Ph.D. in Electronic and Computer Engineering
Dept. of Electrical and Electronic Engineering
University of Cagliari

Knowledge Management: the Issue of Multimedia Contents

Filippo Eros Pani

Advisor: Prof. Giulio Concas
Curriculum: ING-INF/05 SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

XXV Cycle
April 2013

Ph.D. in Electronic and Computer Engineering
Dept. of Electrical and Electronic Engineering
University of Cagliari

Knowledge Management: the Issue of Multimedia Contents

Filippo Eros Pani

Advisor: Prof. Giulio Concas
Curriculum: ING-INF/05 SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

XXV Cycle
Aprile 2013

Dedicated to the most important

people in my life

I would like to thank Prof. Giulio Concas, Prof. Michele Marchesi and the Agile Group of University of Cagliari for their valuable support and encouragement during the writing of this thesis.

Contents

List of Figures

List of Tables

List of Acronyms

AI Adobe Illustrator

AIS Associazione Italiana Sommelier (Italian Sommelier Association)

API Application Programming Interface

ASAS Analytic Sound Archive of Sardinia

ASF Advanced Systems Format

AVCHD Advanced Video Codec High Definition

AVI Audio Video Interleave

BSD Berkeley Software Distribution

BU Bottom-Up

CMS Content Management System

DAML DARPA Agent Markup Language

DARPA Defense Advanced Research Projects Agency

DC Dublin Core

DCF Design rule for Camera File system

DIN Deutsches Institut für Normung

DL Description Logic

DN Digital Negative

EP Elementary Product

ES Encapsulated PostScript

Exif Exchangeable image file format

FV Flash Video

GIF Graphics Interchange Format

GPS Global Positioning System

HDV High Definition Video

HP Hewlett-Packard

HTML HyperText Markup Language

HTTP HyperText Transfer Protocol

ICT Information and Communication Technology

IIM Information Interchange Model

INDD INDesign Document

INDT INDesign Template

IPA International Phonetic Alphabet

IPTC International Press Telecommunication Council

IR Institutional Repository

ISO International Standard Organization

JPEG Joint Photographic Experts Group

KB Knowledge Base

KM Knowledge Management

KMS Knowledge Management System

MDS Multimedia Description Schemes

MIR Music Information Retrieval

MIT Massachussetts Institute of Technology

MMS Multimedia Messaging Service

MOM Multimedia Ontology Manager

MPEG Moving Picture Experts Group

MWG Metadata Working Group

NAA Newspaper Association of America

OAI-PMH Open Archives Initiative - Protocol for Metadata Harvesting

OWL Ontology Web Language

P2 Professional Plug-in

PBS Project Breakdown Structure

PDF Portable Document Format

PNA Personal Navigator Assistant

PNG Portable Network Graphics

POI Point Of Interest

PS PostScript

PSD PhotoShop Document

PVC PolyVinyl Chloride

RDF Resource Description Framework

RSS Really Simple Syndication

SWF ShockWave Flash

TD Top-Down

TIFF Tagged Image File Format

UCF Universal Container Format

UGC User Generated Content

URI Uniform Resource Identifier

URL Uniform Resource Locator

W3C World Wide Web Consortium

WAV WAVeform audio file format

WMA Windows Media Audio

WMV Windows Media Video

XML eXtensible Markup Language

Chapter 1

Introduction

1.1 Knowledge Management

All the information of the real world belong to a well-defined structure which human beings have determined and made up in thousands of years. Our brain links one or more meanings to each single word; these can change according to the context and the sentence in which the word can be found and used.

Given their nature, human beings need to classify, define and schematize every kind of information, also those ones linked to a single word, to establish an order and accelerate the research. This need brought to the developing of always more complex tools which automatically try to simulate human's reasoning and learning. The concept of *Knowledge Management* (KM) came into existence: the term was coined in the first decade of the 1970s, with the aim of indicating the set of strategies adopted by companies to identify, create, represent and arrange knowledge - this one seen both as baggage of information owned by each person belonging to the company, and as a set of processes and practices carried out by the company itself in its daily activities [1]. The KM is about those organizational processes which join human beings' potentiality of combining data and elaborating information to creativity and the ability to innovate.

KM derives from 50 years of research in the field of Artificial Intelligence (AI). In the summer of 1956 Simon, Minsky, Shannon and other researchers met in New Hampshire to discuss the possibility of simulating humans' learning and reasoning, using recently invented machines: computers. With AI the first studies about knowledge - and especially its representation - were born, producing a series of formalisms which gave the bases for the most recent applications of KM techniques in the field of business and the even more recent "vision" of Semantic Web [2][3].

