologie di analisi del metadato per la ricerca e la comparazione di risorse informative al fine di selezionare le più appropriate rispetto alle esigenze dell'utente. In particolare la tesi affronta il problema della complessità del metadato e vuole supportare l'utente nella definizione dei criteri di selezione. L'analisi del metadato viene affrontata considerando due approcci differenti: un approccio basato sull'analisi visuale e un approccio basato sull'analisi semantica. L'approccio visuale è indispensabile per coinvolgere l'utente nella ricerca e permettergli di scoprire i criteri di selezione più convenienti alle sue necessità. L'approccio semantico supporta la navigazione e selezione delle risorse tenendo conto della conoscenza dell'utente esplicitamente formalizzata aiutando a far fronte alla complessità del metadato derivante da ambiti complessi e specialistici.

Acknowledgements

I wish to thank my advisors Monica De Martino and Bianca Falcidieno. Monica De Martino for all the encouragements and the patience she has demonstrated supervising me in this work. Bianca Falcidieno for the confidence placed in me. I know that being in their shoes was as difficult as being in mine.

I am grateful to all the co-authors who have contributed with their suggestions and pieces of work in my publications: Alessio Bertone, Francesco Robbiano, Laura Papaleo, Urska Demsar, Hans Hauska, Simone Marini, Franca Giannini, Marina Monti, Michela Spagnuolo, George Vasilakis and Marios Pitikakis. In particular, I have to thank Elena Camossi for reading part of this thesis and for all the Saturday mornings she has spent with me working on granularities and the new promising research prospectives.

I wish to express my appreciation to all of my colleagues at CNR-IMATI, for their friendship, their suggestions, and all that has contributed to make me grow up in the last five years.

This work has been supported by the European projects INVISIP and AIM@SHAPE. I wish thank all the partners who have stimulated my research activity. In particular, Sara Bini, Laura Citernesi (Gheogoi), Michele Piccazzo, Flavio Maragon (D'APPOLLONIA), Stefan Göbel and Jörg Haist and Uwe Jasnoch (IGD-Fraunhofer), Peter Klein, Harald Reiterer, Frank Müller, Tobias Limbach (University of Constance), Alicja Anna Malczewska, Kajetan Hemzaczek, Gabriela Surowiec, Igor Podolak (AKADEMIA ROLNICZA).

Last but not least I must thank "the source of my strengths", namely my family and all the friends who have supported me outside of my work life: Sandro, Maurizio, my parents, my ninety years old grandmother Mariuccia, Marina Merlo, Stefania&Emanuele, Sir&Sofia, Cristiano&Patrizia, Maura Bongianni, Rita&Gianni, Andrea Bertucci, Daniela&Emilio together with their beautiful daughters Valentina and Giulia, Lorella&Andrea and their daughter Elisa who will be soon a part of our group.

Table of Contents

Introduction

In the last thirty years, the importance of collecting and sharing information and knowledge has been rapidly increasing. Despite a huge headway in the technology to produce and store electronic resources, the explosive growth of the information is determining a paradox: the more information is available the less we are able to take advantage from it.
The effects of this paradox are grievous. There are many activities in our ordinary work life characterized by a process of searching and selecting information resources, and sooner or later, the overwhelming volume of information will hinder the successfully progress of our work.
The thesis aims at providing a framework to prevent this effect. It draws up a solution considering that the search and selection of information resources is a painstaking activity. To decide which are the most suitable resources for specific and severely constricted tasks requires a careful comparison of the available information resources. A "google-like" list of ranked items and some web pages excerpts are probably enough for a typical internet seeker, but they are not sufficient to decide which resources provided by third parties best fit a given task.
Metadata, namely data about data, is adopted to characterize information resources in terms of their important features. The standardization of metadata faces with the resource heterogeneity defining the description independently of formats, acquisition/production processes adopted by different resource providers. The description of resources' features enables to search through metadata instead of resources, making unnecessary the transmission of the entire resources. However, information resources in specialized domains are particularly complex: they are characterized by many features which need to be considered to their selection. The consideration of all of these features ends up in metadata that have a harsh complexity.

The thesis conceives the metadata analysis to deal with this challenging metadata complexity and with the constraint selection of the information resources.
Two kinds of metadata analysis are proposed: a visual metadata analysis and a semantic metadata analysis.

Visual metadata analysis. It relies on Information Visualization and Visual Data Mining techniques. It aims at involving as much as possible the seeker in the research activity. The criteria to search and select information resources are seldom known by the seeker in the early stage of the activity. Thus, the seeker has to be supported in the identification of selection criteria. To this purpose, different visualization and interaction techniques have been devised. They provide a summarized view of available information resources facilitating the comparison of resources and enabling the discovery of implicit patterns among them.

Semantic metadata analysis. It relies on an explicit representation of the seeker's or provider's knowledge. Different relations among information resources cannot be outlined by only adopting the visual metadata analysis because they arise from the relations known by the experts rather than from data. For this reason, different semantic methods are developed to suggest criteria momentarily forgotten by the seeker as well as novel criteria coming from the re-elaboration of the seeker's background knowledge.

The thesis deals with the search and selection adopting an interdisciplinary approach. It combines concepts, principles and techniques belonging to different fields of Knowledge Management which are emerging in the recent years: Information Visualization, Visual Data Mining, Ontology, Semantic Modeling, Metadata Management. Although, the intrinsic difficulties in considering fields that are quite new and continuously evolving, this thesis conceives a first framework to analyze the metadata making headway in the search and selection of information resources.

Contributions

The work presented in this thesis contributes to the search and selection of information resources which are produced by third parties and are aimed at accomplishing some specialized work tasks. My contributions fall in these four major categories:

Problems in the search and selection of information resources. The thesis extends the search problems outlined by Spoerri [Spoe 04c] and the Anomalous State of Knowledge (ASK) defined by Belkin et al. [Belk 82a] considering the specific difficulties pertaining to the search and selection of information resources. The metadata is mandatory to face with the heterogeneity and the volumes of available information resources. As a consequence, the search and selection of information resources is entirely based on the analysis of their metadata. However, the thesis realizes that a large set of features has to be included in the metadata in order to reasonably characterize the resources. Thus, special methods of metadata analysis are required to face with the search and selection problem as well as with the emerging metadata complexity.

Framework facing with the problems of search and selection. The thesis develops a conceptual framework where different methods to analyze metadata are defined to solve the search and selection problems. The metadata analysis supports the seeker in the query refinement process facilitating the deep comparison of the resource characteristics and the discovery of novel selection criteria. It stresses the need to involve as much as possible the seeker in the search and selection activity and the integration of the visual and semantic methods.

Validation in different application domains. The most of the problems dealt within this thesis have arisen from the research activities carried out within European projects. In particular, the interest for metadata analysis tools has been shown in the domain of geographical information as well as in the management of multi-dimensional media. The interest within European projects has outlined the great potential of the undertaken research.

Methods to exploit knowledge in the ontologies. The thesis adopts ontologies to represent the users' background knowledge and to take advantage of such knowledge during the metadata analysis. We have recognized a lack of instruments to take full advantage from the knowledge encoded in ontologies. As a consequence the thesis proposes methods which facilitate the organization and comparison of information resources directly contributing in the research fields of ontology and semantics.

Overview

Before going through the thesis, some general remarks about its structure are provided.

Chapter 1 presents the basic concepts this thesis relies on: information resources (section 1.1), implicit and explicit semantics (section 1.2), ontologies (section 1.3), metadata (section 1.4).

Chapter 2 introduces the scenario, the problems and the metadata analysis this thesis is addressing. The section 2.1 details the scenario where the information resources are searched and selected. The section 2.2 provides the first contribution: it describes the problems pertain to the search and selection activity and motivates the need for metadata analysis. Starting from the outlined problems, the section 2.3 introduces the requirements for the metadata analysis. Then, the section 2.4 describes the conceptual framework which is the second contribution of this thesis. The contribution is detailed afterward in the chapters describing the visual and metadata analysis.

Chapter 3 describes the visual metadata analysis. It refers to my experience in the European project INVISIP (IST-2000-29640) pertaining to the analysis of geographic metadata. Firstly, section 3.1 provides a brief theoretical overview of the Information Visualization and Visual Data Mining concepts. Then, section 3.2 introduces the search and selection of geographical information resources. It provides part of the third contribution of this thesis demonstrating that the assumptions and problems discussed in chapter 2 can be encountered in real case studies. The section 3.3 shows the visual metadata analysis tool developed within the European project INVISIP. In particular, the overall evaluation of tools developed in INVISIP is presented in section 3.3.4. It argues the visual approach is positively perceived by the user. Finally, the section 3.4 analyzes the contributions of the visual analysis with respect to the metadata analysis requirements and section 3.5 concludes remarking the intrinsic limitations of a visual approach.

Chapter 4 introduces the semantic metadata analysis (the last contribution of this thesis). It relies on the experience made within the European Network of Exellence AIM@SHAPE (IST NoE NO 506766) to underlie the importance of semantic methods and to overcome the limitations of instrument based only on visual tools. The section 4.1 motivates the choice of ontologies to encode the seeker's background knowledge, and it introduces the roles that ontologies can play in the organization of metadata. Three semantic methods for metadata analysis are proposed: two are related to the semantic similarity evaluation (sections 4.2 and 4.3) and one to the semantic granularity evaluation in an ontology driven metadata (section 4.4). Finally, section 4.5 analyzes the contributions of semantic analysis with respect to the metadata analysis requirements and section 4.6 discusses the overall results obtained.

Chapter 5 discusses the overall support and limitations pertaining to metadata analysis proposed in this thesis.

The conclusion summarizes the results of this thesis and describes the emerging research activities to be carried out in the future.

Moreover two appendixes are included:

Appendix A illustrates part of the activity I have carried out within AIM@SHAPE. It provides an excerpt of metadata we defined to describe digital shapes. It gives the flavor of the problems to be faced in the characterization of such as complex information resources.

Appendix B examines the potentials of applying existing Information Visualization tools in the World Wide Web and Semantic Web. It considers the problems of search in a slightly different perspective with respect this thesis. Actually, it does not face directly with information resources and metadata. Anyway, it is included because it shows as visual tools might be supporting in the general search activity.

Chapter 1

Background

This chapter aims to introduce the theoretical background fundamental in the understanding of my research activity. The thesis relies on concepts which have been arising in different fields of computer science. Concepts such as information resources, semantics, ontologies and metadata deserve a brief introduction as they will be extensively employed in the rest of the thesis.

1.1 Information Resources

According to [Architec] *information resource* is an electronic artifact which conveys information as its main aim. For example, the electronic version of this thesis is an information resource: it consists of words and punctuation symbols and graphics and other elements that can be encoded with different degrees of fidelity into a sequence of bits. In principle, all its essential information can be transferred in a digital representation.

Although it is conventional on the Web to describe Web pages, images, product catalogs, etc as resources, the term *resource* is often used in broader sense for whatever may be identified by a Uniform Resource Identifier (URI) [URI 06][1]. An information resource is a special case of resource which has the property that all of its essential characteristics can be conveyed in a digital representation.

Example of information resources are: documents, datasets but also resources which digitally represent real entities. A typical kind of information resource in this sense is the digital shape, i.e. multi-dimensional media characterized by a visual appearance in a space of 2, 3, or more dimensions [Albe 05d] (such as pictures, images, 3D models, videos, animations, geographic map). On the contrary, examples of resources are the actors who come into play during the information resources processing as well as hardware and software tools to produce the information resources (scanner, camera, sensor, satellite, etc).

[1] An URI is a compact sequence of characters that identifies an abstract or physical resource.

According to [Architec], the difference between a resource and an information resource is related to the content of the real entity which could be lost in the process of digitalization transforming the real entity in a digital model.
There are entities, such as car and person that are resources but not information resources because their core essence can not be represented by information: it is possible to describe some characteristics of a car or a person in a sequence of bits, but the sum of those information will invariably be an approximation of the essence of these resource.
On the contrary, the digitalized model of a real entity is usually an information resource if considered within a specific context: even if it does not represent all the essence of the represented entity, it provides all the information needed for the application context. Considering that the informative content of a digital model essentially depends on the context and the motivations behind its creation, there are plenty of contexts where a digitalized model provides all the essential information needed to accomplish a study in the specific context. For example, in the geographic domain, a cartographic map of a particular geographic area may provide all the needed information for a site planning although it does not represent all the essence of the geographic area. In the industrial design, a 3D shape model of a car may provide all the stylistic information needed to the designer even if the model is not completely equivalent to the car. In academic domain, the information related to research staff members provides sufficient information for the recruitment process even if the researcher description is not completely representative of the person essence.

This thesis relies on the concept of information resource referring to all the electronic artifacts relevant to perform the information search and selection activity. *Information resources* are expected to have a URI and to be digitalized. Information resources are not necessary textual, they can be images, video, picture, digitalized maps, they can be expressed in different formats (PDF, DOC, XML, JPG, AVI, MPEG), and they are not necessary available for free. They can be digitalized models representing real or non real entities such as persons, projects, objects or whatever is pertinent to convey information useful to carry out an effective activity. The term *resource* will be sometimes adopted as an abbreviation of information resource. It is not misleading because in this thesis each resource, which is not digital but which is relevant to the search process, is expected to have a digitalized counterpart which is an information resource representing its description.

1.2 Semantics

Semantics refers to the aspects of meaning that are expressed in a language, code, or other form of representation[2]. It is a topic of interest in several scientific disciplines, both in Computer Science and outside of it. There are several definitions of semantics varying according to the acceptation that the word meaning assumes in the different disciplines, for example:

[2] http://en.wikipedia.org/wiki/Semantics

in linguistics, semantics is the study of language meaning[3]. In the study of language, semantics concerns with the meaning of words, expressions and sentences, often in relation to the truth;

in logics and formal languages, semantics is defined in contrast to *syntax*. The syntax represents the possible configurations of symbols which form correct formulas or formal expressions. The *semantics* determines the piece of world which the formula or the formal expression is referring to [Russ 02];

in information systems, *data semantics* is referred to the implied meaning of data. Used to define what entities mean with respect to their roles in a system[4].

In this thesis, the concept of *information resource semantics* is adopted to refer to the meaning of information provided by the resource. It includes the relationships the information resource has with other resources, and the role it can assume in a system. The semantics is fundamental in the search and selection of information resources because it provides the knowledge to identify whether an information resources is suitable for a given problem or an activity.
I distinguish between implicit and explicit semantics in coherence to Sheth et al. [Shet 05].
Implicit semantics refers to the kind of semantics, which is not represented explicitly in any strict machine processable syntax. It arises from some kinds of patterns and relations among the information resources. Examples of implicit semantics are co-occurrence of terms, dependencies, and unknown relations among information resources, which can be discovered by statistical tools and Data Mining techniques applied on collection of information resources.
Formal or explicit semantics is machine-processable semantics which is encoded in some formal language which represents relations and facts that someone knows. Examples of explicit semantics are: the semantics of “is-a” (subsumption in Description Logics), which reflects the human tendency of categorizing by means of broader or narrower descriptions, and the semantics of “part of”(partonomy) accounting for what is part of an object.

1.3 Ontologies

According to Gruber an *ontology* is a formal specification of a shared conceptualization of a domain of interest [Grub 95]. Antoniou and van Harmelen [Anto 04] comment the Gruber’s definition explaining that in general, and ontology formally describes a domain of discourse. Typically, an ontology consists of a finite list of terms, and the relationships between these terms. The *terms* denote important *concepts* (*classes* of objects) of the domain. For example, in an ontology pertaining to the university setting, staff members, students, courses and

[3] http://wordnet.princeton.edu/perl/webwn
[4] http://www.sedris.org/glossary.htm

disciplines are some important concepts.
The *relationships* typically include hierarchies of classes. Apart from the subclass relationships, ontology might include information such as *properties* (e.g. X teaches Y), *values restriction* (e.g. only a faculty member can teach courses), disjointness statements (e.g. students and staff are disjoint) and specification of logical relationships between objects (e.g. every department must include al least 10 staff members).

Despite the Gruber's definition is one of the most cited, and it emphasizes the need of formality in the specification, the level of formality required by ontologies in the real application is still subject of much debate. The level of formality adopted to express an ontology might range in a quite large spectrum as pointed out by [McGu 03, Lass 03].
Figure 1.1 illustrates the *ontology spectrum*. The different level of formality can be identified in the following classes listed from the very basic level of formality, i.e. the Catalogs, to more advanced formalism such as General Logic constraints:

- *Catalogs* can provide an unambiguous interpretation of terms because every use of a term, can be denoted by exactly the same identifier. So, we can distinguish among the different acceptations of a term according to distinct identifiers.

- *Controlled vocabulary* is one of the simplest notions of a possible ontology i.e., a finite list of terms.

- *Glossary* is a list of terms and meanings where the meanings of each term are specified typically as natural language statements.

- *Thesauri* provide some additional semantics in the relations between terms. They provide information such as synonym relationships.

- *Informal is-a* as supported in early web specifications of term hierarchies. For examples Yahoo provides a basic notion of generalization and specialization which can be considered as informal is-a.

- *Formal is-a* is supported by systems where the concept of superclass between two generic classes A and B adheres to the following rule: "if a class C is a subclass of B and A is a superclass of B, then necessarily C is a subclass of A as well".

- *Formal instances* is supported by systems where the following rule holds: "if an object is an instance of B and A is a superclass of B, then necessarily the object is also an instance of A".

- *Frames* have classes including property information. For example, the "Apparel" class may include properties of "price" and "isMadeFrom".

- *Value restrictions* on what can fill a property. For example, a "price" property might be restricted to have a filler that is a number (or a number in a certain range).

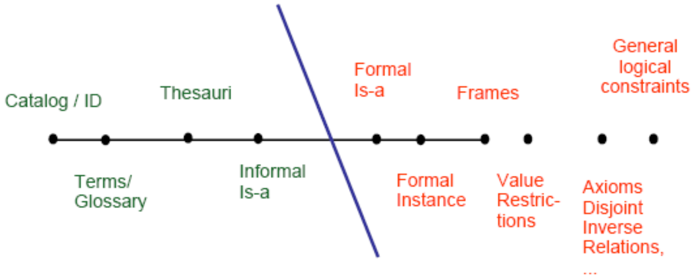


Figure 1.1: *Ontology Spectrum [Lass 03]*

- *Axioms, disjointness, inverse relations* are provided by some languages. They allow to specify more detailed properties about classes as well as relations such as disjoint classes, disjoint coverings, inverse relationships, part-whole relationships.

- *First order logic constraints.* Very expressive ontology languages allow ontologists to specify arbitrary logical statements between terms.

As ontologies need to express more information, their expressive requirements grow. For example, we may want to fill in the value of one property based on a mathematical equation using values from other properties.

The ontology spectrum shown in figure 1.1 is divided in two parts, on the left side there are the characteristics of an *informal ontologies* whereas on the right side there are the characteristics of *formal ontologies.* As illustrate in figure 1.2 the level of formality needed might depend on the kind of task the ontology is supporting. Formal ontologies are mandatory for reasoning and consistency checking, but less formal ontologies are suitable for navigation, sharing of knowledge by humans.

Ontologies encode explicit semantics. In general, ontologies provide a formalization of a conceptualization, which is the structure of reality as perceived by human or artificial agents. In this sense, they can be adopted to formalize some of the background knowledge owned by the agents.

1.4 Metadata

The most traditional definition of *metadata* is “data about data”, it clearly points out that the main characteristic of metadata is its referential nature. Metadata basically deals with the “what”, “who”, “where”, “why”, “when”, “where”, and “how” of the data. The usual adoption of metadata is to answer to questions like:

- Why/how a resource has been created?

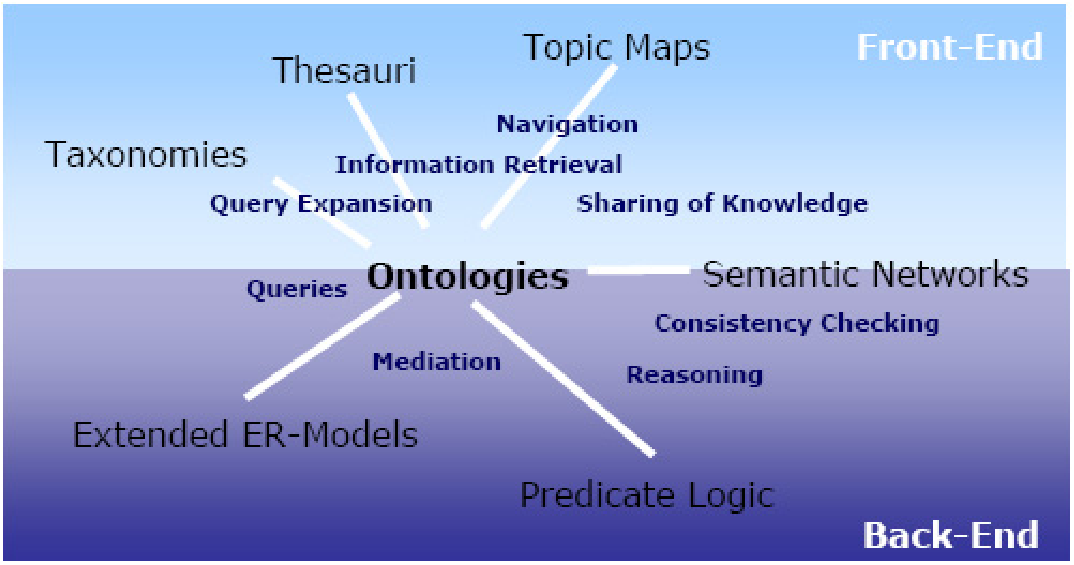


Figure 1.2: *Different kind of ontology according to the task to be supported.*

- How much reliable is a resource?
- How can one access to a resource?
- Who has produced the resource?
- What do the resources describe?

However, the classical definition "data about data" does not capture neither the vast range of purposes with respect to the metadata can be adopted nor the different levels of expressiveness with respect to metadata can be defined. The definition seems to suggest that metadata predicates only about other data, but especially in the recent period metadata tends to be adopted to predicate about anything, as long as the predicated object is associated to an Unique Resource Identifier (URI). For example, the project Friend Of An Friend (FOAF) [FoafUri 06] provides metadata of a person in terms of the relations of friendship with other people. The popularity reached by this project is the most evident proof of how much this trend is becoming relevant in the Web community [Staa 05].
FOAF shows the metadata is emerging as the standard "lingua franca" to describe the information resources as well as resources.

Steamed from the classical metadata definition, different extensions and kinds of metadata have appeared. Greenberg defines metadata as structured data about an object that supports functions associated with designated object [Gree 02]. According to Sicilia [Sici 06] this definition introduces two important aspects: structural organization and functional dependency. The structure in metadata entails that information is organized systematically; this is an aspect that is far from being controversial, especially due to the fact that metadata in many domains is nowadays subject to standardization. The term *metadata schema* is often used to refer to such specific organization. The terms *metadata entry* and *metadata item* refer to a piece of metadata which describes a single resource. It is important to note that all metadata entries are information resources themselves: if the metadata is defined appropriately, the metadata entries provide the information relevant for a set of applications and they can be subject of the search and selection activities as well as of the other information resources. Nonetheless, the fact that metadata is created to support specific functions and applications is sometime overlooked or vaguely acknowledged.

According to Sheth et al. [Shet 02] metadata takes two forms: syntactic and semantic. *Syntactic metadata* describes non-contextual information about the content, such as a language, a bit rate, and format and it offers no insight into a document's meaning. By contrast, *semantic metadata* describes domain specific information about the content. For example in an academical domain, the relevant semantic metadata might include as entities the lecturers, the classes, and the students. It is evident that semantic metadata assumes the relation between the lecturers and the classes where they teach, or among the students and the classes they attend. The assumption of such relations counts on some common understanding about what is a lecturer, a class and a student and which are the relations among these entities. The common understanding can be implicit or explicit.

Ontologies provide the means to make explicit such a common understanding. They are adopted to provide a context for semantic metadata [Shet 02], but also, they are employed to organize the metadata in schema, a formalization of how the metadata features have to be grouped identifying the main entities and their relations. Metadata structured in an ontology is referred to as "*ontology-driven metadata*".

Different kinds of metadata require specific languages having different levels of expressiveness. For example, syntactic metadata can be expressed by the Extensible Markup Language (XML) [XML 06], while as long as metadata get more semantic more sophisticated ontology languages such as the RDF/RDFS [RDF 06], OWL [OWL 06] are needed.

The definition of metadata depends on the kind of information resource and the target application considered. For some kinds of information resources long standardization processes have been carried out to determine a metadata vocabulary to represent the set of features which should be used as well as their structure. Examples of standardized metadata are the following:

- ISO 19115 defines the standard schema to describe geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data [ISO19115 03];

- the Dublin Core is a standard (NISO Standard Z39.85-2001) for cross-domain information resource description. In other words, it provides a simple and standardized set of conventions to describe on-line things in a way that makes them easier to be found. Dublin Core is widely used to describe digital materials such as videos, sounds, images, texts, and composite media like web pages [dc 06];

- MPEG-7, formally named “Multimedia Content Description Interface”, is a standard for describing the multimedia content data that supports some degree of interpretation of the information meaning, which can be passed through, or accessed by, a device or a computer code. MPEG-7 is not aimed at any particular application; rather, the elements that MPEG-7 standardizes support a range of applications as broad as possible [mpeg7 04].

Some examples of ontology driven metadata are outgoing in the Network of Excellence AIM@SHAPE: Advanced and Innovative Models And Tools for the development of Semantic-based systems for Handling, Acquiring, and Processing knowledge Embedded in multidimensional digital objects [AIM]. The Mission of AIM@SHAPE is to foster the development of new methodologies for modeling and processing the knowledge related to digital shapes. This knowledge is concerned with the geometry (the spatial extent of the object), the structure (object features and part-whole decomposition), attributes (colors, textures), semantics (meaning, purpose), and has interaction with time (morphing, animation). Ontology driven metadata are used as a knowledge formalization mechanism for linking semantics to shape or shape parts. The Appendix A provides a more detailed example of an ontology driven metadata resulting from AIM@SHAPE.

Chapter 2

Scenario and Motivation

Nowadays information resources representing data, information and knowledge provided by third parties are crucial to many activities of our work life. However, the explosive growth of the amount of information resources as well as their complexity and heterogeneity make extremely difficult to search and select those are appropriate for a specific activity.
In general, the search and selection of information resources requires to cope with a large range of problems. Some of them can be entirely faced by the computer side (e.g. how to collect the information resources stored in a distributed repository? How to rank the resources according to some specified criteria?), whereas others require a strong involvement of the user (e.g. how to refine the criteria when the resources obtained are not satisfying? How to replace the searched information resource when it results unavailable?).
The current web search engines (e.g. Google, Yahoo) face with the first class of problems. They are quite adequate for retrieving textual information resources, such as web pages and documents. They start from few keywords and they might reduce the number of information resources to be considered especially in the first stage of the search and selection activity. However, the current search engines do not deal with the latter class of problems. In particular, they do not support to a deep comparison of the information resources, despite this activity is mandatory to figure out which information resource best fits complex requirements.

The research illustrated in this thesis aims to face with the second class of problems considering that the information resources are rather heterogeneous, often very complex, and not only textual. It starts from the metadata of information resources which are a careful description of all the factors affecting the suitability of the resources for a given problem. Then, it recognizes that in the ordinary work life, the research and selection activity is often performed by a person with a one-time goal and a one-time query. Thus, the search criteria are different from time to time, and it is not possible to select the information resources on the basis of a user profile.
Considering that profiling techniques cannot be applied, the seeker has to realize as quickly as possible which search criteria provide the most suitable information resources. The thesis

aims at developing methods to support in the criteria discovery, as well as in an effective and an efficient communication between the seeker and the system. Since information resources are described by their metadata, the thesis proposes the *metadata analysis* in order to improve their search and selection. This kind of support is important and it might determine the success of the work activity the user is carrying on.

The chapter is organized as follows. The scenario assumed as preliminary conditions to the search and selection of information resources is outlined in section 2.1. It illustrates how the information resources, or more precisely their metadata, are delivered to the seekers. Section 2.2 discusses the motivations which have brought to conceive the metadata analysis framework. Section 2.3 summarizes the requirements which should be satisfied by a metadata analysis. Finally, a conceptual framework is introduced in section 2.4 providing the general overview of the methods proposed in this thesis.

2.1 The Scenario

This paragraph explains the scenario where the search and selection of information resource is usually performed.

Two main types of actors are expected to come into play: the providers and the seekers. The providers are the third parties who make available information resources. The seekers are the actors who are searching for information resources.

The information resources are provided according to the provision process depicted in Figure 2.1. Each information resource (Figure 2.1 (b)) represents information about real world entities (Figure 2.1 (a)). The providers are in charge of enclosing a description of their own information resources, i.e. the metadata. The metadata is compliant to a vocabulary of the features, which should be considered for a given purpose and application domain, namely a metadata standard (Figure 2.1 (c)). The metadata provides information pertaining to the resources characteristics (e.g. their types, formats and costs, contents, copyrights, and where it is possible to retrieve or at least to order a copy of the information resources), as well as, the information related to the resources acquisition or design (e.g. tools, constraints, rules adopted to produce the information resources, actors involved). Finally, the metadata is published by the providers in a repository (Figure 2.1 (d)) which is accessible to the seekers.

The scenario remarks that the seekers search and select the information resources relying only on their metadata.

The providers and seekers can belong to the same organization, as in the example of the designers who are selecting products previously produced by their company; to different institutions/universities, as in the selection of scientific papers; or even to different kinds of professional fields, as in the case of site planning where engineers, architects, lawyers, traffic analysts have to collaborate together. The kind of relation between the providers and the seekers affects the way of storing the information resources and their metadata.

The metadata of information resources can be stored in local or distributed repositories or

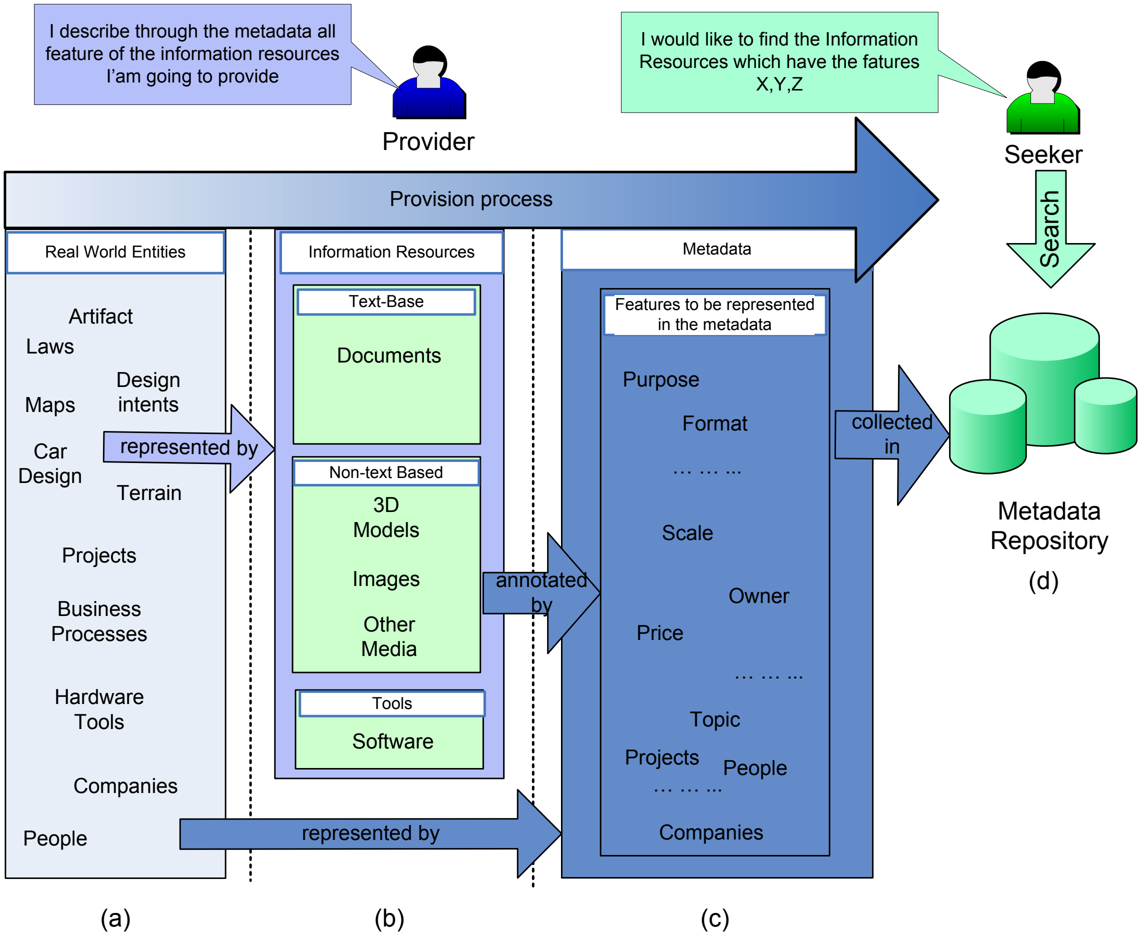


Figure 2.1: *Information Resources Provision and Search. (a) Real world entities modeled/involved in the information resources; (b) Information resources produced; (c) Features encode in metadata to describe the information resources including also the descriptions of companies tools and people involved; (d) Metadata Repository.*

even to be crawled by the World Wide Web or be available as semantically enriched resources in the semantic web.
The research results presented in this thesis are independent on where the metadata are stored. It supports in determining which candidates best fit the seekers' information needs. Therefore, I am assuming that independently from the architecture working below, the appropriate method to collect the information resources will be available.

In particular, some additional hypotheses are assumed in the scenario to ensure that the seekers have a sufficient freedom of selecting the resources:

1. the search and selection of information resources takes place in a market where providers are competing each other;

2. there is no repository master neither to establish a harmonization of the information resources nor to define a strict trade license, which regulates the provider license, to control the typology of their provision.

An environment where these hypotheses hold will be referred from now on as *Open Environment*.

The first hypothesis assumes the providers earn an advantage if their resources are selected: in terms of money as in the case of geographical information resources which are usually available for sale, but also in terms of prestige as in the case of scientific publications where the more a scientist is known and cited, the more his ideas are falsified [1] and he/she can affect the scientific enterprise.

The latter hypothesis is laid down by pragmatic considerations coming from real world experience. Firstly, if only a producer is allowed to produce a specific type of resource, there is no space for competition then the hypothesis related to the market is rejected. Secondly, a deepen harmonization among the resource requires a strong control from the master. This control requires expansive efforts which usually produce a delay in the resources publication. In general, whenever there are many providers or the resource content is expected to cover a large extension of a knowledge field, it is quite impossible to realize a so centralized control.

The hypotheses do not ensure that for each seeker's need there will be a proper information resource, but they ensure that whenever there is the chance to earn an advantage the producers will try to do their best in order to provide updated information. As result some seekers will face with many candidates resources suitable to their need, other will not obtain any candidate.

In synthesis the competition brings to improve the quality and quantity of the information

[1] According to Karl Popper and his work published in *The Logic of Scientific Discovery* (1932), the falsification is the process which can validate a scientific hypothesis. as a result the more are the attempts to falsify a scientific theory the more the theory can be considered valid.

resources, however the ever increasing volume of resources makes impossible or at least very expensive to have repositories that are contemporary harmonized and up to date.

2.2 Motivations behind the Metadata Analysis

In this section, the motivations which have drawn me to design the metadata analysis are outlined.

I will first describe the problems pertaining to the search and selection of information resources. In particular, I distinguish between two kinds of problems:

- problems arising from the open environment;
- problems related to the cognitive abilities of the involved actors.

Concerning the first class of problems, the hypotheses stated in an open environment determine the following drawbacks:

Information resources heterogeneity. In an open environment, different providers describe their information resources according to different characteristics. They also adopt different formats making more difficult to compare information resources.

Data deluge. Information resources, especially when they are non textual, can be really expansive in term of the amount of byte needed to represent their content. For example, an information resource which represents a digital terrain or a 3D model can take more than one gigabyte. As a consequence, if we consider the usual network bandwidth and the hard disk capacity of a personal computer, to download and compare 50 information resources of this type become rather critical.

Information resources inaccessibility. Information resources can be not directly accessible: sometimes the providers want to earn money from their information resources, thus many of them are available only for sale and it is not possible to obtain a copy to examine their contents.

Concerning the actors' cognitive abilities, it is well known the seeker is affected by the *Anomalous State of Knowledge* (ASK) pointed out by Belkin et al. [Belk 82a, Belk 82b, Belk 00]. According to the ASK, the seeker has only a vague knowledge about what can fill his information needs. Due to this lack of knowledge, the seeker provides some criteria in the early stage of search and selection but he needs to modify and refine them as long as he gets more familiar with the repository content.

Moreover, the seeker and the providers have different cognitions of the same repository (see figure 2.2). That is because each actor has his own perception of what the information resources represent and of their relationships. The differences of knowledge and perception

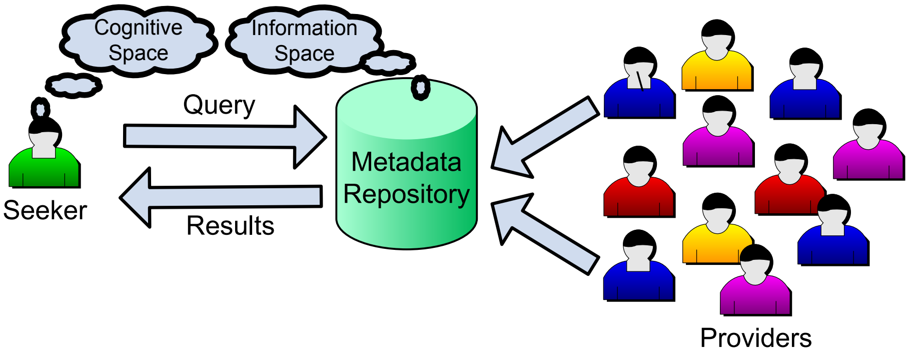


Figure 2.2: *Cognitive Space vs Information Space*

between the information providers and the seekers are modeled in terms of *informative space* and *cognitive space*. The former is defined as a set of objects and relations among them held by the system, whereas the latter is defined as a set of concepts and relations held by the individual [Kim 03]. The less these two spaces overlap, the more it is difficult to achieve an appropriate selection of the information resources.

The search and selection of information resources is performed through their metadata. The choice of metadata as a kind of "lingua franca" represents a headway with respect to the drawbacks arising from the adoption of the open environment. Metadata pave the way for an uniform characterization of the information resources attenuating their heterogeneity. Metadata are much lighter than the information resources easing the data deluge. Metadata represent all the important features needed to select the information resources, so, it does not matter if the information resources are not directly accessible. In this sense, whenever the selection is performed in an open environment, the adoption of metadata is mandatory. However, it does not solve all the problems.

The seeker has to compare many information resources (i.e. metadata item). The characterization of complex information resources requires the adoption of many features (i.e. metadata properties)[2]. The *metadata complexity* arising from many items and many features, together with the problems related to the seeker's cognitive abilities make extremely difficult to get a successful search and selection. This thesis proposes the *metadata analysis* to face with the metadata complexity and to extend the user abilities in searching and selecting the information resources.

[2] The metadata standard for geographical information, ISO 19115, includes more than fifty features.

2.3 Requirements for the Metadata Analysis

According to the discussion made in the previous section, an approach to analyze metadata should meet the following requirements:

The exploitation of both implicit and explicit semantics. The poor overlapping between the Cognitive and Informative spaces causes the most of the problems pertaining to the search and selection of information resources. Methods to make more explicit the semantics are important to increase the alignment of the two spaces . In particular, the implicit semantics is useful to discover the hidden relations in the information space, whereas the explicit semantics can be adopted to formalize the relations in the cognitive space.

The support to an interactive query refinement. Due to the ASK the seeker is not able to completely characterize his needs at the beginning of the search activity (Belkin et al. [Belk 82a]). As a consequence, he needs to alternate phases where he queries the metadata repository to phases where he browses the metadata repository content. The selection cannot be performed without a strong involvement of the individuals. The ASK forces the seeker to enter into dialogue with the IR system engaging in a query refinement process.

Spoerri recalls the query refinement problems in the context of the World Wide Web search [Spoe 04c]. I have also discussed in [Albe 05a] their relevance to the search activity in the Semantic Web (see Appendix B). More generally, I think these problems come up whenever the search is performed in an open environment. As a consequence, the support in the Spoerri's problems is part of the second requirement. An adaptation of Spoerri's problems is discussed in the following:

Problem of query formulation: "How to precisely communicate the query criteria to the system?" This problem is strongly related to the choice of the language adopted to specify the query. The formal languages are usually precise but they result unfriendly for seekers who have an inappropriate background. On the other hand, the natural language is considered friendlier but it is often not enough precise. To make the selection successful it is necessary to provide some friendly means to precisely express the selection criteria the seeker has in his mind. That might also require the increasing of query language expressiveness providing new constructs to manage for instance the level of abstraction of the information resources and their similarity.

Problem of vocabulary: "Which term to use?" The difference of knowledge and perception between the information providers and the seeker modeled in terms of informative space and cognitive space brings to problems of vocabulary. Whenever the cognitive space is not coherent to the information space, the seeker and the providers will use different terms to identify the same concepts. In this case the query formulated by the seeker may fail and the seeker needs to identify the correct term(s).

Problem of retrieved resources exploration: "How to explore many retrieved information resources?". Due to the ASK, the queries formulated by the seeker might not correspond to a proper representation of his needs, thus the order induced by ranking measures might be misleading. [3] Moreover, because of the huge amount of resources that are available, even supposing to have adopted the right criteria, the seeker has to face with a huge amount of query results. The seeker needs to be supported in the analysis of results both to choose the most suitable for his purpose and to refine his query.

Problem of query coordination: "How to query?". Human behavioral studies show that the seeker is lazy, usually he tends to create short queries and rarely adopts boolean expression in his query criteria [Spin 01]. On the other hand, he is forced to a deeper search activity whenever he is the only one who can define the searching criteria and the search results seriously affect the success of his work. Thus, methods to discover criteria and combine them have to be provided.

Problem of database selection: "Which search engine to select?". The seeker has to decide which search engine he will use. The problem is well known in the WWW because the actual search engines are able to cover a limited portion of the web resources. Similarly, the seeker would need to consider and to compare the information resources with respect to the repository which is providing the metadata. For example, whenever the same resources could be available in more then one repository according to different policies, the seeker could be interested in determining where the suitable information resources are at the best conditions (i.e. price, license policy).

2.4 A conceptual framework for Metadata Analysis

This thesis proposes different methods of metadata analysis which should be seen as parts of a conceptual framework. The primary goal of the conceptual framework is to facilitate the seeker to move in a flexible manner through large information spaces. It is designed to understand the results provided by a search engine. In particular, it supports the user in the query refinement taking into account the implicit and explicit semantics of information resources. Figure 2.3 shows how the metadata analysis is conceptually correlated to the repositories and to a search engine. The metadata analysis is characterized by three main phases:

[3] That is independent from the sophistication of the ranking measure adopted. For example, Semantic Search [Guha 03] improves both the proportion of relevant material actually retrieved (recall of results) and the proportion of retrieved material that is actually relevant (precision of results) but if the query is based on a wrong criteria it fails as well as the other kinds of ranking.

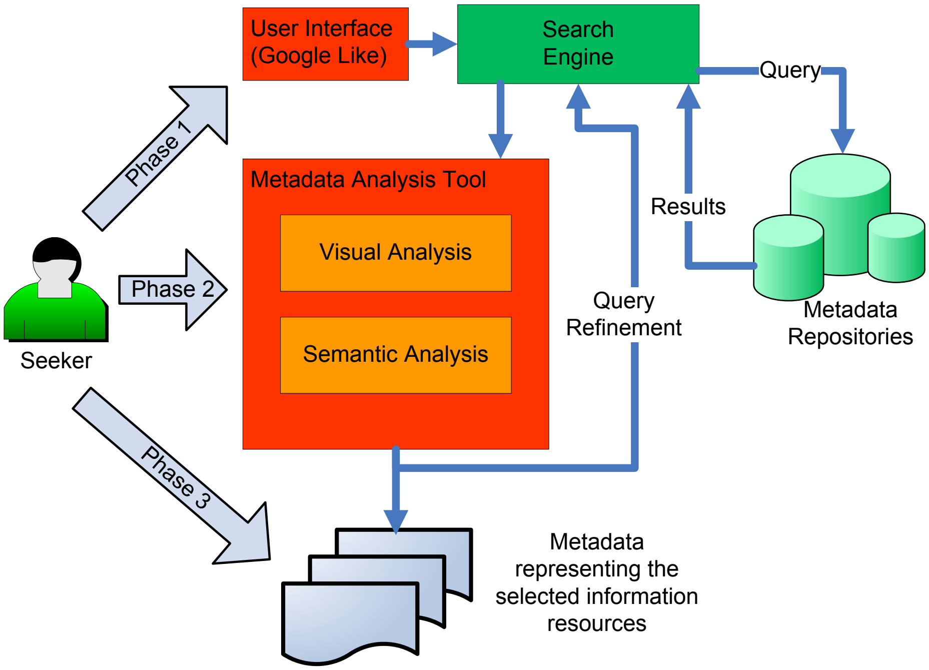


Figure 2.3: *Metadata Analysis Framework*

- phase 1: the seeker poses a preliminary query by a google like interface;
- phase 2: the seeker performs an appropriate query refinement applying the metadata analysis on the retrieved metadata items;
- phase 3: the seeker gets the list of the most suitable datasets.

The metadata analysis relies on two kinds of analysis:

- The **visual metadata analysis**. It is based on the application of visualization techniques and interaction functionalities. It aims to involve the seeker in the search activity providing an intuitive way to pose the query and a support to get a summarized view of the retrieved metadata items. Moreover, it takes advantage from implicit semantics discovering interesting and a priori unknown patterns.
- The **semantic metadata analysis**. It is based on the exploitation of the explicit semantics encoded in ontologies. It has been originally developed to face with the vocabulary problem, and successively it has been extended with the context dependent similarity and the semantic granularity. The similarity provides a context dependent way to sift among information resources, whereas the semantic granularity paves the way for browsing them at different levels of detail.

The two kinds of metadata analysis are illustrated in the following respectively in chapters 3 and 4.

Chapter 3

Visual Metadata Analysis

The chapter introduces the visual approach for the analysis of metadata of information resources. In particular, I will refer to the search for geographical information resources as specific application domain for the proposed approach.
As discussed in the previous chapter the seeker is the only who can provide the proper search criteria, but the definition of these criteria is strongly affected by his limited knowledge. According to Belkin et al. this is a result of the Anomalous State of Knowledge (ASK) [Belk 82a, Belk 82b]. They argue that the need for new information arises whenever the knowledge needed to carry out an activity is incomplete. In this case, the user decides to complete his knowledge by searching for new information but he is not able to precisely specify the information that he needs. The anomalous state of knowledge forces the seeker to enter into dialogue with the systems engaging a query refinement process. As result, a complete involvement of the seeker during the search process is one of the key aspects for a successful search activity.

This chapter aims to demonstrate as the visual metadata analysis can be adopted to obtain this involvement. This research started within the European research project "INformation VIsualization for SIte Planning" (INVISIP IST-2000-29640) whose main objective was to developing new mechanisms to analyze geographical metadata. The visual metadata analysis relies on Information Visualization and Visual Data Mining techniques. It is designed to increase the seeker abilities in the query formulation supporting him in the exploration of unfamiliar spaces and amplifying the cognition of the results.

The contribution of the chapter to the purposes of the thesis is fourfold. Firstly, it demonstrates that visual metadata analysis is useful in real application to search and compare the information resources. Secondly, due to the crucial role that geographical information resources play in numerous business and government application, it demonstrates that the resolution of the issues addressed by this thesis might have a great importance and social impact. Thirdly, it discusses the contribution of the visual analysis with respect to the requirements outlined in the section 2.3. Finally, the human evaluation made within the INVISIP project shows that the visual tools are positively perceived in the search activity. Beside INVISIP focuses of the geographical information resources similar tools can be applied to other application domains. Anyway, the visual metadata analysis cannot solve all the problems unless considering the seekers'/providers' background knowledge. This remark has brought to develop the semantic metadata analysis, which will be described in the next chapter.

The chapter is organized as follows. It first discusses the basics concepts concerning the Information Visualization and Visual Data Mining techniques (section 3.1). It considers the search and selection of geographical information resources as a case study (section 3.2). Then the potentiality of visual metadata analysis will be demonstrated through the experience performed within the European project INVISIP [INV] (section 3.3). Finally, it discusses the contribution of the proposed approach with respect to the metadata analysis requirement in section 3.4 and the conclusion of chapter in section 3.5.

3.1 Information Visualization & Visual Data Mining

3.1.1 Definitions

The Information Visualization (IV) is a field in computer science which aims to improve the interaction between humans and digitalized information. It stems from the awareness that the real problem is not increased access to information, but greater efficiency in finding useful information. According to Card et al. [Card 99] "Information Visualization is the use of computer-supported, interactive, visual representation of abstract data to amplify cognition". Hearst [Mart 03] defines Information Visualization as "the depiction of information using spatial or graphical representations, to facilitate comparison, pattern recognition, change detection, and other cognitive skills by making use of the visual system". Information Visualization has encountered an increasing interest due to its attractive promise: to improves human intelligence by increasing the rate at which people can find and use relevant information. Information Visualization paradigm takes advantage from the human ability to process information pre-attentively namely without to focus the attention. The processing which takes less than 200 - 250ms is qualifies as *pre-attentive*. It does not imply much eye movements which take at least 200ms. More in general, if a decision takes a fixed amount of time regardless of the number of distractors, it is considered to be pre-attentive. A limited set of visual properties are processed pre-attentively as demonstrated by the experiments shown by [G 97]. As a consequence it is important to distinguish between what has to be perceived at glance and what can be recognized later in order to design effective visualizations.

Visual Data Mining (VDM) stems from the fields of Information Visualization and Data Mining. Visual Data Mining aims at integrating the human in the data mining process and applying his abilities to the large data sets available in today's computer systems. For this purpose techniques which provide a good overview of the data and use the possibilities of visual representation for displaying large amounts of multidimensional data are especially important. These visualization techniques are used in the process of hypotheses generation, where the user is guided by the feedback provided by the visualizations and learns more quickly about the properties of data in the database [Keim 96]. The Visual Data Mining adapts the Information Visualization techniques adding some data mining algorithm as post- or pre-processing tool. As result Information Visualization provides techniques to sift information whereas Visual Data Mining extends the Information Visualization to show unknown and novel implicit patters.

Based on the balance and sequence of the automatic (i.e. data mining algorithms) and the interactive (visual) part in the knowledge discovery process, Ankerst [Anke 01] proposes the following classification of the existing Visual Data Mining approaches:

Visualization of the data mining result. (Figure 3.1a) An algorithm performs the data mining task by extracting patterns, which are then visualized to be more interpretable. Based on the result of the visualization, the user can rerun the data mining process with different parameters.

Visualization of an intermediate result. (Figure 3.1b) An intermediate result of the data mining algorithm is visualized, from which the user retrieves the interesting patterns and then potentially reruns the algorithm. One of the basic motivations for this approach is to make the algorithmic part independent from an application. A complete data mining algorithm can be very useful in a certain domain but may have severe drawback in another one. Since there is not a data mining algorithm suitable for all applications, the core part is performed and serves as multipurpose basis for further analysis directed by the user.

Visualization of the data. (Figure 3.1c) Data is visualized without applying any Data Mining algorithm. The user has a possibility to control the search completely from the beginning of the process.

In this chapter I am not defining brand new techniques in the fields of Information Visualization and Visual Data Mining. The chapter aims at demonstrating as visual techniques are useful to meet part of the requirements mentioned in section 2.3 in a specific and relevant context as the geographical domain. Thus a complete state of art of all visualization, interaction and pre-processing techniques is merely out of the purpose of this thesis. The analysis of the state of art in Information Visualization and Visual Data Mining we have performed at the beginning of our activity in INVISIP is presented in the project deliverables [Hais 02]and [De M 02]. Moreover, we discussed how the existing visual tools can ease the problems pertaining the search in World Wide Web and Semantic Web in [Albe 05a]. This discussion has been included in this thesis in the Appendix B. In the next section, I am providing a flavor of

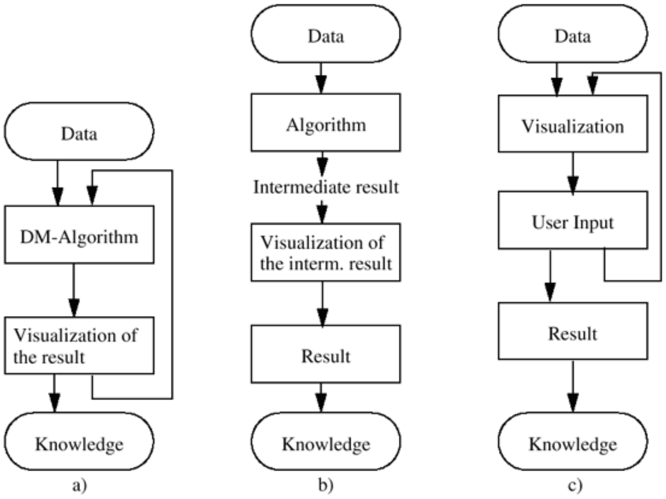


Figure 3.1: *Three Visual Data Mining approaches according to [Anke 01]*

the most representative techniques starting from the well known state of arts made by Keim [Keim 02b], Keim et al. [Keim 02a] and Börner et al. [Born 03].

3.1.2 Overview of Information Visualization and Visual Data Mining techniques

This section introduces an overview of the main techniques provided by Information Visualization and Visual Data Mining. They are classified with respect to four factors:

- data which the techniques are able to visualize;
- visualization techniques;
- interaction and distortion techniques;
- pre/post processing algorithms.

I start from the point of view of Keim illustrated in [Keim 02b]. According to Keim the different techniques adopted by Information Visualization and Visual Data Mining can be classified only with respect to three factors: the data they are able to visualize, the visualization techniques and finally, their interaction/distortion techniques. Keim maps these factors in different orthogonal axes (see figure 3.2) emphasizing that they are independent features of the visualization techniques. However, according to my experience, the independence is

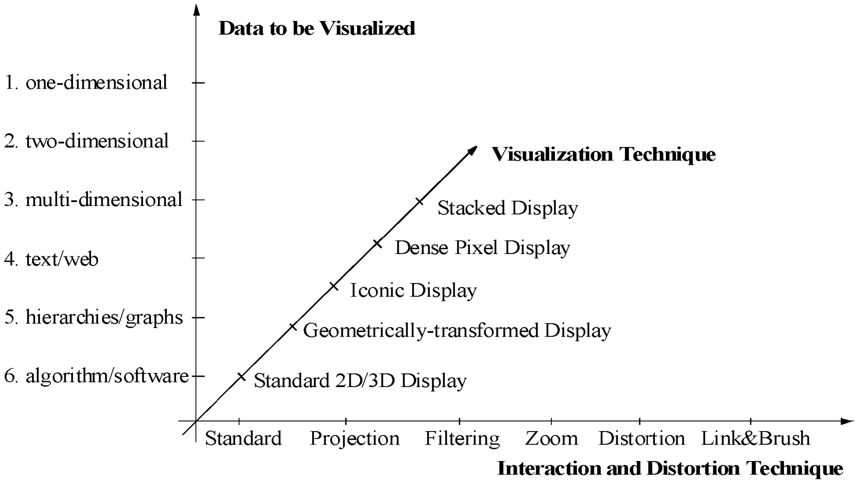


Figure 3.2: *Classification of Information Visualization techniques [Keim 02b]*

not completely true, at least not all the matches among data, visualizations and interaction/distortion techniques result equally satisfying.

The *algebraic properties* of data is one of the main factors influencing the worth of each match. For example, since it is intrinsically easer to map numbers rather than text in 2D-3D space, the standard 2D and 3D visualizations might result more suitable to represent numerical data rather than textual data. According to Mackinlay [Mack 86] the algebraic properties of data induce the distinctions in categorical (also named nominal), ordinal and quantitative domain set: a domain set is categorical when it is a collection of unordered items, such as {Jay, Eagle, Robin}. A domain set is ordinal when it is an ordered tuple, such as (Monday, Tuesday, Wednesday). A domain set is quantitative when it is a range, such as [0,273][1].

Of courses, supposing to have two dimensional categorical data set {DM, AI, CG} and {Full Journal, Book, Proceeding}, it is possible to map the categorical values in two distinct axes and represent the data occurrences as dot in a 2D space (e.g. first axis represents "DM", "AI","CG" placed in equally spaced positions and second axis represents "Full Journal", "Book", "Proceeding"). However, doing that, we force relations among the categorical values that are not necessarily true. In fact, we implicitly state that the "DM" is more similar to the "AI" than the "CG" as well as that the similarity between "DM" and "AI" is equal to the similarity between "CG" and "AI".

This simple example recalls that the axes mentioned by Keim are orthogonal supposing to have proper pre-post processing methods which maps, reduces, re-elaborates the data. Due to the relevance of these methods in order to set up the visual metadata analysis they will be

[1] The text can be seen as a complex case of categorical.

considered in the overview as a kind of fourth virtual axis. In the below I provide a description of the axes illustrated in figure 3.2 including this fourth virtual axis.

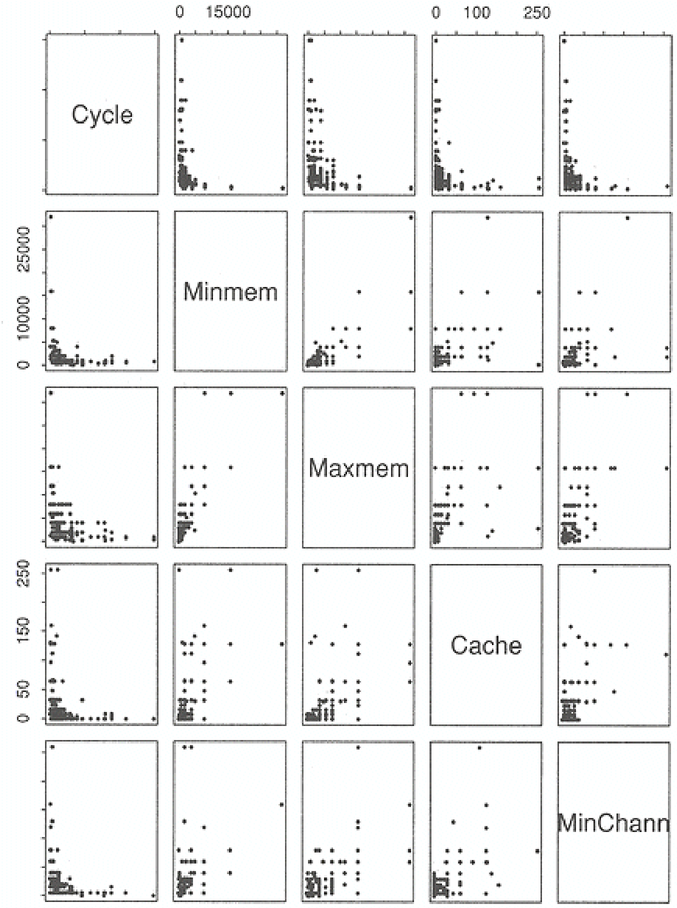


Figure 3.3: *A scatter plot matrix for data with 5 variables.*

Data to be visualizes. Data is characterized according to their *dimensionality* and *density* of their values. The dimensionality intuitively corresponds to the number of axes that are needed to represent them, and the density of the values. Keim subdivides the data to be visualized in the following categories:

- **one-dimensional data**, it has usually on dense dimension. A typical example of one dimensional data is the temporal data;
- **two dimensional data**, it has two distinct dimensions. A typical example is geographical data, where the two distinct dimensional are longitude and latitude. X-Y plots are the usual method to represent the two dimensional data;

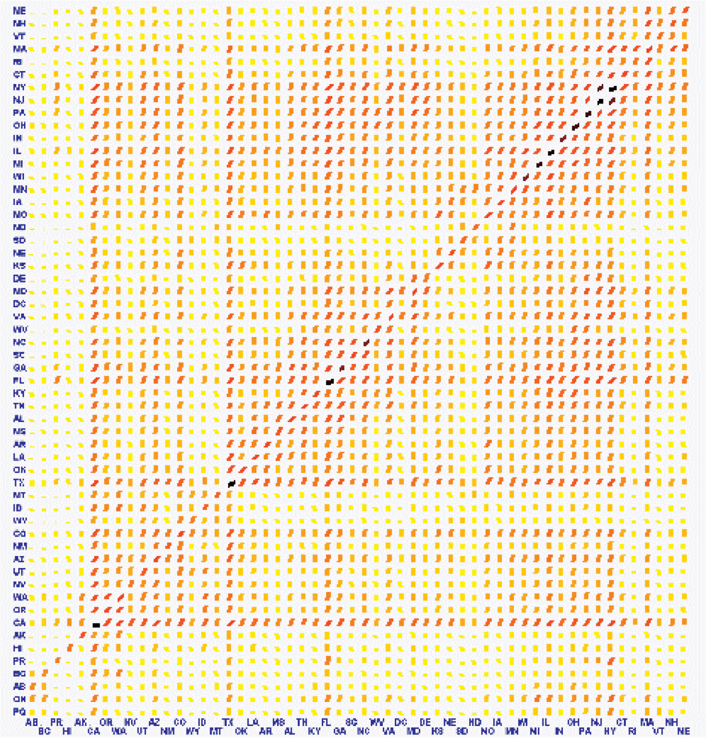


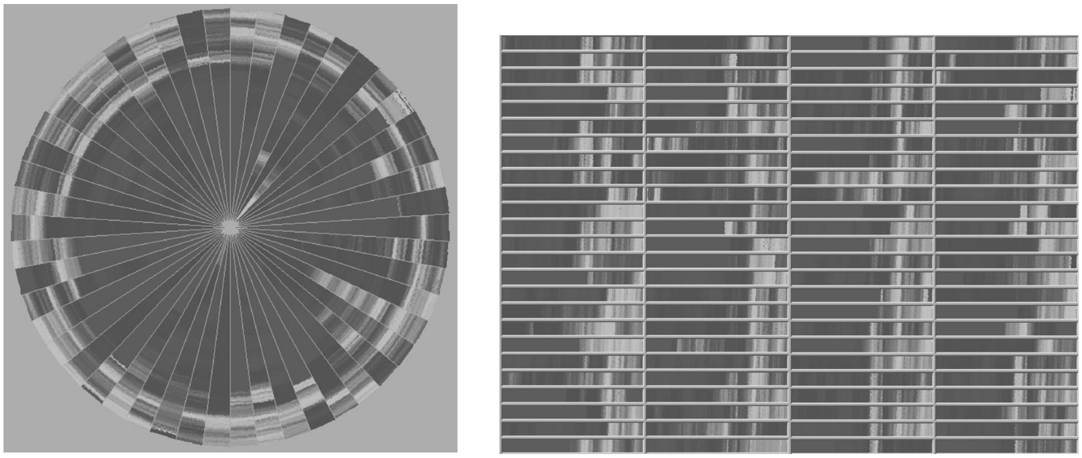
Figure 3.4: *An example of needle grid view [Abel 02].*

- **multidimensional data**, it has more then two dimensions. Examples are tables in relational databases. Since there is no simple mapping between multidimensional data to two dimensions of the screen, more sophisticated visualization techniques are needed;

- **text and hypertext**, it cannot be described easily in terms of dimensionality nor mapped in numbers. In most of the cases, it needed to be transformed into the description vectors to be visualized;

- **hierarchy and graph**, which can be used to represent the relationships among data items. Examples are the e-mail interrelationships among people, their shopping behavior, and the structure of a hard disk as well as hyperlinks in the World Wide Web. Graphs are usually used to represent these interdependencies. They are made of a set of nodes representing objects, edges which are connection between nodes.

- **algorithm and software**, which need to be visualized in order to support in the ever larger software projects. The goal of visualization is to ease the software development by understanding written code, showing the flow of information in a program and so on. A interesting overview is illustrated in [Stas 98].



(a) (b)

Figure 3.5: *Dense pixel displays [Keim 02b]: (a) circle segments technique; (b) recursive pattern technique;*

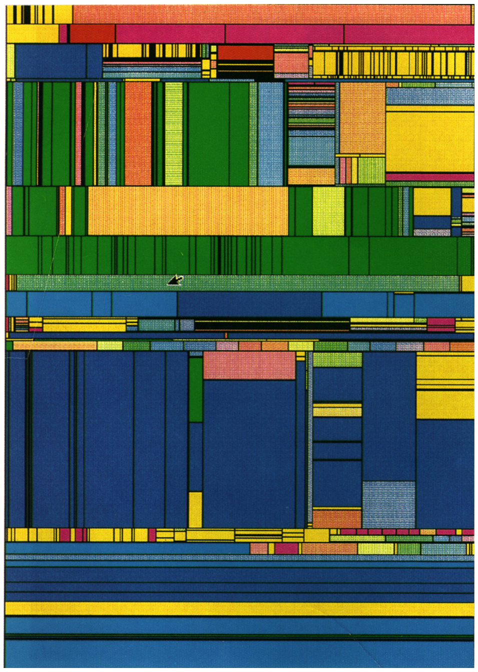


Figure 3.6: *Example of staked display: treemaps [Shne 92].*

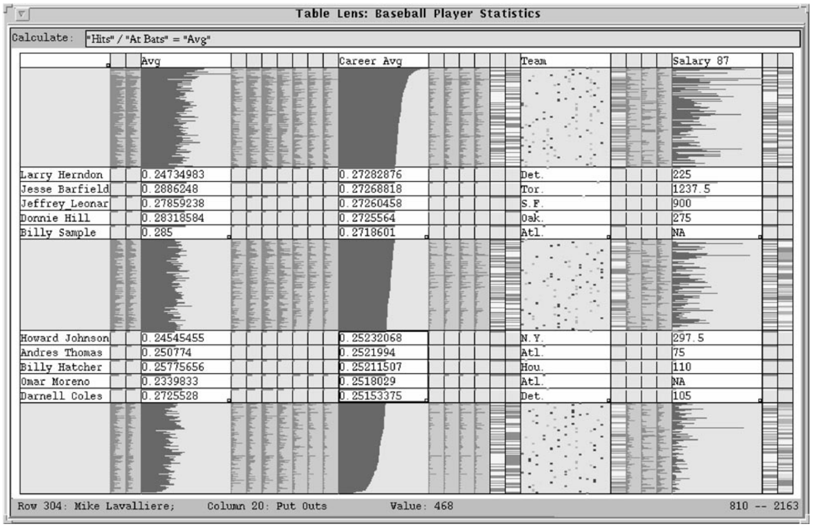


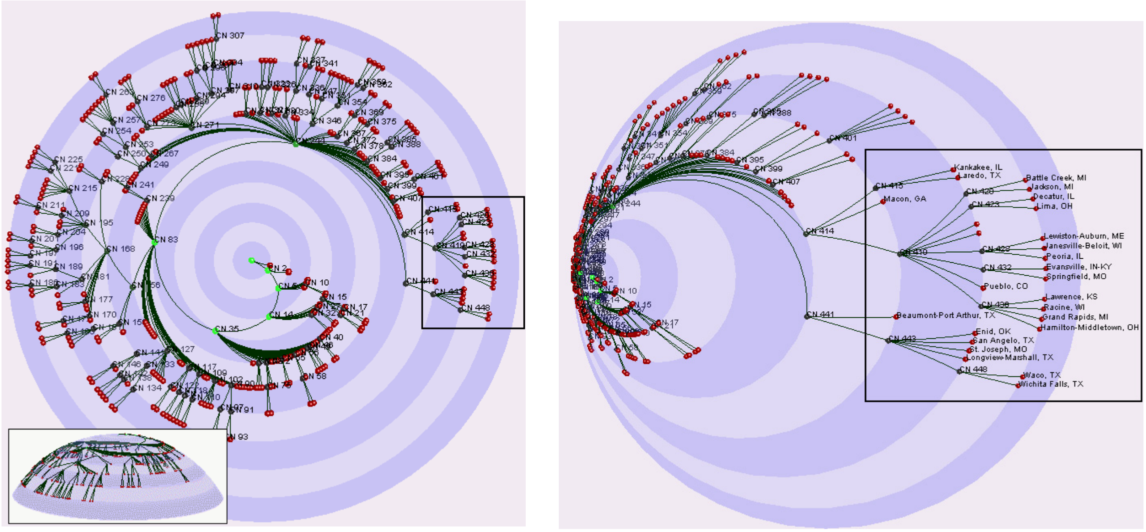
Table Lens: Baseball Player Statistics

Calculate: "Hits" / "At Bats" = "Avg"

	Avg	Career Avg	Team	Salary 87
Larry Herndon	0.24734983	0.27282876	Det.	225
Jesse Barfield	0.2886248	0.27268818	Tor.	1237.5
Jeffrey_Leonar	0.27859238	0.27260458	S.F.	900
Donnie Hill	0.28318584	0.2725564	Oak.	275
Billy Sample	0.285	0.2718601	Atl.	NA
Howard Johnson	0.24545455	0.25232068	N.Y.	297.5
Andres Thomas	0.250774	0.2521994	Atl.	75
Billy Hatcher	0.25775656	0.25211507	Hou.	110
Omar Moreno	0.2339833	0.2518029	Atl.	NA
Darnell Coles	0.2725528	0.25153375	Det.	105

Row 304: Mike Lavalliere; Column 20: Put Outs Value: 468 810 -- 2163

Figure 3.7: *Table Lenses [Rao 94].*



(a) (b)

Figure 3.8: *(a) Complex hierarchy; (b) Complex hierarchy with enlarged focus [Kreu 02].*

Visualization techniques. There is a large number of visualization which can be used to visualized data. They are grouped in classes corresponding to the basic visualization principles which can be combined in order to implement specific visualization systems.

- **Standard 2D-3D display**, which are widely adopted to represent two or three dimensional data, examples are x-y or x-y-z plots, bar charts, line graph (see figure 1 in [Stol 02]).
- **Geometrical transformed displays**, which aim at finding interesting transformation of multidimensional data sets. They include techniques from exploratory statistics such as scatter plot matrices [Andr 72, Clev 94] (see figure 3.3) and techniques which can subsumed under the "projection pursuit" [Hube 85]. Also the well known Parallel Coordinate diagram [Inse 90] is included in this technique. It will be illustrate afterward in the section 3.3.2 pertaining to the prototype we have developed.
- **Iconic displays**, which map the multidimensional data item to features of an icon. Icon can be little faces [Cher 73], needled icon (see figure 3.4 where each axis represents the states of the US and density of phone calls made between pairs of states is represented with a needle with multiple visual cues: color, angle, and length [Abel 02]), star icons [Ward 94], stick figure icon [Pick 88], title bars [Hear 95].
- **Dense pixel displays**, which map each dimension to a colored pixel and group the pixels belonging to each dimension into adjacent areas [Keim 00] (see figure 3.5(a)). Examples of these techniques are recursive techniques such as [Keim 95] and circular segment techniques such as [Anke 96]. The first are based on a generic recursive arrangement, where an attribute is represented with respect to its natural order. For example, considering the visualization of financial data, the recursion can be adopted to show the variations in daily price of 100 stock for a time slice by organizes each stock as a rectangles of pixels, where each pixels is a single variation (see figure 3.5(a)). The latter organizes the pixels from the center of a circle and continues outside by plotting on a line orthogonal to the segment having line in back forth manner (see figure 3.5(b)).
- **Stacked displays** are tailored to present data in hierarchical fashion. Whenever it is used to represent multidimensional data, the data dimension used for partitioning data and building the hierarchy has to be carefully selected. Examples of stacked display techniques are Treemaps [Shne 92, John 91] (see figure 3.6, where are visualized 850 files at four levels with color coding by title type. Final name pops up when cursor rests on a file), Cone Trees [Robe 91].

Interaction and distortion techniques. Beside the visualization techniques, interaction and distortion techniques are needed in order to support the exploration of data. Interaction techniques allow to the analyst to change the visualization according to his exploration objective. They are also adopted to relate each other different visualizations which are contemporaneously displayed. Distortion techniques help in data exploration process by providing means for focusing on details while preserving the overview of data.

- **Interactive projection**, which changes the projection in order to explore a multidimensional data set. An example is provided within GRandTour [Asim 85] which projects multidimensional data set in a series of scatter plots. The number of the projections is exponential in the number of the dimensions. It results intractable for large dimensionality. Then some techniques to consider a subset of all the possible combinations have been developed (see [Sway 92, Tier 90, Carr 97]).
- **Interactive filtering**, which allows to reduce the available data focusing on its subsets. This can be done directly by selecting the interesting items or specifying the properties of desired subset. The direct selection can be done by clicking on the visualized counterpart of the data item, but it is quite uncomfortable for reducing in large data set. Otherwise it is possible to write a query, but also this can be difficult. Therefore, a number of techniques have been developed to improve the interaction in filtering. Examples are Magic lenses [Bier 93], where magnifying glasses are adopted to performs the filtering directly in the visualization. The data under the magnifying glass is processed by the filter and displayed differently if according to the filtering process. Other examples of interactive filtering tools are InfoCrystal [Spoe 93] and dynamic queries [Gold 94].
- **Interactive zooming**, which allows to display data object larger and also to change the representation of data presenting more details. A remarkable example of this technique is TableLens [Rao 94] (see figure 3.7). In this technique each data item is represented as one pixel height row in the visualization and it can be magnified on demand showing an increasing level of attribute details. Other examples are DataSpace [Anup 95], PAD++ [Bede 94].
- **Interactive distortion**, which shows portions of the data with high level of detail while the others are shown with decreasing level of detail. Hyperbolic and spherical distortion are rather popular they are shown respectively in the Scalable Framework proposed in [Kreu 02] (see figures 3.8(a) and 3.8(b)) and [Lamp 95]. And overview of distortion techniques is provided in [Leun 94].
- **Interactive brushing and linking**, which combines different visualization methods to overcame the shortcoming of single technique. The points that are brushed in a visualization are automatically highlighted in all the active visualizations. That eases the discovering of dependences and correlation among data. It will be illustrate afterward (section 3.3.2) describing the prototype we have developed. Examples of this technique are provided in Xmdv [Ward 94] and the scalable framework [Kreu 02].

(Pre-post) processing algorithms. The increasing volumes of information resources as well as the intrinsic difficulties in mapping non quantitative values in 2D-3D spaces force to adopt appropriate processing algorithms. Different kind of techniques are adopted, they can roughly grouped in methods to work out the similarity among information resources, methods to reduce the dimensionality of their information space and automated data mining techniques to discover implicit patterns.

- **similarity** According to Kreuseler et al. [Kreu 02] it is important to define adequate measures to work out similarity because they are prerequisite for implementing many pre-processing approaches. Similarity has to be computed on values with different algebraic properties. Euclidean and Minkowski distances can be adopted when values are quantitative, but in order to face with non quantitative attributes, like text, more complex techniques are needed. For example, in context as the analysis of scientific paper citation, where the most of variables are textual, co-occurence similarity and vector space model have been successfully adopted [Born 03]. The most common *co-occurrence similarities* are co-term, co-classification, author co-citation, and paper co-citation. Two of the more common similarity formulas used with co-occurrence are the simple cosine and Jaccard norms. Each counts the number of attributes common between two units (e.g., the number of terms in common between two articles).
  The *Vector Space Model* (VSM) can be adopted to manage the text. It was originally developed for information retrieval by Salton et al. [Salt 75]. It is a widely used framework for indexing documents based on term frequencies. Each document (or query) is represented as a vector in a high dimensional space. Dimensionality is determined by the number of unique terms in a document corpus. Non-significant words are removed from the document vector. Terms are weighted to indicate their importance for document representation. Most of the weighting schemes (e.g. the inverse document frequency [Salt 75]) assume that the importance of a term is proportional to the number of documents the term appears in. The similarity between documents (or between a query and a document) can be subsequently determined by the distance between vectors in a high-dimensional space. The most popular similarity measure is the cosine coefficient, which defines the similarity between two documents by the cosine of the angle between their two vectors. It resembles the inner product of the two vectors, normalized (divided) by the products of the vector lengths (square root of the sums of squares).
  Other techniques are specifically designed to map categorical values in multidimensional spaces. Rosario et al. [Rosa 04, Rosa 03] propose an approach named distance-quantification-classing to investigate both how to assign an order and spacing among the nominal values, and how to reduce the number of distinct values to display. Once the mapping is done implicitly the similarity is obtained.

- **Dimensionality reduction.** The dimensionality of multivariate data needs to be reduced for displaying on the 2D/3D dimensional spaces. The problem is tackled by applying mathematical dimensionality reduction techniques to map n-dimensional data into a 2D or 3D space [Born 03]. According to [Keim 02a] examples of these techniques are:

- **Factor Analysis** such as *Principal Components Analysis* (PCA), which can transform a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components [Clif 87].
- **Multidimensional Scaling**, which represent a class of optimization techniques where lower dimensional representations approximating the object's distances in information space are generated. The specific algorithms differ in the type of stress function used, which is based on the dissimilarities of the information objects and on the type of optimization strategy [Cox 94, Cox 94].
- **Spring Models** [Thei 98] make use of a similar optimization approach. In this case, however, predefined feature points (one for each attribute) are fixed in the target space. An information object is mapped into the target space by calculating the effect of virtual springs attached to the information object and each feature point. The effect of the springs depends on the distance of the feature point from the information object in the information space. The resulting representation resembles the semantic distances of the information objects but highly depends on the selection of the feature points.
- **Kohonen Networks** (self-organizing maps, SOM [Koho 97]) provide another approach for generating a spatial layout of information objects. Kohonen networks lend from biological models and provide both, a dimension reduction of the attributes to two dimensions and a classification of the data objects based on their features in the information space.

Pertaining to high-dimensionally textual spaces, Bömer et al. [Born 03] mention the Latent Semantic Analysis (LSA), also called Latent Semantic Indexing (LSI). This technique can be adopted to ease the vocabulary problem [Deer 90, Land 98]. LSA handles synonymy (variability in human word choice) and polysemy (same word has often different meanings) by considering the context of words. It uses an advanced statistical technique, singular value decomposition (SVD) to extract latent terms. A latent term may correspond to a salient concept that may be described by several keywords.

- **Automated data mining techniques.** They are adopted in order to discover implicit patterns arising from the information space. *Association rules*, *classification* and *clustering* are the typical data mining techniques considered. A complete introduction of these techniques can be found in [Fayy 96], in the following a summarization of their purposes and example of how they can be combined with visualization techniques are given according to [Keim 02a].
  - **Association rules** are statistical relation between two or more items. Association rules tell us that the presence of an item implies the presence of

other items. The application of supermarket basket is the classical example adopted to explain the use of association rules. In this context, the association rules reveal if two distinct items are bought together. For example, they could show that the 70% clients buying the milk are used to buying the bread. This information might result extremelly useful to positionate the two products increasing the clients' satisfaction. Example of tool combining association rules and appropriate visualization are Mosaic and Double Decker Plots [Hofm 00] and SGI MineSet's Rule Visualizer [Brun 97].

- **Classification** is the process of developing a classification model based on a training data set with known class labels. The attributes of the training data set are analyzed and an accurate description or model of the classes based on the attributes available in the data set is developed. The class descriptions are used to classify data for which the class labels are unknown. Classification is sometimes also called supervised learning because the training set is used to teach the system how to classify the data. Examples of how such technique can be combined with visualization are the decision tree visualizer in SGIs MineSet system [Brun 97] and the visual classification proposed in [Anke 00].
- **Clustering** is the process of finding a partitioning of data set into homogeneous subsets called clusters. Unlike classification, clustering is often implemented as a form of unsupervised learning. This means that the classes are unknown and no training set with class labels is available. Examples of techniques to visualize clustering are OPTICS (Ordering Points To Identify the Clustering Structure) [Anke 99] and HD-Eye system [Hinn 99].

3.2 Searching for geographical information resources in an open environment

In this section, I present my applied research to the geographic domain whose resources play important role in many applications. The geographical domain is significant for the purpose of this thesis as it is a domain where the assumptions of "open environment" and "complexity of metadata" are satisfied. In particular:

- the search and selection of geographical information resources is performed in an Open Environment relying on the metadata;
- the metadata analysis is needed because of the complexity of metadata and the number of entries the user has to consider during his search.

Geographical information resources are indispensable to support public administrators during the definition of national policies, or to evaluate the environmental impact of political choices.

For these purposes, vast collections of heterogeneous geographic data are generated from numerous providers. In addition to publishing the data on CDs and other media, providers are shifting toward making the information available on the Web following the explosive growth of Internet and its users.

The importance of sharing and collecting geographic data is rapidly increasing: usually each country relies on private or public structures to maintain updated geographic information at a regional and at a national level. However, the request for sharing and collecting geographic information crosses the countries border. Spatial Data Infrastructure (SDI) beyond the country borders are arising to integrate geographic information services which allow to identify and access geographic information from a wide range of sources (see [Smit 02], [Scho 98], [Jone 02]).

At European level the importance to access to geographic information has been recognized as essential to ease the definition of coherent European policies [EIONET]. INSPIRE [Smit 02] initiative is proposed to make accessible the resources to each European country by defining a framework for the gradual creation of a harmonized spatial information infrastructure. Other initiatives like SPIRIT (Spatially-Aware Information Retrieval on the Internet) [Jone 02] propose a worldwide access by getting geographic information directly by surfing in Internet. They usually aim to design and implement a high level of intelligence web search engine to find documents and datasets on the web relating to places or regions referred to in a query. In a SDI, the quantity and the heterogeneity of data raise the problem to define instruments to manage a large amount of distributed data: the concept of metadata has been introduced to describe geographic data. Digital archives of metadata such as Metadata Information System (MIS) and Catalogue Service (CS) are developed to manage such information. In particular, the initiative like INSPIRE and Spirit spend a huge effort to solve the problems of publishing, delivery of data and metadata, retrieval of distributed resources. However they pose less attention to the comparison and to the exploration of geographical information resources for data selection activity. Such a problem is broadly faced in my research by the activity performed within the European Project INformation VIsualization for SIte Planning (INVISIP IST-2000-29640) [INV] where the Information Visualization techniques have been adopted in order to visually explore and query the available geographical information resources (namely geographical data set).

3.2.1 The seeker's problem in geographical information resources searching

The search for geographical information resources is a particular kind of search activity which aims to select the most appropriate data for a specific application. Two main aspects are critical in the selection of geographic information resources:

- Usually it is not possible to access to the geographic resources to have a look and to realize at a glance the information that they contain since geographic data are resources that can be heavy in terms of Kbytes, or that can be not available for free.

- To compare different geographic resources requires strong efforts since data are available in a huge kind of variety. They differ in characteristics like Scale, Reference System, Geographic Extension, Themes, Quality, Fees and so on.

Metadata are adopted to overcome these drawbacks providing a detailed description of the geographic information resources according with a specific standard. They represent a first level of data integration and allow to compare resources provided by different organizations. Moreover they represent a mean to choose geographic data without resource download.

The vast collections of geographic resources determine the generation of a large set of metadata. Furthermore the complexity of geographic data forces metadata to be characterized by many attributes and to be represented in a multidimensional information space. Instruments able to manage this large set of metadata and their multidimensionality are needed. A lot of Metadata Information System (MIS) and Catalogue Systems (CS) are generated to organise and manage metadata. [Gobe 98] gives an overview of MIS and CS for geographic data and provides more details about the metadata concept and the related initiatives. In particular, different initiatives have been carrying out to define metadata standard (ISO 19115 [ISO19115 03], FGDC[FGDC 98], CEN/TC 287 ENV 12657 [centc 98]) and to facilitate the searching of metadata (UDK [Swob 99], [Stei 97]). They propose browsing tools for metadata that provide the results as a list of textual information. Considering the multidimensionality and the quantity of metadata, such list of information overwhelms user abilities of comprehension. Visual tools are needed to support the user in the comprehension of the searching results.

The visualization of metadata might appears as a first step toward the solution of such problem [Alpe 96]. However, a visual based approach should also ensure a direct and strong involvement of the seeker who is the only allowed to judge on the suitability of the resources for his needs. As consequence of the strong human involvement different factors should be carefully taken into account:

Seeker knowledge. The seeker has often only a perception of his information needs: he has a limited knowledge of what he is looking for (see the discussion in chapter 2 pertaining to ASK problem).

Seeker and provider relationship. Seekers and information providers have different skill levels and different domains of knowledge. Moreover there is usually no direct interaction between them (the user cognitive space differs from the information space as well as the providers' cognitive spaces).

Seeker anxiety. The gap between what the seeker understands and what he thinks he should understand generates anxiety. This happens whenever information does not fulfill his needs.

These different factors get the user into some problems when he searches for geographic information resources. In particular, my direct interaction with different INVISIP[2] partners experienced in the management and use of geographical information resources has allowed to verify that the complexity of the ISO 19115 standard and the number of attributes that characterize it, may lead to two particular problems [Albe 04, Albe 05c]:

Unfamiliarity with attributes: The searching criteria that the seeker is able to perform might not be enough to successfully end its selection activity. Therefore he needs to refine his criteria using attributes he is not familiar with. In other terms he needs to perform his selection in an unfamiliar information space consisting of metadata attributes and their values.

Data missing: the metadata database might not contain the data the seeker is looking for. Hence he is forced to define new criteria to find similar data.

3.3 INVISIP: INformation VIsualization for SIte Planning

The search for geographical information resources is considered as a case study where demonstrate the Visual Metadata Analysis need and usefulness. The case study has been developed within the project "INformation VIsualization for SIte Planning" (INVISIP IST-2000-29640) in which I have been actively participating during the first part of my PhD activity. INVISIP is a project founded by the European Commission within the Fifth Programme Framework which was aimed to support all involved parties in the site planning process: municipal authorities and departments, planning offices, data suppliers and citizens. Information Visualization techniques are used to improve search and analysis tasks, and to facilitate the decision-making process based on an existing metadata information system (MIS) for geographic data. A basic problem in the site planning process is the search for actual and expressive data and their analysis. In particular, spatial data are needed to analyze and realize planning objectives. INVISIP provides a technical platform as an aid to facilitate information access and data handling for the site planning process (time-saving, intuitive analysis).

The INVISIP framework provides mechanisms based on Information Visualization techniques to support the search for appropriate geographical information resources and to face with the aforementioned problems. INVISIP assumes that each geographical information resource is characterized according to the ISO 19115 geographic metadata standard. Focusing on ISO 19115 metadata standard, the attributes can be represented by categorical values or full text values and bounding box is used to express spatial extent attributes. Different visual analysis

[2] The partners experienced in the management and use of geographical information are two Municipalities (of Genoa (Italy) and Wiesbaden (Germany)) and three SMEs (D'Appollonia S.p.A (Italy), INREGIA (Sweden), THALES (Germany))

approaches are proposed within INVISIP in accordance with the types of attributes they consider or the type of representation used for the spatial extent.

Full text values the German partners of University of Konstanz [Klei 02, Limb 03b, Klei 03] provide the Visual Metadata Browser (VisMeB). VisMeB relies on SuperTable, a visualization approach to solve the problem of the exploration of metadata working mainly on attributes whose value is expressed as full text. They provide an entry point to pose the initial term based query and perform some visual refinement. This approach combines different visualizations into a so called SuperTable. The SuperTable has been realized and evaluated in two design variants: GranularityTable and LevelTable.

Spatial extent the German parnters of IGD Fraunhofer [Gobe 03] provide GeoCrystal system which focuses on the spatial extent and lets the user compose complex queries and visualize search results in a 3D space for geographic data.

Categorical values I and my colleagues together with the Polish and Swedish partners [Albe 03a, Albe 03b] propose an approach applied to the categorical attributes implemented in the Visual Data Mining tool (VDM). In particular, I was involved in the design of the visual approach to solve the data missing and unfamiliarity with attributes problems [Albe 04].

Categorical attributes play an important role both since they are numerically relevant (more than twenty metadata attributes are defined as categorical) and they represent important information such as maintenance attribute, progress; type of spatial representation, resolution, theme classification, etc. In the next paragraph I discuss our approach focusing on the categorical attributes. It combines the functionalities of automatic visualization and graphical interaction to enable users to uncover and extract hidden relationships in large data sets.

3.3.1 An approach for visual categorical metadata analysis.

In this paragraph I describe the visualization-based approach, we have defined within INVISIP, to analyse categorical metadata. The main idea is to simultaneously use visualization techniques, graphic interaction and a dynamic link among the visualization themselves using Brushing and Linking techniques [Keim 02a]. The approach is characterised by three iterative phases as depicted in figure 3.9: a visualization phase, an exploration phase, and a query-building phase.

Visualization phase. During the first phase different representations of metadata attributes and values are provided in order to give the user a compact and human understandable view of the available data. Different visualization techniques are provided and can be applied at the same time. They are classified according to the number of attributes they can display: single attribute and multi attribute visualization.

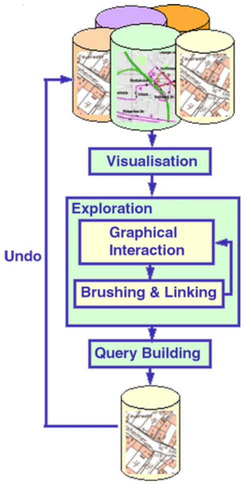


Figure 3.9: *Visual Analysis of categorical attributes-*

Exploration phase. The second phase is based on the analysis of the visualizations previously displayed to extract knowledge about metadata. In particular, single attribute visualizations provide the knowledge of the available values and quantitative information of metadata attributes, whereas multi attribute visualizations provide the knowledge on metadata attributes and the existing relationships. This task is performed using both interaction functionalities with the element displayed in a visualization, and brushing and linking to combine different visualization methods. The result of the exploration phase assists the user in the choice of both attribute and its values to define new query criteria.

Query building phase. In the third phase the criteria are completed and the query is generated. Finally the attribute values are graphically selected to express the query and the starting subset is reduced. As soon as the query is performed, all displayed visualizations are updated showing the new (sub)set of data.