KM is a very important topic in business and in academy research [4][5]. There are many fields of applications for KM, including Cognitive Science, Sociology, Management Science, Information Science, Knowledge Engineering, AI and Economics [6][7][8]. Many studies on different aspects of KM have been published, becoming common in the early 1990s [9][10][11].

There are different professions interested in KM which want to present and interpret what KM is about from their own perspective; these people also want to define the future direction of KM as it fits the traditions and perspectives of their own profession [6][12][13][14]

[15][16]. The different professions interested in knowledge study and analyze it basing themselves on their own perspective [6][7][13][15][17][18]. This spread approach to the problem makes it difficult to find a universal consensus on some of the key issues of KM, including conceptualizations, processes, goals and purposes of KM [16][17][18][19][20][21][22][23][24] [25][26].

There is no universally accepted definition of KM. Daykir [6] says: *"An informal survey identified over 100 published definitions of knowledge management, and of these, at least 72, for their purpose, could be considered very good! Clearly, KM is a multidisciplinary field of study that covers a wide ground. This finding should not be surprising, for applying knowledge to work is integral to most business activities. However, the field of KM does suffer from the Three Blind Men and an Elephant syndrome. In fact, there are probably more than three distinct perspectives on KM, and each leads to a different extrapolation and a different definition."*

KM represents the methodology of managing knowledge, namely methods and software tools that allow for identifying and making use of knowledge so that it may be properly organized and shared.

The processes that create KM are:

1 acquisition;

2 representation;

3 elaboration;

4 sharing;

5 usage of knowledge.

Knowledge can therefore be seen, from an operational point of view, as a valid certainty which improves the abilities of a man to undertake efficient actions.

In the Information Technology context a definition is adopted more frequently. Knowledge is considered as the information stored in human minds, thus an interpreted and subjective information concerning facts, procedures, concepts. Therefore, knowledge is not radically different from information but, as for this definition, we could say that information becomes knowledge when it is processed by a person's mind.

The purpose of KM is, therefore, to keep at the whole company's disposal all the competences acquired by each of its member, so that knowledge becomes a shared, usable and protected over time "asset". An opportunely managed knowledge can be used to easily find answers to problems already dealt by other employees, and for which the company has already invested resources, as well as to draw information to be addressed to new members' education [5].

Just a few years after its birth, KM drew the attention of the academic world, becoming a matter of intense study as the scientific community has always felt the need of making its members cooperate through the exchange and the reciprocal fruition of information.

From our point of view KM can be initially defined as the process of applying a systematic approach to capture, structure and manage knowledge, and to make it available for sharing and reuse [27][28][29][30].

Many approaches to information tend to use sophisticated search engines to retrieve the content. KM solutions have demonstrated to be the most successful in capturing, storing, and consequently making available the knowledge that has been rendered explicit, particularly lessons learned and best practices. Many KM efforts have been largely concerned with capturing, codifying, and sharing the knowledge in organizations.

We want to study a process to identify and locate knowledge and knowledge sources within the domain, paying attention to multimedia objects. Valuable knowledge is then translated into explicit form through formalization and codification of knowledge, in order to facilitate the availability of knowledge. We start from the concept of the *domain knowledge base*. The fundamental body of knowledge available on a domain is the knowledge valuable for the knowledge users. We need to represent and manage this knowledge, to define a formalization and codification of the knowledge in the domain. After this formalization we can manage this knowledge using knowledge repositories.

1.2 The approach to knowledge formalization

In recent years the development of models to formalize the knowledge has been studied and analyzed. The ontologies - explicit formal specifications of the terms in the domain and relations among them [31] - take an important part in these formalization approaches. Ontologies have become common on the World Wide Web at the end of 2000. In the Web range there are many directory services of Web sites: the most famous is Yahoo. These directory services are large taxonomies which organize Web sites in categories. Other systems categorize products for e-commerce purpose: the most famous is Amazon. They use an implicit taxonomy to organize the products for sale by type and features. The World Wide Web Consortium W3C[1] has developed the Resource Description Framework (RDF) [32], a language for encoding knowledge on Web pages to make it understandable to electronic agents searching for information, as main foreground concept of the Semantic Web. The Defense Advanced Research Projects Agency (DARPA)[2], in cooperation with the W3C, has developed DARPA Agent Markup Language (DAML) by extending RDF with more expressive constructs aimed at facilitating agent interaction on the Web [33].

Many disciplines develop standardized formalization of the knowledge which domain experts can use to share information in the form of reusable knowledge.

Many people use ontology [34]: *”to define a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them.*

Why would someone want to develop an ontology? Some of the reasons are:

- *To share common understanding of the structure of information among people or software agents*
- *To enable reuse of domain knowledge*
- *To make domain assumptions explicit*
- *To separate domain knowledge from the operational knowledge*
- *To analyze domain knowledge*

[1] World Wide Web Consortium (W3C), www.w3.org

[2] Defense Advanced Research Projects Agency (DARPA), http://www.darpa.mil

Sharing common understanding of the structure of information among people or software agents is one of the most common goals in developing ontologies [31][35]. For example, suppose that several different Web sites contain medical information or provide medical e-commerce services. If these Web sites share and publish the same underlying ontology of the terms they all use, then computer agents can extract and aggregate information from these different sites. The agents can use this aggregated information to answer user queries or as input data to other applications.“

Enabling reuse of domain knowledge was one of the driving forces behind our studies. Analyzing domain knowledge is possible once a formal specification of the terms and their structure is available.

The basic concepts of our approach are the following:

1 there is no “correct“ way or methodology to develop ontologies [34] and in general to analyze and codify/formalize knowledge;

2 the main goal is to make the knowledge of a specific domain available and reusable for specific purpose;

3 the formalized knowledge is not all the knowledge in the domain, but only the interesting information for the specific problem. We don’t want to formalize the world;

4 not-formalized information have to be inventoried (they can be included in the multimedia objects);

5 the formalized information will be represented using metadata;

6 the structure of the knowledge of the specific domain has to be analyzed using a top-down approach;

7 the information in the object of the specific domain have to be analyzed using a bottom-up approach;

8 the analysis process will be iterative mixing top-down and bottom-up approaches.

In fact, we want to represent knowledge through a mixed-iterative approach, where top-down and bottom-up analyses of the knowledge domain which has to be represented are applied: these are typical approaches for this kind of problems. In this case, they are applied following an iterative approach which allows, through further refinements, for the efficient formalization able to represent the domain’s knowledge of interest.

Top-down phase

When our knowledge or our expectations are influenced by perception, we refer to schema-driven or top-down (TD) elaboration. A schema is a model formerly created by our experience. More general or abstract contents are indicated as higher level, while concrete details are indicated as lower level.

The TD elaboration happens whenever a higher level concept influences the interpretation of lower level information. Generally, the TD process is an information process based on former knowledge or acquired mental schemes; it allows us to make inferences: to “perceive“ or “know“ more than what can be found in data. TD methodology starts, therefore, by identifying a target to reach, and then pinpoints the strategy to use in order to reach the established goal.

Our aim is, therefore, to begin by formalizations of the reference knowledge (ontology, taxonomy, metadata schema) to start classifying the information on the reference domain.

The model could be, for instance, a formalization of one or more classifications of the same domain, formerly made in a logic of metadata. Therefore, the output of this phase will be a table with all the elements of knowledge formalized through the definition of the reference metadata.

Bottom-up

With this phase the knowledge to be represented is analyzed by pinpointing, among the present information, the ones which are to be represented together with a reference terminology for data description.

When an interpretation emerges from data, it is called data-driven or bottom-up (BU) elaboration. Perception is mainly data-driven, as it must precisely reflect what happens in the external world. Generally, it is better if the interpretation coming from a system of information is determined by what is effectively transmitted at sensory level rather than what is perceived as an expectation. Applying this concept, we analyze the interesting object in the domain (multimedia objects, Web sites, documents, annotations) containing the information of the domain of interest, both information whose structure needed to be extrapolated and the information in them were pinpointed. Typically, reference objects for that information domain are selected, namely the ones which users mainly use to find information of their interest over the domain itself.

Primary information, important ones, already emerge during the phase of objects analysis and gathering: during a first skimming phase, the minimum, basic information necessary to well describe our domain can be noticed. Then, important information are extrapolated by choosing fields or keywords which best represent the knowledge in order to create a knowledge base (KB). In this phase, one of the limits could be the creation of the KB itself, because each object can have a different structure and a different way of presenting the same information. Therefore, it will be necessary to pinpoint the present information of interest, defining and outlining them.