Furthermore, if necessary, a new step of the process can be performed starting from the previous visualizations or activating new visualizations. Otherwise, if the results obtained does not satisfy user requirements, it is possible to delete some selections previously performed and return to an “old” data set (a so called Undo).

3.3.2 The Visual Data Mining Tool

This section describes the main functionality of the VDM tool, its architecture and an example of its application.

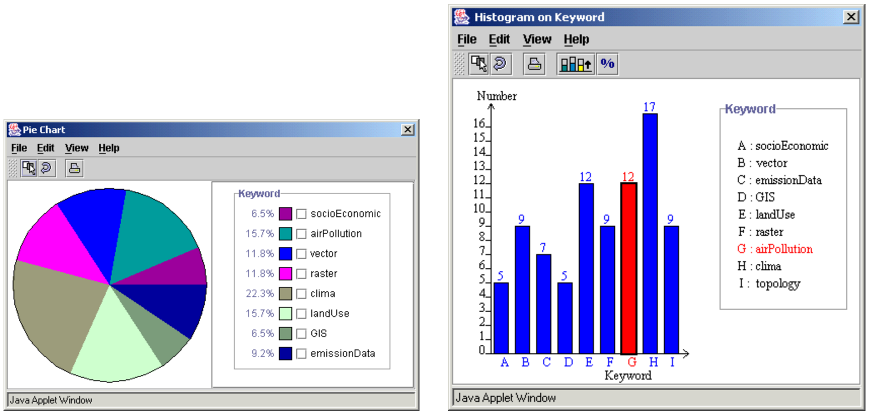
3.3.2.1 The Functionality of the VDM Tool

There are two different types of functionality provided within the VDM tool: the visualization techniques and the interaction functionalities.

The *visualization techniques* include two different types of visualization techniques:

- visualizations of one attribute (such as a pie chart and a histogram see figure 3.10) and
- visualizations of multiple attributes (such as table and a parallel diagram see figure 3.11).

Other visualization techniques can be included in the VDM tool in the future.



(a) (b)

Figure 3.10: *Visualizations of one attribute: piechart (a), histogram (b)*

A brief description of each visualization follows:

A *pie chart* (Figure 3.10(a)) shows the proportional size of categories that make up a data series, which represents values of one chosen attribute. It always shows only one data series and is useful when the user wants to recognize a significant element within the data series.

A *histogram* (Figure 3.10(b)) shows the number of objects for each specified value of the chosen attribute. Values are shown on the x-axis, numbers of objects on the y-axis. The histogram is useful to recognize the distribution of data objects and can help to identify potentially suspicious objects. These can be removed from further analysis by appropriate selection.

A *table visualization* allows the user to choose one or several attributes and visualize them in a table of values. Each attribute is represented by a column in the table. Each row of

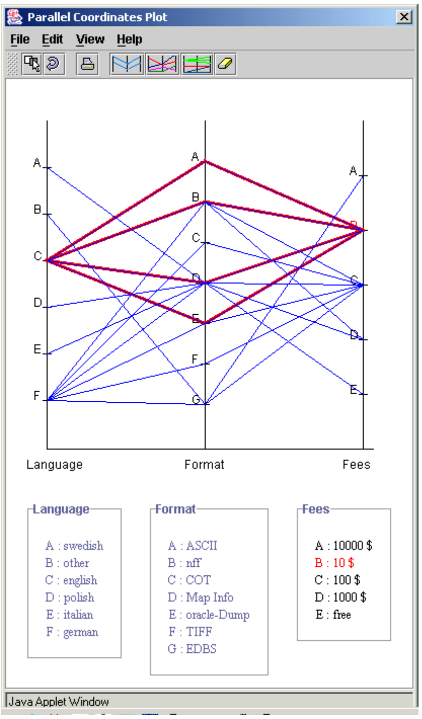


Figure 3.11: *Visualization of multiple attributes*

the table represents a data object. It is not a graphical visualization and it does not provide a very compact representation of data but nevertheless it is a common way to show the search results. From the users point of view, it can be useful to visualize a large number of attributes when the data reduction has already been performed by previously applying some other visualization technique.

A *parallel diagram* or *a parallel coordinates plot* (Figure 3.11) maps the attributes of the dataset onto vertical axes. Each data object in the dataset is represented as a piecewise linear line connecting the axes. The line intersects the vertical axes at the points that correspond to its attribute values. Since the line representing an object connects different attributes, it is necessary to select at least two attributes for a non-trivial plot.

There are two different kinds of *interaction functionality* implemented in the VDM tool:

- interaction between a single visualization and the user,
- interaction among different visualizations.

The first kind of interaction, which is between a visualization and the user, enables the user to explore the content of the visualization and to graphically extract selected subset of metadata from the metadatabase. The functionality of graphical selection for data exploration exists in all the visualization techniques, but varies according to the type of each technique. It links each graphic entity to the attribute which it represents. Graphical entities in question can be angular segments of a pie chart, bars of a histogram, rows of a table or lines in a parallel diagram. The values of the attributes are shown in the legend next to the visualization. The user can select a desired subset of objects by clicking on the graphical entities that represent their attribute values or by choosing one or several values in the legend. A special type of graphical selection is implemented in the parallel diagram. Unlike in the other visualizations the polygonal lines represent the correlation among different attributes rather than one attribute only. Figure 3.11 shows an example: the selected polygonal lines in red show the correlation between the specified cost of the datasets (10€) with their format and their language. The datasets with this cost can be obtained in four different data formats, but they are all in English.

The second kind of interaction helps to discover correlation among graphic entities represented in different visualizations. All the visualizations are interconnected according to the concept of *brushing and linking*. Brushing is an interactive selection process, while linking connects the selected data from the current visualization to other open visualizations. If the user has several different visualizations open (of one type or of more different types) and decides to perform a selection of objects in one of them, the graphical entities that represent this selection and correspond to the same subset of selected data objects are highlighted in each visualization, providing a better visual impression. When the selection is performed, all other graphical entities disappear from each open visualization.

3.3.2.2 The Architecture of the VDM Tool

The VDM tool consists of three main components designed as a Java applet in order to easily handle web-based explorations:

- the data manager that connects the VDM tool to different resources of metadata,
- the control panel which integrates all components and designed as a Graphical User Interface and
- the visualization wrapper component that provides a common template for the different visualization techniques.

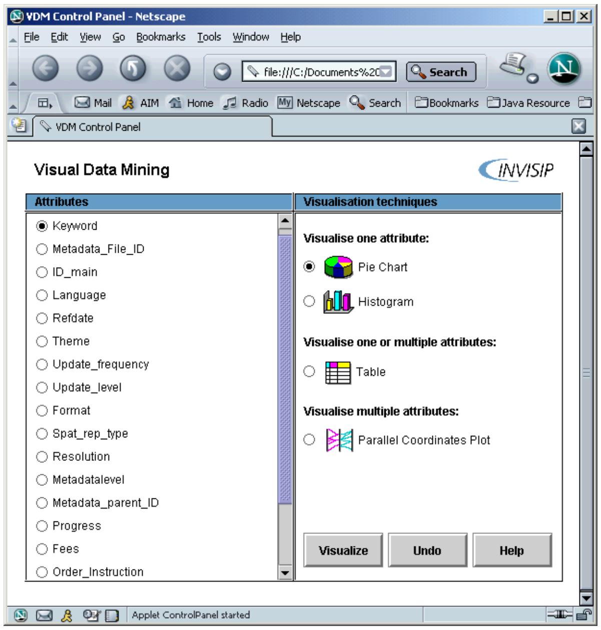
The *data manager* handles input and output of data. It is based on a table that contains all metadata that can be visualized. The data connection implemented so far is based on ODBC. The VDM tool is therefore open to integrate other database managers such as Oracle, SQL server and so on.

The *control panel* is the main component of the VDM tool. It provides a Graphical User Interface (GUI) as shown in Figure 3.12. The left side of the control panel shows the list of metadata attributes, while the right side shows all available visualizations. This gives the user the possibility to apply different types of visualizations to the selected metadata attributes. The control panel activates the visualizations contained in the data manager and manages the general layout of the different visualizations.

All the visualization techniques are based on the *visualization wrapper* component. In Java programming, a visualization wrapper is an abstract class that all visualizations have to extend. It provides the interface between all the visualizations and the control panel, as well as functionalities to activate and to update the graphs contained in the wrapper. It is also responsible for the look & feel (colours, character fonts, etc.) of all visualization techniques and for the common characteristics such as toolbar and menu.

3.3.3 Illustrative Examples

This paragraph provides examples of how to exploit the approach to search for geographic information, in particular to solve the "Query formulation" and "Retrieved results comprehension" problems. Once demonstrated these two it shows some examples illustrating how the approach can ease the "unfamiliarity with attributes" and "data missing".


VDM Control Panel - Netscape
File Edit View Go Bookmarks Tools Window Help
file:///C:/Documents%20
Search
Mail AIM Home Radio Netscape Search Bookmarks Java Resource
VDM Control Panel
INVISIP
Visualize
Undo
Help

Visual Data Mining

Attributes

- ● Keyword
- ○ Metadata_File_ID
- ○ ID_main
- ○ Language
- ○ Refdate
- ○ Theme
- ○ Update_frequency
- ○ Update_level
- ○ Format
- ○ Spat_rep_type
- ○ Resolution
- ○ Metadatalevel
- ○ Metadata_parent_ID
- ○ Progress
- ○ Fees
- ○ Order_Instruction

Visualisation techniques

Visualise one attribute:

- ● Pie Chart
- ○ Histogram

Visualise one or multiple attributes:

- ○ Table

Visualise multiple attributes:

- ○ Parallel Coordinates Plot

Applet ControlPanel started

Figure 3.12: *The VDM Control Panel*

3.3.3.1 Example 1: "Query formulation" problem

In the metadata analysis framework the queries are performed in a visual manner. Datasets used in the formulation of the query criteria are visualized in appropriate visualizations and starting from them, the query is formulated by a sequence of graphical selection and reduction. The available interactions are mapped as different logical operators: the selection is mapped into an OR ($\vee$) between the elements which are clicked on, whereas the reduction is mapped as an AND ($\wedge$) between two different selections. The query language supported by the framework corresponds to the subset of the conjunctive normal form as defined in [Mend 97], where the literals are equalities between attributes and desired values, and the negation of literals are not allowed. Let suppose to have the following scenario: the seeker has to acquire geographic data dealing with climatology content, written in English or in Italian and having a MapInfo file format. Since Language, Theme and Format are metadata attributes the query the seeker has to perform is quite simple and looks like:

$$\begin{gathered}(\text{Language=Italian} \vee \text{Language=English})\wedge \\ \wedge(\text{Theme=Climatology})\wedge(\text{Format=MapInfo})\end{gathered}$$

The query can be formulated applying a Parallel Coordinate Plot (PCP) (see Figure 3.13) to visualize the information related to Language, Theme and Format. The PCP provides a compact overview of the available features and the relations among the attributes. This technique maps the attributes of a dataset onto vertical axes and represents each data object as a piecewise linear line, and this polygonal line intersects the vertical axes at one point that corresponds to its attribute value. The query is performed by the following interactions:

1. Two selections on the histogram to formulate the preferences on Language. This is performed by clicking on "Italian" and "English", and as shown in Figure 3.13 the properties related to Italian and English dataset highlighted.

2. A reduction of the dataset according with the criteria defined by the Language selection. It is performed by clicking the toolbar reduction button in Figure 3.13.

3. A selection to formulate a preference on the Theme. It is performed clicking on "climatology" in the PCP .

4. A reduction of the dataset according with the criteria defined by the Theme selection.

5. A selection to formulate the preference on the format. It is performed by clicking on MapInfo in the PCP.

6. A reduction of the dataset according to the adopted criteria.

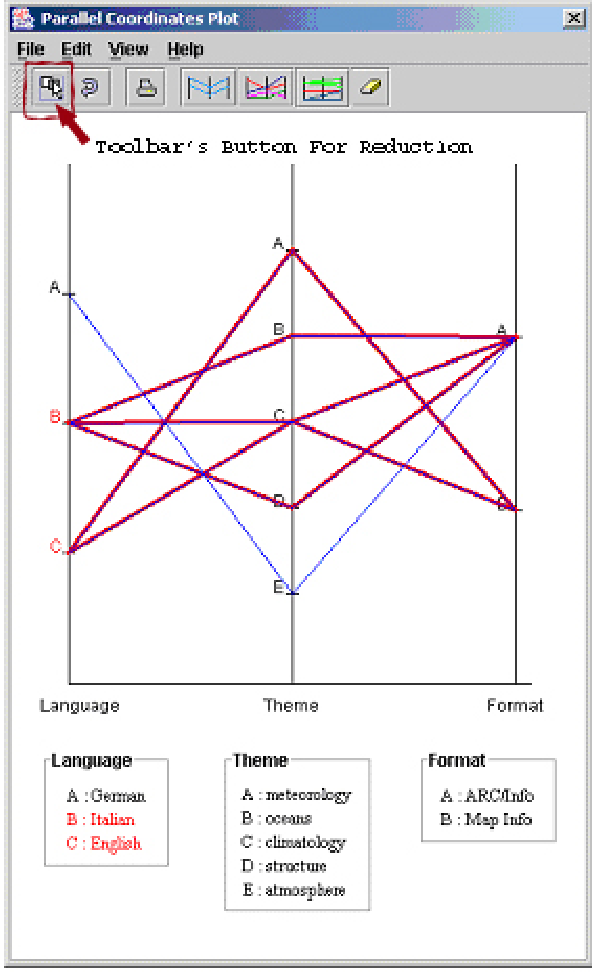
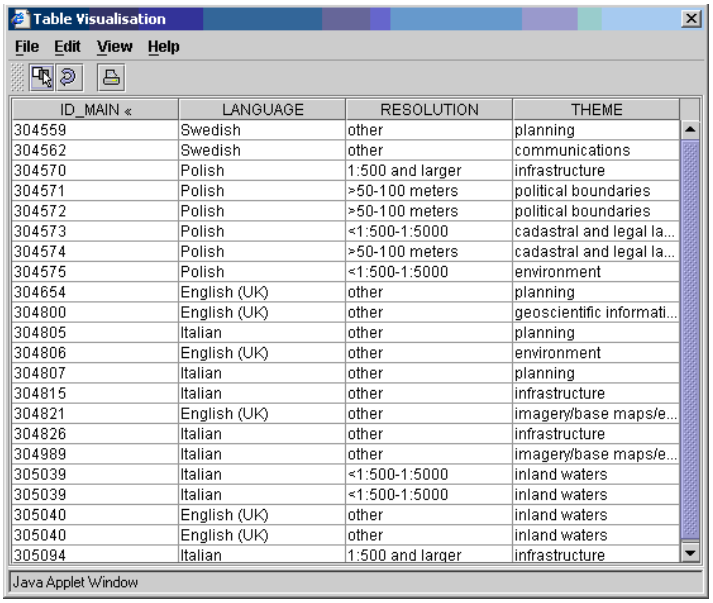


Figure 3.13: *PCP visualization of Language, Theme and Format.*

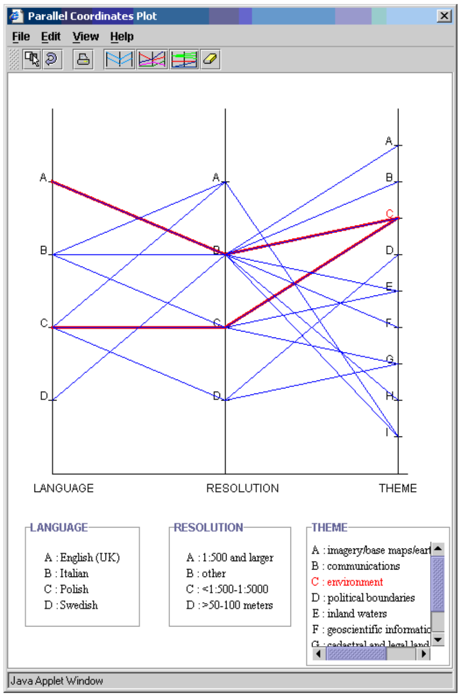
3.3.3.2 Example 2: "Retrieved Results Comprehension" problem

Let suppose to have the following scenario: the seeker is looking for geographic information related to specific criteria. He formulates a query specifying his preferences and then he obtains a large set of metadata related to the geographic data satisfying the query. The seeker needs to comprehend the retrieved results to choose the most appropriate one(s) for his needs. Search engines usually return the results set as textual list of items (e.g. table representation as in Figure 3.14(a)). It is a poor and redundant manner to represent the information; it results inappropriate for many tasks of analysis needed in the query refinement process. For example if the user is interested in "How are related the attributes language, resolution to the theme Environment?", the tabular representation forces the user to examine carefully each rows. On the contrary, a representation which visually correlates more information



ID_MAIN «	LANGUAGE	RESOLUTION	THEME
304559	Swedish	other	planning
304562	Swedish	other	communications
304570	Polish	1:500 and larger	infrastructure
304571	Polish	>50-100 meters	political boundaries
304572	Polish	>50-100 meters	political boundaries
304573	Polish	<1:500-1:5000	cadastral and legal la...
304574	Polish	>50-100 meters	cadastral and legal la...
304575	Polish	<1:500-1:5000	environment
304654	English (UK)	other	planning
304800	English (UK)	other	geoscientific informati...
304805	Italian	other	planning
304806	English (UK)	other	environment
304807	Italian	other	planning
304815	Italian	other	infrastructure
304821	English (UK)	other	imagery/base maps/e...
304826	Italian	other	infrastructure
304989	Italian	other	imagery/base maps/e...
305039	Italian	<1:500-1:5000	inland waters
305039	Italian	<1:500-1:5000	inland waters
305040	English (UK)	other	inland waters
305040	English (UK)	other	inland waters
305094	Italian	1:500 and larger	infrastructure

(a) A tabular representation of a result set.



(b) The PCP representation of a result set.

Figure 3.14: *Example of improved retrieved result comprehension.*

(e.g. the Parallel Coordinate Plot in Figure 3.14(b)) is more suitable. Figure 3.14(b) clearly shows what features are available and the relations among the attributes. Interacting with the visualization, the user can easily understand the pattern which occurs in the result set. Moreover a selection on the theme "environment" automatically highlights the language and resolution properties of the selected dataset.

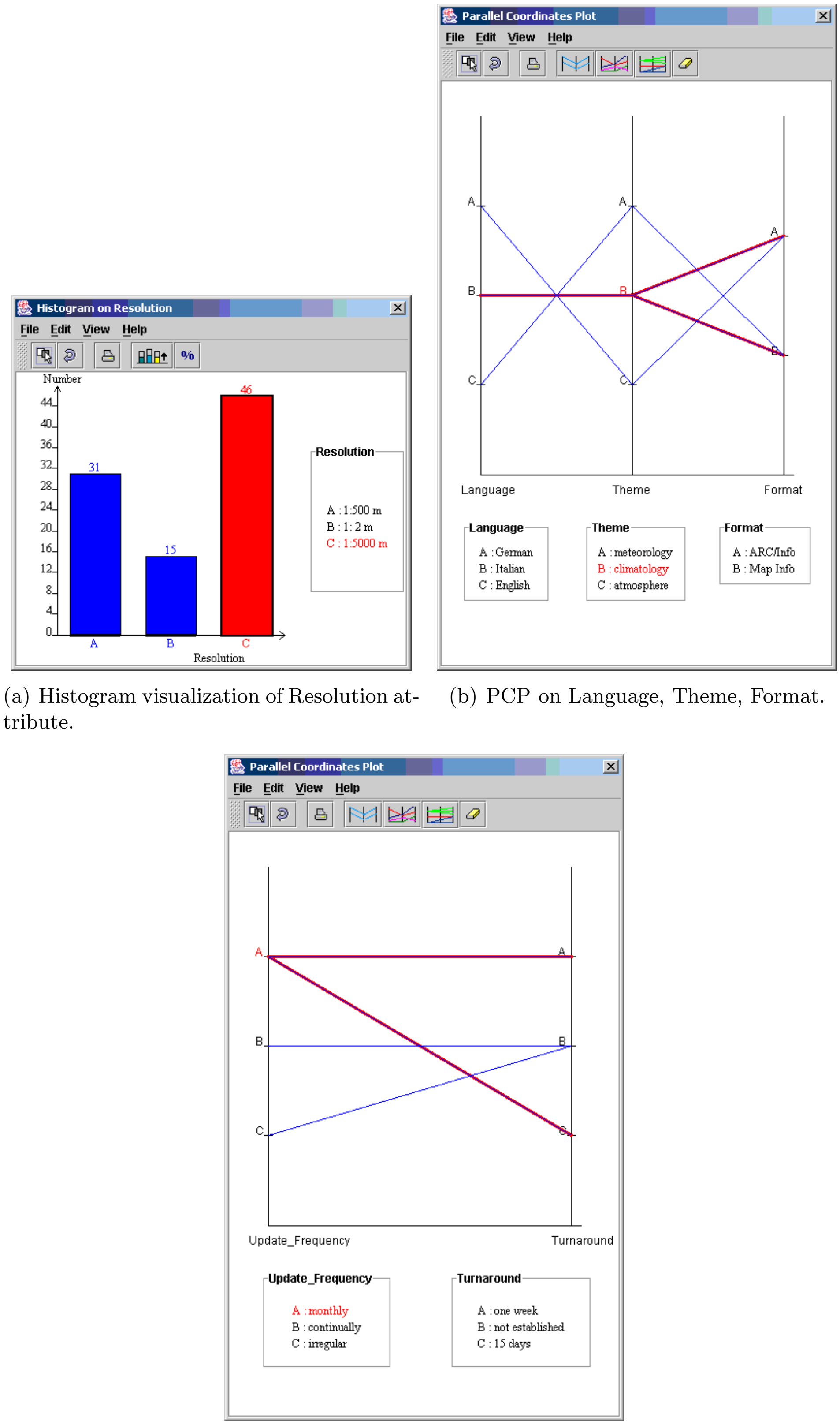
3.3.3.3 Data missing and unfamiliarity with attributes working scenario

Let suppose to have the following scenario: the user has to acquire geographic about climatology content, having 1:5000 resolution, in MapInfo format and written in English. To investigate which are the available data, the seeker can apply the following steps:

1. to display the Histogram related to the Resolution attribute (Figure 3.15(a)). He selects and reduces the dataset according with the available resolution measure he is more interested in.
2. to display a PCP on Language, Format and Theme to analyze the relations among these attributes.
3. to select the "climatology" value of Theme attribute Figure 3.15(b) in the PCP. It outlines that data dealing with this theme, in a MapInfo format and written in English are missing.
4. the seeker recalls that some information resources pertaining to "metereology" theme could be suitable for replacing those about "climatology" . Moreover this allows to select data according to seeker needs, since a line on the PCP connects English, "meteorology" and MapInfo.
5. to refine the query criteria. Since the amount of the results is still too huge, to successfully complete his task the user has to refine his criteria using attributes he is not familiar with. He analyzes other attributes as Update Frequency and Turnaround by using a PCP visualization (Figure 3.15(c)). Since the data that are "continually" updated have a not established turnaround (in other words it is not clear when they will be available), he might decide to select the data that are "monthly" updated and having "one week" as turnaround.

3.3.4 INVISIP evaluation

A human evaluation of the visual tool proposed within INVISIP has been worked out during the project. The methodology adopted and the results obtained are detailed in the project deliverables [Limb 02, Limb 03a, Limb 04]. In general, this evaluation demonstrates as the Information Visualization tools support the seeker.



(a) Histogram visualization of Resolution attribute.

(b) PCP on Language, Theme, Format.

(c) PCP on Update_Frequency and Turnaround.

Figure 3.15: *Solving data missing and unfamiliarity with attributes.*

The Visual Data Mining Tool (VDM) and the Visual Metadata Browser (VisMeB) softwares have been tested. Different test scenarios were developed:

- Visual Data Mining tool (VDM) has been tested to search for information resources on a database using categories;
- Visual Metadata Browser has been tested to search for information resources on a database using query terms;

The following questions have been posed to the users:

1. Is the Layout of the system clear?
2. Could you orientate yourself fast?
3. Was the use of color appropriate?
4. Was the navigation intuitive?
5. Was important information highlighted in some way?
6. Was the terminology understandable throughout the site?
7. Were text and graphics presented in a visually aesthetic manner?
8. Was the amount of visual information adequate?
9. Could you find searched objects quickly?
10. Were you able to navigate through the site without the feeling of getting lost?
11. Was there to much information on individual pages?
12. Was there to little information on individual pages?
13. Was the information grouped consistently?
14. Were the graphics clear and sharp?

Ten users have been participating to the evaluation. The first six questions have been tested for both VisMeB and VDM, the second group from the seventh to fourteenth has been applied only for VisMeB. The results of the first and the second group of questions are depicted respectively in figure 3.16 and 3.17. Overall the results are good. As the VDM was to most participants an unknown concept, the results seem to be even better. Question number two, "is the layout of the system clear?" has a very high rating. As this is a very relevant question, especially considering that the Visual Data Mining tools were mostly unknown to

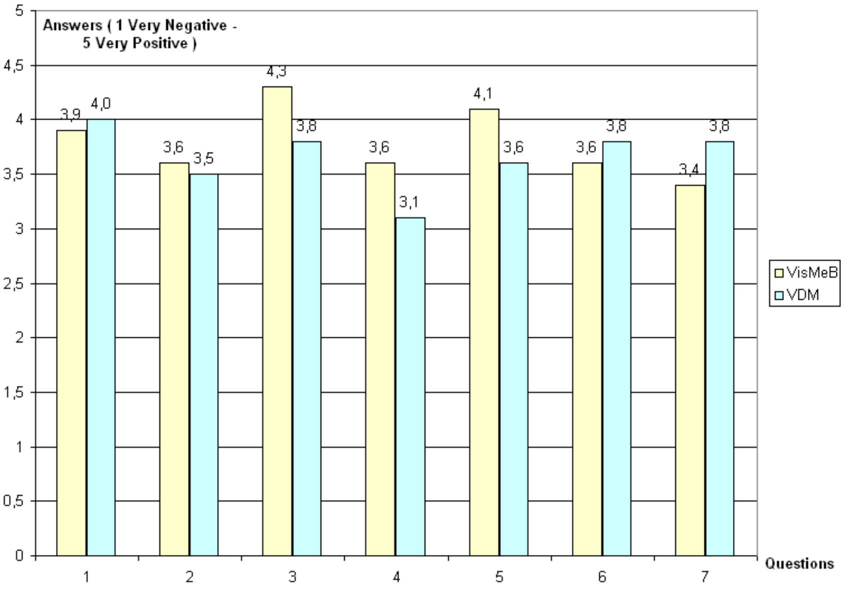


Figure 3.16: *Evaluation VisMeB and VDM: questions 1-7*

the participants. It exemplifies the overall good design quality of the VDM tool. Figure 3.18 shows the result of the post test in comparing the three visualizations: the Level Table and Granularity Table are part of the VisMeB. The users had to give a rating on a scale ranging from 1 to 10, where 1 means negative, and 10 a positive rating. The LevelTable was the highest rated visualization, followed closely by the GranularityTable. The distance between the two table visualizations and the VDM tool is significant. Nevertheless, a rating is only a subjective glimpse at a situation and should not be overweighted, especially if one looks at differences between rated objects. Regardless of that it is save to say, that all three visualizations were rated quite positive and users were satisfied working with them and that both the VisMeB and the VDM provide techniques to deeply compare complex metadata entries.

3.3.5 Results

An approach for metadata analysis is introduced to support users during the search for geographical information resources. It is based on well-known visualizations and powerful graphic interaction techniques. The direct interaction with the INVISIP users let us point out the problems of “unfamiliarity with attribute” and “data missing”. The approach facilitates the user in the comprehension of the results of a browsing search as well as to discover

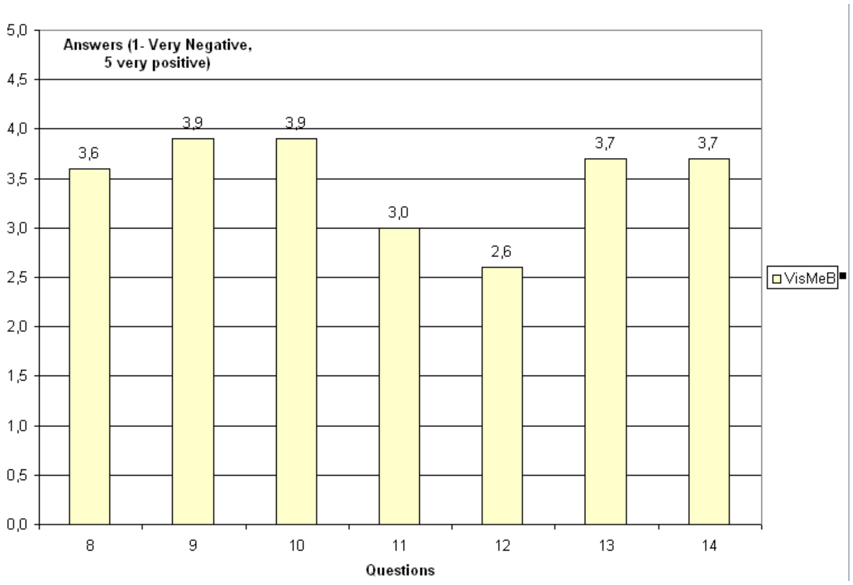


Figure 3.17: *Second Evaluation VisMeB: question 8-14*

relationship among data facing with the aforementioned problem.

The overall evaluation of tools developed in INVISIP demonstrated as a visual approach is positively perceived by the user at least with respect to the usual text based research. However, beside the user perception, an experimentation to assess how much the proposed visual approach make faster the selection of information resources has not yet addressed.

Related Publications

- R. Albertoni, A. Bertone, M. De Martino, "Visual Analysis of Geographical Metadata in a Spatial Data Infrastructure", Fifteenth International Workshop on Database and Expert System Application, University of Zaragoza, Zaragoza, IEEE Computer Society Press, pp. 861-865, 2004.

- R. Albertoni, A. Bertone, U. Demsar, M. De Martino, H. Hauska, "Knowledge Extraction by Visual Data Mining of Metadata in Site Planning", Proc. ScanGIS2003, Scandinavian Research Conference on Geographic Information Science, Espoo, Finland, pp. 119-130, 2003. (best paper /presentation)

- R. Albertoni, A. Bertone, U. Demsar, M. De Martino, H. Hauska, "Visual And Automatic Data mining for Exploration of Geographical Metadata", Proc. of 6th AGILE Conference on Geographic Information Science, Lyon, France, pp. 479-488, 2003.

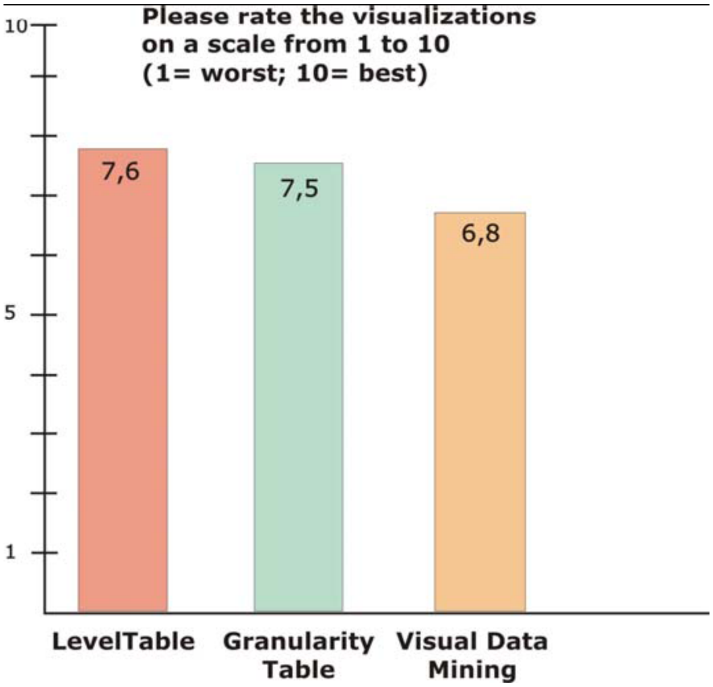


Figure 3.18: *Mean results comparison of the overall tools appreciation*

- R. Albertoni, A. Bertone, M. De Martino, "A Visualization-Based Approach to Explore Geographic Metadata", WSCG POSTERS proceedings WSCG'2003 - The 11-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic, 2003.

- R. Albertoni, A. Bertone, V. Contursi, M. De Martino, M. Franceschetti, F. Marangon, M. Piccazzo, V. Procopio, "INVISIP: Uno strumento integrato per l'accesso alle informazioni e il trattamento dei dati geografici", Sesta Conferenza Nazionale Asita, Perugia, Italy, pp. 95-100, 2002.

- R. Albertoni, A.Bertone, M. De Martino, "Information visualization and interactive geodata mining in site planning process", Primo Convegno Eurographics Italian Chapter, Milano, Italy 2002.

3.4 Contributions with respect to the requirements for metadata analysis

The experience performed within INVISIP shows that instruments based on the visual analysis are appreciated by the seekers. The evidences supporting this sentence is the positive feedback the users gave during the evaluation to questions such as “Is the Layout of the system clear?”, “Could you orientate yourself fast?”, “Was the navigation intuitive?”, “Was important information highlighted in some way?”. The visual tools are perceived by the seeker as headway at least with respect to the usual text based search engines. However, the visual tools discussed in this chapter and pertaining to INVISIP partially meet the requirements addressed in the section 2.3. The support they provide is summarized in Table 3.1. The examples shown in paragraph 3.3.3.1 and 3.3.3.2 illustrate the support to ease the *query formulation* and the *retrieved result comprehension*. They are limited to the categorical attributes. Anyway, similar support has been obtained by the other tools for the textual attributes [Klei 02] and attributes on spatial extension [Gobe 03]. All these tools have been integrated in order to provide a seamless analysis of geographical metadata.

Requirement	**Visual Metadata Analysis in INVISIP**
Query Formulation	good
Retrieved result comprehension	good
Query coordination	partial
Database problem	not applicable
Vocabulary problem	partial

Table 3.1: *Visual metadata analysis contribution to the requirements.*

The *database problem* is not faced within the research activity performed in INVISIP, actually in the project the information resources were collected in a unique repository, and as consequence there was no need to compare the result coming from different repositories, search engine or database.

The *query coordination* is partially supported through the Visual Data Mining tool which eases the discovery of query criteria. It allows to select the most suitable information resources. However, except from some simple functionalities provided by the other INVISIP tools Geocrystal [Gobe 03] and SuperTable [Klei 02], there is no support in comparing the results obtained by different queries.

The *vocabulary problem* is not directly addressed in INVISIP. Even if the Visual Data Mining tool supports in discovery pattern among categorical attributes which might ease in such a problem. For example, supposing that whenever a geographical information resource has the value of the metadata for theme equal to “climatology” it assumes also the value “environmental”, that is a pattern which can be pointed out through the VDM tool. These kind of pattern might be useful to realize that there are no environmental information resources out from climatology; thus climatology and environmental are “synonymous” in the repository the seeker is considering. This is a relation among terms which arise from the information spaces and belong to the implicit semantics of information resources. However, the repository seldom indicates all the synonyms in term of implicit patterns. More often these kinds of relations have to be obtained by explicit semantics. An example is provided in the scenario in section 3.3.3.3 where the seeker recalls that some information resources pertaining to "metereology" theme could be suitable for replacing those about "climatology". In this case the seeker is using a relation among terms which is induced by his background knowledge, which pertains to the explicit semantics.

3.5 Conclusions

The research activity applied to the geographical information resources results to be particularly suitable for the purpose of this thesis. It provides a real example where the hypotheses made in the first chapter are satisfied.

- The search and selection of geographical information sources has be done within an Open Environment because
  - the concurrency among the providers is determined by the economic interests pertaining to the selection of the geographical information resources,
  - the adoption of a centralized and strongly harmonized repository is unreasonable due to the vast number of stakeholders interested in producing and consuming geographical information resources.
- The search has to rely on metadata because
  - the information related to who, how and under which condition the geographical information is produced are indispensable for determining a correct use of the resources,
  - the geographical resources are often not available for free and they might require large band of network to be transmitted or large amount of disk space to be stored.
- The metadata analysis is needed because
  - the intrinsic complexity of geographical information in terms of formats, sources and domain specific features force to have a complex metadata,
  - potentially a large number of metadata entries have to be considered during the search.

Moreover, the crucial role which geographical information resources play in numerous business and government application demonstrates that the issues addressed by this thesis might have a great importance and social impact.

The experience made within INVISIP has clearly shown as the adoption of metadata standard is just the first step to solve the problems that the seeker has to face searching for geographical information resources. It demonstrates the worth of caring about the metadata analysis.

As demonstrated in INVISIP the visual analysis eases the user in the selection of geographical information resources: Information Visualization and Visual Data Mining techniques are perceived by the seeker as a progress in the search and selection activity. They support in the resolution of unfamiliarity with attribute and data missing.
With respect to the usual textual search engine it supports in the problems of:

- *query formulation* providing a more efficient and precise formulation of selection criteria;
- *retrieved result comprehension* providing a more compact representation of the result obtained;
- *query coordination/definition of selection criteria* providing new selection criteria on the base of patterns arising from implicit relation among information resources.

Although the satisfying evaluation achieved, the tool developed within INVISIP do not meet some of the requirements discussed in section 2.3
Independently from the techniques developed within INVISIP, the *query coordination* as well as *database* selection problems are brilliantly solved by Spoerri who introduced the MetaCrystal [Spoe 04d, SPOE 04b]. MetaCrystal is a visualization based on the Venn diagram which allows to compare the results of both different query as well as search engine. It can be adopted to achieve a complete support.
Beside the positive feedback given by users during the evaluation of tools developed in INVISIP, an assessment of how much these tools improve the task of search and select the information resources should be undertaken.

3.5.1 Remarks and Research Issues

The research experience on the visual analysis of categorical attributes has pointed out some interesting remarks:

- The application of Information Visualization techniques as well as Visual Data Mining to effectively browse unfamiliar information spaces rely on some kind of relationships among the space elements such similarity or a partial order. In the case all the metadata attributes are numerical, the information space is mapped in some metric space and these relationships are ensured by the metric space properties. Otherwise, it is not so easy to ensure that such relations are available, and this may hamper the complete exploitation of visual techniques.
- The background knowledge of the user affects the way the selection is performed. The vocabulary problem is a typical example: on the base of his background knowledge the seeker might modify the selection criteria i.e. replacing a term with an other which he relates to (see example in the scenario presented in section 3.3.3.3).

These remarks bring to interesting research issues:

- How to represent the background knowledge?
- How to take advantage from the background knowledge to suggest possible query refinement?
- Does background knowledge induce relations of partial order or similarity among the information space element?

These research issues are object of investigation in the chapter 4.

Chapter 4

Semantic Metadata Analysis

As discussed in chapter 3 the visual analysis of metadata facilitates the user during the search and selection of information resources. The techniques developed in the fields of Information Visualization and Visual Data Mining make friendlier the way the queries are posed, visualization techniques summarize the query results improving the results comprehension, and the combination of interaction and visualization techniques enables a deeper comparison of the resource features. The visual metadata analysis is fundamental to involve the user in the search activity. It amplifies the users' cognition facilitating the discovering of new selection criteria. However, there are problems which require some kind of representation of the seekers'/providers' background knowledge. The vocabulary problem is one of the typical cases where the lack of background knowledge hampers the effectiveness of the visual analysis: resources related in the reality but described with different terms cannot be properly visualized because the relations among terms are unknown to the visual metadata analysis. The background knowledge strongly affects the selection process and non taking it into account might inhibit a complete exploitation of visual metadata analysis or even worse it might result in misleading visualizations.
This chapter presents different methods which take advantage from the representation of the background knowledge to ease the comparison and exploration of information resources. The background knowledge is represented by ontologies. The methods are classified as "semantic" because they support in determining how information resources semantics affects the selection process. Three different approaches to support the semantic metadata analysis are proposed:

- Asymmetric Semantic Similarity among metadata Categorical Values (ASSCV) to face with the vocabulary problem;
- Asymmetric and Context Dependent Semantic Similarity (ACDSS) to compare information resources;
- Semantic Granularity (SG) to browse information resources with respect to categorical

features and at different levels of abstraction.

All the methods have been developed considering problems and applications arising from the European project INVISIP [INV] and the network of excellence AIM@SHAPE [AIM]. They both face with information resources characterized by metadata: INVISIP focuses on geographical dataset characterized by the Geographical Metadata Standard ISO 19115 [ISO19115 03], whereas, AIM@SHAPE focuses on digital media resources i.e. resources whose essential characteristic is to have a shape, and which are additionally characterized by ontology driven metadata. In particular, in AIM@SHAPE the definition of a framework of ontology driven metadata, extending the result obtained by MPEG specification [mpeg7 04], is part of the expected results. I have been actively participating in the definition of these ontology driven metadata [Albe 05d, Papa 05, Albe 07, Albe 06c]. An example of the ontology driven metadata developed in AIM@SHAPE and the issues pertaining its design are illustrated in Appendix A.
The motivations behind the adoption of ontologies and what can be represented by them are described in section 4.1. The semantic similarity among categorical values, which is presented in section 4.2, has been conceptualized as an extension of the VDM tool developed within INVISIP. As a consequence, it is exemplified in the context of geographical metadata. The asymmetric and context dependent similarity as well as the semantic granularity stem from the experience of AIM@SHAPE. Since the final definition of the AIM@SHAPE's ontologies has not yet released they are demonstrated referring to the domain of academic research respectively in the sections 4.3 and 4.4.

4.1 Ontologies and background knowledge

The following two subsections present respectively the reasons why ontologies are adopted to represent the background knowledge and the role that ontologies can play during the matadata analysis.

4.1.1 Why ontologies to represent the background knowledge?

This section presents the reasons behind the adoption of the ontology to represent the background knowledge.

Ontology are suitable to formally represent shared conceptualization. As defined by Gruber [Grub 95] and as already discussed in the first chapter, "an ontology is an explicit and formal specification of a conceptualization". By the early 1980s, researchers in AI and especially in Knowledge Representation had realized that to work in Ontology was relevant to the process of describing the world for intelligent systems to reason about and act in. This awareness and integration grew, and spread into other areas until, in the latter half of the final decade of the 20th century, the term "ontology" actually became a buzzword, as enterprise modeling, e-commerce, emerging XML meta-data standards, and knowledge management, among others, reached the top of many businesses strategic plans. In addition, an emphasis on "knowledge sharing" and interchange has placed an emphasis on ontology as an application area itself [Welt 01]. Such a popularity and the vast employment the ontologies have for sharing knowledge are the first factors suggesting their adoption to represent the background knowledge.

Ontologies play an important role in the Semantic Web. Ontologies will play an important role in the Semantic Web which is an extension of the well known World Wide Web. In the semantic web, ontologies provide a shared understanding of a domain. Such a shared understanding is necessary to overcome differences in terminology [Anto 04]. The number of initiatives and the efforts spent by World Wide Web consortium (W3C)[1] defining a web ontology language [OWL 06] is the most prominent evidence of the importance ontologies are obtaining within the WWW. The development of the Semantic Web has a lot of industry momentum, and governments are investing heavily: the US government has established the DARPA Agent Markup Language (DAML) Project, and the Semantic Web was among the key action lines of the European Union 6th Framework Programme [Anto 04]. Even if the Semantic Web is still far from replacing the current web, the importance of ontologies at different level of formalization in the next development of the web is almost universally accepted. The current web is the most remarkable example of infrastructure where is possible to exchange information resources under the hypotheses this thesis made defining the open environment. The fact that the ontologies are promising in the next web development, and the importance that web has in obtaining the open architecture, are the second factor influencing the choice of ontologies as background knowledge representation.

Metadata about complex information resources are expressed by ontologies. As already discussed in the first chapter, a careful and successful selection of information resources requires to consider a complex set of resources' features including entities and actors contributing in the resources creation. Considering the richness and the complexity pertaining to information, it is reasonable to anticipate that their metadata will be more and more often encoded by ontologies. As a consequence, entities and relations which today are part of the background knowledge will be encoded in the ontology schema and the choice of an ontology as background representation will be quite natural.