Iterations of phases

After these analyses of gathering of information, a classification is made and it has to reflect, in the most faithful way, the structure of the knowledge in itself, respecting both its contents and hierarchy.

In this phase we will try to reconcile these two representations of knowledge of the domain, as represented in the previous phases.

Thus, we want to pinpoint, for each single TD's metadata, where the information can be found in the metadata representing the knowledge of each object (which, for us, represents the knowledge we want to represent, considering the semantic concept and not the way to represent it, absolutely subjective for every knowledge object).

Starting from this KB, further iterative refining can be made by re-analyzing the information in different phases:

1 with a TD approach, checking if the information which are not represented by the chosen formalization can be formalized;

2 with a BU approach, analyzing if some information of the Web sites can be connected to formalized items;

3 with the iterations of phases by which these concepts are reconciled.

This is obviously made only for the information to be represented. The knowledge we want to represent is the one considered of interest by the users for the domain: for this reason, the most important are chosen.

At the end of this analysis we will define a formalization, in form of ontologies, taxonomies, metadata schema, able to represent the knowledge of interest for this domain.

The final result of these phases will be a formalized knowledge able to be represented, reused and managed trough knowledge management repositories, where the knowledge of interest is available.

1.3 Outline

Then, in this thesis we present four different formalization and management of knowledge for multimedia contents, using our proposed approach:

1 User-Generated Contents (UGCs) from famous platform (Flickr, YouTube, etc.);

2 audio recordings regarding linguistic corpus and information added to that corpus with annotations;

3 knowledge associated with construction processes;

4 descriptions and reviews of Italian wines.

This thesis is organized as follows:

- in Chapter 2, we propose an approach for the management and categorization of UGC which comes from different sources such as popular digital platforms. This study has been described in some scientific publications (see [36], and [37]). The diffusion and rise in popularity of software platforms for the UGC management, especially multimedia objects. These platforms handle a large amount of unclassified information. UGC Web sites (e.g. YouTube and Flickr) do not force the users to perform classification operations and metadata definitions, leaving space to a logic of free-tags (folksonomies). We analyzed the standards used in UGC Web sites for the management of the multimedia contents and their metadata. We defined an ontology to represent the semantics of these multimedia contents, so that in turn the metadata classification can give an unambiguous meaning. In order to unify metadata coming from different sources we defined all rules of mapping towards a structure defined by sources such as YouTube and Flickr. The innovation is in the approach for the formalization of Web semantics for multimedia content: we used standards such as Dublin Core, Exif, IPTC and in particular the Adobe XMP standard as a starting point of this domain. With the proposed approach, one can categorize and catalog all non-standard and unclassifiable information inside the ontology, using pre-made schemas;

- in Chapter 3, we propose an approach for formalization and management of knowledge, in the domain of audio recordings in a linguistic corpus. This study has been described in some scientific publications (see [38], [39], and [40]. Within the Analytic Sound Archive of Sardinia (ASAS) project, which aims to create an institutional archive with a linguistically and musically annotated electronic corpus, this work proposes a new approach for KM. Our KB is represented not only by the texts of the corpus, but especially by the metalanguage and linguistic annotations that enhance them. For each audio clip, a set of metainformation describing the content is needed in order to enable the search and retrieval of data by local author, title, date of recording, to more particular features like linguistic variety or singing type. Each audio clip is also enhanced by a set of linguistic annotations. The purpose of this study is to offer an original way to associate multilevel musical and linguistic annotations (information associated to specific text portions) to the corpus by treating them as metadata;

- in Chapter 4, we analyze the elements constituting the construction process. The growing complexity of the construction sector - due to the proliferation of products, techniques and needs related to side, not secondary, aspects of objects (environmental impact, energy efficiency, durability, safety, etc.) - shows that the current management styles in construction processes are no longer appropriate to their context. Therefore, the construction sector faces an inevitable process of growth in which knowledge is an indispensable resource. The purpose of this case study is to show how knowledge associated with construction processes can be represented using KM techniques. The analysis of such knowledge uses a mixed TD and BU approach, which can formalize it and make it ready for an easy access and search. The underlying goal is the rational organization of large amounts of data using the knowledge that characterizes the various stages of a construction process. Elementary Products could be the core concepts that can group the objects associated with such process, guiding the management of relevant information and knowledge involved in construction processes. The formalization was used to define a prototype implementation of the Knowledge Management System using DSpace;

- in Chapter 5, we analyze the knowledge on the domain of descriptions and reviews of Italian wines which can be found on the Internet. The knowledge we want to represent is the one considered of interest by the users for the domain: for this reason, the most important and looked up Web sites are chosen. At the end of this analysis we will define a taxonomy able to represent the knowledge of interest for this domain, which may also not have items from the taxonomy (or ontology from which we started in the TD analysis), but may have items which did not exist in it, emerged from the BU analysis. The final result of this phase will be a reference taxonomy, where, for each item, there is a linked information about where the knowledge of interest can be found on each Web site. The spread of the Social Web is influencing the evolution of Semantic Web: the way of producing and consulting information changes, as well as the way people relate themselves with the Internet and the services it gives. Users will participate at first hand to the developing of the Web which therefore becomes interactive. This study considers this feature, trying to link the worlds of Social Media and Semantic Web, with the aim of proposing a semantic classification of the information coming from the Web, which do not always follow a well-defined order and organization. Starting from a precise analysis of the information of the Web through an accurate and meticulous study on how these are presented and used, in order to give a sorted and easily usable data structure, this approach wants to define a taxonomy able to represent knowledge through an iterative combined approach, where TD and BU analyses are applied on the knowledge domain we want to represent.

The thesis concludes with the Chapter 6, where we report some considerations about the research findings and future works.

Chapter 2

The Issue of User Generated Content

In recent years, many software platforms managing a large quantities of multimedia content have risen in popularity within the Web 2.0. User-generated content (UGC) in particular, the most famous of which are YouTube[1], Flickr[2], Del.icio.us[3], Zooomr[4], Picasa[5], owe their great success to a spread of digital technology accessible by a mass, paralleled by the quantity and quality of the services offered. The prominent features of such platforms are their ease of use, the possibility for users to create and manage their own spaces (personal channels or pages), carrying and sharing any kind of multimedia content from various sources, the implementation of efficient content research and localization methods, the definition of access and usage types for them, and storage of information about legal restrictions and rights management.

With the evolution of the Web in its semantic form named Web 3.0, issues about application interoperability and management of shared information arose in UGC Web sites. For this reason we consider worthwhile to move on to a more effective representation of knowledge.

We analyzed the standards used in UGC Web sites for the management of the multimedia contents and their metadata. We defined an ontology to represent the semantics of these multimedia contents, so that in turn the metadata classification can give an unambiguous meaning. In order to unify metadata coming from different sources we defined all rules of mapping towards a structure defined by sources such as YouTube and Flickr. The innovation is in the approach for the formalization of Web semantics for multimedia content: we used standards such as Dublin Core, Exif, IPTC and in particular the Adobe XMP standard as a starting point of this domain. With the proposed approach, one can categorize and catalog all non-standard and unclassifiable information inside the ontology, using pre-made schemas.

We choose to formalize this knowledge using an ontology, an explicit formal specifica-

[1] YouTube, http://www.youtube.com

[2] Flickr, http://www.flickr.com

[3] Del.icio.us, https://delicious.com

[4] Zooomr, http://www.zooomr.com

[5] Picasa, http://picasa.google.com

tion of how to represent the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that exist among them and a formal, explicit specification of a shared conceptualization. Our ontology is conceived as a tool able to exploit pre-made schemas in order to represent content belonging to various types and coming from different sources. Such schemas are typical of standards and were used as means to model the domain.

The purpose of our ontology is to associate semantic value to all non-standard, mappable tags as well as to store information found in non-mappable tags, not to represent all properties of multimedia content.

The chapter is organized as follows: after an overview about ontologies and referential standards, the domain of interest concerning the study will be described; then, the semantic model developed to represent multimedia contents will be shown.

2.1 Related concepts

2.1.1 The ontologies

An ontology can be defined as a formal representation of a set of concepts inside a domain, as well as the relationships among those concepts. In theory, an ontology is "the formal, explicit specification of a shared conceptualization" [41]. It provides a shared vocabulary, which can be used to model a domain, the type of objects and concepts, their properties and relationships. At the same time it is a powerful tool for the knowledge formalization and sharing; in fact, it is able to provide a clear and efficient explication of the conceptualization of every field of knowledge. The term "conceptualization" can be seen as the formal definition of a certain part of reality, as it is perceived and organized by an entity, regardless of the terms taken into consideration for its description and of the recurrence of facts or precise events. An ontology is, therefore, an explicit specification of a conceptualization which, in turn, is based on a collection of objects, concepts, other entities and relationships among them which are known. Ontological relationships can be divided as follows:

- taxonomic relationships;
- axiomatic relationships.