[1] http://www.w3.org/

4.1.2 What can be represented through an ontology?

In the section 4.1.1 I have already discussed the choice of ontologies to represent the background knowledge. Here I introduce the way the ontologies and metadata can be assembled. In figure 4.1 a schema of how the real world entities, the resource features and the ontologies, take part into metadata definition is depicted. Starting from box (a) on the left side of figure 4.1, different kinds of real entities are represented. All these real entities or at least the digitalized counterparts of them can be considered as information resources which someone could need to select. In the figure 4.1 (b) some example of features which can be used to select these resources are given. Of course representing the features might require to handle also the relations kept by entities and features. Moving to the figure 4.1 (c), the features and relations are organized in metadata. Different representations are possible according to the kind of metadata chosen. In case of *plain metadata* figure 4.1 (c.1) the features might be represented in terms of XML attributes and only simple relations among attributes can be inserted by nesting the attributes. Alternatively, the metadata are represented through an ontology figure 4.1 (c.2): the features will be grouped into meaningful entities, attributes and relations. Both *plain* and *ontology-driven* metadata might have attributes in order to represent the properties of an information resource. The attribute values can range in different domains. They can be numerical, textual, categorical or even represent spatial and time extensions. If the attributes are numerical, textual, spatial or temporal there are explicit encoding of relations among the attribute values. Computers handle numbers quite well and techniques to manage spatial and time reference have been developed. For example, in [Guti 00] an algebraic framework for modelling moving points and regions has been formalized, by proposing abstract data types that can be integrated in relational and object relational models. Chomicky and Revesz in [Chom 01] discuss the closure properties of a set of geometric spatio-temporal objects (rectangle and convex polygons) with respect to base algebra operators (e.g. closure with respect to intersection) have been discussed. However, Computers are not so successfully handling with categorical attributes, i.e. attributes whose domain is made of a set of prefixed terms. As illustrated in figure 4.1 (d), ontologies conceptualizing the domain where the categorical attributes range can be adopted. They can represent part of the users' background knowledge in order to make explicit the relations among categorical values. According to the above discussion it is possible to distinguish two main roles ontologies can play:

Role of domain conceptualization to formalize relationships among categorical values;

Role of metadata schema to organize the metadata describing resources in specific domains.

I would remark that the semantic metadata analysis proposed in this thesis goes extensively through these two roles. In fact, the experience within INVISIP and AIM@SHAPE have forced us to conceive our methods according to different metadata set-ups. The ISO standard 19115 geographical metadata is encoded within INVISIP as plain metadata, and the semantic

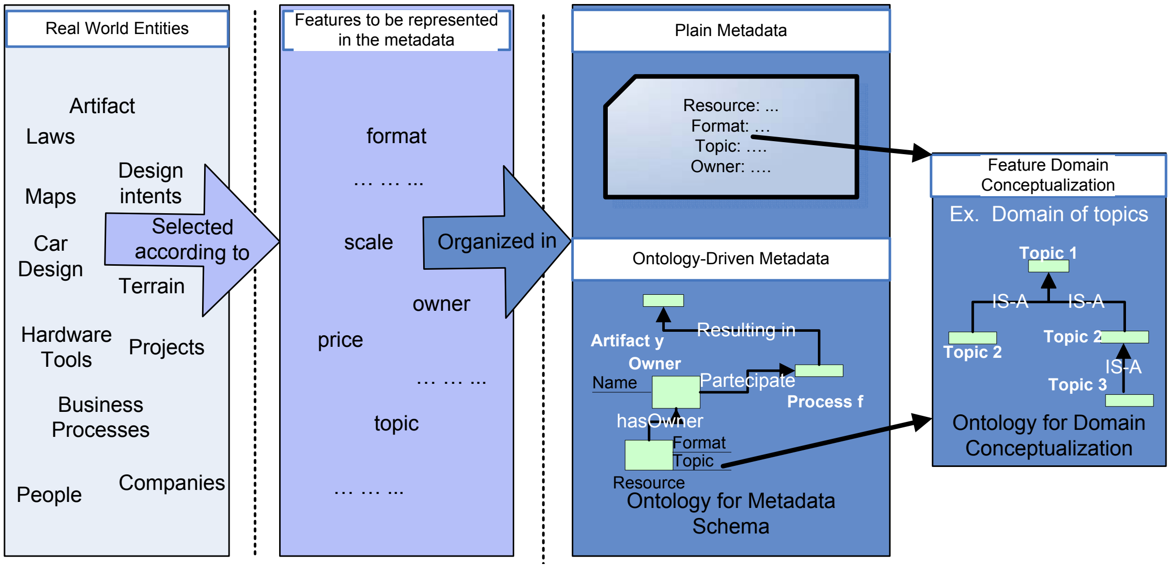


Figure 4.1: *How metadata can be assembled. (a) represents examples of real world entities; (b) represents example of features; (c) shows two way to specify a metadata schema; (d) shows an example of ontology to represent the conceptualization for the feature "topic"*

similarity among categorical values has been designed mainly taking advantage from ontologies to represent attribute domain conceptualization (see section 4.2). On the contrary, within the AIM@SHAPE network of excellence, the metadata are encoded as ontology driven metadata. Thus, the asymmetric and context dependent similarity among information resources relies on the ontology adopted to represent the metadata schema (see section 4.3). Moreover, the semantic granularity relies on an ontology which mixes the two roles up, representing both the metadata schema and the hierarchical organization of the qualities used to define the granularity (see section 4.4).

4.2 Asymmetric Semantic Similarity among metadata Categorical Values (ASSCV)

Asymmetric Semantic Similarity among Categorical Values is the first metadata analysis technique designed within my research activity. The aim is to support the seeker when the criteria used to formulate the query fail: the system is able to provide a background knowledge pertaining to similar terms that could be adopted in the reformulation of the query. The semantic relations are detected by applying similarity criteria among data and it is provided to the user through visualization techniques to amplify his cognition and to facilitate the interpretation of the query results. The rest of the section is organized as follows: section 4.2.1 provides the reason why this approach has been developed, section 4.2.2 introduces how the ontologies and an adaptation of the Matching-Distance Measure for Semantic Similarity (MDMS)[Rodr 04] are adopted to work out the semantic similarity among ontology classes, section 4.2.3 introduces the approach we propose; and section 4.2.4 shows how the approach can deal with the vocabulary problem.

4.2.1 Motivations

Semantic similarity among categorical values stems from the experience made within the EU project INVISIP. It has provided a visual metadata analysis framework which allows the seeker to move through large information spaces. An exploration approach is defined: it is characterized by a reasoning activity based on the integration of visualization techniques, graphic interaction and brushing and linking functionality. The approach is widely illustrated in section 3.3 and an example of how it can be used to solve some problems of geographical information search is described in section 3.3.3. However, the approach has some limitations: it does not take into account of any user background knowledge, for example it ignores the "semantic" relations among the categorical values which represent precious hints for the formulation of new search criteria.

Categorical values or nominal values are data that can be separated into different categories according to some non-numeric characteristics. Their exploration is challenging because the values they assume provide information that can be easily understood by a human agent but cannot be trivially managed in an automatic way. Some approaches to visualize categorical values have been proposed [Kola 01],[Rosa 03]. They visualize implicit relations arising from the data. For example, assuming that there is a pattern of dependency between two categorical values x, y (e.g. "if a resources has nominal value 'x' then it has also value 'y' "), these techniques visualize the resources taking advantage from such an implicit relation. However none of them is based on explicit semantics: they do not consider the relations existing in the seekers'/providers' cognitive space, so they are not able to follow faithfully the user's interpretation. The approach based on semantic similarity among categorical values is proposed in this thesis to overcome this limitation.

4.2.2 Preliminary assumptions

In this paragraph I introduce some concepts concerning the representation of semantic relations among data needed to design our approach: the hypothesis pertaining the ontology expressiveness and the adopted similarity are provided.

It is assumed to play a role of domain conceptualization (see section 4.1.2). The ontology is composed by class entities (named classes) representing the most important concepts of the domain, instances representing specific elements of classes, and slots which can be attributes characterizing the classes or relations representing types of interaction among concepts. The classes can be related by is-a or part-of relations.

Semantic similarity facilitates the comparison among the class entities and allows to handle those which are semantically similar. In semantic similarity among categorical attributes the Matching-Distance Measure for Semantic Similarity (MDMS) [Rodr 04] is considered and it is defined in terms of slots comparison. Slots are classified according to three different types of features called distinguishing features: *function* features which are relations with specific properties describing what is done to or with a class, *part* features (a part-of relation) describing structural elements of a class and *attribute* features, which represent class properties. Two entities are more or less similar according to the number of slots belonging to the same kind of distinguishing features they share each other.
A formal definition of "global similarity" is based on the definition of two measures: the "slots importance" and the "slots similarity".

Definition 1 (function α of slots importance) *Let us call c_1, c_2, two class entities, d the distance function between the two class entities and lub the immediate super-class that subsumes both classes. α is the function that evaluates the importance of the difference between the two class entities in term of their slots and it is defined by:*

$$\alpha(c_1, c_2) = \begin{cases} \frac{d(c_1, lub(c_1, c_2))}{d(c_1, c_2)} & d(c_1, lub(c_1, c_2)) \leq d(c_1, c_2) \\ 1 - \frac{d(c_1, lub(c_1, c_2))}{d(c_1, c_2)} & d(c_1, lub(c_1, c_2)) > d(c_1, c_2) \end{cases} \tag{4.1}$$

where $d(c_1, c_2) = d(c_1, lub(c_1, c_2)) + d(c_2, lub(c_1, c_2))$.

It is important to note that the computation of the distance d considers both is-a and part-of relations to determine the immediate super-class lub.

Definition 2 (slots similarity) *Given two class entities c_1 (target) and c_2, (base), t one type of distinguishing features (part, function, attribute) and C_1 and C_2 the sets of features of type t respectively of the class entities c_1 and c_2. The similarity value of c_1, c_2 is:*

$$S(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + (1 - \alpha(c_1, c_2))||C_1/C_2| + \alpha(c_1, c_2))||C_2/C_1|} \tag{4.2}$$

Definition 3 (global similarity) *Given two class entities c_1 and c_2, w_p, w_f, w_a the weights of the respective importance of parts, functions and attributes, the global similarity function S between two class entities c1 and c2 is the weighted sum of the similarity values of parts (S_p), functions (S_f) and attributes (S_a):*

$$S(c_1, c_2) = \omega_p \cdot S_p(c_1, c_2) + \omega_f \cdot S_f(c_1, c_2) + \omega_a \cdot S_a(c_1, c_2) \tag{4.3}$$

The sum of weights is expected to be equal to 1 and the value of each of them is calculated according to contextual information [Rodr 04].

4.2.3 The approach to analyze categorical attributes

An approach developed in this thesis to analyze categorical attributes of metadata exploiting the semantic relations among the categorical values is described in [Albe 05b]. The ontology is playing a role of domain conceptualization to express the relations and to make them machine understandable, the MDMS similarity [Rodr 04] is adopted to work out similarities among the concepts. Moreover, Information Visualization is applied to improve similarity cognition of the seeker.
The analysis is performed in three phases:

1. definition of the ontology,
2. exploitation of the ontology to explicit the relations of the categorical values,
3. exploration of semantic relations among categorical values.

The ontology definition is obtained by mapping the categorical values of a metadata attribute into class entities. A class hierarchy is built adding is-a relations. Each metadata entry whose values are mapped into a class entity, is defined as instance of the class. Moreover, as the similarity among classes is defined in terms of slots comparison, a careful definition of slots for each class is needed. Slots grouped according to their functions, parts and attributes should be associated to each class entity also taking into account the intended similarity assessment: in other words, if two classes are expected to be similar they should be forced to share some slots.
The exploitation of the ontology to work out the semantic similarity is based on the MDMS similarity: it adopts the asymmetric similarity, which seems to be more appropriate to support a query refinement based on the distance among concepts [Rodr 04]. The MDMS allows to make explicit the similarity among categorical values, which is a semantic relationship, usually not available from the ontology design. In the exploration of semantic relations the main issue is to choose appropriate visualization techniques to display the similarity measure providing useful hints in the query refinement. Visualizations such as the cluster maps [Stuc 04] are developed to support the query refinement but they still do not visualize any information about class similarity. Other ontology visualizations are mainly based on a graph representation where a node can be either an instance or a class entity and edges can be either properties or relations. The graph visualization helps the seeker to analyse and better understand the domain described in the ontology, but it does not provide any support in the query refinement process. For example Protégé [Muse 01] offers different visualizations [Erns 05]: figure 4.2 shows the interactive and graph-based visualization to browse an ontology provided by protg plug-in TgVizTab [TGVizTab 04]. This visualization is inadequate to provide an explicit representation of the semantic relations: it shows the structure of the ontology where each node is a theme and each line is an “is-a” or “part-of” relation. It does not provide any interpretation of classes in terms of similarity. It supports engineer during the ontology

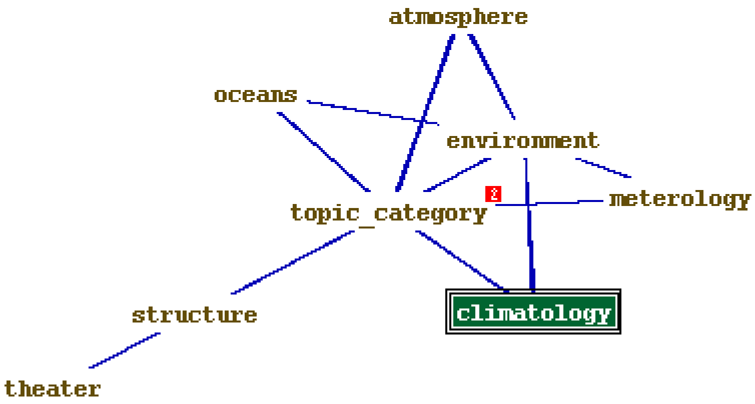


Figure 4.2: *Ontology visualization with Protégé plug-in "TgVizTab".*

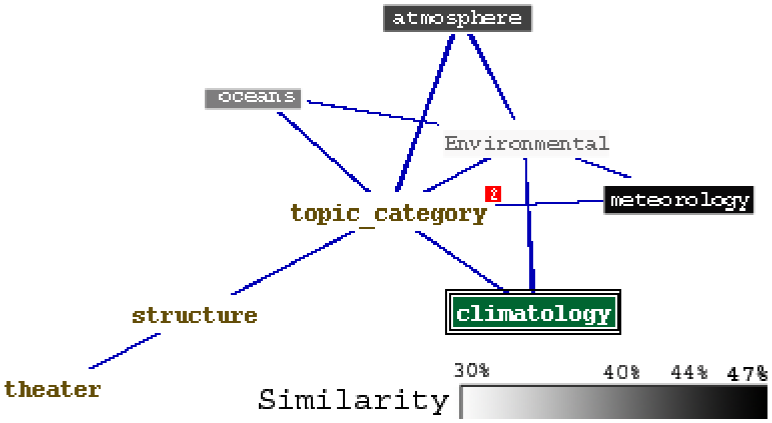


Figure 4.3: *Ontology and Similarity Visualization*

design rather than the seeker in the query refinement.

In this thesis I propose a visualization aimed also to facilitate the seeker in recognizing similarities among the categorical values. It is characterized by the integration of a graph visualization to show the overall structure of the ontology and graphic techniques to represent similarity information. Figure 4.3 depicts the visualization of an ontology enhanced using similarity information. The target entity (first variable in the similarity function) is the class entity "climatology", the similarities are worked out with respect to it and it is marked by a double-squared rectangle. The other entities considered as a base in the similarity (second variable in the similarity function) and whose similarity measures are different from zero are represented by rectangles with different grey levels. The grey level is brought down proportionally to the measure of similarity between target and base: the more similar are climatology and the class entity, the darker is the box surrounding the class in the visualization.

4.2.4 A practical example

This paragraph aims to clarify the semantic based analysis through a practical example. The example pertains to the search for geographical information resources and it is applied at the categorical attributes of the ISO 19115 metadata standard [ISO19115 03]. In particular, the attribute "topicCategory", which describes a high-level classification for geographic data themes, is analyzed.

The first phase of the approach is to develop the ontology of the data themes with the identification of the distinguishing feature for each class entity. In this example I presents a simple ontology (Figure 4.4). The definition of a complete ontology of theme would require a long interactive design process where experts of the domain have to be directly involved. It requires efforts out of the purpose of this example.

Figure 4.4 shows a subset of the categorical values defined in the metadata specification as possible themes and their organization in ontology in terms of is-a and part-of relations. Table 4.1 shows the distinguishing features of each data theme. Note that the entity "topic category" is an abstract class that cannot be instantiated and does not have parts, functions and attributes. It is represented in the ontology mainly for technical reasons, it is an explicit reference to the metadata attribute, which is considered, and it can be useful to contextualize the ontology in the overall metadata schema.

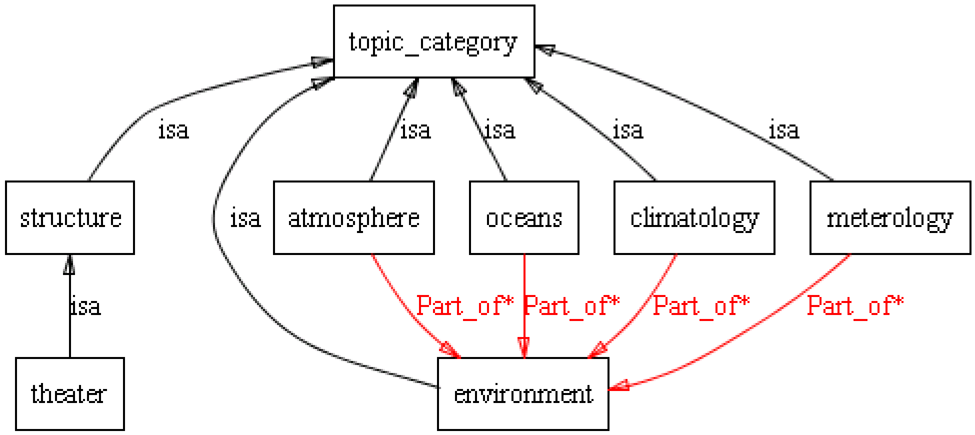


Figure 4.4: *The "topic_category" ontology example.*

The second phase of the approach concerns the ontology exploitation to explicate the semantic relations among the categorical values. Let us suppose to analyze the semantic relations among the theme "climatology" and the other themes. Table 4.2 shows the similarity measure between "climatology" and the other entities belonging to the ontology applying the MDMS. The similarity values are calculated considering the ontology graph in Figure 4.4 and the

Class entity	**Parts**	**Functions**	**Attributes**
Topic Category			
Environment	Climatology Atmosphere Meteorology Oceans	Environmental assessment Climate phenomena analysis Monitoring environmental risk	Ex_GeographicExtent Ex_TemporalExtent Variable
Atmosphere		Air quality analysis Climate phenomena analysis	Ex_GeographicExtent Ex_TemporalExtent Atmosphere layers Temperature
Meteorology		Weather forecast Climate phenomena analysis	Ex_GeographicExtent Ex_TemporalExtent Precipitation
Climatology		Climate phenomena analysis	Ex_GeographicExtent Ex_TemporalExtent Temperature Precipitation Wind
Oceans	Sea life	Tidal wave forecast Tide analysis Climate phenomena analysis	Ex_GeographicExtent Ex_TemporalExtent Temperature Wind Water composition Sea level
Structure			Material Location
Theatre	Foundation Roof Ticket office	Perform Present Recreate	Material Location Height

Table 4.1: *Distinguishing features of the ontology entities.*

distinguishing features in Table 4.1. To provide a simple example the global similarity function S(a,b) has been calculated considering all the weights w_t equal to one third. The result shows that the topic “climatology” is more similar to “environmental” than to “structure” or “theatre”. The same happens for “meteorology”, “atmosphere” and “oceans”. Furthermore, Table 4.2 also quantifies the similarity between the themes: “climatology” is more similar to “meteorology” than “atmosphere”, “atmosphere” than “oceans”, “oceans” than the generic environment.

The third phase of the approach concerns the presentation of similarity information to the seeker. In this example the simple visualization illustrated in the previous paragraph is applied. The TGVizTab visualization depicted in Figure 4.2 can provide a compact overview about the domain, which is useful to get the contextual information. On the contrary, the visualization proposed in Figure 4.3 provides also similarity information through grey level colours of the box surrounding the entities. For example “meteorology” has a surrounding black box since “climatology” is more similar to the “meteorology” than “atmosphere”, “oceans” and “environmental” whereas the color of surrounding box of “atmosphere”, “oceans” and “environmental” discolors according to decreasing of the similarity decrease. Moreover, entity classes like “structure” and “theater” do not have any surrounding box since the similarity is equal to zero. Even if it can appear a trivial representation, this kind of visualization provides a useful support in the query refinement. In other terms, the similarity exploitation in the metadata analysis makes machine understandable the fact that when the user is searching for data having “climatology” as theme and he gets unsatisfying results, the system suggests him to refine his query. It can be adopted to ease the *vocabulary problem* presented in section 2.3. Let us suppose to have the following scenario: the seeker has to acquire geographic data dealing with “climatology” content. He formulates his query, but it fails since he adopts different terms from those exploited by data providers to indicate the same concepts. Hence he is forced to identify the same or similar terms used by the data providers for an effective search. The semantic based reasoning proposed supports the seeker in this problem. The ontology visualization allows to identify the relationships among the themes. For instance the selection of “climatology” by a simple click on the ontology graph, allows to understand that “meteorology” is more similar to “climatology” than to “atmosphere”. The gray level color of the box surrounding the entities clearly shows this situation in Figure 4.3. So in or-

b	α	S_p(a,b)	S_f(a,b)	S_a(a,b)	S(a,b)
Environment	0,00	0,00	0,33	0,67	0,33
Meteorology	0,50	0,00	0,67	0,75	0,47
Atmosphere	0,50	0,00	0,67	0,66	0,44
Oceans	0,50	0,00	0,50	0,72	0,40
Theater	0,33	0,00	0,00	0,00	0,00
Structure	0,50	0,00	0,00	0,00	0,00

Table 4.2: *Similarity measures between the theme “a”, “Climatology”, and the theme “b”.*

der to overcome the vocabulary problem "meteorology" is the first attempt. Moreover, entity classes like "structure" and "theater" do not have any surrounding box since the similarity to these classes is equal to zero.

4.2.5 Results

The semantic similarity among categorical values is the first methods of semantic metadata analysis proposed in this thesis. It proposes an approach to combine a new ontology visualization, a fragment of the seeker background knowledge expressed as ontology and the Matching -Distance Measure for Semantic Similarity. The contribution is twofold, on the first hand the method eases the seeker in solving the *vocabulary problem*. On the other hand, the similarity and its visualization extends the visual metadata analysis framework described in chapter 3. The encoding of background knowledge in an ontology might result a costly task. As a consequence future extensions could integrate the approach with WordNet[2]. WordNet provides a lexical database where some common sense relation between terms are already considered. Then the ontology definition could encode only those semantic relations specific to a domain application (i.e. geographical domain) and the relations induced by the common sense could be inherited from WordNet.

Related Publications

- Albertoni R., Bertone A., De Martino M. Visualization and Semantic Analysis of Geographic Metadata. In: 2005 Workshop on Geographic Information Retrieval (GIR '05) (Bremen, Germany, 4 November 2005). Proceedings, pp. 9-16. C. Jones, R. Purves (eds.). ACM Press, 2005.

- Albertoni R., Bertone A., De Martino M. Semantic Analysis of Categorical Metadata to Search for Geographic Information. In: 16th International Workshop on Database and Expert Systems Applications (DEXA'05) (Copenhagen, Denmark, 22-26 August 2005). Proceedings, pp. 453-457. IEEE Computer Society, 2005.

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

4.3 Asymmetric and Context Dependent Semantic Similarity among information resources (ACDSS)

The asymmetric and context dependent similarity among information resources is the second semantic method to analyze the metadata I propose in this thesis.

While the previous method relies on ontologies to conceptualize the domain of categorical attributes, here, the ontology is considered as a metadata schema (see section 4.1.2). All the features to be included as metadata are encoded according to the ontology schema and the metadata of each single resource is encoded as an ontology instance.

The similarity plays an important role in information systems as it supports the identification of objects that are conceptually closed but not identical. Similarity assessment is particularly significant in different areas of the Knowledge Management (such as data retrieval, information integration and data mining) because it facilitates the comparison of the information resources in different types of domain knowledge [Schw 05b, Wang 02]. Being the resources' metadata represented as ontology instances, the definition of a method for assessing the semantic similarity among instances becomes essential for the purpose of this thesis.

The assessment of similarity is affected by the human way of perceiving the similarity as well as by the application domain. Its evaluation cannot ignore some cognitive properties related to the way human beings perceive the similarity. Three main aspects have to be highlighted. Firstly, considering that in the naive view of the word, similarities defined in terms of a conceptual distance are frequently asymmetric, the formulation of similarity should for many applications provides an asymmetric evaluation. Secondly, it should be flexible and adaptable to different application contexts, which affect the similarity criteria. Thirdly, the similarity evaluation should be able to exploit as much as possible all the hints that have already been expressed in the ontology. That is because ontologies represents part of a domain knowledge as it is perceived by the experts and their definitions require time consuming and costly processes.So far, most of the research activity pertaining to similarity and ontologies has been carried out within the field of ontology alignment or in order to assess the similarity among concepts. Unfortunately, these methods produce results that are inappropriate for the similarity among instances. On the one hand, similarities for ontology alignment strongly focus on the comparison of the structural parts of distinct ontologies and their application to assessing the similarity among instances might give misleading results. On the other hand, similarities among concepts mainly deal with the lexicographic database ignoring the comparison of the values of the instances. Few methods for assessing similarities among instances have been proposed. Unfortunately, these methods rarely take into account the different hints hidden in the ontology and they do not consider that the ontology entities concur differently in the similarity assessment according to the application.

To overcome the limitations mentioned above obtaining a similarity suitable for the exploration of information resources a brand new and asymmetric semantic similarity among instances is proposed in this thesis.

The similarity is asymmetric to stress the principle of "containment" between information resources (whether a resources provides more/less characteristic than the others). Moreover, the proposed similarity evaluates the “overlapping” between resources: it is greater the less the resources differ and the more the resources share characteristics.

The section is organized as follows. In the section 4.3.1, I illustrate the motivations and the scenarios that have driven me to the similarity definition. Then, after providing some useful assumptions (section 4.3.2), I discuss the main principle of the approach (section 4.3.3). A formalization of the similarity criteria induced by the context is proposed (section 4.3.4). The section 4.3.5 is devoted to the definition of the similarity functions that characterize our approach. Notwithstanding the method is general and can be applied to the comparison of any kinds of information resources, it is discussed considering the description pertaining to researchers as information resources. In particular, the method is demonstrated in section 4.3.6 by two experiments and an evaluation of the results analyzing the researchers which belonging to the section of Genova of the Institute of Applied Mathematics and Information Technology which belong to the Italian National Council of Research. At the end, I evaluate related works underlining how they have been useful as a starting point for our research but how, contrary to the proposed framework, they do not fulfill the requirements and goals I address by this contribution.

4.3.1 Motivations

The need of similarity among information resources is fully motivated in this thesis by the aim of supporting search for information resources. I am providing here further motivations, underlying the need of a similarity evaluation among ontology instances that takes into account the hints hidden in the ontology as well as the dependence on the context. In particular, taking advantage from the experience made within in the European founded Network of Excellence AIM@SHAPE [AIM], my goal is to answer to the following questions:

- Why to define a semantic similarity among ontology instances?
- What is the role of the implicit knowledge expressed by the ontology setting up a similarity assessment?
- What is the role of the application context in the similarity evaluation?

Why define a semantic similarity among ontology instances? Defining a semantic similarity among ontology instances represents a challenging priority in future research as it will pave the way for the next wave of knowledge intensive methods that will facilitate the intelligent browsing as well as information analysis. Here I do not refer to the similarity as a tool for identifying possible mapping or alignment among different ontologies. Rather I address a different problem related to the comparison of the ontology instances. I realize the importance of solving this problem from our direct research experience working in the AIM@SHAPE [AIM]. Within this network of excellence, ontologies have been adopted to organize the metadata of complex information resources. Different ontologies are integrated to describe 3D / 2D models (i.e. models of mechanical objects, digital terrains or artefacts from cultural heritage) as well as the tools for processing the models [Falc 04]. I have been actively participating in the definition of these ontologies [Albe 05d, Papa 05, Albe 07, Albe 06c]. From this experience, I have realized that the ontology driven metadata definition turns out to be outrageously expansive in terms of man-month efforts needed, especially whenever the domain that is expected to be formalized is complex and compound. The "standard ontology technology" provides reasoning facilities that are very useful in supporting querying activity as well as in checking ontology consistency, but the current technology lacks an effective tool for comparing the resources (instances). Although, the efforts necessary to formalize the ontology, domain experts are often quite willing to provide the domain knowledge required to characterize their resources. However, they are disappointed when their efforts do not result in any measure of similarity among the resources. Aware of this shortcoming, I address our research efforts toward investigating how to better employ the information encoded in the ontology and to provide tools that exploit as much as possible the result of the aforementioned efforts [Albe 06b, Albe 06a].

What is the role of the knowledge expressed by the ontology in setting up a similarity assessment? An ontology reflects the understanding of a domain, which a community has agreed upon. Gruber defines an ontology as "the specification of conceptualizations, used to help programs and humans share knowledge" [Grub 95]. There is a strong dependence between the knowledge provided by the domain expert in order to define the ontology and his expectation of the results of the semantic similarity. Actually, the domain expert will perceive a similarity that is based on the knowledge he has provided. The main ontology components (concepts, relations, instances) as well as its structure are representative of the domain knowledge conceptualized in the ontology. Therefore, they provide the base on which to set up the different hints to define the similarity. Classes provide knowledge about the set of entities within the domain. Properties, namely relations and attributes, provide information about the interactions between classes as well as further knowledge about the characteristics of concepts. Moreover the class structure within the ontology is also relevant as the attributes and relations shared by the classes as well as their depth in the ontology graph are representative of the level of similarity among their instances. In our proposal, the similarity assessment takes advantage of all of these ontology entities, which are usually available in the most popular ontology languages. Other entities could be considered as long as more specific ontology languages are adopted.

What is the role of the application context in the similarity evaluation? The definition of a similarity explicitly parameterised according to the context is essential because the similarity criteria depend on the application context. Two instances may be more closely related to each other in one context then in another since humans compare the instances according to their characteristics but the characteristics adopted vary with the context. In particular, as consequence of the explicit parameterisation of the similarity with respect to the application context it is possible:

- to use the same ontology for different application contexts. The ontology design usually ignores the need to tailor the semantic similarity according to specific application contexts. In this case, to assess the similarity between two different applications, two distinct ontologies need to be defined instead of simply defining two contexts.
- to provide a tool for context tuning that supports the decision-making process of the ontology user. The user often has not clearly defined in his mind the set of characteristics relevant for the comparison of the instances or his specification does not match the result induced by the information system. A parameterisation of the semantic similarity measurement supports a refinement process of the similarity criteria. The parameterisation provides a flexible and adaptable way to refine the assessment toward the expected results, and therefore it reduces the gap between user-expected and system results.

4.3.1.1 Framework scenarios

I have identified two main scenarios where the proposed similarity framework is relevant: the scenario 1 refers to a similarity evaluation in different application contexts exploiting the same ontology, the scenario 2 refers to the iterative criteria refinement process used to properly assess the similarity according to the expectations of the domain expert. In both scenarios I assume that I have an ontology describing the metadata of the resources in a complex domain and that the different resources are already annotated according to this ontology driven metadata. Two actors play important roles in the two scenarios:

- The user, who is the domain expert and who is looking for the semantic similarity. He has the proper knowledge to formulate the similarity criteria in the domain.
- The ontology engineer, who is in charge of defining the similarity assessment on the basis of the ontology design and the information provided by the domain expert. He plays the role of communication channel for the requests of the domain expert with the system defining the application context to properly parameterize the similarity assessment.

Scenario 1: two different application contexts. The figure 4.5 illustrates the first scenario, which highlights the dependence of the similarity result on the similarity criteria induced by the application. The domain expert user formulates different similarity criteria in two different application contexts. The two sets of criteria are formalized by the ontology engineer according to the system formalization and the assessment performed. Two different results of the similarity assessment are provided by the system and represented by similarity matrices. It is evident in this scenario how two application contexts induce two different similarity matrices just by exploiting the same ontology.

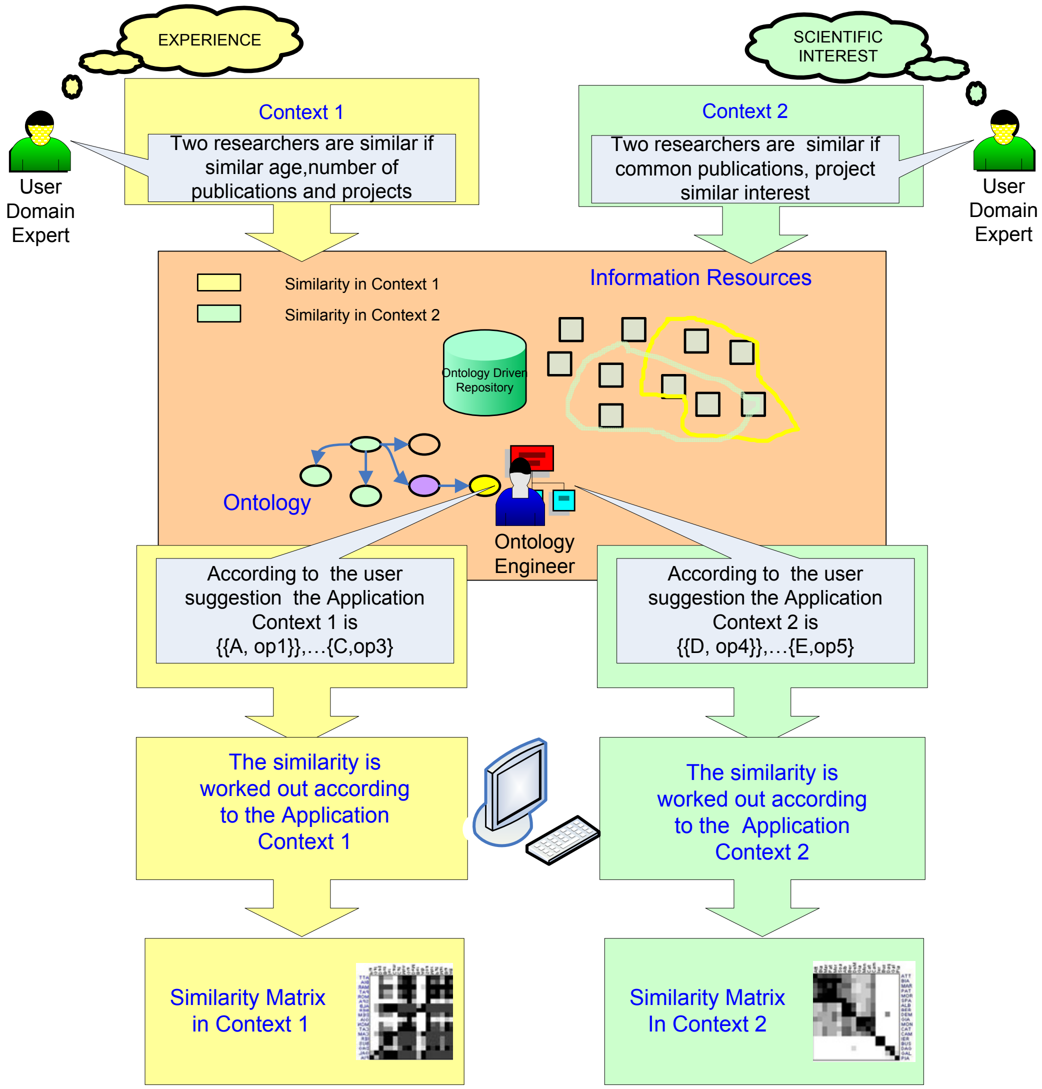

EXPERIENCE
SCIENTIFIC INTEREST
User Domain Expert
User Domain Expert
Context 1
Two researchers are similar if similar age,number of publications and projects
Context 2
Two researchers are similar if common publications, project similar interest
🟨 Similarity in Context 1
🟩 Similarity in Context 2
Information Resources
Ontology Driven Repository
Ontology
Ontology Engineer
According to the user suggestion the Application Context 1 is {{A, op1}},…{C,op3}
According to the user suggestion the Application Context 2 is {{D, op4}},…{E,op5}
The similarity is worked out according to the Application Context 1
The similarity is worked out according to the Application Context 2
Similarity Matrix in Context 1
Similarity Matrix In Context 2

Figure 4.5: *Scenario 1: similarity evaluation according to different application contexts.*

Scenario 2: similarity criteria refinement The scenario depicted in figure 4.6 is characterized by an interactive exchange of information between the two actors. The domain expert browses the repository looking for similar resources. He relies on his domain of knowledge to compare the resources, perceives the similarities among resources (which are not provided directly from the standard ontology reasoning technology) and provides some informal sim-

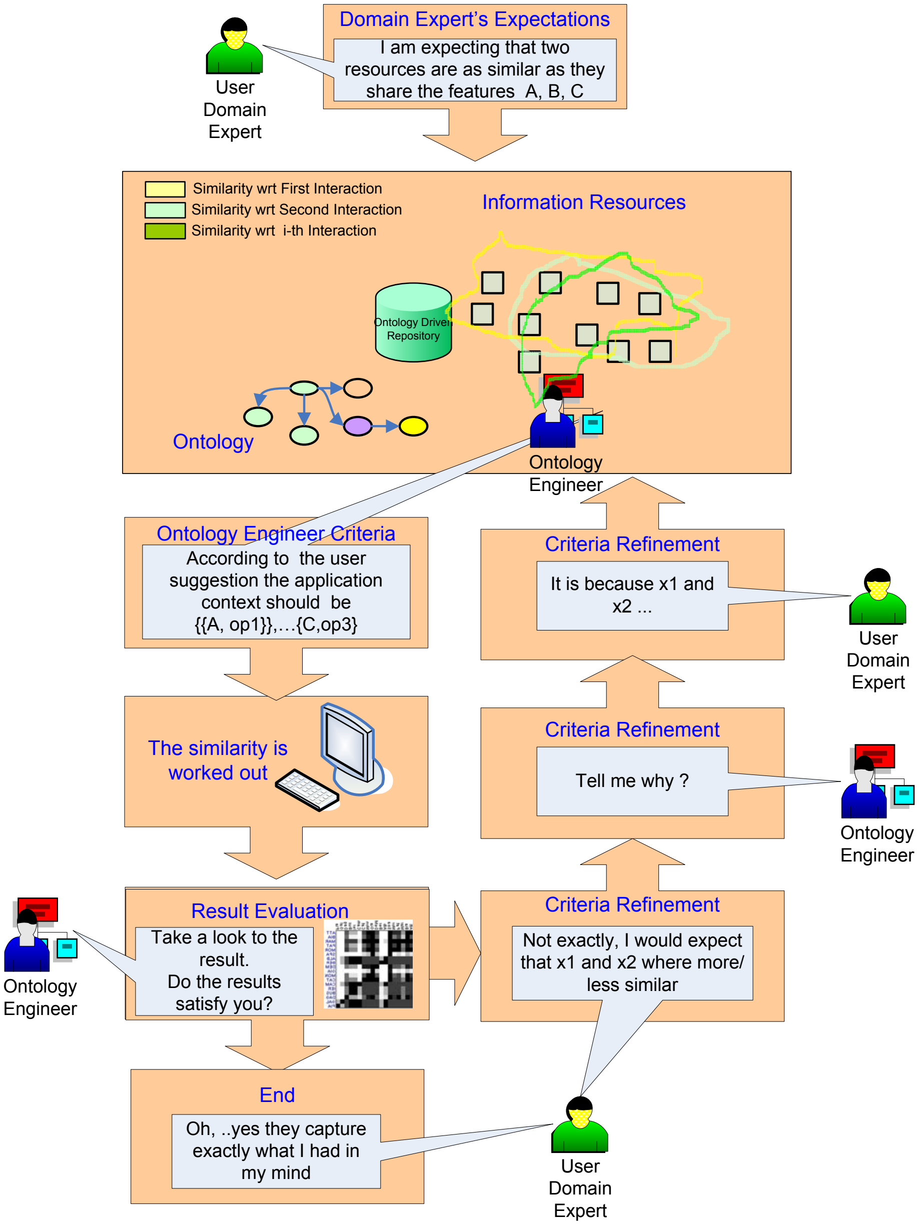

User Domain Expert
Domain Expert's Expectations
I am expecting that two resources are as similar as they share the features A, B, C
🟨 Similarity wrt First Interaction
🟩 Similarity wrt Second Interaction
🟩 Similarity wrt i-th Interaction
Information Resources
Ontology Driven Repository
Ontology
Ontology Engineer
Ontology Engineer Criteria
According to the user suggestion the application context should be {{A, op1}},…{C,op3}
The similarity is worked out
Result Evaluation
Take a look to the result. Do the results satisfy you?
Ontology Engineer
End
Oh, ..yes they capture exactly what I had in my mind
User Domain Expert
Criteria Refinement
Not exactly, I would expect that x1 and x2 where more/ less similar
Criteria Refinement
Tell me why ?
Ontology Engineer
Criteria Refinement
It is because x1 and x2 ...
User Domain Expert

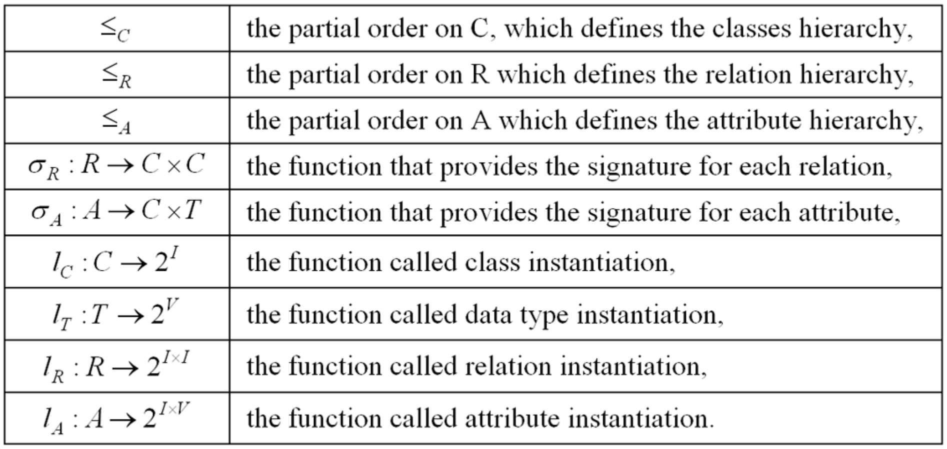
Figure 4.6: *Scenario 2: similarity criteria refinement.*

ilarity criteria to be adopted in the similarity evaluation. The ontology engineer translates the user requests to the system: he figures out which ontology entities are relevant and how to use them during the similarity assessment. The ontology engineer runs the similarity evaluation I have proposed and he shows the result to the domain expert. Analyzing the result, the domain expert might point out some unexpected result to the ontology engineer. Then the ontology engineer refines the similarity criteria interacting with the domain expert, until the results are considered correct. In this scenario, I assume that usually the domain expert is so familiar with the domain conceptualized in the ontology that his expectations about similarities are often implicit. Thus, he does not provide to the ontology engineer a complete set of information concerning the criteria of similarity to be used. Under this assumption the criteria definition process require further iterative refinement. In this scenario the framework supports the iterative criteria refinement process to precisely adapt the similarity assessments to the user expectations.

4.3.2 Preliminary assumptions

I propose a semantic similarity among instances taking into account the different hints hidden in the ontology. As the hints that can be considered largely depend on the level of formality of the ontology model adopted, it is important to state clearly to which ontology model a similarity method is referring. In my proposal, the ontology model with data type defined by Ehrig et al.[Ehri 05] is considered.

Definition 4 (Ontology with data type.) *An Ontology with data type is a structure* $O := (C, T, \leq_C, R, A, \sigma_R, \sigma_A, \leq_R, \leq_A, I, V, l_C, l_R, l_A)$ *where C,T,R,A,I,V are disjointed sets, respectively, of classes, data types, binary relations, attributes, instances and data values, and the relations and functions are defined as follows:*



$\leq_C$	the partial order on C, which defines the classes hierarchy,
$\leq_R$	the partial order on R which defines the relation hierarchy,
$\leq_A$	the partial order on A which defines the attribute hierarchy,
$\sigma_R : R \rightarrow C \times C$	the function that provides the signature for each relation,
$\sigma_A : A \rightarrow C \times T$	the function that provides the signature for each attribute,
$l_C : C \rightarrow 2^I$	the function called class instantiation,
$l_T : T \rightarrow 2^V$	the function called data type instantiation,
$l_R : R \rightarrow 2^{I \times I}$	the function called relation instantiation,
$l_A : A \rightarrow 2^{I \times V}$	the function called attribute instantiation.

A symmetric normalized similarity is a function $S : I \times I \rightarrow [0, 1]$, which satisfies the following axioms:

$$\begin{array}{ll} \forall x, y \in I \;\; S(x, y) \geq 0 & \textit{Positiveness} \\ \forall x \in I, \forall y, z \in I \;\; S(x, x) \geq S(y, z) & \textit{Maximality} \\ \forall x, y \in I \;\; S(x, y) = S(y, x) & \textit{Symmetry} \end{array}$$

An asymmetric normalized similarity is a normalized similarity $S : I \times I \rightarrow [0, 1]$ that does not satisfy the symmetric axiom. The preference between symmetric and asymmetric similarity mainly depends on the application scenario; in general, there is not an a-priori reason to formulate this choice. A complete framework for assessing the semantic similarity should be provided by both of them. I suppose to design an asymmetric similarity to determine when a resource can be used as a replacement for another. For this purpose, I am stressing the relation of containment between the information resources. The information resources are described by ontology driven metadata: therefore each resource is assumed to be an instance and the similarity is defined among pairs of instances.

Definition 5 (Containment between two information resources/instances.) *Given two information resources x, y (represented as instances in the ontology) and their sets of characteristics (coded as instance attributes and relation values), x is contained in y if the set of characteristics of x is contained in the set of characteristics of y.*

I assume that instance similarity behaves coherently with the concept of containment. Given two instances x, y their similarity is sim(x,y)=1 if and only if the set of characteristics of x is contained in the set of characteristics of y. On the contrary, unless y is contained in x, the similarity between y and x is sim(y,x)<1. The similarity value between x and y tends to decrease as long as the level of containment of their sets of properties decreases. Of course, the containment has to consider also the inheritance between the classes: if x belongs to a sub-class of the class of y, the asymmetric evaluation is performed relying on the idea that humans perceive similarity between a sub-concept and its super-concept as greater than the similarity between the super-concept and the sub-concept. Considering a browsing activity, this design of similarity is useful for determining which resource can be used as a replacement for another, sim(x,y) =1 means that y can be used as a substitute for x; on the other hand, the lower the similarity value, the more is lost in performing the replacement.

4.3.3 The approach of context dependent semantic similarity

The proposed approach adopts the schematisation of the similarity framework defined by Ehrig et.al. [Ehri 05]: the similarity is structured in terms of *data*, *ontology* and *context* layers plus the *domain knowledge* layer which spans all the other. The *data layer* measures the similarity of entities by considering the data values of simple or complex data types such as integer and string. The *ontology layer* considers the similarities induced by the ontology entities and the way they are related to each other. The *context layer* assesses the similarity according to how the entities of the ontology are used in some external contexts. The framework defined by Ehrig et al. is suitable for supporting the ontology similarity as well as instances similarity. Our contribution with respect to the framework defined by Ehrig et al. is mainly in the definition of a context layer including an accurate formalization of the criteria in order to tailor the similarity with respect to a context and in the definition of an ontology layer explicitly parameterized according to these criteria. Concerning the data and domain knowledge layers, my proposal adopts a replica of what is illustrated in [Ehri 05]. The formalization of the criteria of similarity induced by the context is employed to parameterize the computation of the similarity in the *ontology layer*, forcing it to adhere to the application criteria. The overall similarity is defined by the following amalgamation function (Sim) which aggregates two similarity functions defined in the ontology layer named *external similarity* (ExternSim) and *extensional similarity* (ExtensSim). The external similarity performs a structural comparison between two instances $i_1 \in l_c(c_1)$, $i_2 \in l_c(c_2)$ in terms of the classes c_1, c_2 that the instances belong to, whereas the extensional similarity performs a comparison of the instances in terms of their attributes and relations.

$$\mathrm{Sim}(i_1, i_2) = \frac{w_{\mathrm{ExternSim}} * \mathrm{ExternSim} + w_{\mathrm{ExtensSim}} * \mathrm{ExtensSim}}{w_{\mathrm{ExternSim}} + w_{\mathrm{ExtensSim}}} \tag{4.4}$$

$w_{\mathrm{ExternSim}}$ and $w_{\mathrm{ExtensSim}}$ are the weights used to balance the importance of the functions. By default they are equal to 1/2. In the sections below the context layer is described as well as the two similarities ExtensSim and ExternSim.

4.3.4 Context layer

The context layer, according to Ehrig at al. [Ehri 05], describes how the ontology entities concur in different contexts. Here I adopt the same point of view. However, I aim to formalize the application context in the sense of modelling the criteria of similarity induced by the context. This design choice does not hamper the eventual definition of a generic description of context followed by an automatic determination of which criteria would have been suitable for a given context. Rather, it allows us to calculate directly the similarity acting on the criteria, especially when it is necessary to refine them. In the following I underscore the importance of this formalization.

4.3.4.1 Motivation behind the application context formalization

The application context provides the knowledge for formalizing the criteria of similarity induced by the application. The criteria are context-dependent as the context influences the choice of classes, attributes and relations that are considered in the similarity assessment and the operations used to compare them. I describe the motivation behind the proposed formalization through an example based on the domain of academic research, considering as resources to be compared the researchers of a research institution. I chose this domain instead of a more specific area related to our research experience in the AIM@SHAPE project (such as solid modelling, 3D model reconstruction, virtual humans, etc.) because the ontologies within AIM@SHAPE are still under development (the ontology schema and the instances population are about be finalized within the end of 2007) and anyway the Researcher and academic research are without doubt more familiar fields to the readers. Let us consider a simplified version of the ontology KA[3] that defines concepts from academic research (Figure 4.7) and focus on the two applications: "comparison of the members of the research staff according to their working experience" and "comparison of the members of the research staff with respect to their research interest". Two distinct application contexts may be induced according the applications:

- "Exp" induced by the comparison of the members of the research staff according to their working experience. The similarity among the members of the research staff (instances of the class *ResearchStaff*[4] is roughly assessed by considering the member's age (the attribute age inherited by the class Person) and the number of projects and publications a researcher has worked on (the number of instances reachable through the relation publication and the relation *workAtProject* inherited by Staff).
- "Int" induced by the comparison of the members of the research staff with respect to their research interest. The researchers can be compared with respect to their interests (instances reachable through the relation interest), and again their publications (instances reachable through the relation publications) and the projects (instances reachable through the relation *workAtProject*).

The following points need to be considered when analyzing these examples:

1. The similarity between two instances can depend on the comparison of their related instances: the researchers are compared with respect to the instances of the class Publication connected through the relation publications.
2. The attributes and relations of the instances can contribute differently to the evaluation according to the context: the attribute age of the researchers is functional in the first application but it might not be interesting in the second; the relations publication and *workAtProject* are included in both application contexts but using different operators

[3] http://protege.stanford.edu/plugins/owl/owl-library/ka.owl

[4] The italics is used to explicit the references to the entities (attributes, relations,classes) of the ontology in Figure 4.7.

of comparison-in the first case just the number of instances is important whereas in the latter case the related instances have to be compared.

3. The ontology entities can be considered recursively in the similarity evaluation: in the context "Int" the members' research topic (instances of *ResearchTopic* reachable navigating through the relation *ResearchStaff->interest*[5] ed and their related topics (instances of *ResearchTopic* reachable via *ResearchStaff->interest->relatedTopic*) are recursively compared to assess the similarity of distinct topics.

4. The classes' attributes and relations can contribute differently to the evaluation according to the recursion level of the assessment: in the second application the attribute *topicName* and the relation *relatedTopic* can be considered at the first level of recursion to assess the similarity between *researchTopics*. By navigating the relation *relatedTopic* it is possible to apply another step of recursion, and here the similarity criteria can be different from the previous ones. For example, in order to limit the computational cost and stop the recursion, only the *topicName* or the instances identifier could be used to compare the *relatedTopic*.

As pointed out in the second remark, different operations can be used to compare the ontology entities, such as:

- Operation based on the "cardinality" of the attributes or relations: the similarity is assessed according to the number of instances the relations have, or the number of values that an attribute assumes. For example in the first context "Exp", two researchers are similar if they have a similar "number" of publications.

- Operation based on the "intersection" between sets of attributes or relations: the similarity is assessed according to the number of elements they have in common. For example in the context "Int", the more papers two researchers share, the more their interests are similar.

- Operation based on the "similarity" of attributes and relations: the similarity is assessed in terms of the similarity of the attributes values and related instances. For example, in the context "Int", two researchers are similar if they have "similar" research topics.

The example shows that an accurate formalism is needed to properly express the criteria that might arise from different application contexts. The formalization has to model the attributes and relations as well as the operations to compare their values. Moreover, as stated in the fourth remark, the level of recursion of the similarity assessment also has to be considered.

[5] The arrow is used to indicate the navigation through a relation, for example *A->B->C* means that starting from the class *A* I navigate through the relations *B* and *C*

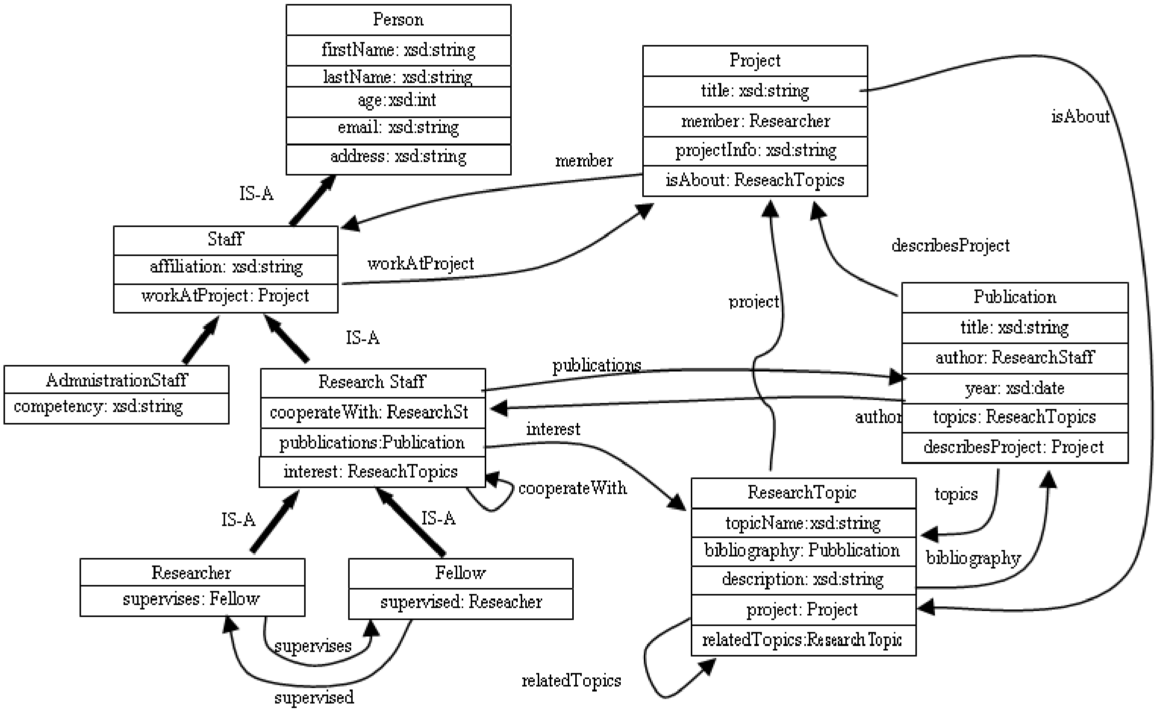


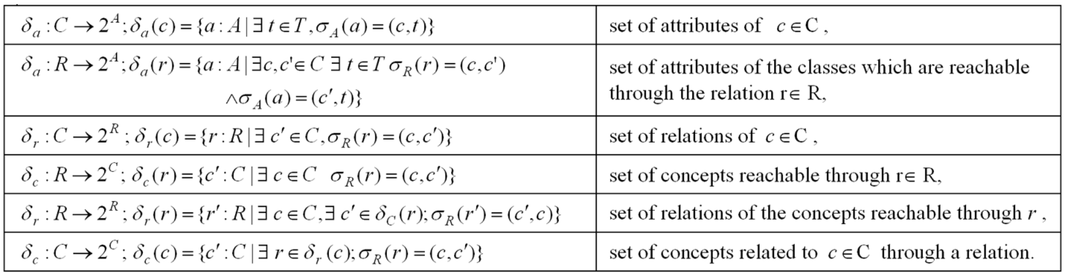
Figure 4.7: *Ontology defining concepts related to the academic research.*

4.3.4.2 Application context formalization

The formalization provided here represents the restrictions that the application context must adhere to. An ontology engineer is expected to provide the application context according to specific application needs. The formalization relies on the concepts of a "sequence of elements belonging to a set X", which formalizes generic sequences of elements, and a "path of recursion of length i" to track the recursion during the similarity assessment. In particular, a "path of recursion" represents the recursion in terms of the sequence of relations used to navigate the ontology. The application context function (AC) is defined inductively according to the length of the path of recursion. It yields the set of attributes and relations as well as the operations to be used in the similarity assessment. The operations considered are those described in the previous section and named, respectively, Count to evaluate the cardinality, Inter to evaluate the intersection, and Simil to evaluate the similarity.

Definition 6 (Sequences of a set X.) *Given a set X, a sequence s of elements of X with length n is defined by the function* $s : [1, .., n] \rightarrow X, n \in N^{+}$ *and represented in a simple way by the list [s(1),..,s(n)].*

Let $S_X^n = \{s | s : [1, n] \rightarrow X\}$ be the set of sequences of X having length n.
Let $\cdot : S_X^n \times S_Y^m \rightarrow S_{X \cup Y}^{n+m}$ be the operator "concat" between two sequences.
In Table 4.3 the polymorphism functions, which identify specific sets of entities in the ontology model are defined.



$\delta_a : C \rightarrow 2^A ; \delta_a(c) = \{a : A \mid \exists\, t \in T, \sigma_A(a) = (c,t)\}$	set of attributes of $c \in \mathrm{C}$,
$\delta_a : R \rightarrow 2^A ; \delta_a(r) = \{a : A \mid \exists c, c' \in C\ \exists\, t \in T\ \sigma_R(r) = (c,c')$ $\wedge \sigma_A(a) = (c',t)\}$	set of attributes of the classes which are reachable through the relation r∈ R,
$\delta_r : C \rightarrow 2^R ; \delta_r(c) = \{r : R \mid \exists\, c' \in C, \sigma_R(r) = (c,c')\}$	set of relations of $c \in \mathrm{C}$,
$\delta_c : R \rightarrow 2^C ; \delta_c(r) = \{c' : C \mid \exists\, c \in C\ \ \sigma_R(r) = (c,c')\}$	set of concepts reachable through r∈ R,
$\delta_r : R \rightarrow 2^R ; \delta_r(r) = \{r' : R \mid \exists\, c \in C, \exists\, c' \in \delta_C(r); \sigma_R(r') = (c',c)\}$	set of relations of the concepts reachable through r ,
$\delta_c : C \rightarrow 2^C ; \delta_c(c) = \{c' : C \mid \exists\, r \in \delta_r(c); \sigma_R(r) = (c,c')\}$	set of concepts related to $c \in \mathrm{C}$ through a relation.

Table 4.3: *List of functions defining specific sets of elements in the ontology model*

Definition 7 (Path of Recursion.) *A path of recursion p with length i is a sequence whose first element is a class and whose other elements are relations recursively reachable from the class:* $p \in S_{C \cup R}^i \mid p(1) \in C \wedge \forall j \in [2, i]\ p(j) \in R \wedge p(j) \in \delta_r(p(j-1))$.

For example, a path of recursion with length longer than three is a path that starts from a class p(1) and continues to one of its relations as the second element p(2) and then to one of the relations of the class reachable from p(2) as the third element p(3), and so on. In general, a path of recursion p represents a path that is followed to assess the similarity recursively. The recursion expressed in the previous section in the context "Int" as *ResearchStaff->interest->relatedTopic* is formalized with the path of recursion [ResearchStaff,interest,relatedTopic]. Let P^i be the set of all paths of recursion with length i and P be the set of all paths of recursion $P = \bigcup_{i \in N} P^i$.

Definition 8 (Application Context AC.) *Given the set P of paths of recursion, L=*{*Count, Inter, Simil*} *the set of operations adopted, an application context is defined by a partial function AC having the signature* $AC : P \rightarrow (2^{A \times L}) \times (2^{R \times L})$*, yielding the attributes and relations as well as the operations to perform their comparison.*

In particular, each application context AC is characterized by two operators $AC_A : P \rightarrow (2^{A \times L})$ and $AC_R : P \rightarrow (2^{R \times L})$, which yield, respectively, the parts of the context AC related to the attributes and the relations. Formally $\forall p \in P\ AC(p) = (AC_A(p), AC_R(p))$ and $AC_A(p)$ and $AC_R(p)$ are set of pairs $\{(e_1, o_1), (e_2, o_2), ..., (e_i, o_i), ..., (e_n, o_n)\}\ n \in N$ where e_i is, respectively, the attribute or the relation relevant to define the similarity criteria and $o_i \in L$ is the operation to be used in the comparison. I provide two examples of AC formalization referring to the two application contexts "Exp" and "Int" mentioned in the previous section.

Example 1 Let us formalize the application context "Exp" with AC_{Exp} to assess the similarity among the members of a research staff according to their experience. I consider the set of paths of recursion {[ResearchStaff],[Research],[Fellow]} and I compare them according to age similarity and the numbers of publications and projects. Thus AC_{Exp} is defined by:

$$\begin{aligned}
\text{[ResearchStaff]} &\overset{AC_{Exp}}{\rightarrow} \{\{(\text{age,Simil})\},\{(\text{publications,Count}),(\text{workAtProject,Count})\}\} \\
\text{[Researcher]} &\overset{AC_{Exp}}{\rightarrow} \{\{(\text{age,Simil})\},\{(\text{publications,Count}),(\text{workAtProject,Count})\}\} \quad (4.5)\\
\text{[Fellow]} &\overset{AC_{Exp}}{\rightarrow} \{\{(\text{age,Simil})\},\{(\text{publications,Count}),(\text{workAtProject,Count})\}\}
\end{aligned}$$

An example of AC_R is {(publication,Count), (workAtProject , Count)} while an example of AC_A is {(age,Simil)}. Note that [Researcher] and [Fellow] belong to the set of paths of recursion considered in AC_{Exp} because their instances are also instance of *ResearchStaff*. The application context can be expressed in a more compact way assuming that whenever a context is not defined for a class but is defined for its super class, the comparison criteria defined for a super class are by default inherited by the subclasses. According to this assumption AC_{Exp} can be expressed by:

$$\text{[ResearchStaff]} \overset{AC_{Exp}}{\rightarrow} \{\{(\text{age,Simil})\},\{(\text{publications,Count}),(\text{workAtProject,Count})\}\} \quad (4.6)$$

Example 2 Let us formalize the application context "Int" to assess the similarity among the members of a research staff according to their research interest. The similarity is computed considering the set of paths of recursion {[ResearchStaff], [ResearchStaff,interest]}. The researchers are compared considering common publications, common projects or similar interests. A compact formalization for "Int" is defined by AC_{Int}:

$$\begin{aligned}
\text{[ResearchStaff]} &\overset{AC_{Int}}{\rightarrow} \{\{\phi\},\{(\text{publications,Inter}),(\text{workAtProject,Inter}),(\text{interest,Simil})\}\} \\
\text{[ResearchStaff,interest]} &\overset{AC_{Int}}{\rightarrow} \{\{(\text{topicName,Inter})\},\{(\text{relatedTopics,Inter})\}\} (4.7)
\end{aligned}$$

In general, the operator *Count* applied to attributes or relations means that the number of attribute values or related instances is considered in the similarity assessment. For example, according to the context formalized in equation 4.6 (second row), two researchers, who are represented as instances of *Researcher*, are similar if they have a similar numbers of instances of *Publication* reachable through the relation *publications*. The operator *Inter* applied to attributes or relations means that common attributes values or related instances are considered in the similarity assessment. For example, according to the context formalized in equation 4.7 (first row) two researchers are considered as similar if they have common project instances. When applied to an attribute, the operator *Simil* determines that the attribute values of two instances will be compared according to a datatype similarity provided by the data layer (see the example in equation 4.6, first row, attribute *age*). When it is applied to a relation, it determines a step of recursion, in the sense that the instances related through the relation have to be considered during the similarity assessment. How these related instances have to be compared is specified by the value provided by the context function for the corresponding recursion path. Note that the researchers are compared recursively in the context expressed by equation 4.7. In fact the relation *interest* is included with the operator *Simil* in the first row of equation 4.7. This means that the instances of *ResearchTopic* associated with the researcher via *interest* have to be accessed and compared recursively when the researchers' similarity is worked out. Actually, [ReseachStaff,interest] is the path of recursion to navigating the ontology from *ResearchStaff* to *ResearchTopic* via the relation interest. Once the assessment has accessed the related instances, it compares them as indicated by the second row of equation 4.7. The interests are compared with respect to both their *topicName* and their *relatedTopic*; thus, two *ResearchTopic*(s) having distinct *topicNames* but some *relatedTopic* in common are not considered completely dissimilar.

The image of an AC function can be further characterized by the following:

1. For a path of recursion p, AC has to yield only the attributes and relations belonging to the classes reached through p. For example, considering the ontology in Figure 4.7and the path of recursion [ReseachStaff,interest] it is expected that only the attributes and relations belonging to the class *ResearchTopic* reachable via [ReseachStaff,interest] can be identified by AC([ReseachStaff,interest]). Attributes or relations (such as *age*, *publications*, etc) which do not belong to *ResearchTopic* define an incorrect application context.

2. Given a path of recursion p, an attribute or a relation can appear in the context image at most one time. In other words, given a path of recursion it is not possible to associate two distinct operations with the same relation or attribute. For example the following application context definition is not correct as *interest* is specified twice

$$\text{[ResearchStaff]} \rightarrow \{\{\phi\}, \{\text{(publications,Inter)},\{\text{Interest,Simil}\}\{\text{Interest,Inter}\}\} \quad (4.8)$$

4.3.5 Ontology layer

The ontology layer defines the asymmetric similarity functions *ExternSimil* and *ExtensSimil* that constitute the amalgamation function (equation 4.4). The "external similarity" *ExternSimil* measures the similarity at the level of the ontology schema computing a structural comparison of the instances. Given two instances, it compares the classes they belong to, considering the attributes and relations shared by the classes and their position within the class hierarchy. The "extensional similarity" *ExtensSimil* compares the extension of the ontology entities. The similarity is assessed by computing the comparison of the attributes and relations of the instances. At the ontology layer additional hypotheses are assumed:

- All classes defined in the ontology have the fake class *Thing* as a super-class.

- Given $i_1 \in l_c(c_1)$, $i_2 \in l_c(c_2)$, if c_1,c_2 do not have any common super-class different from *Thing*, their similarity is equal to 0.
- The least upper bound (*lub*) between c_1 and c_2, is unique and it is c_2 if c_1 IS-A c_2, or c_1 if c_2 IS-A c_1, otherwise the immediate super-class of c_1 and c_2 that subsumes both classes.

The aim is to force the *lub* to be a sort of "template class" that can be adopted to perform the comparison of the instances whenever the instances belong to distinct classes. Referring to the ontology in Fig. 4.7, it can be appropriate to compare two instances belonging respectively to *AdministratorStaff* and *ResearchStaff* as they are both a kind of staff and *Staff* is their *lub*. However, it does not make sense to evaluate the similarity between two instances belonging to *Publication* and to *Staff*, because they are intimately different: in fact, there is not any *lub* available for them. Whenever a *lub* x between two classes exists, the path of recursion [x] is the starting path in the recursive evaluation of the similarity.

4.3.5.1 External similarity

The external similarity (*ExternSimil*) performs the structural comparison between two instances i_1, i_2 in terms of the classes c_1, c_2 that the instances belong to: more formally $\mathit{ExternSimil}(i_1, i_2) = \mathit{ExternSim}(c_1, c_2)$ where $i_1 \in l_c(c_1), i_2 \in l_c(c_2)$.
In my proposal the external similarity function is defined starting from the similarities proposed by Maedche and Zacharias [Maed 02] and Rodriguez and Egenhofer [Rodr 04]. The structural comparison is performed by two similarity evaluations:

Class Matching which is based on the distance between the classes c_1, c_2 and their depth with respect to the hierarchy induced by $\leq_C$.

Slot Matching which is based on the number of attributes and relations shared by the classes c_1, c_2 and the overall number of their attributes and relations. Then two classes having many attributes/relations, some of which are in common, are less similar than two classes having fewer attributes but the same number of common attributes/relations.

Both similarities are needed to successfully evaluate the similarity with respect to the ontology structure. For example, let us consider the ontology schema in Fig. 4 and let compare an instance of the class D with an instance of the class E. They are quite similar with respect to class matching but less similar with respect to slot matching. In fact, the sets of IS-A relations joining the classes D and E to *Thing* are largely shared. However, from the point of view of the slots, D and E share only the attribute A_1 and the relation $\underline{C_1}$, and they differ with respect to the others. Likewise it would be easy to show an example of two classes that are similar with respect to slot matching and dissimilar according to class matching.

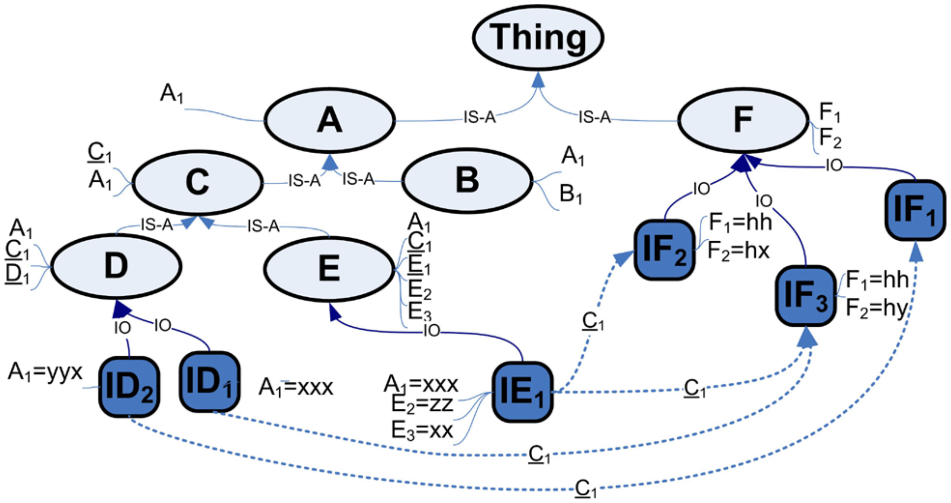


Figure 4.8: *Class hierarchy example: A, B, C, D, E, F are classes, $A_1, B_1, E_2, E_3, F_1, F_2$ are attributes, $\underline{C_1}, \underline{D_1}, \underline{E_1}$ are relations, $ID_1, ID_2, IE_1, IF_1, IF_2, IF_3$ are instances*

Definition 9 (ExternSim similarity.) *The similarity between two classes according to the external comparison is defined by:*

$$ExternSim(c_1, c_2) = \begin{cases} 1 & \text{if } c_1 > c_2 \\ \frac{w_{SM}*SM(c_1,c_2)+w_{CM}*(c_1,c_2)}{W_{SM}+W_{CM}} & \text{otherwise} \end{cases} \tag{4.9}$$

where SM is the Slots Matching, CM is the Classes Matching and w_{SM} , w_{CM} the respectively weights in the range [0,1].
For the purpose of this section, wSM and wCM are defined as equal to 1/2.

4.3.5.1.1 Class Matching Classes Matching is evaluated in terms of the distance of the classes with respect to the IS-A hierarchy. The distance is based on the concept of Upwards Cotopy (UC)[Maed 02]. I define an asymmetric similarity adapting the symmetric definition of CM in [Maed 02] .

Definition 10 (Upward Cotopy (UC).) *The Upward Cotopy of a set of classes C with the associated partial order $\leq_C$ is:*

$$UC_{\leq_C}(c_1) := \{c_j \in C | (c_i \leq_C c_j) \vee c_i = c_j\} \tag{4.10}$$

It is the set of classes composing the path that reaches from c_i to the furthest super-class (*Thing*) of the IS-A hierarchy: for example considering the class D in figure 4.8 $UC_{\leq_C}(D) = \{D, C, A, Thing\}$.

Definition 11 (Asymmetric Class Matching.) *Given two classes c_1, c_2 and the Upward Cotopy $UC_{\leq_C}(c_i)$, the asymmetric Class Matching is defined by:*

$$CM = (c_1, c_2) := \frac{|UC_{\leq_C}(c_1) \cap UC_{\leq_C}(c_2)|}{|UC_{\leq_C}(c_1)|} \tag{4.11}$$

CM between two classes depends on the number of classes they have in common in the hierarchy. Let us note that the class matching is asymmetric, for example referring to figure 4.8, CM(B,D)=2/3 but CM(D,B)=2/4. Moreover, it is important to note that CM(A,D)=1; the rationale behind this choice of design pertains to the property of containment between instances: the instances of D fit with the instances of A, and they can replace the instances of A at the class level.

4.3.5.1.2 Slot Matching

Slot Matching is defined by the slots (attributes and relations) shared by the two classes. I refer to the similarity proposed by Rodriguez and Egenhofer [Rodr 04] based on the concept of distinguishing features employed to differentiate subclasses from their super-class. In their proposal, different kinds of distinguishing features are considered (i.e. functionalities, and parts) but none coincides immediately with the native entities in our ontology model. Of course it would be possible to manually annotate the classes, adding the distinguishing features but I prefer to focus on what is already available in the adopted ontology model. Therefore only attributes and relations are mapped as two kinds of distinguishing features.

Definition 12 (Slot Matching.) *Given two classes c_1, c_2, two kinds of distinguishing features (attributes and relations) and w_a, w_r, the weights of the features, the similarity function between c1 and c2 is defined in terms of the weighted sum of the similarities S_a and S_r, where S_a is the slot matching according to the attributes and S_r in the slot matching according to the relations.*

$$SM(c_1, c_2) := w_a * S_a(c_1, c_2) + w_r * s_r(c_1, c_2) \tag{4.12}$$

The sum of the weights is expected to be equal to 1, and by default I assume $w_a = w_r = 1/2$. The two slot matching similarities S_a and S_r rely on the definitions of *slot importance* as defined in the following.

Definition 13 (Function of "Slot Importance" α.) *Let c_1, c_2, be two distinct classes and d be the class distance $d(c_1, c_2)$ in term of the number of edges in an IS-A hierarchy, then α is the function that evaluates the importance of the difference between the two classes.*

$$\alpha(c_1, c_2) = \begin{cases} \frac{d(c_1, lub(c_1, c_2))}{d(c_1, c_2)} & d(c_1, lub(c_1, c_2)) \leq d(c_1, c_2) \\ 1 - \frac{d(c_1, lub(c_1, c_2))}{d(c_1, c_2)} & d(c_1, lub(c_1, c_2)) > d(c_1, c_2) \end{cases} \tag{4.13}$$

where $d(c_1, c_2) = d(c_1, lub(c_1, c_2)) + d(c_2, lub(c_1, c_2))$.

α(c1, c2) is a value in the ranges [0,0.5]. Referring to the figure 4.8, α(D,C) is equal to zero because the lub between D and C is C itself, d(C,D)=1 and d(C,C)=0. Whereas α(D,E) is equal to 0.5 because the lub is still C, and d(D,E)=2.

Definition 14 (Slot Matching according to the kind of distinguishing feature t.) *Given two classes c_1 (target) and c_2 (base) and t a kind of distinguishing feature (t=a for attributes or t=r for relations), let C_1^t and C_2^t be the sets of distinguishing features of type t respectively of c_1 and c_2; the Slot Matching $S_t(c_1, c_2)$ is defined by*[6]

$$S_t(c_1, c_2) = \frac{|C_1^t \cap C_2^t|}{|C_1^t \cap C_2^t| + (1 - \alpha(c_1, c_2))||C_1^t / C_2^t| + \alpha(c_1, c_2)||C_2^t / C_1^t|} \tag{4.14}$$

According to the ontology in figure 4.8, considering the classes D and E their sets of distinguishing features of type relation are $D^r = \{\underline{C_1}, \underline{D_1}\}$ and $E^r = \{\underline{C_1}, \underline{E_1}\}$ and $\alpha(E, D) = 0.5$; then $S_r(E, D) = 0.5$. Furthermore, this formulation of the class matching is coherent to the containment property: considering the classes A and E, their sets of distinguishing features of type attribute are respectively $A^a = \{A_1\}$, $E^a = \{A_1, E_2, E_3\}$ and $\alpha(A, E) = 0$, so that $S_a(A, E) = 1$. This means that the instances of E can replace the instances of A because they have some quality more rather than less similar. The contrary is not true: in fact $\alpha(E, A) = 0$ and $S_a(E, A) = 0.33$. In general, whenever $\alpha = 0.5$ the differences between features of both classes are equally important for the matching: for example this happens when the classes are sisters as for D and E. In the case of $\alpha = 0$ only the features that are in c_1 and not in c_2 are important for the matching.

4.3.5.2 Extensional similarity

The extension of entities plays a fundamental role in the assessment of the similarity among the instances: it is needed to perform a comparison of the attribute and relation values. For example, in the ontology in figure 4.8 relying only on the structural comparison it is not possible to assess that ID_1 is more similar to IE_1 than to ID_2. The main principle of the proposed extensional similarity between two instances is to consider the lub x of their classes as the common base for comparing them when the instances belong to different classes: it is adopted to define the path of recursion [x] from which starts the recursive assessment induced by an application context. For example, considering the instances ID_1 and IE_1 in figure 4.8, the class C is their lub. Then the initial path of recursion from which to start the similarity assessment is [C]. Let us suppose I have already defined an application context as the following $[C] \rightarrow \{\{(A_1, Iter)\}, \{(\underline{C_1}, Simil)\}\}$; $[C, \underline{C_1}] \rightarrow \{\{(F_1, Simil)\}, \{\}\}$. The computation starts from the values of the attribute A_1 for the instances ID_1 and IE_1. The

[6] The formulation is slightly different from that provided by Egenhofer and Rodriguez: the parameters of the similarity have been reversed to be coherent with the relation between instances containment and the similarity value equal to 1.

operation *Inter* induces a comparison, which focuses on the common values between ID_1 and IE_1 for the attribute A_1. Then as the operation *Simil* is associated with the relation $\underline{C_1}$ by the context, the new path of recursion $[C, \underline{C_1}]$ is considered to compare the instances related to IE_1 and ID_1. The context associated with the new path of recursion induces a similarity comparison of the values of the attribute F_1. The extensional comparison is characterised by two similarities functions: a function based on the comparison of the attributes of the instances and a function based on the comparison of the relations of the instances.

Definition 15 (Extensional Asymmetric Similarity.) *Given two instances $i_1 \in l_c(c_1)$, $i_2 \in l_c(c_2)$, $c = lub(c_1, c_2)$ and p=[c] a path of recursion defined in the application context AC[7], let $Sim_a^p(i_1, i_2)$ and $Sim_r^p(i_1, i_2)$ be the similarity measurements between instances considering respectively their attributes and their relations. The extensional similarity with asymmetric property is defined by:*

$$ExtensSim(i_1, i_2) = \begin{cases} 1 & i_1 = i_2 \\ Sim_I^p(i_1, i_2) & otherwise \end{cases} \tag{4.15}$$

where $Sim_I^p(i_1, i_2)$ is defined by:

$$Sim_I^p(i_1, i_2) = \frac{\sum\limits_{a \in \delta_a(c)} Sim_a^p(i_1, i_2) + \sum\limits_{r \in \delta_r(c)} Sim_r^p(i_1, i_2)}{|AC_A(p)| + |AC_R(p)|} \tag{4.16}$$

Note that the index p is a kind of stack of recursion adopted to track the navigation of relations whenever the similarity among instances is recursively defined in terms of the related instances. and are defined by a unique equation as follows.

Definition 16 (Similarity on Attributes and Relations.) *Given two instances $i_1 \in l_c(c_1)$, $i_2 \in l_c(c_2)$, $c = lub(c_1, c_2)$, p=[c] a path of recursion, X a placeholder for the "A" or "R", $x \in A \cup B$, then let:*

- $i_A(i) = \{v \in V | (i, v) \in l_A(a), \exists y \in C \ s.t. \ \sigma_A(a) = (y, T) \wedge l_T(T) = 2^V\}$*the set of values assumed by the instance i for the attribute a,*
- $i_R(i) = \{i' \in l_c(c') | \exists c \ i \in l_c(c) \ \exists \ c' \ s.t. \ \sigma_R(r) \in (c, c') \wedge (i, i') \in l_R(r)\}$ *the set of instances related to the instance i by the relation r,*
- *AC be the application context defined according to the restrictions defined in section 4.3.4.2*
- $F_X = \{g : i_X(i_1) \rightarrow i_X(i_2) | g \ is \ partial \ and \ bijective\}$

[7] Note that $|AC_A(p)| + |AC_R(p)| \neq 0$ whenever the context AC specifies at least a relevant attribute or relation for the recursion path p.

The similarity between instances according to their attributes or relations is:

$$Sim^{p}_{x}(i_1,i_2) = \begin{cases} 1 & \textit{if } i_X(i_1) \textit{are empty set} \\ 0 & \textit{if } (i_X(i_1) \neq \phi \wedge i_X(i_2) = \phi) \\ \frac{|i_X(i_2)|}{max(|i_X(i_1)|,|i_X(i_2)|)} & \textit{If}(x, \textit{Count}) \in AC_X(p) \\ \frac{|i_X(i_1) \cap i_X(i_2)|}{|i_X(i_1)|} & \textit{if } (x, \textit{Inter}) \in AC_X(p) \\ \frac{max_{f \in F_A} \sum\limits_{v \in i_A(i_1)} sim^{a}_{T}(v, f(v)))}{min(|i_A(i_1)|,|i_A(i_2)|)} * Fact^{i_1,i_2}_{A} & \textit{if } (x = a) \wedge (a, \textit{Simil} \in AC_A(p)) \\ \frac{max_{f \in F_R} \sum\limits_{v \in i_R(i_1)} sim^{pNew}_{T}(v, f(v))}{min(|i_R(i_1)|,|i_R(i_2)|)} * Fact^{i_1,i_2}_{R} & \begin{array}{l} \textit{if } (x = r) \wedge (r, \textit{Simil} \in AC_R(p)) \\ pNew = p \cdot s, s \in S \in S^{1}_{R}, S(1) = r \end{array} \end{cases} \tag{4.17}$$

where $Fact^{i_1,i_2}_{X}$ *is defined as*

$$Fact^{i_1,i_2}_{X} = (1 - max(0, \frac{|i_X(i_1)| - |i_X(i_2)|}{|i_X(i_1)|})) \tag{4.18}$$

These equations are designed to be asymmetric and to respect the properties of containment among instances: if an instance i_2 has at least the same attribute and relation values as i_1, then the extensional similarity between i_1 and i_2 is equal to one. The approach computes Sim^{p}_{x}, selecting one of the above equations according to the definition of AC

- In the first case the similarity is 1 if the set of the property values of the first instance is empty, because an instance having no characteristics is contained in all the other instances.

- In the second case the similarity is 0 if the first instances having at least a property value are compared with an instance that does not have any value.

- The third equation is adopted if AC yields a relation or attribute associated with the operation *Count*.

- The fourth equation is adopted if AC yields a relation or attribute associated with the operation *Inter*.

- The fifth equation is adopted if AC yields an attribute with the operation *Simil*.

- The last equation is adopted if AC yields a relation with the operation *Simil*. It is important to note that each time the similarity is assessed in terms of related instances (whenever $(r, \textit{Simil}) \in AC_R(p)$), the relation r followed to reach the related instances is added to the path of recursion. Thus, during the recursive assessment, the AC is always worked out on the most updated path of recursion.

In the last two equations, the comparison of the attribute values relies on the function Sim_T^a, which defines the similarity for the values of the attribute a having data type T. Sim_T^a is provided by the data layer as suggested by [Albe 06a]. The set of partial functions in F_X are employed to represent the possible matching among the set of values when the instances have relations or attributes with multiple values. For example, the instance IE_1 has IF_3 and IF_2 related via $\underline{C_1}$ and ID_1 has IF_3. When IE_1 and ID_1 are compared, two possible partial and bijective functions f_1 and f_2 can be considered between the instances related to IE_1 and ID_1: $f1 : IF_2 \rightarrow IF_3$ and $f_2 : IF_3 \rightarrow IF_3$.

Moreover, note that, in the last two equations, the number of properties of i_1 that are not shared in i_2 (if there are any) also affects the similarity evaluation. It is represented by the two factors: $Fact_A^{i_1,i_2}$ and $Fact_R^{i_1,i_2}$. These factors yield 1 if i_1 is contained in i_2; otherwise they yield the ratio between the number of properties of i_1 and the number of properties of i_2.

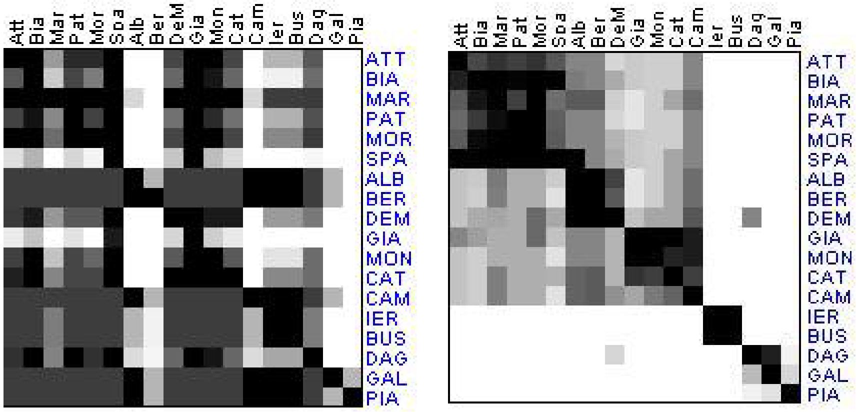
4.3.6 Experiments and evaluations

I evaluated our approach for the similarity assessment among the research staff working at the Institute (CNR-IMATI-GE). An experiment was performed to demonstrate both the need for the content-dependent similarity and the importance of defining an asymmetric similarity based on the containment to select similar resources.

4.3.6.1 Experiments

Two experiments are performed considering the contexts “Exp”, “Int” mentioned in section 4.3.4.1. Eighteen members of the research staff are considered; the information related to their projects, journal publications and research interests are inserted as instances in the ontology depicted in Figure 4.7 according to what is published at the IMATI web site[8]. The ontology is expressed in OWL ensuring that only the language constructs consistent with the ontology model considered in definition 4 are adopted. The resulting ontology is available at the web site [TestOnto]. Our method is implemented in JAVA and tested on this ontology. Using the formalization of the two application contexts AC_{Int} and AC_{Exp} previously defined (equations 4.6, 4.7) I have computed the similarity through the proposed framework. The results are represented by the similarity matrices in figure 4.9: (a) is the result related to the context “Exp” and (b) is the result related to the context “Int”. Each column j and each row i of the matrix represent a member of the research staff (identified by the first three letter of his name). The grey level of the pixel (i,j) represents the similarity value (Sim(i,j)) between the two members located at row i and columns j: the darker the colour, the more similar are the two researchers.

[8] http://www.ge.imati.cnr.it, accessed the 12/05/2006



(a) (b)

Figure 4.9: *(a) Similarity matrix for context "Exp"; (b) Similarity matrix for context "Int"*

Analysing the similarity matrices I can make the following statements. It is easy to see that they are asymmetric: for example sim(Dag,Bia)=1 while sim(Bia,Dag)<1. This confirms that the proposed model assesses an asymmetric similarity. The asymmetry result is particularly useful for comparing researchers because it behaves according to the property of containment previously defined. For example the two results sim(Dag,Bia)=1 and sim(Bia, Dag)<1 in figure 4.9(a) mean that if Bia has at least the experience of Dag, then Dag can replace Bia. The inverse is not true and if the domain expert decides to choose Dag instead of Bia, the similarity value provides a hint about the loss inherent in this choice [for example, if sim(Bia, Dag)=0.85, then the loss is 15%]. The comparison of the two matrices shows how they are different; it is evident that the two contexts induce completely different similarity values. For example, "Dag" results are very similar to "Bia" with respect to their experience (black pixel in figure 4.9(a)), but they are no similar with respect to their research interests (white pixel in figure 4.9(b)). Moreover during the test process I realised that the approach provides a sort of tool for context tuning supporting us in the decision making process to formulate the similarity criteria. From the similarity results I were able to learn and refine our criteria to obtain the expected results.

4.3.6.2 Evaluations

Two kinds of evaluations of the results concerning the similarity obtained with respect to the research interest (figure 4.9(b)) are performed. The first evaluation is based on the concept of recall and precision calculated considering the same adaptation of recall and precision made by [Rodr 03]. More precisely, considering an entity x the recall and precision are defined respectively as $(A \cap B)/A$ and $(A \cap B)/B$ where A is the set of entities expected to be similar to x, and B is the set of similar entities calculated by a model. A critical issue in the similarity evaluation is to have a ground truth with respect to comparing the results obtained. I faced this problem in referring to the research staff of our institute when considering as "similar" two members of the same research group. In fact at IMATI researchers and fellows are grouped into three main research groups, and one of those is composed of three further sub-groups. Therefore, I considered the research staff as split into five groups. For each member i, A is the set of members of his research group while B is composed of the first n members retrieved by the model. I have calculated recall and precision for each group considering "n" as the smallest number of members needed to obtain a recall of 100%, and then I have evaluated the precision. The average recall was estimated to be equal to 100% with a precision of 95%. These results are quite encouraging: a recall equal to 100% demonstrates that, for each research group, the similarity is able to rank all the expected members, while a precision equal to 95% means that the average number of outsiders that need to be included to rank all group members is equal to 5%. I have performed a second evaluation according to the context "Int" using a data mining application. For each researcher and fellow I have computed his similarity with respect to the other members applying our method. In this way, I associate with each research staff member a string of values, which correspond to his relative distance from the other members. The strings correspond to the rows of the similarity matrix (figure 4.9(b)). Then I have applied a tool to perform hierarchical clustering among the genetic microarray [Hierarch 04] to the set of strings, considering each string as a kind of researcher genetic code. The dendrogram obtained is shown in figure 4.10. It recognizes the five clusters that resemble the research group structure of our institute.

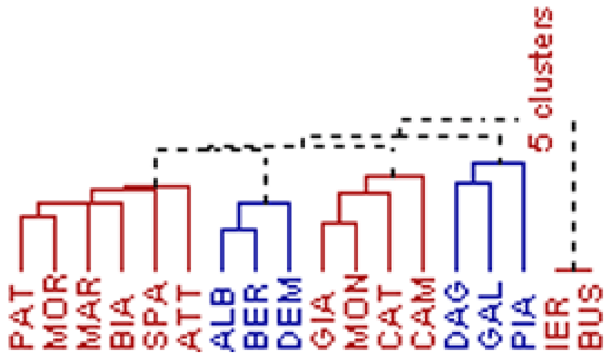


Figure 4.10: *The dendrogram obtained through the hierarchical gene clustering*

4.3.7 Discussion and related work

Semantic similarity is employed differently according to the application domain where it is adopted. Currently it is relevant in ontology alignment [Euze 04b, Euze 04a] and conceptual retrieval [Schw 05a] as well as in Semantic Web service discovery and matching

[Usan 05, Hau 05]. It is expected to increase in relevance the framework for metadata analysis [Albe 05b]. I discuss here related works according to their purpose and the ontology model they adopt.

Similarities in the Ontology alignment. There are many methods for aligning ontology, as pointed out by Euzenat et al. [Euze 04a]. Semantic similarity is adopted in this context to figure out relations among the entities in the ontology schema. It is used to compare the name of classes, attributes and relations, determining reasonable mapping between two distinct ontologies. However, the method proposed in this thesis is specifically designed to assess similarity among instances belonging to the same ontology. Some similarities adopted for ontology alignment consider quite expressive ontology language, (e.g., [Euze 04b] focus on a subset of OWL Lite) but they mainly focus on the comparison of the structural aspects of ontology. Due to the different purposes of these methods, they turn out to be unsuitable for properly solving the similarity among instances.

Concepts similarity in lexicographic databases. Different approaches to assess semantics similarity among concepts represented by words within lexicographic databases are available. They mainly rely on edge counting-base [Rada 89] or information theory-based methods [Li 03]. The edge counting-base method assigns terms that are subjects of the similarity assessment as edges of a tree-like taxonomy and defines the similarity in terms as the distance between the edges [Rada 89]. The information theory-based method defines the similarity of two concepts in terms of the maximum information content of the concept that subsumes them [Resn 95, Lin 98]. Recently new hybrid approaches have been proposed: Rodriguez and Egenhofer [Rodr 04] takes advantage from the above methods and adds the idea of features matching introduced by Tversky [Tver 77]. Schwering [Schw 05a] proposes a hybrid approach to assess similarity among concepts belonging to a semantic net. The similarity in this case is assessed by comparing properties of the concept as feature [Tver 77] or as geometric space [Gard 04]. With respect to the method presented in this thesis Rada et al. [Rada 89], Resnik [Resn 95] and Lin [Lin 98] work on lexicographic databases where the instances are not considered. If they are adopted as they were originally defined to evaluate the similarity of the instances, they are doomed to fail since they ignore important information provided by the instances, attributes and relations. Moreover, Rodriguez and Egenhofer [Rodr 04] and Schwering [Schw 05a] use the features or even conceptual spaces, information that is not native in the ontology design and would have to be manually added. Instead our approach aims at addressing the similarity, as much as possible, by taking advantage of the information that has already been disseminated in the ontology. Additional information is considered only to tune the similarity with respect to different application context.

Similarities that rely on ontology models with instances. Other works define similarity relying on ontology models closer to those adopted in the Semantic Web standards. On the one hand, Hau et al. [Hau 05] identify similar services measuring the similarity between their descriptions. To define a similarity measure on semantic services explicitly refers to the ontology model of OWL Lite and defines the similarity among OWL objects (classes as well as instances) in terms of the number of common RDF statements that characterize the objects. On the other hand, Maedche and Zacharias [Maed 02] adopt a semantic similarity measure to cluster ontology based metadata. The ontology model adopted in this similarity refers also to IS-A hierarchy, attributes, relations and instances. Even if these methods consider ontology models, which are more evolved than the taxonomy or terminological ontology, their design ignores the need to tailor the semantic similarity according to specific application contexts. Thus to assess the similarity investigated in this thesis, two distinct ontologies need to be defined instead of simply defining two contexts as I do.

Contextual-dependent similarity. Some studies combine the context and the similarity. Kashyap and Sheth [Kash 96] use the concept of semantic proximity and context to achieve the interoperability among different databases. The context represents the information useful for determining the semantic relationships between entities belonging to different databases. However they do not define a semantic similarity in the sense I am addressing and the similarity is classified as some discrete value (semantic equivalence, semantic relevance, semantic resemblance, etc). Rodriguez and Egenhofer [Rodr 04] integrate the contextual information into the similarity model. They define as the application domain the set of classes that are subject to the user's interest. As in our proposal, they aim at making the similarity assessment parametric with respect to the considered context. Moreover, in contrast with our methods they formalise the context rather then the similarity criteria induced by the context. This discussion of related works shows that, apart from the different definitions of semantic similarity proposed by different parties, these definitions are far from providing a complete framework as intended in our work. They often have different purposes, they consider a simpler ontology model, or they completely ignore the need to tailor the similarity assessment with respect to a specific application context. Of course, some of the works mentioned have been particularly important in the definition of our proposal. As already mentioned, both of the paper Maedche and Zacharias [Maed 02] and Rodriguez and Egenhofer [Rodr 04] have strongly inspired the part related to structural similarity. However, to successfully support our purposes the class slots have been considered as distinguishing features. Furthermore, the methods proposed by Maedche and Zacharias [Maed 02] for class matching defines a similarity that is symmetric, thus I have adapted the original in order to make it asymmetric. The similarity framework proposed in this thesis contributes, along with related work, toward paving the way to a tool that each ontology engineer can adopt

- to define different similarities among instances on the same ontology according to different application contexts;
- to refine the similarity criteria as long as new instances are inserted or the obtained

result does not satisfy the user domain expert.

The explicit parameterization of the similarity assessment with respect to the application contexts yields a precise definition of the hints to be considered in similarity assessment as well as complete control of the recursive comparison needed to work out the similarity.

4.3.8 Results

This thesis proposes a methods for assessing semantic similarity among instances within an ontology. It combines and extends different existing similarity methods, taking into account, as much as possible, the hints encoded in the ontology and considering the application context. A formalization of the criteria induced by the application is provided as a means of parameterizing the similarity assessment and to formulate a measurement more sensitive to the specific application needs. The framework is expected to bring great benefit in the analysis of the ontology driven metadata repository. It provides a flexible solution for tailoring the similarity assessments according to the different applications: the same ontology can be employed in different similarity assessments simply by defining distinct criteria, and it is not necessary to build a different ontology for each similarity assessment. The formalization of the application contexts in terms of explicit similarity criteria paves the way to an iterative and interactive process where the ontology engineer and the domain experts can perform fine-tuning of the resulting similarity. Nevertheless, some research and development issues are still open: the formalization of the application context affects only the similarity defined by the extensional comparison. It would be interesting to determine if the context results also in external comparison similarity. The ontology model considered is quite representative of how ontology are intended in many application but it do not cover all the expressiveness provided by the Web Ontology Language (OWL). It would be worth to extend the similarity to ontology models closer to OWL. Moreover, complex use cases should be considered in order to test and evaluate it. New use cases and the related community of end user are indispensable to perform a more careful human evaluation of the similarity obtained and stands out the impact of this method.

Related Publications

- Albertoni R., De Martino M. Semantic Similarity of Ontology Instances Tailored on the Application Context. In: 5th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2006) (Montpellier (France), 31 October - 2 November 2006). Proceedings, pp. 1020-1038. R. Meersman, Z. Tari et al. (eds.). (Lecture Note in Computer Science, vol. 4275). Springer, 2006.

- Albertoni R., De Martino M. Semantic Similarity Of Ontology Instances Tailored on the Application Context. Poster and Demo proceedings of the 15th International Conference

On Knowledge Engineering and Knowledge Management (EKAW 2006), (Podebrady, Czech Republic, 2-6 October 2006), ISBN 80-86742-15-6.

4.4 Semantic Granularity to browse information resources (SG)

In this section I describe the third semantic method to analyze the information resources' metadata. Analogously to the context dependent similarity, it stems from the experience made within the AIM@SHAPE NoE [AIM]. It represents a first result demonstrating how ontologies can be adopted to define semantic granularities. Granularities allow exploring data according to different levels of detail, enhancing the flexibility in the information representation and retrieval. Moreover, ontologies are fated to play a central role in the search for information resources. However, to build them is a costly and time consuming activity. Thus I consider a complete exploitation of such a precious artifact as a primary research issue. This work can be conceived as a step in this direction: ontologies representing *qualities* which describe information *resources* and the relations among them provide the starting point from which to build the granularity system. The ontology plays both the role of metadata schema and domain conceptualization (see section 4.1.2). Actually, in this method, the ontology formalizes the schema of metadata and the domains of at least one of the features, i.e. quality, represented in the schema. The domain of the quality is assumed to be made of categorical values hierarchically defined. The method proposed follows a two-phase process. In the first phase, namely *quality filtering*, each quality is evaluated with respect to its capability of abstracting sources. This evaluation is given taking into account both the relations in which the quality is involved, and the sources defined for it. Then, the qualities which provide a satisfactory facility to abstract sources become *granules* of some granularity. The second phase, namely *granularity building*, distributes the granules returned by the filtering phase among different granularities. This phase returns the set of granularities to employ for the repository navigation. The contribution of the method is twofold. First, I propose semantic granularities as a mean to browse the repository of information resources at different levels of detail. Secondly, I provide an approach to automatically define semantic granularities extending to ontologies and repository navigation previous research results presented for topics identification [Lin 95] and automatic discourse structuring [Pike 03].

In the following, I consider browsing a repository of papers to exemplify the proposed methodology and its objectives. Scientific papers are considered as the sources to be navigated, whereas their topics are the quality of interest. The section is organized as follows. First, I provide an overview of the motivations and related work to semantic granularities section 4.4.1. Then, in section 4.4.2 I formalize semantic granularities. In section 4.4.3 and 4.4.4 I discuss how semantic granules and granularities are built, describing quality filtering and granularity building phases. In section 4.4.5 I provide an example of repository navigation. Finally, section 4.4.6 concludes the part of the thesis pertaining to the methods, outlining future research directions.

4.4.1 Motivations

As already discussed in section 2.2, the difficulties pertaining to information access are mainly due to a poor overlap between the information model employed by the seeker (i.e., the Cognitive Space) and the model defined by the information provider (i.e., the Information Space) [Newb 01]. Techniques of semantic searching as those proposed in [Guha 03] can be useful to improve the retrieval mechanism. However, in order to increase the aforementioned overlapping, every search activity has to be characterized by a highly interactive process [Belk 82a]: the seeker refines the selection criteria according to the results he/she obtains alternating querying and browsing activities.

Automatic techniques such as clustering and classification can be employed to organize repositories and to ease their browsing [Cutt 92, Albe 03b, Baru 06]. The clustering only relies on the model emerging from the Information Space, whereas the classification relies on a set of classes which are expected to be meaningful as belonging to the user's Cognitive Space. As a consequence, the first might result in non meaningful grouping whereas the second might result in empty classes.

In order to overcome these drawbacks the semantic granularity takes into account both the spaces. It considers parts of the Cognitive Space represented in the ontology as well as the Information Space by balancing the sources at a given granularity according to their occurrence.

In the area of Information Systems, granularities have been already studied for both the temporal and the spatial domains [Bett 00, Stel 98, Camo 06]. However, in this field, granularities are static and embedded in the data model or in the database schema. The recent research has focused mainly on cognitive issues pertaining to the perception of vagueness, indeterminacy, imperfection, roughness, etc. [Bitt 03]. Moreover, some attempts to define semantic granularities have been made with respect to terminologies by Fonseca et al. [Fons 02]. They introduced the term semantic granularity exploiting the casting mechanism employed in the object-oriented paradigm to represent ontology instances at different levels of detail.

In this work semantic granularities are dynamically defined, according to the data model (represented by an ontology schema) as well as to data (represented as ontology instances).

4.4.2 The approach of semantic granularities

Semantic granularities aim at structuring a repository at different levels of detail taking into account both its conceptual structure and its content. Assuming the following entities are available:

S, the set of sources subject of the user exploration;

Q, a quality represented as set of nominal values according to which the sources are organized;

O, an ontology providing a data schema, where the sources and the quality are represented in terms of ontology entities;

the framework generates a sequence of granularities $\mathbf{G} = < G_1, G_2, ..., G_n >$, such that each granularity G_i groups the sources the repository contains at increasing levels of detail, i.e., G_{i+1} gives a finer view on the sources in $\mathbf{S}$ than G_i (G_{i+1} is said to be finer than G_i). A granularity G_i is defined as a set of granules which provide a discrete view of the quality. A generic granule in G_i is denoted by g_j^{Gi}, and is described by a unique textual label $l_j^{Gi} \in \mathbf{Q}$. The granule labels are expected to be semantically meaningful for the user.
Let g_j^{Gi} be a granule at granularity Gi, the sources in S defined for g_j^{Gi} are denoted with s_j^{Gi}, whereas the sources defined for all the granules of granularity G_i are denoted with s^{Gi}. Let's consider for example a granularity G_i identifying scientific fields, such that the labels of its granules are l_1^{Gi}= "COMPUTER SCIENCE", l_2^{Gi}= "MATHEMATICS", l_3^{Gi}= "PHILOSOPHY", etc., and a repository $\mathbf{S}$ of scientific papers classified with respect to their topics. Then, s_1^{Gi} denotes the set of papers on Computer Science, s_2^{Gi} the set of papers on Mathematics, and so on. Given two granules of the same granularity G_i I do not require that the corresponding sets of sources are disjoint. For instance, $s_1^{Gi} \cap s_2^{Gi}$ gives the set of papers on topics shared by Mathematics and Computer Science.
The approach I propose assumes the repository is organized according to an ontology which recalls the pattern depicted in figure 4.11. In figure 4.11, IO is the relation Instance-Of; Is-A and Part-Whole are partial order relations; entities with capital initials are classes; $s_1, s_2, s_3, q_1, q_2, .., q_n$ are instances; F_1, F_2, A_1, A_2 are other class properties, which might be employed to further characterize the sources and the qualities. In particular, I assume the

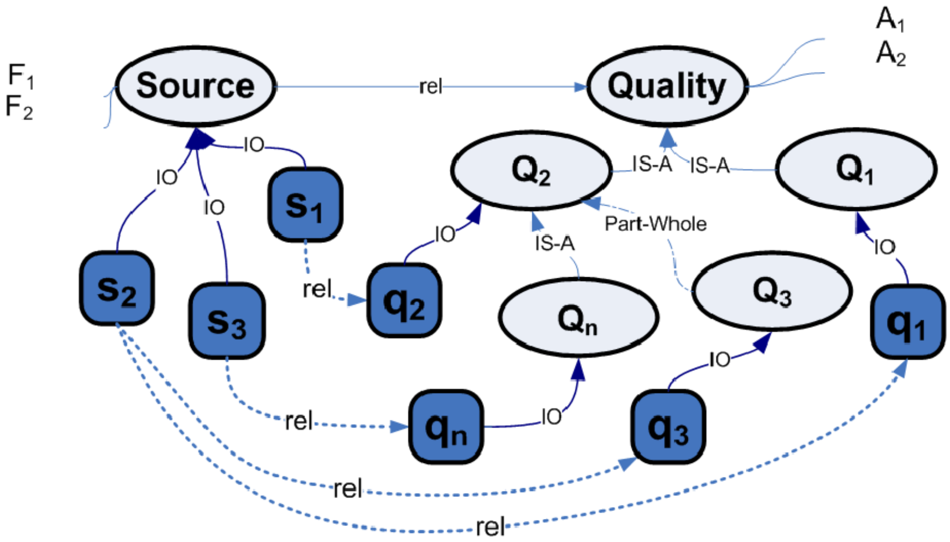


Figure 4.11: *The Ontology reference schema*

existence of a class grouping all the sources in **S** the user is going to browse (**Source** in figure 4.11); a class representing the quality **Q** with respect to structuring the repository (**Quality** in 4.11); and a relation, which joins the sources s_1, s_2, s_3, etc. to the instances $q_1, q_2, .., q_n$ representing the quality values (**rel** in figure 4.11). Given for instance the granularity G_i representing scientific fields I defined above, and the set S of scientific papers on those fields, the papers in S are sources to browse, which will be reorganized according to a quality topics (i.e., the granules in G_i). The quality **Topic** is defined for the source **Paper** through the relation **hasTopic**.

Futhermore, I assume a hierarchy H^Q is induced by the relations Is-A and Part-Whole relating the qualities $Q_1, Q_2, ..., Q_n$ in figure 4.11. Part-Whole and similar relations among parts and wholes have been widely investigated by the scientific literatures (e.g., [Wins 87, Varz 96, Arta 96]). The literature demonstrates that distinct parthood relations can be identified depending on how the parts differently contribute to the structure of the whole. A complete treatment of the issues concerning parts and wholes is beyond the scope of this work, thus I restrict the possible interpretations of the relation Part-Whole assuming the parthood among the qualities in **Q** adheres to the following properties defined in [Wins 87]:

transitivity, i.e., parts of parts are parts of the whole;

reflexivity, i.e., every part is part of itself;

antisymmetricy, i.e., nothing is a part of its parts;

homeomericity, i.e., parts are of the same kind of things as their wholes.

The first three properties induce the partial order needed to preserve the hierarchical structure of the qualities. The fourth property instead ensures the parts in the hierarchy are still a quality as their whole (e.g., a topic can be part of another topic).

I observe that the entities in figure 4.11 can be represented differently according to the ontology design choices. The solution adopted mainly depends on the expressiveness of the ontology language employed and the needs pertaining to the reasoning. For simplicity, the paper assumes classes in the hierarchy can not be used as relation values, i.e., the relations have only instance values. Similar assumptions have been made also by OWL-DL[9], one of the ontology language most adopted. As consequence of the latter assumption, the qualities are represented both as classes and instances, and each class Q_n has exactly one instance q_n. I introduce these restrictions to simplify the presentation of our method, but the approach could be easily adapted to different ontology designs.

Given the ontology schema in figure 4.11, the qualities in **Q** are related by a partial order $\leq_Q$ induced by H^Q according to both Is-A and Part-Whole. Since the granule labels correspond to qualities in **Q**, $\leq_Q$ is defined also on labels. Note that not all the values in **Q** become granule labels for some G_i in **G**, but only those resulting from the phase of quality filtering

[9] http://www.w3.org/2004/OWL/(Accessed in July 2006)

described in section 4.4.3. Given a quality Q in **Q**, the set of sources s^Q are the instances of S associated via rel to Q, while the set of sources s^{Q*} are the instances of **S** associated via rel to Q and to each quality Q' such that $Q' \leq_Q Q$. For example, considering the quality Q_2 in figure 4.11, s^{Q2} corresponds to $\{s_1\}$, whereas s^{Q_2*} corresponds to the set of instances $\{s_1, s_2, s_3\}$. Given the granule g_j^{Gi} at granularity G_i, such that $Q = l_j^{Gi}$, note that s_j^{Gi} is equivalent to s^{Q*}. Let G_1 and G_2 be two granularities belonging to G, such that G_2 is finer than G_1. Given g_j^{G1} a valid granule of G_1, I denote with $G_2(g_j^{G1})$ the set of granules of G_2 which labels are related to l_j^{G1} through $\leq_Q$, i.e., $\{g_k^{G2} | l_k^{G2} \leq_Q l_j^{G1}\}$. Analogously, given g_k^{G2} a valid granule of G_2, I denote with $G_1(g_k^{G2})$ the set of granules of G_1 $\{g_j^{G1} | l_k^{G2} \leq_Q l_j^{G1}\}$. Through this notation, I can move from a granularity to a different one. For instance, let's consider G_i the granularity for scientific fields defined above, g_1^{Gi} the granule for Computer Science, and G_{i+1} a granularity finer than G_i; $G_{i+1}(g_1^{Gi})$ results in all the granules representing the sub-fields in which Computer Science can be expanded into (e.g., Database, Artificial Intelligence, Computer Graphics, etc.).

In the following sections, I will detail how granularities to browse the sources are built starting from information sources defined with respect to a set of qualities. Firstly, the phase *quality filtering* (described in section 4.4.3) evaluates which qualities can be employed as granule labels. Then, in the phase *granularity building* (described in section 4.4.4), granules representing qualities are assigned to different granularities.

4.4.3 Quality filtering: selection of semantic granules

The filtering selects the quality values to be adopted as granule labels. In the proposed approach, a quality value is considered as a granule label whenever it ensures a good level of abstraction or it is involved in Part-Whole. The idea I adopt to evaluate the abstraction capability of qualities is borrowed from existing techniques applied in the area of Natural Language Processing for topic identification and generalization [Lin 95, Pike 03]. Lin [Lin 95] introduced the notion of degree of informativeness and summarization of a concept C in a lexical taxonomy as a measure of the capability of C to generalize its specializations, i.e., the children in the taxonomy, according to the terms occurrence in a corpus. According to Lin, the more the children of C have a similar number of occurrences, the more the concept C is a good generalization. The Lin algorithm is based on the assumption that the occurrences of the corpus are associated only to the leaves in the noun taxonomy.

Pike and Gahegan [Pike 03] extend the Lin's approach to identify and to abstract arguments of a discourse allowing the intermediate concepts of the taxonomy to have their own occurrences associated. Both works consider only the relation Is-A for structuring concepts, and ignore the occurrence of the concept C in the corpus for the evaluation of its degree of informativeness.

Given the schema of figure 4.11, to evaluate the degree of informativeness of a quality Q aiming at the definition of *semantic granularities*, I consider each source related through rel to Q as an occurrence of Q. I extend the method proposed by Pike and Gahegan to Part-Whole structures, taking into account also the influence of the occurrences of each quality under evaluation. Consider for instance the situation in figure 4.12, where I report a portion of a possible classification of the topics pertaining to the Data Mining research field. The values reported have been retrieved by querying the ACM digital library and considering qualities as paper keywords.

In figure 4.12, Clustering has four unbalanced sub-topics, i.e., Document Clustering, K-Nearest Neighbors Based, K-Mean, and Hierarchical Clustering. According to both Lin-Pike's approach, Clustering does not provide a high level of generalization of its sub-topics. However, I observe that for Clustering a considerable amount of instances has been retrieved (14923), and this value is much bigger than the values reported for its sub-topics. In this situation, I would expect that this concept is eligible to be included among the most meaningful topics the repository refers to. Given the aforementioned observations, I say that

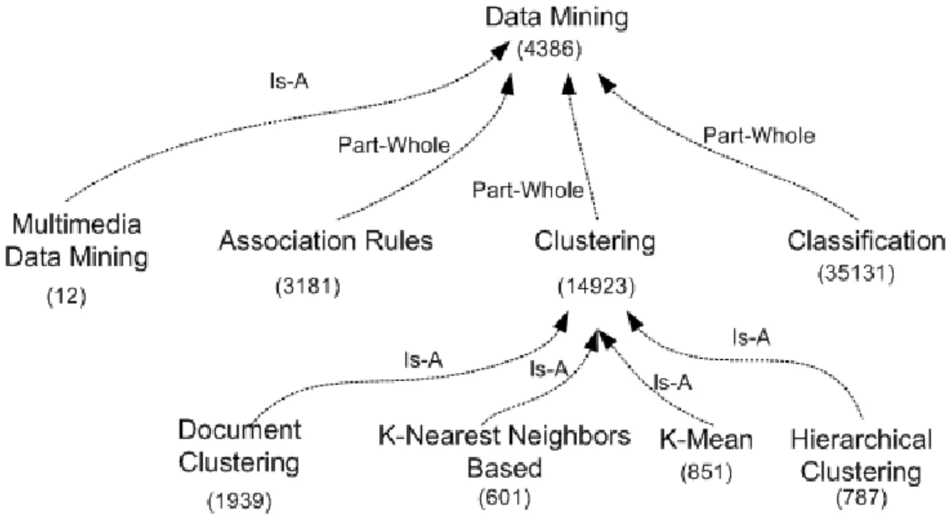


Figure 4.12: *Data Mining topics classification*

a quality Q is a *good abstraction* for its direct sub-qualities (including, as I specify above, both the qualities reachable through Is-A and Part-Whole) if the ratio R_Q between the maximum numbers of occurrences in the repository defined for its sub-qualities and the recursive number of occurrences defined for Q in the repository (i.e., including both its own occurrences and the recursive occurrence of its immediate sub-qualities) is less then a given a threshold R_t. R_Q is defined in equation 4.19 and its value ranges in [0,1]. Leafs in the hierarchy have R_Q equal to 0. I denote with Q the non-reflexive and non-transitive relation induced by H^Q (e.g., referring to figure 4.12, Multimedia Data Mining $\prec_Q$ Data Mining, Association Rules $\prec_Q$ Data Mining, K-mean $\prec_Q$ Clustering). R_Q is defined as follows:

$$R_Q = \frac{max_{\{Q'|Q' \prec_Q Q\}}|s^{Q'*}|}{\sum_{\{Q'|Q' \prec_Q Q\}}|s^{Q'*}| + s^Q} \tag{4.19}$$

Let us consider in figure 4.12 the topics Association Rule, Classification and Clustering, which are related to Data Mining through the relation Part-Whole. Conversely to Is-A, I consider Part-Whole as a landmark for granule identification, because it intrinsically discriminates two separate levels of abstraction. Thus whenever two qualities are directly related by Part-Whole they are considered as granules of distinct granularities. In the next section, I adopt the predicate **isGranule(Q,R_t)** to determine if the quality Q is promoted to be a granule according to the quality filtering phase. The predicate **isGranule(Q,R_t)** is true whenever Q is involved in a Part-Whole, or it is a good abstraction, i.e. $R_Q \leq R_t$ where R_Q is defined according to equation 4.19.

4.4.4 Granularity building: distribution of granules among granularities

The granularity building phase aims at determining the distribution of the granules resulting from quality filtering among the semantic granularities in $\mathbf{G}$. The partial order $\leq_Q$ induced by the relations Is-A and Part-Whole leads such a distribution. In particular, considering two distinct granule labels a and b, the principles I follow are:

(1) if $b \leq_Q a$, then the two granules have to belong to distinct granularities,

(2) if $b \leq_Q a$, b Part-Whole a holds and the whole granule with label a belongs to the granularity G_i, then the part granule with label b has to belong to the granularity G_{i+1}, such that G_{i+1} is finer than G_i.

Given for instance the hierarchy of figure 4.12, if Data Mining, Clustering and Document Clustering satisfy the predicate isGranule for a given value of R_t, and Data Mining belongs to granularity G_i, Clustering must belong to granularity G_{i+1}, while Document Clustering must belong to a granularity G_j, with $j > i+1$. The granularity building phase is performed according to the algorithm in figure 4.4.4 The algorithm returns the sequences of granularities $\mathbf{G}$. It performs a breath first visit of H^Q, inserting the granules in distinct granularities according to the principles stated in (1) (2). It terminates whenever the visit has reached all the H^Q leaves.

In figure 4.4.4, **ds** is the starting level (from the root) in H^Q; R_t is the ratio threshold for the evaluation of the degree of informativeness of qualities; **next(Q)** returns the child of Q in H^Q; **node(l)** returns the qualities laying at the level l in the hierarchy; **+** and **-** are the set operators for union and difference.

4.4.5 Example

In this section I refer to the repository of scientific papers provided by the ACM Digital Library, considering papers as the information sources to be browsed and the associated keywords as sources qualities. In the ACM Digital Library each paper is classified according

```
i = 0; Gi  = {};
NodeToConsider = Node(ds);
While (! empty(NodeToConsider)){
  For each Q belonging to NodeToConsider {
    If  isGranule(Q, Rt) Gi   += { Q };
    else If Q is not leaf NodeToConsider += next(Q);
    NodeToConsider -= { Q };
  }
  i++; Gi  = {};
  For each Q belonging to Gi-1 {
    For j=1 to i-1 {  // check for multi-inheritance
      For each Q1 belonging to Gj
        If ((Q1 part-whole Q) or (Q1 is-a Q)){
   Gj -= { Q1 }; Gi += { Q1 };
 }
      If Q is not leaf
           NodeToConsider += next(Q);
    }
  }
}
```

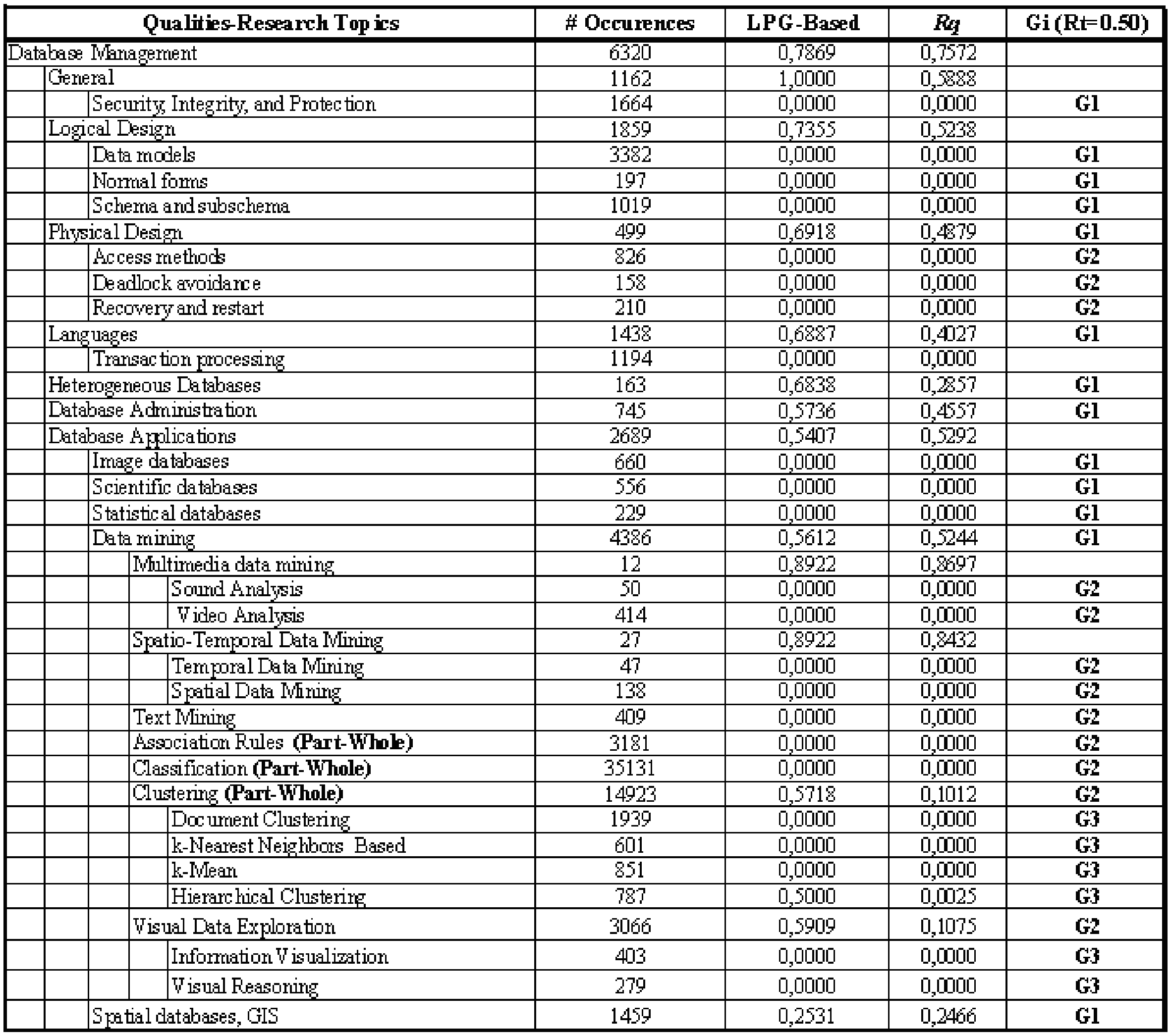
Figure 4.13: *Granularity building algorithm*

to the ACM Computing Classification System, which is a taxonomy depicting the broadest research fields in computer science. To provide an example of semantic granularities extracted through its application, I extend the ACM taxonomy deepening the fields I am familiar with. A portion of this taxonomy has been already shown in figure 4.12. The classes in the resulting taxonomy are mainly related through the relation Is-A. Moreover, the relation Part-Whole has been adopted whenever a research field is classified with respect to its important sub-parts, like techniques applied in the field (e.g., Clustering and Data Mining in figure 4.12).

In Table 4.4 a meaningful portion of the quality hierarchy I consider and the results obtained are shown. The first column of the table reports the quality values (the indentation resembles the H^Q structure); the second gives the number of papers referring to each quality in the ACM repository; the third column reports the degree of informativeness obtained by applying an approach based on what proposed in [Lin 95, Pike 03], while the fourth reports the values obtain according to the evaluation presented in section 4.4.3; finally, the last column gives the granules distribution according to the phase granularity building described in section 4.4.4.

The threshold R_t applied in this example is 0.50. With this evaluation of R_t the qualities Database Management, General, Logical Design, Transaction processing, Database application, Multimedia Data Mining and Spatio-Temporal Data Mining are not considered as granules because their degree of informativeness is greater than the threshold. Note that the granularities preserve the partial order given by the hierarchy among qualities related by Is-A and Part-Whole, but not the general order as stated by the hierarchy. In particular, qualities at the same level in the hierarchy can be labels for granules that belong to different granularities.

Let's suppose a user wants to navigate the sources in the repository. At the first step, the labels of the granules at granularity G_1 give him/her a broad idea of the most meaningful arguments represented in the repository (l_1^{G1} is the label for the quality "Security, Integrity and Protection", l_2^{G1} is the label for "Data Models", etc). Let's suppose the user chooses to navigate the resources related to Data Mining. The labels of granules at granularity G_2 such that $G_2(g_{12}^{G1})$, where $g_1^{G1}2$ is the granule with label Data Mining, are retrieved. As formally defined in section 4.4.2, the conversion is based on the partial order induced by the hierarchy H^Q. Thus, the set of labels l_5^{G2}= "Sound Analysis", l_6^{G2} = "Video Analysis", l_7^{G2} = "Temporal Data Mining", l_8^{G2} = "Spatial Data Mining", l_9^{G2} = "Text Mining", l_{10}^{G2} = "Association Rules", l_{11}^{G2} = "Classification", "l_{12}^{G2} = "Clustering", and l_{13}^{G2} = "Visual Data Exploration" is retrieved. The same happens for all the levels for which a granularity has been built, according to the algorithm I described in section 4.4.4. Once the user chooses the set of instances is interested in by the browsing of labels, the corresponding set of sources is retrieved. Let the user be interested in Document Clustering, represented by the granule g_1^{G3}. Then, the set of sources s_1^{G3} is returned to be processed by applying, for instance, existing navigation techniques (e.g., see the Information Visualization tools surveyed in [Albe 05a]).



Qualities-Research Topics	# Occurences	LPG-Based	*Rq*	Gi (Rt=0.50)
Database Management	6320	0,7869	0,7572	
General	1162	1,0000	0,5888	
Security, Integrity, and Protection	1664	0,0000	0,0000	**G1**
Logical Design	1859	0,7355	0,5238	
Data models	3382	0,0000	0,0000	**G1**
Normal forms	197	0,0000	0,0000	**G1**
Schema and subschema	1019	0,0000	0,0000	**G1**
Physical Design	499	0,6918	0,4879	**G1**
Access methods	826	0,0000	0,0000	**G2**
Deadlock avoidance	158	0,0000	0,0000	**G2**
Recovery and restart	210	0,0000	0,0000	**G2**
Languages	1438	0,6887	0,4027	**G1**
Transaction processing	1194	0,0000	0,0000	
Heterogeneous Databases	163	0,6838	0,2857	**G1**
Database Administration	745	0,5736	0,4557	**G1**
Database Applications	2689	0,5407	0,5292	
Image databases	660	0,0000	0,0000	**G1**
Scientific databases	556	0,0000	0,0000	**G1**
Statistical databases	229	0,0000	0,0000	**G1**
Data mining	4386	0,5612	0,5244	**G1**
Multimedia data mining	12	0,8922	0,8697	
Sound Analysis	50	0,0000	0,0000	**G2**
Video Analysis	414	0,0000	0,0000	**G2**
Spatio-Temporal Data Mining	27	0,8922	0,8432	
Temporal Data Mining	47	0,0000	0,0000	**G2**
Spatial Data Mining	138	0,0000	0,0000	**G2**
Text Mining	409	0,0000	0,0000	**G2**
Association Rules **(Part-Whole)**	3181	0,0000	0,0000	**G2**
Classification **(Part-Whole)**	35131	0,0000	0,0000	**G2**
Clustering **(Part-Whole)**	14923	0,5718	0,1012	**G2**
Document Clustering	1939	0,0000	0,0000	**G3**
k-Nearest Neighbors Based	601	0,0000	0,0000	**G3**
k-Mean	851	0,0000	0,0000	**G3**
Hierarchical Clustering	787	0,5000	0,0025	**G3**
Visual Data Exploration	3066	0,5909	0,1075	**G2**
Information Visualization	403	0,0000	0,0000	**G3**
Visual Reasoning	279	0,0000	0,0000	**G3**
Spatial databases, GIS	1459	0,2531	0,2466	**G1**

Table 4.4: *Method evaluation with* $R_t = 0.50$

4.4.6 Results

In section 4.4 I have presented a framework for the dynamic definition of semantic granularities aiming at the effective representation of information resources repository at different levels of detail. The method is inspired by existing work on topic identification for discourse structuring. With respect to existing methods in the literature, I extend them to the browsing of any kind of information sources described with respect to a set of qualities represented in ontologies. I encompass the generalization performed according to a taxonomy, dealing also with qualities related by Part-Whole, toward a full ontology support. Furthermore, I provide a definition of semantic granularity which is dynamic, providing a bridge among the conceptual model of the user (i.e., the Cognitive Space) and the model structuring the repository (i.e., the Information Space).
Semantic granularities provide a way of supporting the user in the repository browsing at different levels of detail. For each granularity only the meaningful granule labels for the specific repository are represented. Thus, the sources having qualities the user is not interested in can be discarded since the very first steps of the browsing. Semantic granularity is defined with respect to ontologies, aiming at fully exploiting the information this formal conceptualization can provide. A crucial extension is related to the inclusion of properties of ontology classes and their values in the evaluation of the degree of informativeness of qualities. I am also planning an experimental evaluation performed on multimedia sources (text, shape, etc.). Finally, I am going to integrate the method with traditional techniques of resource browsing (e.g., Information Visualization).

Related Publications

- Albertoni R., Camossi E., De Martino M., Giannini F., Monti M. Semantic Granularity for the Semantic Web. In: Second IFIP WG 2.12 & WG 12.4 International Workshop on Web Semantics (SWWS'06) in conjunction with OnTheMove Federated Conferences (OTM'06) (Montpellier (France), 1-2 November 2006). Proceedings, pp. 1863-1872. R. Meersman, Z. Tari, P. Herrero et al. (eds.). (Lecture Note in Computer Science, vol. 4278). Springer, 2006.

4.5 Contributions with respect to the requirement for metadata analysis

In this section I summarize the contribution provided by the semantic methods described in this chapter with respect to the requirements posed in section 2.3. The support they provide is summarized in Table 4.5.

The *query formulation* is supported by all the three semantic methods proposed. In fact, all exploit the background knowledge enhancing the expressiveness of the query language em-

Requirement	ASSCV	ACDSS	SG
Query Formulation	good	good	partial
Retrieved result comprehension	none	good	good
Query coordination	partial	partial	partial
Database problem	none	none	none
Vocabulary problem	good	none	none

Table 4.5: *How the semantic metadata analysis supports the analysis requirements.*

ployed. In particular, supposing a first interesting resource have been selected, the similarity among information resources (section 4.3) makes possible to ask for similar items. The asymmetry has been designed to emphasizes the containment between resources thus it is easer to find out if similar items are more or equally interesting. The semantic similarity among categorical values (section 4.2) adds the possibility to query for similar terms, whereas, the semantic granularity (section 4.4) enables to query resources at different levels of abstraction.

The problems of *retrieved result comprehension* is faced with the similarity (section 4.3) among resources. The definition of multiple contexts which differently explore the same resources provides a valuable tool to understand the obtained results. On the other hand, the semantic granularity might be used to properly sifting resources: it eases the *retrieved result comprehension* reorganizing the result in a way more meaningful for the seeker.

The *vocabulary problem* is resolved by the semantic similarity among categorical values (see section 4.2.4). Of course, the proposed method works if the distinct terms can be related according to someone's background knowledge. If the providers and the seeker adopt totally unrelated terms there is no way in my opinion to solve this problem.

The *query coordination* problem is eased by the proposed semantic methods because they can remind the seekers about relations useful to build the query. Often, the seekers are so focused on the search that they forget what is trivial. Moreover, the proposed semantic methods elaborate user's background knowledge. This elaboration can bring to unexpected knowledge that can induce new selection criteria.

Finally, the *databases* problem is not faced with the proposed semantic methods. As already discussed in the visual metadata analysis [Spoe 04d, SPOE 04b] brilliantly face with this problem. Moreover, some additional support could be obtained adopting techniques of ontologies and instances alignment [Euze 04a, Euze 04b].

4.6 Conclusions

The chapter has introduced the semantic analysis of metadata describing information resources. It shows as it is possible to support the search for information resources taking advantage from the background knowledge formalized in ontologies. Three methods have been described:

- Asymmetric Semantic Similarity among metadata Categorical Values (ASSCV);
- Asymmetric and Context Dependent Semantic Similarity among information resources (ACDSS);
- Semantic Granularity among information resources (SG).

The first method has been inspired by the research experience made within INVISIP. It demonstrates the use of background knowledge and similarity to ease the *vocabulary* problem. By contrast, the last two methods stem from the AIM@SHAPE research activity. The asymmetric and context dependent similarity among information resources enables the browsing of information resources according to different context dependent views. Whereas the semantic granularity is the first step toward the definition of granularities from an ontology to browse a repository of information resources at different levels of abstraction. Both support the user in easing *query formulation* and *retrieved result comprehension.* AIM@SHAPE aimed at defining ontologies to characterize the multidimensional information resources (e.g. Computer aided model, 3D mesh and so on). The definition of ontologies describing so complex information resources is fundamental to properly retrieve them. However, it takes months of painstaking work interacting with the experts in the field of Computer Graphics. Thus, both context dependent similarity and granularity have as ultimate goals to exploit as much as possible the costly artifacts like ontologies defined in AIM@SHAPE.
The methods proposed in this chapter have been demonstrated considering resources belonging to geographical domain or pertaining to the academical world. An experimentation on multidimensional information resources will be worked out once the ontologies developed within AIM@SHAPE are fully finalized and populated by real instances.

4.6.1 Remarks and research issues

The overall activity on semantic metadata analysis has pointed out some interesting remarks:

- Ontologies are useful to represent user background knowledge: they can be adopted to organize metadata as well as to define the domain of categorical metadata attributes.
- Ever-expressive ontology language are emerging from the research pertaining to ontology. However, the methods which takes advantage from ontologies should consider the context to be effectively tailored on the specific analysis purposes.

These remarks have brought to plan the following future activities:

- How to define a context which parameterizes the semantic granularity with respect the specific application purpose?
- How to extend the proposed methods to consider more expressive ontology languages?

The semantic similarity among categorical values and information resources already considers some kind of context. Then it would be worth to see which concept of context can be adopted for the semantic granularity.

Concerning the extension toward more expressive ontology languages, it would be worth to consider axioms which can be provided by description logics. In particular, the transitivity, symmetry and cardinality constraints pertaining to relations might be considered to cover a greater extent of OWL.

Chapter 5

Visual vs Semantic Metadata Analysis

This chapter summarizes the contribution of the visual and semantic metadata analysis to the resolution of the search problems illustrated in section 2.3 and it critically discusses some limitations of my research activity.

5.1 Overall contribution to the metadata analysis requirements

The contribution of the visual and semantic metadata analysis to Spoerri's problems is summarized in table 5.1 on the basis of the discussions already illustrated in the sections 3.4 and 4.5. Table 5.1 points out that the proposed metadata analysis succeed in the problems of *query formulation* and *retrieved result comprehension*. It shows that *vocabulary problem* is eased more by the semantic analysis than the visual analysis. This is not surprising if we consider that the vocabulary problem has an intrinsic relation with the interpretations that humans give to the terms. Moreover, the table indicates that the proposed metadata analysis

Requirement	Visual Analysis	Semantic Analysis
Query Formulation	good	good
Retrieved result comprehension	good	good
Query coordination	partial (good)	partial
Database problem	none (good)	none
Vocabulary problem	partial	good

Table 5.1: *It shows the contribution of visual and semantic metadata analysis to meet the search problems. Between round brackets has been indicated the support which would be obtained including the MetaCrystal [Spoe 04d, SPOE 04b] in the proposed metadata analysis.*

partially faces with the *query coordination* problem and it does not support at all in the *database* problem.

The lack of support for the last two problems deserves a more detailed discussion. The metadata analysis provides a good support to a part of the *query coordination* problem: it relies on visual exploration and on the exploitation of the implicit and the explicit semantics to discover the most suitable selection criteria. However, it does not provide methods to compare the results from different queries. For this reason, the overall support has been rated as partial. The *database* problem neither has been faced in INVISIP nor in AIM@SHAPE because in both the projects dedicated repositories have been built to contain the information resources to be selected. Beside that, I want to stress these limitations are not so relevant as they could seem at first sight. The techniques developed in MetaCrystal [Spoe 04d, SPOE 04b] by Spoerri already succeed in the comparison of the results obtained by different queries as well as different repositories. Thus, it is sufficient to integrate the Spoerri's techniques into the proposed conceptual metadata analysis framework to overcome these limitations. The support provided with the integration is illustrated in the table 5.1 between round brackets.

5.2 Advantages in integrating Visual and Semantic Analysis

The visual and semantic approaches are conceived starting from the results which have been obtained in different fields of computer science. Of course in principle they could be adopted separately. But in this thesis, I have stressed the importance of adopting both contemporaneously. The motivations behind this choice are two: the intrinsic nature of metadata and the need to involve as much as possible the seeker in the selection process.

Concerning the intrinsic nature of metadata it is important to bear in mind that, unlike the usual dataset, the most of information encoded in metadata is non-numerical and it is thought to be human-interpretable rather than machine processable. Such a characteristic makes extremely challenging to visualize the metadata items. The most of interaction and visualization techniques relies on algebraic properties among data in order to properly represent them. Unfortunately, these algebraic properties are built-in for the numbers, but they are not natively available for other kinds of datatype such as categorical data. Thus, to overcome this drawback the semantic methods are proposed as pre/post processing visualization techniques. In particular, the thesis proposes methods to evaluate the similarity and the granularity among information resources. The similarities intuitively indicates how to represent the metadata items “close” in a visualization. Whereas the granularity provides different levels of abstraction which can be employed to implement a kind of semantic zooming among information resources.

On the other hand, if semantic metadata analysis was employed without any visual analysis, both the sufficient involvement of the seeker and the implicit semantic exploitation would not be obtained. As a consequence, the user would not be supported in the discovery of the selection criteria and part of the metadata analysis requirements would not be fulfilled.

5.3 Limitations

Beside all the contributions some issues might be pointed out as limitations in this thesis. The thesis assumes that both metadata and ontologies are available to characterize the information resources. A first objection to my work can be that this is not always true. However, this assumption in my opinion is not unreasonable: a huge effort to automatically extract the metadata and automatically defines ontologies are currently carried out from the research community. I am aware that the above activities will come up with semi-automatic methods and they will be restricted to specific application domains. However, I am also convinced that in many domains, providers appropriately motivated (e.g., economically) are quite willing to compile the metadata. For example, in the geographical domain, public administrations as well as private providers have started to acknowledge the importance to supply metadata together with the information resources.

An other aspect which results touchy is how to obtain a wider evaluation of the proposed methods. The thesis considers a human evaluation for the visual metadata analysis. Concerning the semantic analysis, it considers an evaluation of the asymmetric and context dependent semantic similarity based on recall and precision and it provides only some examples of how the other proposed methods work. The human evaluation performed within INVISIP demonstrates that the seekers perceive the metadata analysis as useful, the experimentation with precision and recall and the examples in specific domain show that the proposed methods work. However, the thesis does not quantify how much the metadata analysis improves the search for information resources. A more careful evaluation of this aspect would require an expansive and dedicated research activity. The definition of specific case studies to test objective advantages is tricky: it requires to collaborate with experienced partners who are able to limit the bias produced by misleading tests. Due to the lack of collaborations with such a kind of partners this activity could not be undertaken in this thesis. We are considering specific research projects to perform it in the future.

Conclusions

This thesis highlights the importance of defining new mechanisms for the analysis of metadata to supports the search and selection of information resources. Two different approaches dealing with the complexity of the metadata of large amount of heterogeneous information resources are proposed: visual and semantic metadata analysis. Both are required in order to support the seeker in the selection of information resources. The visual metadata analysis is based on the application of Information Visualization and Visual Data Mining techniques. In particular, Information Visualization is adopted to amplify the cognition of the seeker and to involve him more in the search activity, whereas the Visual Data Mining allows to exploit the implicit semantics, i.e., the patterns arising from the collection of resources, to discover unknown and novel selection criteria. The semantic metadata analysis is proposed as preprocessing methods in visual metadata analysis. It relies on explicit semantics to suggest the user some criteria of selection arising from the formalization of his background knowledge. It does not recall only criteria momentarily forgotten because of the painstaking research activity, but in addition, it suggests novel criteria coming from the re-elaboration of the background knowledge.

Contributions

The contributions of this thesis are manifold and range from foundational to applied research. The results obtained in the foundational research can be divided in advances in searching and selecting information resources and improvement in the exploitation of ontologies.

Advances in searching and selecting information resources. A first contribution consists in the identification of the issues pertaining to the search and selection of information resources and the analysis of their relevance. Actually, during the first stage of my working activity within European projects I have had the chance to identify new research issues interesting for the European research community and for the team at CNR-IMATI where I am working in. In particular, the recognition of metadata complexity as a research problem as well as the relevance of exploiting both implicit and explicit semantics stored in the metadata structure have been the first significant result. The main contribution of this thesis is the development of appropriate methods to support in the search and selection of information resources. In particular, it proposes different metadata analysis methods which, once integrated, meet the requirements discussed in section 2.3:

- it remarks the importance of taking advantage of the implicit and explicit semantics because, in the real world, both kinds of semantics might suggest useful selection criteria;
- it stresses the need to involve the seeker in the search activity: the seeker has to discover the selection criteria actively taking part in the search and selection process because the proper criteria seldom are known since the early stage of the search;

Improvement in the exploitation of ontologies. The thesis adopts ontologies in order to represent user's background knowledge and to exploit the information resources semantics. To this purpose, it introduces three methods:

- the Asymmetric Semantic Similarity among metadata Categorical Values, which is designed to face with the vocabulary problem;
- the Asymmetric and Context Dependent Semantic Similarity, which is proposed to compare the metadata of information resources and to highlight the "containment" between information resources;
- the Semantic Granularity, which is defined to browse repositories of information resources at different levels of abstraction and with respect to their categorical qualities.

The first method assembles an existing semantic similarity and an enhanced ontology visualization. Instead, the other two methods are new proposal which advance the current state of art in the field of ontologies and semantics. The proposed methods can be adopted within the visual metadata analysis. In this sense, they extend the state of art on pre-processing algorithms and indirectly contribute in the field of information visualization.

From the point of view of applied research the most part of the research presented in this thesis stems from the European projects INVISIP (IST-2000-29640) and AIM@SHAPE (IST NoE NO 506766). It confirms the relevance of the problems faced in the thesis at a European level or at least in the application domains of the two projects: the geographic domain and the field of multidimensional media. In particular, parts of the methods proposed have been demonstrated within the geographical domain. Due to the importance the search and selection of geographical information resources have at a European level, the success in this application domain can be considered as a first proof of the economical and social potential impact of the proposed metadata analysis.

Future work

The work described in this thesis can be extended in different directions. Both the visual and the semantic metadata analysis pave the way for new challenging research insights.

Visual metadata analysis. The visual metadata analysis presented in chapter 3 considers the research experience carried out within the INVISIP project. It is a starting point which demonstrates the importance to set up visual tools to search and select information resources. It proposes a set of visual techniques from the state of art but it does not care about which is the minimal subset of visual techniques that best faces with the search problems. It would be interesting to experiment further techniques apart from those I have considered identifying which is the best set that properly supports the search. Next steps in this direction includes to face with the temporal evolution [Andr 03] as well as the results emerging from the new research fields of Knowledge Visualization [Terg 05] and Visual Analytics [Keim 06].

Semantic metadata analysis. In chapter 4, I have proposed different semantic metadata analysis methods. They are designed to suggest selection criteria which cannot be easily obtained by implicit semantics. Part of the short-term future works pertains to their extension.

- Asymmetric Semantic Similarity among Categorical Values requires the categorical values are represented into an ontology. However, the definition of an ontology for each attribute can be quite costly. In the case of a background knowledge which is not arising from a specialized application domain but from common sense, WordNet could be investigated to replace or to support the task of knowledge representation.
- Asymmetric Context Dependent Semantic Similarity adopts both external and extensional parts of the ontology to assess the similarity. However, the formalization of application context is currently affecting only the extensional part of similarity. Whether or not the context should be considered also in the external part deserves further investigations. Other interesting developments might arise considering the extension of the ontology model which has been considered so far. Up to now, the ontology model I have adopted considers the expressiveness which is common to most popular ontology languages. However, it can be extended to take into account particular features of a specific ontology language. For example, considering OWL, it allows the definition of transitive and symmetric relations which are not included in our ontology model. It would be interesting to study if these kinds of relations require a particular precautions during the similarity assessment.
- Semantic Granularity organizes the resources in granularities considering the link between the resources and qualities, the hierarchy of qualities and the qualities degree of informativeness. Currently the method considers resources as a whole

linked to the qualities, but it does not look into resources nor qualities to exploit potential hints associated to attributes. Some hints more concerning the expected granularities might be obtained analyzing the resource and quality attributes. An other interesting issue is given by the assumptions made behind the degree of informativeness. It assumes that the more a quality is a good abstractor the more it should be considered as a granule. Unfortunately, sometimes the opposite is true. For example, there are cases where it is more interesting to point out the qualities that are more specific and less common instead of those which are good abstractors. In general, which assumptions can be applied for a given specific domain and an analysis goal has to be investigated. Some kind of context should be considered also in the method of semantic granularity.

Other future works are foreseen considering the application domains. In the geographical domain, it will be possible to take full advantage from the methods proposed in this thesis as soon as the representation of metadata moves from the standard ISO 19115 to ontology driven representations. The research community has been already acknowledging the importance of considering explicit semantics, but from an application point of view ontology driven standards are not defined yet. In the domain of 3D modeling, ontology driven metadata are emerging from AIM@SHAPE results. Stable and populated version of the ontologies will be released by the end of this year. The advance in the definition of stable ontologies will pave the way for a more careful human evaluation of the proposed methods outlining further contributions.

Finally, longer term research activities are foreseen considering the work made within this thesis as the first step in the definition of “software prostheses”. The need of direct involvement of the seeker to successfully search and select information resources shows how in some situations the matter is not to obtain a fully automated system but to provide the user with the proper extensions. “Software prosthesis” might provide these extensions. They should be designed to be transparently adopted by the user and rely on “thought processes” such as context-dependent similarity and granularity to ease the interpretation of the sheer volume of information. Of course, the realization of the metadata analysis framework as “software prostheses” results quite challenging and in my opinion it deserves further and deeper investigation. It would require to determine if there are other kinds of “thought processes” which can be automatically performed, their dependency with respect to the background knowledge and context. Moreover, other kinds of knowledge formalization (e.g., conceptual spaces [Gard 00, Gard 04]) might be considered and integrated in the current metadata framework.

Appendix A

Ontology Driven Metadata for Digital Shape Acquisition and Reconstruction

This appendix aims to illustrate the activity I have performed within the research programme of the Network of Excellence AIM@SHAPE: Advanced and Innovative Models And Tools for the development of Semantic-based systems for Handling, Acquiring, and Processing knowledge Embedded in multidimensional digital objects. Objective of the NoE is to advance the research in the direction of semantic-based shape representations and semantic-oriented tools to acquire, build, transmit, and process shapes with their associated knowledge. For this purpose ontology driven metadata have been used as mechanism to formalize the shape knowledge In particular I have contributed to archived one of the NoE results: the design of an ontology driven metadata for Digital Shape Acquisition and Reconstruction.

The success of the scientific enterprise largely depends on the ability of sharing different kind of informative resources among the scientific community. As pointed out by Hendler ([Hend 03]) researchers may need to find and explore results at different levels of granularity, from other perspectives in a given field or from a complete different scientific field. This problem is particularly relevant in the field of Computer Graphics and Vision, which is based on a large spectrum of fundamental fields. Recently, the area has reached a state where each individual fundamental domain is well understood and exploited. A fast evolution of it is now conditioned by how research teams will be able to intercommunicate, in particular concerning the sharing of the basic kind of resources, i.e. the digital shapes. This resources should preserve as much meaningful information as possible, in order to allow and improve collaborative research and complete understanding of complex tasks.

But where is this meaningful information? Some of it is intrinsic of the shape (e.g. appearance, topological structure and geometry) and so it is naturally preserved in the digital world, some

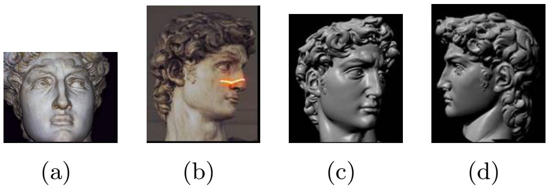


Figure A.1: *The Michelangelo's David. (a)-(b) two pictures of the real statue. (c)-(d) two different views of the triangulated digital models (Images thanks to the Michelangelo Project - Stanford University www.stanford.edu)*

other is in the context (e.g. environmental conditions, location, ownership) and so is pertinent only to the real world in which the shared object was originally embedded. This kind of information is usually not associated to the digital model. A critical phase is the *Acquisition Phase*, in which the contextual knowledge includes different conditions and properties related to the object to be scanned, to the surrounding environment or even to the knowledge of the scanning experts. Most of this information must be preserved and passed to the other steps of often complex modeling pipelines, in order to improve the quality of the results and to open to new research approaches.

The novelty of our work is to integrate Knowledge Management approaches to Computer Graphics and Vision and, in particular, we aim at preserving information when passing from the real world (real objects) to the digital one (digital shapes). This is a fundamental step for moving knowledge from the human experts to the machines. We foresee a research generation in which Digital Shape Knowledge is explicitly represented and, therefore, can be retrieved, processed, shared, and exploited to construct new knowledge.
In this appendix we analyze the problem of linking Real and Digital, trying to preserve significant information during digitalization, with the specific intent of creating semantically enriched digital replica of real 3D objects. For this reason, we faced the problem of formalizing the Acquisition Process in a domain specific ontology, called Shape Acquisition and Processing Ontology (SAP).

The remainder of this appendix is organized as follows: Section A.1 points out the requirements/observations when passing from the Real World to the Digital one. Section A.2 depicts the informative power of digital shapes when correctly embedded in a specific context, and Section A.3 presents our proposed ontology for Shape Acquisition and Processing, focusing the attention on the acquisition session.

A.1 From Real to Digital

Formally representing objects through models is fundamental in any application field. The term model usually means a mathematical construct which describes objects or phenomena.

The modeling step is done by defining the entities and rules which formally describe the object and its behavior, thus defining a symbolic structure which can be used and queried as if it were the object itself under certain conditions([Falc 98]).

A lot of information about an object is conveyed through the use of models of the object itself, since much of our knowledge about the physical world comes to us in the form of shape information. On the one hand, architects, engineers, product designers have always used physical models and graphical representations for visualizing their hypotheses and to show their projects. On the other hand, a model can represent a real object, and can replicate its intrinsic information value. Examples are relief maps or, in cultural heritage applications, replicas of famous statues, (see A.1).

The use of computers has given further emphasis to the informative purposes of Shape Modeling. At the beginning, this effort gave rise to research in Geometric Modeling, which sought to define the abstract properties describing the geometry of an object (geometric model) and the tools to handle the related symbolic structure. Terminology and definitions for the foundations of Geometric Modeling were first introduced in Requicha's seminal 1980 article ([Requ 80]), whose basic notions have shaped the whole field to this day.

Following his paradigm, considerable research activity has been developed in the two most well known representation schemes: CSG (Constructive Solid Geometry) and BRep (Boundary Representation), which have deeply influenced current commercial geometric modeling systems ([Mant 88]). However, the above representation schemes rely only on geometrical information which is not enough to fully characterize a shape. Additional information should be modeled and associated to digital models and some of this information must be taken from the real world. It is possible to consider an object evolving in time, and to note that several events can occur during its lifecycle. For example, when the object is a real object, it may become a part of another object in a construction process, or some parts of it may be substituted, or may get accidentally broken. When the object is in the digital world, for design purposes it may evolve according to the designer's intent, or may be reused in different contexts; for rendering purposes it can be simplified or optimized for specific hardware; for quality enhancing some tools may be used to smooth a noisy area, to fix unwanted holes, to localize and preserve the edges sharpness, and so on.

In this work we focus on the evolution of the object in its most critical phase: when it passes from the real world to the digital world. In other terms, in the process of its acquisition. The most important question that has to be posed is: what information has been lost in this passage? Which immediately drives to the question: how the acquired geometric model should be augmented to hold the information value that it had?

Two interrelated observations can ease in answering to these questions. The first observation is that the information held by digital shapes cannot be merely geometric: the precise description of the boundary of an object is not all that characterizes it. The human perception goes beyond, as some structural information or even some high-level semantic information are

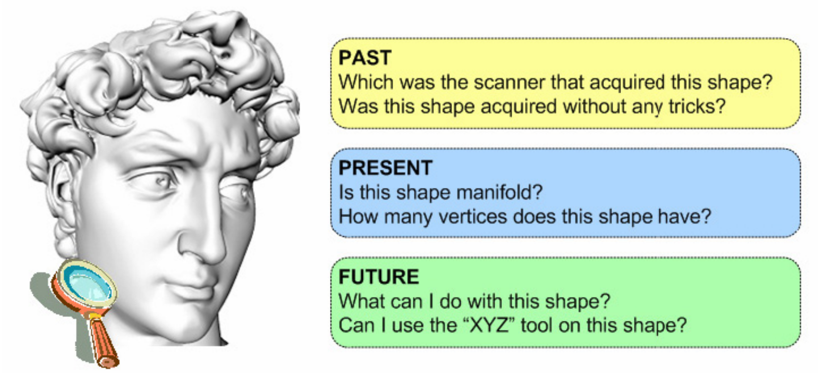

PAST
Which was the scanner that acquired this shape?
Was this shape acquired without any tricks?
PRESENT
Is this shape manifold?
How many vertices does this shape have?
FUTURE
What can I do with this shape?
Can I use the "XYZ" tool on this shape?

Figure A.2: *An expressive characterization of a shape is made up by the information related to its history, the information intrinsically held by the shape itself and the information related to its capabilities*

immediately perceived by the humans approaching the object. The second observation is that when we observe a real object, it is in a specific place and it exists in a specific instant. In the real world the objects are not independent from the context (e.g. space, time, ownership) in which they are embedded.

Therefore, as the representation level should adhere at its best to the reality it describes, also the digital objects must take into account other levels than mere geometric information. We think that augmented information is a key issue in the field of Shape Modeling. In the AIM@SHAPE Network of Excellence ([AIM]) the above observations converge in one simple statement: digital shapes have to be coupled with both intrinsic and contextual semantic information. On the one hand, elaborated techniques are under development to produce and process not only geometric, but also structural and semantic data. On the other hand, domain specific ontologies are in their construction phase, and their aim is to capture contextual information of shapes (Virtual Humans ([Guti 05]), Industrial Design ([Ucel 05]) which are used to deal with resources according to context-dependent views). Thus, coming back to the question "How the acquired geometric model should be augmented to hold the information value that the real object had?", we thought that the key issue was to adequately formalize the contextual information related to the object and, specifically, to its acquisition phase. Then, this information should be kept together with the digital object as an added value. This formalization is achieved through the Shape Acquisition and Processing ontology, which will be further described in Section A.3.

A.2 Scenario, Domain, Applications

In this section we present an exemplificative scenario to illustrate a possible application where the adoption of our ontology can demonstrate its usefulness. For clarity of the explanation, we will consider a scenario (the acquisition of a statue) which will lead to choices concerning the specific kind of digital resources (3D shapes) and the possible applications that could be addressed (real- or digital-based).

Suppose to have a 3D digital model of the Michelangelo's David whose physical counterpart is placed at the Uffizi Museum. Via inspection of the digital shape, we can discover that the quality of the head junction is not sufficient to study the way in which Michelangelo has created the David (e.g. which tools has used and how). In the case in which we can use the digital shape without any additional information, we can decide to re-plan an acquisition, relying only on the expertise of the acquiring researcher. Instead, if we can look at and evaluate appropriate additional information associated to the digital shape, we can reason on how to improve the quality of the planned acquisition (e.g. lighting conditions, logistic conditions, error estimation and so on). We could also understand that, with a given scanning device, the quality cannot be further improved and we should plan a new acquisition with another, more powerful, acquisition device. Thus, appropriated information can catch the expertise and the knowledge in a particular research field (in this case, Computer Vision) and can maintain it for reusing, sharing in other research fields (such as Computer Graphics and Geometric Modeling).

Note that, when passing from real objects to digital shapes, two interrelated macro-classes of applications can be identified:

1. One class is still related to the real world, such as acquisition planning and documentation. An acquisition expert looks at a real object (e.g. the statue) and plans an acquisition, choosing an acquisition system, defining particular lighting conditions and evaluating the existing logistic constraints. For example, in case of an historical statue, maybe, it could be not possible to move it for scanning it elsewhere and some future occlusions may appear. The acquisition process must be reported in details for documenting a research activity.

2. The other class acts only on the digital world, and it includes applications such as shape remeshing, shape enhancing, analysis and structuring. Once a digital shape is obtained (e.g. a digital replica of a statue), it can be analyzed in order to extract characteristic features or can be modified in order to enhance the quality of the model for visualization purposes.

The quality of the applications in the latter macro-class (2) can be improved if information related to the real world is kept together with digital shapes. For example, in documenting the acquisition procedure, information related to a real object can be attached to the obtained digital shape and maintained as a piece of its history: this information will be lost otherwise.

A.3 An Ontology for Shape Acquisition and Processing

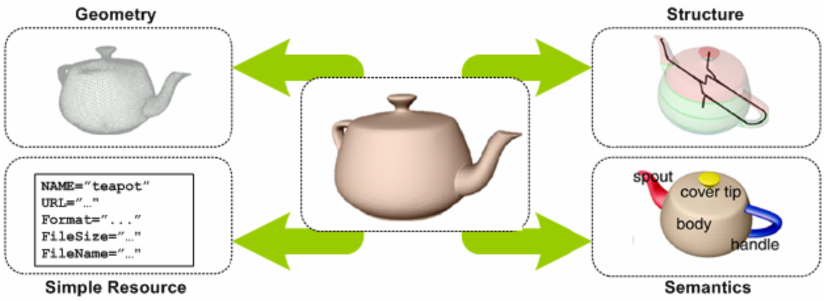
Due to the intrinsic complexity of shapes, an ontology is necessary in order to reach a sufficient level of expressiveness. The ontology entities should provide a thorough characterization of shapes (Figure A.2) by storing: (i) the information related to its history, such as the acquisition devices and techniques for creating it or the tools for transforming it (its past, e.g. for documentation), (ii) the information intrinsically held by the shape itself (its present) and (iii) the information related to its capabilities and potential uses, such as the possible steps that can be performed or the tools that can be used (its future, e.g., for acquisition/process planning).

Our desired ontology should be able also to represent different levels of sophistication describing a shape as a simple resource (e.g. for cataloging) and characterizing it according to its geometry (e.g. for rendering), to its structure (e.g. for matching and similarity), and to what it represents (e.g. for recognition or classification). Figure A.3 gives an example of a digital shape and its intrinsic characteristics: it can be seen as simple resource (e.g. name and URL), or can be considered by its geometric characteristics (e.g. a set of triangles and normals). It has a structure (e.g. the skeleton of a teapot) or it can be seen a teapot composed by a handle, a spout, a body and a tip. It is important also to take into account the different contexts where the shape can be used since the specific application determines relevant characteristics. For example, if the main purpose is to build a teapot, the identification of parts by which a teapot is composed is fundamental, while if the purpose is to let a robot grasp it, the localization of the handle is the only necessary task.

The existing branches of research in the field of Computer Graphics and Vision are interested in one or more of the above mentioned characterizations, but also on the conditions and the tools to pass from one characterization to another. Finally, it is important to note that shapes play a central role in Computer Graphics and Vision, but they do not represent the only kind of resource that must be characterized in the common framework. Every day, scientists work with shapes, tools and publications.

It is important to devise the role of these resources in different conceptualizations, making relationships among them explicit. For example, a scientist may want to evaluate her latest implementation of a method. In this case, it is interesting to figure out which are the tools providing other implementations of the same method, the publications related to the above tools and methods, or the shapes used as tests for the other implementations (e.g. for testing/benchmarking activities).

The domain of our Shape Acquisition and Processing (SAP) ontology is defined as the development, usage and sharing of hardware tools, software tools and shape data by researchers and experts in the field of acquisition and processing of shapes. In the creation of the SAP ontology, the following macro-steps have been considered: *Shape Acquisition* (and Registration) - the phase in which sensors capture measurements from a real object; *Shape Processing*



NAME="teapot"
URL="..."
Format="..."
FileSize="..."
FileName="..."

Figure A.3: *A shape is described as a simple resource, or by its geometry, its structure, its semantics, depending on the application domain (part of the image courtesy of the AIM@SHAPE project)*

- the phase in which all acquired data are merged to construct a single shape and in which further computations may be performed (e.g. smoothing, simplification, enhancement, and so on).

Thus, the ontology target applications are related to acquisition planning, data validation, benchmarking, testing and data enhancement (e.g. automatic recovery).

The conceptualization arises from a process of elicitation which has taken place interviewing the experts within the NoE AIM@SHAPE ([AIM]) and from the assumption that the ontology is intended to be targeted to the scientific community. On the basis of the experts' needs, we have identified a set of formal competency questions, as for example: what are the Acquisition Systems able to scan a transparent real object? What tricks have to be performed in order to be able to scan a light absorbent real object? What Acquisition Systems are available at the DISI institute? In which environmental conditions the model was acquired?
Some of the identified competency questions are strictly related to the Acquisition phase, some others regard information related to real objects and/or digital shape and some other concern tools and algorithms managing digital shapes.

In the following subsection we will present in more details some concepts defined in SAP, presenting also relations and attributes. In particular we will present concepts aiming at modelling the knowledge related to the Acquisition Process. Its development has been made according to the OntoKnowledge approach ([Sure 04]) and the SAP ontology has been expressed in OWL-DL([OWL 06]). Note that, in the following, what is written like THIS is actually an entity in the ontology we have developed using Protégé ([Muse 01]).

A.3.1 Modeling the Acquisition Process

The acquisition process basically deals with an acquisition session which takes place considering a particular real object and producing a digital shape on the basis of certain conditions. In SAP, the AcquisitionSession has been modeled as an entity and an overview of this entity is given in figure A.4. The AcquisitionSession is related to an AcquisitionSystem (which is made up by one or more AcquisitionDevices - e.g. scanners) and to the AcquisitionConditions in which the acquisition is performed.

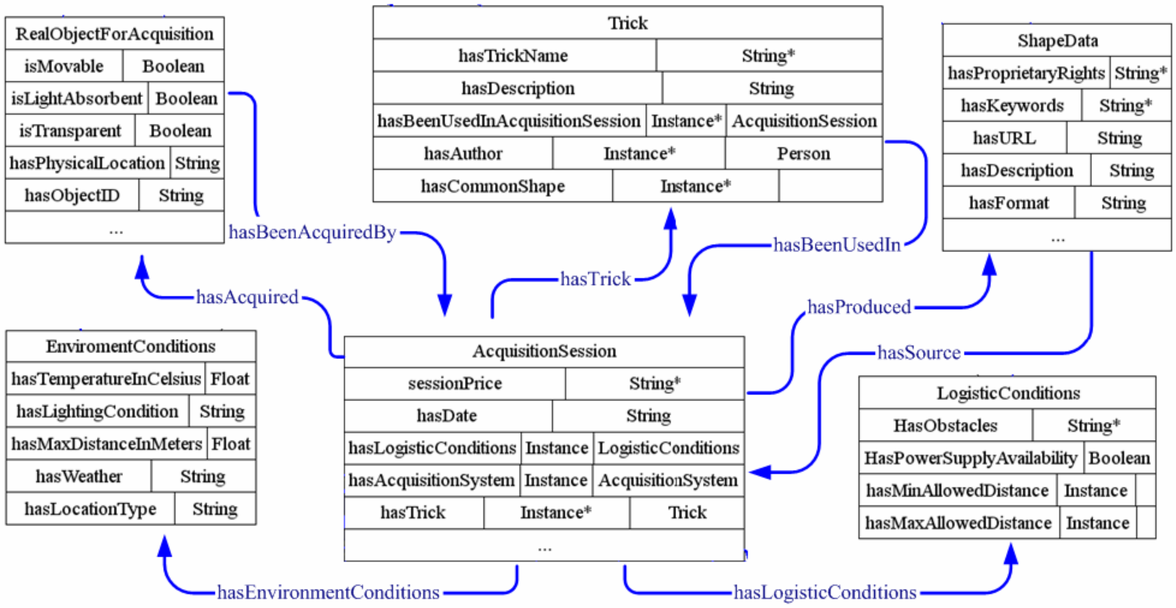

RealObjectForAcquisition
isMovable | Boolean
isLightAbsorbent | Boolean
isTransparent | Boolean
hasPhysicalLocation | String
hasObjectID | String
...
Trick
hasTrickName | String*
hasDescription | String
hasBeenUsedInAcquisitionSession | Instance* | AcquisitionSession
hasAuthor | Instance* | Person
hasCommonShape | Instance*
ShapeData
hasProprietaryRights | String*
hasKeywords | String*
hasURL | String
hasDescription | String
hasFormat | String
...
hasBeenAcquiredBy
hasTrick
hasBeenUsedIn
hasAcquired
hasProduced
hasSource
EnviromentConditions
hasTemperatureInCelsius | Float
hasLightingCondition | String
hasMaxDistanceInMeters | Float
hasWeather | String
hasLocationType | String
AcquisitionSession
sessionPrice | String*
hasDate | String
hasLogisticConditions | Instance | LogisticConditions
hasAcquisitionSystem | Instance | AcquisitionSystem
hasTrick | Instance* | Trick
...
LogisticConditions
HasObstacles | String*
HasPowerSupplyAvailability | Boolean
hasMinAllowedDistance | Instance
hasMaxAllowedDistance | Instance
hasEnvironmentConditions
hasLogisticConditions

Figure A.4: *A zoom on the* AcquisitionSession *entity in our ontology. The most significant relations are highlighted by arrows. Each rectangle represents an entity. The rows in each entity represent a slot which can be either an attribute or a relationship. For each attribute the type is specified, while for each relationship the range is indicated. Whenever a symbol "*" appears next to the name of an attribute or a relationship, the cardinality can be more than 1.*

These AcquisitionConditions can be LogisticConditions (they include the presence of lights, if there exist any obstacle between the real object and the scanning device and so on) or EnviromentConditions (which include the information on is the type of environment - indoor, outdoor or underwater, the level of humidity or even the weather). Moreover, some attributes are directly related to the AcquisitionSession (e.g. the price for renting the technological devices), while others are related to the different entities in the framework (e.g. the person/institute responsible for a scanning system). An AcquisitionSession basically documents the acquisition of a RealObject and the production of a ShapeData (a digital shape), using a particular AcquisitionSystem. A RealObject has also been modeled as an entity, and the knowledge related to it and to its context is thus preserved: in the ontology are recorded the location of the object, the possibility to move it, whether or not it is transparent or light-absorbent, and so on. Note that the mentioned characteristics (e.g. transparency and being or not light-absorbing) have immediate impact on the Acquisition Planning. For instance, a Trick can be used when there is a problem of compatibility between the AcquisitionSystem and the RealObject to be scanned: a light absorbent object and a laser scanner might be incompatible, but if we need to perform the scanning, it is possible to avoid the problem by spreading powder over the object before scanning. Otherwise, it can also be possible to plan the acquisition with another (compatible) AcquisitionSystem. ShapeData (which identifies a digital shape) has been modeled as an entity with some specific properties, such as format, URL, description, source (i.e. the AcquisitionSession that has produced it) and owner (an Institution or a Person). A ShapeData can be based on another (or more than one) ShapeData, or a ShapeData can be used to generate a new

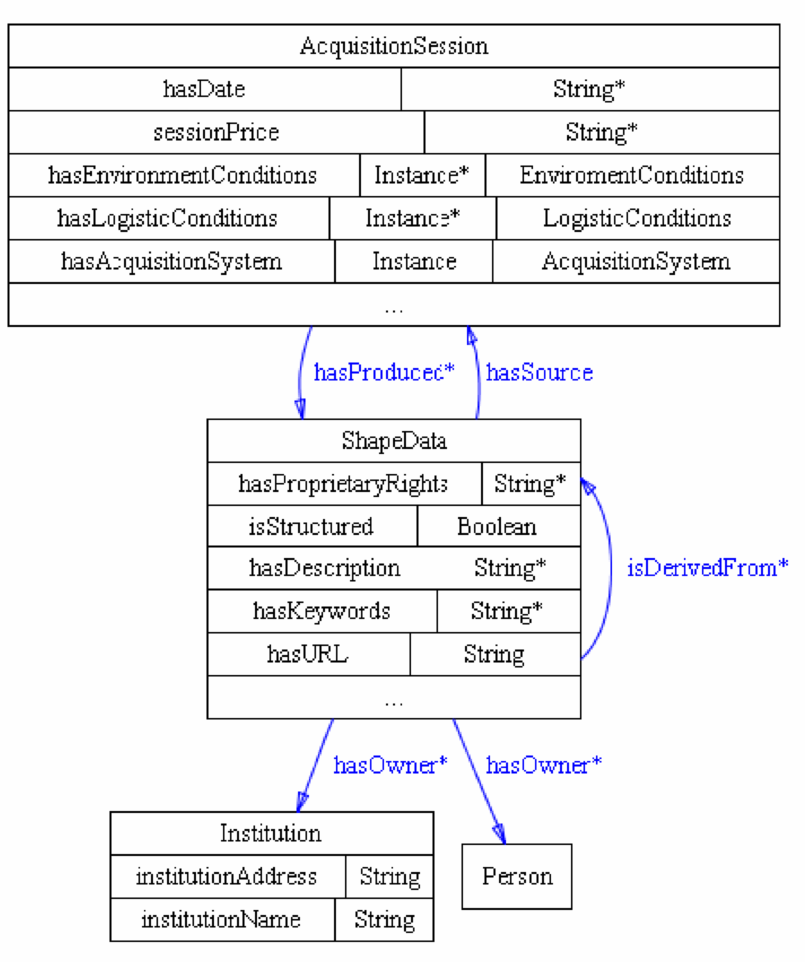


Figure A.5: ShapeData *entity and its relation with the* AcquisitionSession *entity.*

one. The relation isDerivedFrom formalizes the knowledge related to the history of a given shape.

The ontology introduced so far, even if here only partially described, is already sufficient to describe the macro-step of the acquisition of a real object. Such a simple description provides the basics to embed in the digital shapes information that usually gets lost after acquisition. This information might result important for comparing shapes coming from different providers, for improving the assessment about their quality and for better understanding the results arising from further processing.

Related Publications

- Albertoni, R.; Papaleo, L.; Robbiano, F. "Preserving Information from Real Objects to Digital Shapes" In: Eurographics Italian Chapter Conference. De Amicis, Raffaele; Conti, Giuseppe (Eds.) pp.79-86 (2007)
- Albertoni R., Papaleo L., Robbiano F., Spagnuolo M. "Towards a Conceptualization for Shape Acquisition and Processing". In: 1st International Symposium on Shapes and Semantics (Matsushima, Japan, 17 June 2006). Proceedings, pp. 85-90. Genova, Italy, CNR - A.Ri.GE, 2006.
- Papaleo L., Albertoni R., Marini S., Robbiano F., "An ontology-based Approach to Acquisition and Reconstruction", Workshop towards Semantic Virtual Environment, Eurotel Victoria, Villars, Switzerland, pp:148-155, 2005.
- Albertoni R., Papaleo R., Pitikakis M., Robbiano F., Spagnuolo M., Vasilakis G., "Ontology-Based Searching Framework for Digital Shapes", (Lecture Notes in Computer Science Vol.3762), Springer, pp. 896-905, 2005.

Appendix B

Integrating Information Visualization and Semantic Web

This appendix illustrates the research activity described in the paper [Albe 05a]. It examines the potentialities offered by Information Visualization to improve information search in the Semantic Web. In particular, the appendix discussed some of the seeker's problems illustrated in chapter 2 in the Semantic Web. It aims to demonstrate that Information Visualization is effective to solve these problems even if it has not been yet properly adopted in the Semantic Web.

The Semantic Web (SW) is rising as an extension of the current Web to improve the accessibility of web content providing sophisticated and powerful inferences to sift intelligently through this large information space. In particular, Semantic Search [Guha 03] is emerging as the application of the SW designed to improve the search in the Web, since it relies on an explicit representation of semantics about web resources and real world objects. Nevertheless Semantic Search does not exhaustively solve all the problems related to the search. Firstly, every searching activity in the WWW as well as in the SW is characterized by a highly interactive process. The seeker needs to refine his selection criteria according to the obtained results alternating phases where he queries the SW to phases where he browses the SW content. Secondly, since anyone can publish new information sources in the SW, the information sources resulted by a user's query might be redundant or represent different points of view. Even supposing that Semantic Search is able to improve the quality of results, there are stages in the search activity where only the seeker can decide which sources to discharge. The appendix analyzes and argues how Information Visualization may aid SW in a number of fundamental issues concerning the information search. In the first part it presents an analysis of the problems the user has still to face with during the information search in the SW. In the second part an analysis of the most representative visualization-based tools available in the WWW and in the SW is performed: a classification of some functionalities implemented in the tools are identified and a synthesis of their potentiality to solve the mentioned problems in the WWW and in the SW is provided. The appendix does not want to provide a state of the art of the available tools, but it aims to underlay that the potentialities offered by Information Visualization in the WWW could be applied in the SW to improve information search.

B.1 Problems in information search in the Semantic Web

Information search cannot be performed without any involvement of individuals since the "searching skills" are strongly dependent on human factors as the seeker's anxiety whenever the query result does not fulfill his needs, the limited seeker's knowledge, the lack of relationship between the seeker and the information providers. These different factors get the user into some problems even when he searches for information in the SW. For instance, the definition of the search criteria is strongly affected by the limited knowledge of the seeker. According to Belkin this is a result of the Anomalous State of Knowledge (ASK) [Belk 82a] as already discussed in chapter 2 Independently of the considered environment (traditional information retrieval systems, WWW or SW), the ASK forces the seeker to enter into dialogue with the IR systems engaging in a query refinement process.
Spoerri mentioned how the query refinement process is affected by some problems in WWW [Spoe 04c]. In the following, the appendix discusses the problems mentioned by Spoerri and it argues their relevance in the SW context.

Problem of query formulation: "how to precisely communicate the query criteria to the system?" This problem is strongly related to the choice of the language adopted to query. The formal languages are usually precise but they result unfriendly for seekers who have an inappropriate background. On the other hand, the natural language is considered more friendly but it is often not precise. To make the SW successful it is necessary to provide a friendly mean to precisely express the selection criteria the user has in his mind.

Problem of vocabulary: "Which term to use?" The difference of knowledge and perception between the information providers and the seekers is modeled in terms of informative space and cognitive space. The former one is defined as a set of objects and relations among them held by the system, whereas the latter one is defined as a set of concepts and relations held by individual [Newb 01]. Information providers organize resources according to their knowledge and vocabulary concurring to built the "informative space". If the seeker has different knowledge background his cognitive space has a poor overlapping with the information space, and he will use different terms to identify the same concepts. This problem is still relevant in the SW. Even if the use of ontology and of lexical databases like WordNet can ease the vocabulary problem, a mapping between each pair of cognitive space and information space is hardly provided.

Problem of database selection: “which search engine to select?”. The seeker has to decide which search engine to use. The problem is well known in the WWW because the actual search engines are able to cover a limited portion of the web resources. Also SW is affected by similar problems since it will probably have different engines or web services, which differ each other in the technology and in the conceptualisation they rely on.

Problem of retrieved resources exploration: “how to explore many retrieved documents?”. Semantic Search [Guha 03] is based on an explicit representation of semantics about web resources and real world objects. It aims to improve both the proportion of relevant material actually retrieved (recall of results) and the proportion of retrieved material that is actually relevant (precision of results). However, because of the huge amount of resources that will be available in the SW, even adopting the Semantic Search the seeker will have to face with a huge amount of query results. Furthermore, due to the ASK the queries formulated by the seeker might not correspond to a proper representation of his needs, as a result the order induced by ranking measures might be misleading. The seeker needs to be supported in their analysis of results to choose the most suitable for his purpose.

Problem of query coordination: “How to query?”. Human behavioral studies shows that the seeker is lazy, usually he tends to create short queries and rarely adopts boolean expression in his query criteria [Spin 01]. On the other hand, he is forced to a deeper search in the Web as well as in the SW whenever he is the only one who can define the searching criteria and the result of his search can seriously affect the success of his work.

B.2 Tools analysis

An analysis of some visual-based tools in the WWW and SW is illustrated: a comparison of the tools in terms of their functionalities is proposed to identify their complementarities and to underline the potentialities offered by IV to improve information search. The aim is to demonstrate the advantages of adopting IV in the SW rather than to provide a complete state of art.

B.2.1 Tools vs. functionalities

Some of the most representative tools available in the WWW and SW are considered. Concerning the SW there are no specific tools based on visualization designed for Semantic Search; anyway there are well known tools developed in the field of Ontology which could be extended to ease information search. The analysis concerns tools for the Web as Kartoo [Kartoo 04], Grokker [Grokker 04], Web Theme [Whit 02], Aduna AutoFocus [Autofocu 04], MetaCrystal [Spoe 04a], and the tools for visualization and interaction with the ontology as OntoViz [OntoViz 04], TGViz [TGVizTab 04], Jambalaya [Jambalay 04], Spectacle [Flui 02a]. Several other tools are available in the web, but they mainly differ in the implementation or in how they combine the functionalities. The choice of the tools has been performed giving priority to those ones that can be freely downloaded. Only two tools raise an exception: Spectacle and Metaviz. The reason is that they offer functionalities which are not provided by the other tools, and even if it was impossible to make a direct use of them, they had to be included. The analysis concerns the study of the tools to identify the most relevant functionalities useful in the information search. We have grouped them in three main categories: graphical visualization, graphical interaction, and a combination of them. The following functionalities have been outlined:

Hierarchical Visualization to visualize and browse the content according to different levels of granularity (e.g. Grokker).

Clustering Visualization to visualize and group the content according to similarity criteria. The groups are obtained either by applying a clustering algorithm (galaxy view [Whit 02]) or according to properties specified by the user (cluster map [Flui 02b]).

Map Based Visualization to organize the content according to thematic terms or co-occurrence criteria as in the geographical map (e.g. Kartoo).

Venn diagram representation to describe and compare the elements and characteristics of items and to quickly convey a compact view of data (e.g. MetaCrystal, Spectacle).

Visualization Manipulation to re-organize, move and add graphical elements (e.g. Grokker allow to insert a new web site in the displayed graph).

Graphical Selection to select different information sources such as URI, PDF or DOC document in Grokker, Aduna Autofocus, Kartoo or data as in Web Theme[Whit 02].

Highlighting to visualize a selected element and all its related sources (e.g. Aduna AutoFocus, Kartoo and Spectacle allow to highlight the related co-occurring terms, in Grokker the keywords used in filters are highlighted in the web pages visualization).

Co-Occurring Terms Visualization to visualize a statistical thesaurus to expand user queries with other highly frequent terms. They should help the user in discriminating relevant documents ([Peat 91]).

Colored Query Result to set different colors for the query results to facilitate their comparison (e.g. Web Theme [Whit 02]).

Filter Results Representation to apply filters to the contents. For instance, Grokker allows to filter the rank, the domain and the source, whereas Kartoo allows to filter the co-occurring terms.

String Search: to search for a co-occurring word and to navigate the ontology hierarchy.

Choice of the Hierarchy Level Shown to choose the number of levels displayed in the hierarchy exploring it at different levels of details.

Ontology Instances to visualize the instances of a selected class separately or directly in the ontology graph (e.g. OntoViz, Jambalaya and Spectacle).

Ontology Graph Navigation to easily navigate the ontology graph structure (e.g. Jambalaya proposes different layouts and an animated navigation to browse the hierarchy).

		World Wide Web					Semantic Web			
		Grokker	Aduna AutoFocus	Kartoo	Meta-Crystal	Web theme	OntoViz	TGViz	Jambalaya	Spectacle
Graphical Visualization	Hierarchical Visualization	☺	-	-	-	-	😐	😐	☺	😐
	Clustering Visualization	☺	☺	-	-	☺	-	-	-	☺
	Map Based Visualization	-	-	☺	-	☺	-	-	-	-
	Venn diagram representation	-	☺	-	☺	-	-	-	-	☺
Graphical Interaction	Visualization Manipulation	☺	☺	☺	-	-	-	☺	☺	☺
	Graphical Selection	☺	☺	☺	☺	☺	-	☺	☺	☺
Interaction and Visualization	Highlighting	☺	-	-	☺	-	😐	☺	☺	☺
	Co-Occurring Terms Interaction/Visualization	☺	☺	☺	-	-	-	☺	☺	☺
	Coloured Query Result	-	-	-	-	☺	-	-	-	-
	Filter Results Representation	☺	☺	😐	😐	-	-	-	-	-
	String Search	☺	-	-	-	😐	-	☺	☺	☺
	Choice of Hierarchy Level shown	☺	n.a.	n.a.	n.a.	n.a.	-	☺	-	-
	Ontology Instances	n.a.	n.a.	n.a.	n.a.	n.a.	😐	☺	☺	☺
	Ontology Graph Navigation	n.a.	n.a.	n.a.	n.a.	n.a.	-	☺	☺	😐

Table B.1: *Functionalities provided by Information Visualization tools in the WWW and SW*

Table B.1 summarizes the results of the analysis. It shows the associations between the tools (columns) and their functionalities (rows), as following:

- "n.a." the functionality is not applicable;
- "-" the functionality is not implemented;
- "sad face" the functionality is partially implemented or implemented in a trivial way;
- "smiley face" the functionality is fully implemented.

The idea of this table is to identify a set of conceptual functionalities implemented in the considered tools rather than to provide a complete state of art. The evaluation of a functionality for each tool is based on its description reported in the material related to the tools or on its direct usage whenever the tools were available. The choice between "smiley" or "sad face" is performed according to the authors' impression. In general, a functionality is classified as implemented in trivial way (sad face) if its implementation in the considered tool appears less impressive than the implementations provided by the other tools.

B.2.2 Functionalities vs. problems

Table B.2 shows the result of the analysis of the support provided by each functionality in solving the mentioned problems for the WWW as well as for the SW. Functionalities which are not yet implemented for the SW but potentially useful are also outlined. The table has the following legend:

- "n.a." the functionality is not applicable in the considered context (WWW or SW);
- "-" the functionality does not seem to help to solve the problem, independently from its implementation;
- "Empty cell" the functionality could help to solve the problem, but it is not provided by any tool;
- "sad face" the functionality gives a partial support in the resolution of the problem;
- "smiley face" the functionality provides a satisfactory help in the problem solution.

For each problem the table shows the comparison between the contributions that each functionality provides in the WWW and in the SW, respectively represented in the first half-column and second half-column of each problem column. The functionality/problem evaluation is obtained according to the principle that exists at least one of the considered tools which provides the functionality able to solve the problem. Different faces "sad face, smiley face" are assigned according to the quality of the support provided by the functionality to solve the problem.

Analyzing the Table B.2 it is possible to state that:

- Some functionalities ease the problems only if conceptualized and implemented in the SW ("light-gray background");
- Some functionalities are implemented both in the WWW and in the SW, but in the former one they provide better results than in the latter one ("middle-gray background");

		Database		Vocabulary		Query Formulation		Results Comprehension		Query Coordination	
		WWW	SW	WWW	SW	WWW	SW	WWW	SW	WWW	SW
Graphical Visualization	Hierarchical Visualization	-	-	😐	😐	😐	😐	☺	😐	-	-
	Clustering Visualization	-	-	😐		😐	😐	☺	😐	-	-
	Map Based Visualization	-	-	😐		😐		☺		-	-
	Venn diagram representation	☺		-	-	😐	😐	😐	😐		
Graphical Interaction	Visualization Manipulation	-	-	-	-	-	-	☺	😐	-	-
	Graphical Selection	-	-	-	-	☺	😐	-	-	-	-
Interaction and Visualization	Highlighting	-	-	-	😐	-	-	☺	😐	-	-
	Co-Occurring Terms Interaction/Visualization	-	-	☺	☺	☺	☺	-	-	-	-
	Coloured Query Result			-	-	-	-	-	-	☺	
	Filter Results Representation	-	-	-	-	-	-	☺	😐	-	-
	String Search	-	-	-	😐	-	-	-	-	-	-
	Choice of Hierarchy Level shown	-	-	-	😐	-	-	😐	😐	-	-
	Ontology Instances	n.a.	-	n.a.	-	n.a.	😐	n.a.	😐	n.a.	😐
	Ontology Graph Navigation	n.a.	-	n.a.	😐	n.a.	😐	n.a.	😐	n.a.	-

Table B.2: *Information Visualization and information search problems in the WWW and in the SW*

- Some functionalities are implemented only in the WWW (represented in "dark-gray background").

In general, Information Visualization has the potentiality to help the user in the searching task: for each problem there is at least a functionality that provides a useful support in the Web. However as detailed in the following, IV seems still not properly developed for the SW. Concerning the vocabulary problem, a partial support is already provided in the SW. It is important to note that all the comparisons highlighted by light-gray background are in the columns of vocabulary problem. It suggests how SW partially takes advantage from the use of ontologies. However, the IV tools in the SW are far from to completely solve the problem: ontologies contain information about the application domain but patterns induced by the use of domain are not made explicit. Novel IV techniques able to make explicit similarities and patterns among sources have to be developed. Considering the columns related to the problems of query formulation and result comprehension, the middle-gray and the dark-gray background are mainly due to the fact that except for Spectacle the tool in SW are realized to support in the design of ontologies rather than in the search. A significant improvement can be obtained by focusing the same functionalities on the solution of the problem. Regarding the database and query coordination problems, although a support is provided in the WWW, there are no "face" symbols (smiley face/sad face) in correspondence of the columns related to the SW. The adoption of some functionalities coming from the Web (e.g. color query results, filter results representation) may help both in the database and the query coordination problems.

B.3 Final remarks

The appendix examines the potentiality of applying IV into SW to improve information search. Several seekers' problems are outlined and an analysis of their occurrence in the context of SW is provided. A set of visual searching tools are analyzed to show that IV is able to face with these problems. Even if the analysis has been limited to a subset of tools, it has outlined that there are functionalities that according to our perception are able to ease in facing with all the mentioned problems. In spite of that, IV is largely applied in WWW but it is not yet completely exploited in the SW. Future development will address how to adapt the existing IV techniques to the SW. It will pave the way for a conceptual framework which integrates Semantic Search, Ontologies and Information Visualization to solve all the searching problems.

Related Publications

- Albertoni R., Bertone A., De Martino M., "Information Search: The Challenge of Integrating Information Visualization and Semantic Web", 16th International Workshop on Database and Expert Systems Applications (DEXA'05), Copenhagen Business School, Copenhagen, Denmark, August 22-26, 2005, IEEE Computer Society Press, pp:529-533, 2005.

- Albertoni R., Bertone A., De Martino M., "Semantic Web and Information Visualization", Proceedings of First Italian Workshop on Semantic Web Application and Perspective, DEIT- Dipartimento di Elettronica, Intelligenza Artificiale e Telecomunicazioni, Ancona, Italy, pp. 108-114, 2004.

Bibliography

[Abel 02] J. Abello and J. L. Korn. "MGV: A System for Visualizing Massive Multidigraphs.". *IEEE Trans. Vis. Comput. Graph.*, Vol. 8, No. 1, pp. 21–38, 2002.

[AIM] "AIM@SHAPE: Advanced and Innovative Models And Tools for the development of Semantic-based system for Handling, Acquiring, and Processing knowledege Embedded in multidimensional digital objects- IST NoE NO 506766". `http://www.aimatshape.net`.

[Albe 03a] R. Albertoni, A. Bertone, U. Demsar, M. D. Martino, and H. Hauska. "Knowledge Extraction by Visual Data Mining of Metadata in Site Planning.". In: K. Virrantaus and H. Tveite, Eds., *ScanGIS*, pp. 119–130, Department of Surveying, Helsinki University of Technology, 2003.

[Albe 03b] R. Albertoni, A. Bertone, M. D. Martino, U. Demsar, and H. Hauska. "Visual and Automatic Data Mining for Exploration of Geographical Metadata". In: M. Gould, R. Laurini, and S. Coulondre, Eds., *Proceedings of the 6th AGILE Conference on GIScience (Lyon, France, 24-26 April 2003)*, pp. 479–488, Sciences appliqués de l'INSA de Lyon, 2003.

[Albe 04] R. Albertoni, A. Bertone, and M. D. Martino. "Visual Analysis of Geographic Metadata in a Spatial Data Infrastructure.". In: *DEXA Workshops*, pp. 861–865, IEEE Computer Society, 2004.

[Albe 05a] R. Albertoni, A. Bertone, and M. D. Martino. "Information Search: The Challenge of Integrating Information Visualization and Semantic Web.". In: *DEXA Workshops*, pp. 529–533, IEEE Computer Society, 2005.

[Albe 05b] R. Albertoni, A. Bertone, and M. D. Martino. "Semantic Analysis of Categorical Metadata to Search for Geographic Information.". In: *DEXA Workshops*, pp. 453–457, IEEE Computer Society, 2005.

[Albe 05c] R. Albertoni, A. Bertone, and M. D. Martino. "Visualization and semantic analysis of geographic metadata.". In: C. Jones and R. Purves, Eds., *GIR*, pp. 9–16, ACM, 2005.

[Albe 05d] R. Albertoni, L. Papaleo, M. Pitikakis, F. Robbiano, M. Spagnuolo, and G. Vasilakis. “Ontology-Based Searching Framework for Digital Shapes.”. In: R. Meersman, Z. Tari, P. Herrero, G. Méndez, L. Cavedon, D. Martin, A. Hinze, G. Buchanan, M. S. Pérez, V. Robles, J. Humble, A. Albani, J. L. G. Dietz, H. Panetto, M. Scannapieco, T. A. Halpin, P. Spyns, J. M. Zaha, E. Zimányi, E. Stefanakis, T. S. Dillon, L. Feng, M. Jarrar, J. Lehmann, A. de Moor, E. Duval, and L. Aroyo, Eds., *OTM Workshops*, pp. 896–905, Springer, 2005.

[Albe 06a] R. Albertoni, E. Camossi, M. D. Martino, F. Giannini, and M. Monti. “Semantic Granularity for the Semantic Web.”. In: R. Meersman, Z. Tari, and P. Herrero, Eds., *OTM Workshops (2)*, pp. 1863–1872, Springer, 2006.

[Albe 06b] R. Albertoni and M. D. Martino. “Semantic Similarity of Ontology Instances Tailored on the Application Context.”. In: R. Meersman and Z. Tari, Eds., *OTM Conferences (1)*, pp. 1020–1038, Springer, 2006.

[Albe 06c] R. Albertoni, L. Papaleo, F. Robbiano, and M. Spagnuolo. “Towards a Conceptualization for Shape Acquisition and Processing”. In: *1st International Workshop on Shapes and Semantics, Matsushima, Japan*, June 2006.

[Albe 07] R. Albertoni, L. Papaleo, and F. Robbiano. “Preserving Information from Real Objects to Digital Shapes”. In: D. A. Raffaele and C. Giuseppe, Eds., *Eurographics Italian Chapter Conference*, pp. 79–86, feb 2007.

[Alpe 96] N. Alper and C. Stein. “Geospatial metadata querying and visualization on the WWW using Java/sup TM/ applets”. In: *INFOVIS ’96: Proceedings of the 1996 IEEE Symposium on Information Visualization (INFOVIS ’96)*, p. 77, IEEE Computer Society, Washington, DC, USA, 1996.

[Andr 03] N. Andrienko, G. Andrienko, and P. Gatalsky. “Exploratory spatio-temporal visualization: an analytical review”. *Journal of Visual Languages and Computing*, Vol. 14, No. 6, pp. 503–541, 2003.

[Andr 72] D. Andrews. “Plots of High-Dimensional Data”. *Biometrics*, Vol. 28, No. 1, pp. 125–136, 1972.

[Anke 00] M. Ankerst, M. Ester, and H. Kriegel. “Towards an effective cooperation of the user and the computer for classification”. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 179–188, 2000.

[Anke 01] M. Ankerst. *Visual Data Mining.* PhD thesis, Ludwig-Maximilians-Universität, München, 2001.

[Anke 96] M. Ankerst, D. Keim, and H. Kriegel. “Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets”. *Proc. Visualization*, Vol. 96, 1996.

[Anke 99] M. Ankerst, M. Breunig, H. Kriegel, and J. Sander. "OPTICS: ordering points to identify the clustering structure". *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, pp. 49–60, 1999.

[Anto 04] G. Antoniou and F. Van Harmelen. *A Semantic Web Primer.* MIT Press, 2004.

[Anup 95] V. Anupam, S. Dar, T. Leibfried, and E. Petajan. "Research report. DataSpace: 3-D visualizations of large databases". *Proceedings of Information Visualization.*, pp. 82–88, 1995.

[Architec] "Architecture of the World Wide Web, Volume One,W3C Recommendation 15 December 2004". `http://www.w3.org/TR/webarch/`.

[Arta 96] A. Artale, E. Franconi, N. Guarino, and L. Pazzi. "Part-Whole Relations in Object-Centered Systems: An Overview.". *Data Knowl. Eng.*, Vol. 20, No. 3, pp. 347–383, 1996.

[Asim 85] D. Asimov. "The grand tour". *SIAM Journal of Scientific and Statistical Computing*, Vol. 6, No. 1, pp. 128–143, 1985.

[Autofocu 04] "Autofocus, Ver. 2004.1". `http://www.aduna-software.com/solutions/autofocus/overview.view`, 2004.

[Baru 06] Z. Barutçuoglu and C. DeCoro. "Hierarchical Shape Classification Using Bayesian Aggregation.". In: *SMI*, p. 44, IEEE Computer Society, 2006.

[Bede 94] B. Bederson and J. Hollan. "Pad++: a zooming graphical interface for exploring alternate interface physics". *Proceedings of the 7th annual ACM symposium on User interface software and technology*, pp. 17–26, 1994.

[Belk 00] N. Belkin. "Helping people find what they don't know". *Commun.ACM*, Vol. 43, pp. 58–61, 2000.

[Belk 82a] N. Belkin, N. Oddy, and M. Brooks. "ASK for Information Retrieval: Part I. Background and Theory". *Journal of Documentation*, Vol. 38, No. 2, 1982.

[Belk 82b] N. Belkin, N. Oddy, and M. Brooks. "ASK for Information Retrieval: Part II. Result of Design Study". *Journal of Documentation*, Vol. 38, No. 3, 1982.

[Bett 00] C. Bettini, S. Jajodia, and S. Wang. *Time Granularities in Databases, Data Mining, and Temporal Reasoning.* Springer, 2000.

[Bier 93] E. Bier, M. Stone, K. Pier, W. Buxton, and T. DeRose. "Toolglass and magic lenses: the see-through interface". *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pp. 73–80, 1993.

[Bitt 03] T. Bittner and J. G. Stell. "Stratified Rough Sets and Vagueness.". In: W. Kuhn, M. F. Worboys, and S. Timpf, Eds., *COSIT*, pp. 270–286, Springer, 2003.

[Born 03] K. Börner, C. Chen, and K. Boyack. "Visualizing knowledge domains". *Annual Review of Information Science and Technology*, Vol. 37, No. 1, pp. 179–255, 2003.

[Brun 97] C. Brunk, J. Kelly, and R. Kohavi. "MineSet: An integrated system for data mining". *Proceedings of the The Third International Conference on Knowledge Discovery and Data Mining, August*, 1997.

[Camo 06] E. Camossi, M. Bertolotto, and E. Bertino. "A multigranular object-oriented framework supporting spatio-temporal granularity conversions.". *International Journal of Geographical Information Science*, Vol. 20, No. 5, pp. 511–534, 2006.

[Card 99] S. K. Card, J. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think. Series in Interactive Technologies*, The Morgan Kaufmann, 1999.

[Carr 97] D. Carr, E. Wegman, and Q. Luo. "ExplorN: Design Considerations Past and Present". *Center for Computation Statistics Technical Report*, Vol. 137, 1997.

[centc 98] "CEN/TC 287 ENV 12657, ENV:Euro-norme Voluntaire for Geographicinformation Data description- Metadata.". http://www.cenorm.be, 1998.

[Cher 73] H. Chernoff. "The Use of Faces to Represent Points in K-Dimensional Space Graphically". *Journal of the American Statistical Association*, Vol. 68, No. 342, pp. 361–368, 1973.

[Chom 01] J. Chomicky and P. Revesz. "Parametric Spatiotemporal Objects". *Periodico Dell'Associazione Italiana per l'Intelligenza Artificiale*, Vol. 14, No. 1, pp. 41–47, 2001.

[Clev 94] W. Cleveland. "Visualizing Data". *TECHNOMETRICS*, Vol. 36, No. 3, 1994.

[Clif 87] N. Cliff. *Analyzing multivariate data.* Harcourt Brace Jovanovich San Diego, Calif, 1987.

[Cox 94] T. Cox and M. Cox. *Multidimensional Scaling.* Chapman & Hall/CRC, 1994.

[Cutt 92] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. "Scatter/Gather: a cluster-based approach to browsing large document collections". In: *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 318–329, ACM Press, New York, NY, USA, 1992.

[dc 06] "Dublin Core Metadata Initiative". http://dublincore.org/, 2006.

[De M 02] M. De Martino, A. Bertone, and R. Albertoni. "Technical Report of Data Mining". Tech. Rep., European Commission, IST-2000-29640, INVISIP Project Derivable 2.2, Bruxelles, 2002.

[Deer 90] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. "Indexing by latent semantic analysis". *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407, 1990.

[DEXAW 20 05] *16th International Workshop on Database and Expert Systems Applications (DEXA 2005), 22-26 August 2005, Copenhagen, Denmark*, IEEE Computer Society, 2005.

[Ehri 05] M. Ehrig, P. Haase, M. Hefke, and N. Stojanovic. "Similarity for Ontologies - A Comprehensive Framework.". In: *ECIS*, 2005.

[EIONET] "EIONET: European Environment Information And Observation Network". `http://www.eionet.eu.it`.

[Erns 05] N. Ernst, M. Storey, and P. Allen. "Cognitive support for ontology modeling". *International Journal of Human-Computer Studies*, Vol. 62, No. 5, pp. 553–577, 2005.

[Euze 04a] J. Euzenat, T. L. Bach, J. Barrasa, P. Bouquet, J. D. Bo, R. Dieng, M. Ehrig, M. Hauswirth, M. Jarrar, R. Lara, D. Maynard, A. Napoli, G. Stamou, H. Stuckenschmidt, P. Shvaiko, S. Tessaris, S. V. Acker, and I. Zaihrayeu. "State of the Art on Ontology Alignment". `http://www.starlab.vub.ac.be/research/projects/knowledgeweb/kweb-223.pdf`, 2004.

[Euze 04b] J. Euzenat and P. Valtchev. "Similarity-Based Ontology Alignment in OWL-Lite.". In: R. L. de Mántaras and L. Saitta, Eds., *ECAI*, pp. 333–337, IOS Press, 2004.

[Falc 04] B. Falcidieno, M. Spagnuolo, P. Alliez, E. Quak, E. Vavalis, and C. Houstis. "Towards the Semantics of Digital Shapes: The AIM@SHAPE Approach.". In: P. Hobson, E. Izquierdo, I. Kompatsiaris, and N. E. O'Connor, Eds., *EWIMT*, QMUL, 2004.

[Falc 98] B. Falcidieno and M. Spagnuolo. "Invited Lecture: A Shape Abstraction Paradigm for Modeling Geometry and Semantics.". In: *Computer Graphics International*, p. 646, IEEE Computer Society, 1998.

[Fayy 96] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in knowledge discovery and data mining.* American Association for Artificial Intelligence Menlo Park, CA, USA, 1996.

[FGDC 98] "FGDC, Document FGDC-STD-001-1998, Content Standard for Digital Geospatial Metadata,". 1998.

[Flui 02a] C. Fluit, M. Sabou, and F. van Harmelen. "Ontology-based Information

Visualization.". In: *Visualizing the Semantic Web*, pp. 36–48, Springer, 2002.

[Flui 02b] C. Fluit and J. Wester. "Using Visualization for Information Management Tasks.". In: *IV*, pp. 447–, 2002.

[FoafUri 06] "Friends Of A Frieds Project (FOAF)". `http://www.foaf-project.org/`, 2006.

[Fons 02] F. T. Fonseca, M. J. Egenhofer, C. A. Davis, and G. Câmara. "Semantic Granularity in Ontology-Driven Geographic Information Systems.". *Ann. Math. Artif. Intell.*, Vol. 36, No. 1-2, pp. 121–151, 2002.

[G 97] H. C. G. "Perception in Visualization". `http://www.csc.ncsu.edu/faculty/healey/PP/PP.html`, 1997.

[Gard 00] P. Gärdenfors. *Conceptual spaces: the geometry of thought.* MIT Press, 2000.

[Gard 04] P. Gärdenfors. "How to Make the Semantic Web More Semantic". *Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004)*, pp. 17–34, 2004.

[Gobe 03] S. Göbel, J. Haist, and U. Jasnoch. "GeoCrystal: graphic-interactive access to geodata archives". *Proceedings of SPIE*, Vol. 4665, p. 391, 2003.

[Gobe 98] S. Göbel and K. Lutze. "Development of Meta Databases for Geospatial Data in the WWW.". In: R. Laurini, K. Makki, and N. Pissinou, Eds., *ACM-GIS*, pp. 94–99, ACM, 1998.

[Gold 94] J. Goldstein and S. Roth. "Using aggregation and dynamic queries for exploring large data sets". *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, pp. 23–29, 1994.

[Gree 02] J. Greenberg. "Metadata and the World Wide Web". *Encyclopedia of Library and Information Science*, Vol. 72, No. 35, pp. 244–261, 2002.

[Grokker 04] "Grotter, Ver. 2.1". `http://www.grokker.com`, 2004.

[Grub 95] T. R. Gruber. "Toward principles for the design of ontologies used for knowledge sharing?". *Int. J. Hum.-Comput. Stud.*, Vol. 43, No. 5-6, pp. 907–928, 1995.

[Guha 03] R. V. Guha, R. McCool, and E. Miller. "Semantic search.". In: *WWW*, pp. 700–709, 2003.

[Guti 00] R. Güting, M. Bhölen, M. Erwig, C. Jensen, N. Lorentzos, M. Shneider, and M. Vazirgiannis. "A Foundation for Representing and Querying Moving Objects". *ACM Transaction On Database Systems*, Vol. 25, pp. 1–42, 2000.

[Guti 05] M. Gutierrez, D. Thalmann, F. Vexo, L. Moccozet, N. Magnenat-Thalmann, M. Mortara, and M. Spagnuolo. "An Ontology of Virtual Humans: incorporating semantics into human shapes". In: *Workshop Towards Semantic Virtual Environments, Villars, CH*, pp. 57–67, 2005.

[Hais 02] J. Haist, S. Göbel, and F. Limbach, T.and Müller. "Technical Report of Metadata Visualization". Tech. Rep., European Commission, IST-2000-29640, INVISIP Project Derivable 2.4, Bruxelles, 2002.

[Hau 05] J. Hau, W. Lee, and J. Darlington. "A Semantic Similarity Measure for Semantic Web Services". *Web Service Semantics Workshop: towards Dynamic Business Integration, Workshop at WWW 05*, 2005.

[Hear 95] M. A. Hearst. "TileBars: Visualization of Term Distribution Information in Full Text Information Access.". In: *CHI*, pp. 59–66, 1995.

[Hend 03] J. Hendler. "Science and the Semantic Web". *Science*, Vol. 299, No. 5606, p. 520, 2003.

[Hierarch 04] "Hierarchical Clustering Explorer, 3.0". `http://www.cs.umd.edu/hcil/multi-cluster/`, 2004.

[Hinn 99] A. Hinneburg, D. Keim, and M. Wawryniuk. "HD-Eye: visual mining of high-dimensional data". *Computer Graphics and Applications, IEEE*, Vol. 19, No. 5, pp. 22–31, 1999.

[Hofm 00] H. Hofmann, A. Siebes, and A. Wilhelm. "Visualizing association rules with interactive mosaic plots". *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 227–235, 2000.

[Hube 85] P. Huber. "Projection Pursuit". *The Annals of Statistics*, Vol. 13, No. 2, pp. 435–475, 1985.

[Inse 90] A. Inselberg and B. Dimsdale. "Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry.". In: *IEEE Visualization*, pp. 361–378, 1990.

[INV] "INVISIP:INformation VIsualization for Site Planning IST-2000-29640". `http://www.invisip.de`.

[ISO19115 03] "International Organization for Standardization, Technical Committee 21, Geographic information/ Geomatics. ISO 19115:2003,Geographic Information- Metadata6". 2003.

[Jambalay 04] "Jambalaya, Ver.2 2004". `http://www.thechiselgroup.org/`, 2004.

[John 91] B. Johnson and B. Shneiderman. "Tree maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures.". In: *IEEE Visualization*, pp. 284–291, 1991.

[Jone 02] C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. J. van Kreveld, and R. Weibel. “Spatial information retrieval and geographical ontologies an overview of the SPIRIT project.”. In: *SIGIR*, pp. 387–388, ACM, 2002.

[Kartoo 04] “Kartoo, Ver. 4”. `http://www.kartoo.com`, 2004.

[Kash 96] V. Kashyap and A. P. Sheth. “Semantic and Schematic Similarities Between Database Objects: A Context-Based Approach.”. *VLDB J.*, Vol. 5, No. 4, pp. 276–304, 1996.

[Keim 00] D. A. Keim. “Designing Pixel-Oriented Visualization Techniques: Theory and Applications.”. *IEEE Trans. Vis. Comput. Graph.*, Vol. 6, No. 1, pp. 59–78, 2000.

[Keim 02a] D. Keim, W. Müller, and H. Schumann. “Information Visualization and Visual Data Mining; State of the art report”. *Eurographics 2002*, 2002.

[Keim 02b] D. A. Keim. “Information Visualization and Visual Data Mining.”. *IEEE Trans. Vis. Comput. Graph.*, Vol. 8, No. 1, pp. 1–8, 2002.

[Keim 06] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. “Challenges in Visual Data Analysis.”. In: *IV*, pp. 9–16, IEEE Computer Society, 2006.

[Keim 95] D. A. Keim, M. Ankerst, and H.-P. Kriegel. “Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data.”. In: *IEEE Visualization*, pp. 279–, 1995.

[Keim 96] D. Keim and H. Kriegel. “Visualization techniques for mining large databases: a comparison”. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, pp. 923–938, 1996.

[Kim 03] G. Kim and T. Hung. “Information needs in information space”. *Lecture Notes in Computer Science*, Vol. 2911 / 2003, pp. 123–133, 2003. It cites works where Information spaces and Cognitive Spaces are defined.

[Klei 02] P. Klein, F. Müller, H. Reiterer, and M. Eibl. “Visual Information Retrieval with the SuperTable + Scatterplot.”. In: *IV*, pp. 70–75, 2002.

[Klei 03] P. Klein, H. Reiterer, F. Müller, and T. Limbach. “Metadata Visualisation with VisMeB.”. In: E. Banissi, K. Börner, C. Chen, G. Clapworthy, C. Maple, A. Lobben, C. J. Moore, J. C. Roberts, A. Ursyn, and J. Zhang, Eds., *IV*, pp. 600–605, IEEE Computer Society, 2003.

[Koho 97] T. Kohonen. *Self-organizing maps*. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 1997.

[Kola 01] E. Kolatch and B. Weinstein. “CatTrees: Dynamic visualization of categorical data using Treemaps.”. `http://www.cs.umd.edu/class/spring2001/cmsc838b/project`, 2001.

[Kreu 02] M. Kreuseler and H. Schumann. “A Flexible Approach for Visual Data Mining.”. *IEEE Trans. Vis. Comput. Graph.*, Vol. 8, No. 1, pp. 39–51, 2002.

[Kuhn 03] W. Kuhn, M. F. Worboys, and S. Timpf, Eds. *Spatial Information Theory. Foundations of Geographic Information Science, International Conference, COSIT 2003, Ittingen, Switzerland, September 24-28, 2003, Proceedings*, Springer, 2003.

[Lamp 95] J. Lamping, R. Rao, and P. Pirolli. “A focus+ context technique based on hyperbolic geometry for visualizing large hierarchies”. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 401–408, 1995.

[Land 98] T. Landauer, P. Foltz, and D. Laham. “An introduction to latent semantic analysis”. *Discourse Processes*, Vol. 25, No. 2-3, pp. 259–284, 1998.

[Lass 03] O. Lassila and D. McGuinness. “The Role of Frame-Based Representation on the Semantic Web”. *Electronic Transactions on Artificial Intelligence*, 2003.

[Leun 94] Y. Leung and M. Apperley. “A review and taxonomy of distortion-oriented presentation techniques”. *ACM Transactions on Computer-Human Interaction (TOCHI)*, Vol. 1, No. 2, pp. 126–160, 1994.

[Li 03] Y. Li, Z. Bandar, and D. McLean. “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources.”. *IEEE Trans. Knowl. Data Eng.*, Vol. 15, No. 4, pp. 871–882, 2003.

[Limb 02] T. Limbach. “Evaluation Concept”. Tech. Rep., European Commission, IST-2000-29640, INVISIP Project Derivable 7.1, Bruxelles, 2002.

[Limb 03a] T. Limbach. “Evaluation Demonstrator 1”. Tech. Rep., European Commission, IST-2000-29640, INVISIP Project Derivable 7.2, Bruxelles, 2003.

[Limb 03b] T. Limbach, H. Reiterer, P. Klein, and F. Müller. “VisMeB: A Visual Metadata Browser.”. In: M. Rauterberg, M. Menozzi, and J. Wesson, Eds., *INTERACT*, IOS Press, 2003.

[Limb 04] T. Limbach. “Evaluation Demonstrator 2”. Tech. Rep., European Commission, IST-2000-29640, INVISIP Project Derivable 7.3, Bruxelles, 2004.

[Lin 95] C. Y. Lin. “Knowledge-based automatic topic identification”. In: *Proc. of the 33rd Annual Meeting on Association for Computational Linguistics*, pp. 308–310, Association for Computational Linguistics, 1995.

[Lin 98] D. Lin. “An Information-Theoretic Definition of Similarity.”. In: J. W. Shavlik, Ed., *ICML*, pp. 296–304, Morgan Kaufmann, 1998.

[Mack 86] J. Mackinlay. “Automating the design of graphical presentations of relational information”. *ACM Trans. Graph.*, Vol. 5, No. 2, pp. 110–141, 1986.

[Maed 02] A. Maedche and V. Zacharias. "Clustering Ontology-Based Metadata in the Semantic Web.". In: T. Elomaa, H. Mannila, and H. Toivonen, Eds., *PKDD*, pp. 348–360, Springer, 2002.

[Mant 88] M. Mäntylä. *Introduction to Solid Modeling.* WH Freeman & Co. New York, NY, USA, 1988.

[Mart 03] H. Marti. "Information Visualization: Principles, Promise, and Pragmatics". 2003.

[McGu 03] D. L. McGuinness. "Ontologies Come of Age.". In: D. Fensel, J. A. Hendler, H. Lieberman, and W. Wahlster, Eds., *Spinning the Semantic Web*, pp. 171–194, MIT Press, 2003.

[Meer 05] R. Meersman, Z. Tari, P. Herrero, G. Méndez, L. Cavedon, D. Martin, A. Hinze, G. Buchanan, M. S. Pérez, V. Robles, J. Humble, A. Albani, J. L. G. Dietz, H. Panetto, M. Scannapieco, T. A. Halpin, P. Spyns, J. M. Zaha, E. Zimányi, E. Stefanakis, T. S. Dillon, L. Feng, M. Jarrar, J. Lehmann, A. de Moor, E. Duval, and L. Aroyo, Eds. *On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops, OTM Confederated International Workshops and Posters, AWeSOMe, CAMS, GADA, MIOS+INTEROP, ORM, PhDS, SeBGIS, SWWS, and WOSE 2005, Agia Napa, Cyprus, October 31 - November 4, 2005, Proceedings*, Springer, 2005.

[Mend 97] E. Mendelson. *Introduction to Mathematical Logic.* Chapman & Hall/CRC, 1997.

[mpeg7 04] "MPEG-7 Overview (version 10)". `http://www.chiariglione.org/MPEG/standards/mpeg-7/mpeg-7.htm`, 2004.

[Muse 01] M. Musen, R. Fergerson, N. Noy, and M. Crubezy. "Protege-2000: A plug-in architecture to support knowledge acquisition, knowledge visualization, and the semantic Web". *J. Am. Med. Inform. Assoc*, pp. 1079–1079, 2001.

[Newb 01] G. Newby. "Cognitive space and information space". *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 12, pp. 1026–1048, 2001.

[OntoViz 04] "OntoViz, Ver.2 2004". `http://protege.cim3.net/cgi-bin/wiki.pl?OntoViz`, 2004.

[OWL 06] "Web Ontology Language (OWL)". `http://www.w3.org/2004/OWL/`, 2006.

[Papa 05] L. Papaleo, R. Albertoni, S. Marini, and F. Robbiano. "An ontology-based approach to Acquisition and reconstruction". In: *Workshop Towards Semantic Virtual Environments,Villars, CH*, 2005.

[Peat 91] H. J. Peat and P. Willett. "The limitations of term co-occurrence data for query expansion in document retrieval systems.". *JASIS*, Vol. 42, No. 5, pp. 378–383, 1991.

[Pick 88] R. Pickett and G. Grinstein. "Iconographic Displays For Visualizing Multidimensional Data". *Systems, Man, and Cybernetics, 1988. Proceedings of the 1988 IEEE International Conference on*, Vol. 1, 1988.

[Pike 03] W. Pike and M. Gahegan. "Constructing Semantically Scalable Cognitive Spaces.". In: W. Kuhn, M. F. Worboys, and S. Timpf, Eds., *COSIT*, pp. 332–348, Springer, 2003.

[Rada 89] R. Rada, H. Mili, E. Bicknell, and M. Blettner. "Development and application of a metric on semantic nets". *Systems, Man and Cybernetics, IEEE Transactions on*, Vol. 19, No. 1, pp. 17–30, 1989.

[Rao 94] R. Rao and S. Card. "The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information". *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, pp. 318–322, 1994.

[RDF 06] "Resource Description Framework (RDF)". `http://www.w3.org/RDF/`, 2006.

[Requ 80] A. A. G. Requicha. "Representations for Rigid Solids: Theory, Methods, and Systems.". *ACM Comput. Surv.*, Vol. 12, No. 4, pp. 437–464, 1980.

[Resn 95] P. Resnik. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy.". In: *IJCAI*, pp. 448–453, 1995.

[Robe 91] G. Robertson, J. Mackinlay, and S. Card. "Cone Trees: animated 3D visualizations of hierarchical information". *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, pp. 189–194, 1991.

[Rodr 03] M. A. Rodríguez and M. J. Egenhofer. "Determining Semantic Similarity among Entity Classes from Different Ontologies.". *IEEE Trans. Knowl. Data Eng.*, Vol. 15, No. 2, pp. 442–456, 2003.

[Rodr 04] M. A. Rodríguez and M. J. Egenhofer. "Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure.". *International Journal of Geographical Information Science*, Vol. 18, No. 3, pp. 229–256, 2004.

[Rosa 03] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, and M. O. Ward. "Mapping Nominal Values to Numbers for Effective Visualization.". In: *INFOVIS*, IEEE Computer Society, 2003.

[Rosa 04] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, M. O. Ward, and S. Huang. "Mapping nominal values to numbers for effective visualization.". *Information Visualization*, Vol. 3, No. 2, pp. 80–95, 2004.

[Russ 02] S. Russell and P. Norvig. *Artificial intelligence: a modern approach.* Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 2002.

[Salt 75] G. Salton, A. Wong, and C. S. Yang. “A vector space model for automatic indexing”. *Commun. ACM*, Vol. 18, No. 11, pp. 613–620, 1975.

[Scho 98] H. Scholten, A. LoCashio, and B. Bonn. “ESMI, towards a European Spatial Metadata Infrastructure’, presented by Henk Scholten at EOGEO’98”. *ESMI webserver: http://www. geodan. nl/esmi*, 1998.

[Schw 05a] A. Schwering. “Hybrid Model for Semantic Similarity Measurement.”. In: R. Meersman, Z. Tari, M.-S. Hacid, J. Mylopoulos, B. Pernici, Ö. Babaoglu, H.-A. Jacobsen, J. P. Loyall, M. Kifer, and S. Spaccapietra, Eds., *OTM Conferences (2)*, pp. 1449–1465, Springer, 2005.

[Schw 05b] A. Schwering and M. Raubal. “Measuring Semantic Similarity Between Geospatial Conceptual Regions.”. In: M. A. Rodríguez, I. F. Cruz, M. J. Egenhofer, and S. Levashkin, Eds., *GeoS*, pp. 90–106, Springer, 2005.

[Shet 02] A. P. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, and Y. S. Warke. “Managing Semantic Content for the Web.”. *IEEE Internet Computing*, Vol. 6, No. 4, pp. 80–87, 2002.

[Shet 05] A. P. Sheth, C. Ramakrishnan, and C. Thomas. “Semantics for the Semantic Web: The Implicit, the Formal and the Powerful.”. *Int. J. Semantic Web Inf. Syst.*, Vol. 1, No. 1, pp. 1–18, 2005.

[Shne 92] B. Shneiderman. “Tree Visualization with Tree-Maps: 2-d Space-Filling Approach.”. *ACM Trans. Graph.*, Vol. 11, No. 1, pp. 92–99, 1992.

[Sici 06] M. A. Sicilia. “Metadata, semantics, and ontology: providing meaning to information resources”. *International Journal of Metadata, Semantics and Ontologies*, 2006.

[Smit 02] P. Smith, U. Düren, O. Ostensen, L. Murre, M. Gould, U. Sandgren, M. Marinelli, K. Murray, E. Pross, A. Wirthmann, F. Salgé, and M. Konecky. “INSPIRE Architecture and Standard Position Paper”. 2002.

[Spin 01] A. Spink, D. Wolfram, M. Jansen, and T. Saracevic. “Searching the web: The public and their queries”. *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 3, pp. 226–234, 2001.

[Spoe 04a] A. Spoerri. “Coordinated views and tight coupling to support meta searching”. In: *Proceedings of Second International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pp. 39–48, 2004.

[SPOE 04b] A. SPOERRI. “Toward Enabling Users TO Visually Evaluate THE Effectiveness OF Different Search Methods”. *Journal of Web Engineering*, Vol. 3, No. 3&4, pp. 297–313, 2004.

[Spoe 04c] A. Spoerri. “How Visual Query Tools Can Support Users Searching the Internet.”. In: *IV*, pp. 329–334, IEEE Computer Society, 2004.

[Spoe 04d] A. Spoerri. "Metacrystal: visualizing the degree of overlap between different search engines.". In: S. I. Feldman, M. Uretsky, M. Najork, and C. E. Wills, Eds., *WWW (Alternate Track Papers & Posters)*, pp. 378–379, ACM, 2004.

[Spoe 93] A. Spoerri. "InfoCrystal: A Visual Tool for Information Retrieval.". In: G. M. Nielson and R. D. Bergeron, Eds., *IEEE Visualization*, pp. 150–157, IEEE Computer Society, 1993.

[Staa 05] S. Staab, P. Domingos, P. Mika, J. Golbeck, L. Ding, T. W. Finin, A. Joshi, A. Nowak, and R. R. Vallacher. "Social Networks Applied.". *IEEE Intelligent Systems*, Vol. 20, No. 1, pp. 80–93, 2005.

[Stas 98] J. Stasko. *Software Visualization: Programming as a multimedia experience.* MIT Press, 1998.

[Stei 97] D. Stein. "Geospatial Data Sharing through the Exploitation of Metadata". *ESRI International User Conference, San Diego, California, July*, pp. 8–11, 1997.

[Stel 98] J. Stell and M. Worboys. "Stratified map spaces: A formal basis for multiresolution spatial databases". *SDH*, Vol. 98, pp. 180–189, 1998.

[Stol 02] C. Stolte, D. Tang, and P. Hanrahan. "Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases.". *IEEE Trans. Vis. Comput. Graph.*, Vol. 8, No. 1, pp. 52–65, 2002.

[Stuc 04] H. Stuckenschmidt, F. van Harmelen, A. de Waard, T. Scerri, R. Bhogal, J. van Buel, I. Crowlesmith, C. Fluit, A. Kampman, J. Broekstra, and E. M. van Mulligen. "Exploring Large Document Repositories with RDF Technology: The DOPE Project.". *IEEE Intelligent Systems*, Vol. 19, No. 3, pp. 34–40, 2004.

[Sure 04] Y. Sure, S. Staab, and R. Studer. "On-To-Knowledge Methodology (OTKM).". In: S. Staab and R. Studer, Eds., *Handbook on Ontologies*, pp. 117–132, Springer, 2004.

[Sway 92] D. Swayne, D. Cook, and A. Buja. "Users Manual for XGobi, a Dynamic Graphics Program for Data Analysis". *Bellcore, Morristown, NJ*, 1992.

[Swob 99] W. Swoboda, F. Kruse, R. Nikolai, W. Kazakos, D. Nyhuis, and H. Rousselle. "The UDK Approach: the 4th Generation of an Environmental Data Catalogue Introduced in Austria and Germany". *Proceedings of the Third IEEE Meta-Data Conference, Bethesda, Maryland, USA URL: http://computer. org/conferen/proceed/meta/1999/papers/45/wswoboda.html*, 1999.

[Terg 05] S.-O. Tergan and T. Keller, Eds. *Knowledge and Information Visualization, Searching for Synergies [outcome of a workshop held in Tübingen, Germany, May 2004].*, Springer, 2005.

[TestOnto] "TestOntology". `http://www.ge.imati.cnr.it/ima/personal/albertoni/odbase06p.owl`.

[TGVizTab 04] "TGVizTab". `http://www.ecs.soton.ac.uk/~ha/TGVizTab`, 2004.

[Thei 98] H. Theisel and M. Kreuseler. "An Enhanced Spring Model for Information Visualization". *Computer Graphics Forum*, Vol. 17, No. 3, pp. 335–344, 1998.

[Tier 90] L. Tierney. *LISP-STAT: an object oriented environment for statistical computing and dynamic graphics.* Wiley-Interscience New York, NY, USA, 1990.

[Tver 77] A. Tversky *et al.* "Features of similarity". *Psychological Review*, Vol. 84, No. 4, pp. 327–352, 1977.

[Ucel 05] G. Ucelli, R. De Amicis, G. Conti, G. Brunetti, and A. Stork. "Shape Semantics and Content Management for Industrial Design and Virtual Styling". In: *Workshop Towards Semantic Virtual Environments, Villars, CH*, pp. 127–137, 2005.

[URI 06] "Naming and addressing:URIs,URLs,..". `http://www.w3.org/Addressing/`, 2006.

[Usan 05] S. Usanavasin, S. Takada, and N. Doi. "Semantic Web Services Discovery in Multi-ontology Environment.". In: R. Meersman, Z. Tari, P. Herrero, G. Méndez, L. Cavedon, D. Martin, A. Hinze, G. Buchanan, M. S. Pérez, V. Robles, J. Humble, A. Albani, J. L. G. Dietz, H. Panetto, M. Scannapieco, T. A. Halpin, P. Spyns, J. M. Zaha, E. Zimányi, E. Stefanakis, T. S. Dillon, L. Feng, M. Jarrar, J. Lehmann, A. de Moor, E. Duval, and L. Aroyo, Eds., *OTM Workshops*, pp. 59–68, Springer, 2005.

[Varz 96] A. C. Varzi. "Parts, Wholes, and Part-Whole Relations: The Prospects of Mereotopology.". *Data Knowl. Eng.*, Vol. 20, No. 3, pp. 259–286, 1996.

[Wang 02] H. Wang, W. Wang, J. Yang, and P. Yu. "Clustering by pattern similarity in large data sets". *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pp. 394–405, 2002.

[Ward 94] M. Ward. "XmdvTool: integrating multiple methods for visualizing multivariate data". *Visualization, 1994., Visualization'94, Proceedings., IEEE Conference on*, pp. 326–333, 1994.

[Welt 01] C. A. Welty and N. Guarino. "Supporting ontological analysis of taxonomic relationships.". *Data Knowl. Eng.*, Vol. 39, No. 1, pp. 51–74, 2001.

[Whit 02] M. A. Whiting and N. Cramer. "WebTheme: Understanding Web Information through Visual Analytics.". In: I. Horrocks and J. A. Hendler, Eds., *International Semantic Web Conference*, pp. 460–468, Springer, 2002.

[Wins 87] M. Winston, R. Chaffin, and D. Herrmann. "A Taxonomy of Part-Whole Relations". *Cognitive Science*, Vol. 11, No. 4, pp. 417–444, 1987.

[XML 06] “Extensible Markup Language (XML)”. `http://www.w3.org/XML/`, 2006.