Taxonomic relationships, through which it is possible to build hierarchies of concepts, can be expressed through the two following constructions: specialization and generalization (is-a); part-of and composed-by (part-of, has-part) [42].

Axiomatic relationships are added to the is-a relationships to better define the concepts of the ontology. In particular, they represent always true sentences which clarify restrictions over the ontology structure and the relationships among the different classes. The relationships are usually declared in the ontology through a definition which includes the name, the description in natural language, the concept they refer to and the attributes used. An ontology can be composed by one or more taxonomies classifying the terms of the speech and the relationships among them, and a set of axioms defining the other relationships. Particularly, taxonomies have to organize the ontological knowledge by using specification and generalization relationships.

Ontologies are an increasingly popular tool because of the advantages they offer in sharing information. They play a leading role in the representation and utilization of knowledge processes, also in the context of computer and information sciences. In the past, the study of ontology mainly focused on its philosophical context, but recently it has assumed an important role in many different fields of research and industries: ontologies are in fact able to isolate, retrieve, organize and integrate information on the basis of their core feature and their semantic context. For this reason they can be used to provide semantic annotations also for collections of multimedia objects such as images or audio: in this case we talk about multimedia ontologies.

In computer and information science, the ontology term has been proposed and well defined by Gruber [31] as an explicit specification of a conceptualization. In 1997 Swartout offers a new definition: *"an ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base"* [43]. So in 2008 Gruber in [44] defines the ontology as a technical term denoting an artifact that is designed for a purpose, which is to enable the modeling of knowledge about some domains, real or imagined ones.

An ontology points out the use of terms to define and interpret a field of knowledge. It can be used by simple users, database or applications exchanging information about a certain domain of interest. Particularly, IT ontologies consider definitions of the domain's basic concepts and the relationships among them, in a language which is understandable and usable by a computer: such ontologies represent a good model to define and analyze whichever study. Ontologies are also used to model the knowledge on more levels. They include schemes for the management of metadata or simple hierarchies such as taxonomies.

Multimedia ontologies

According to the W3C definition, multimedia ontologies can belong to two types:

- media specific ontologies, that use taxonomies and describe properties of different multimedia;
- content specific ontologies, that describe the subject of resources, such as the setting or participants.

Nowadays the creation of multimedia ontologies has become a crucial component; ontologies have many application fields, including Content Visualization, Knowledge Sharing and Learning [45].

Can a well defined multimedia ontology be built? The construction of multimedia ontologies is rather complex, as it is an iterative process that includes a phase for the selection of concepts to be included in the ontology, a phase to create properties and relations linking them together, and a phase for the maintenance of the ontology. We could mention many attempts to create a multimedia content ontology. In [46] multimedia ontologies were built semi-automatically. Textual information provided in videos were manually extracted and assigned to concepts, properties, or relations within the ontology; it was found that using standard tools for semi-automated construction of ontologies was more helpful in building data-driven multimedia ontologies.

In the last years new methods for extracting semantic knowledge from data were presented. A method for semantics knowledge extraction from annotated images is presented by Benitez and Chang [47]. Perceptive knowledge is built organizing the images in clusters based on their visual and textual features. Semantic knowledge is extracted removing all semantic ambiguity, using WordNet and image clusters. In [48] a Visual Descriptors Ontology and a Multimedia Structure Ontology, respectively based on MPEG-7 Visual Descriptors and MPEG-7 MDS, are used together with a domain ontology so as to support content annotation. In [49] ontologies enhanced with images were introduced to automatically annotate videos. Clip highlights were considered as examples of ontology concepts and were directly related to corresponding concepts, grouped into subclasses based on their perceptive similarity.

Bertini et al. developed MOM (Multimedia Ontology Manager), a complex system according to the principles and concepts of ontologies, enhanced through images [50]. It supports dynamic creation and update of multimedia ontologies and offers functionalities to automatically perform annotations and create extended textual comments. It also allows complex queries on video databases. Based on the same ontology, Jewell et al. provide a so-called OntoMedia ontology: a multimedia ontology based on an information system. Its main purpose was to manage a large amount of multimedia collections using semantic metadata integration techniques [51]. The annotations on multimedia documents were generally developed according to two different routes. Both approaches focused on low-level descriptors.

Dasiopoulou et al. presented a systematic survey of the state of the art MPEG-7 based multimedia ontologies, and highlighted issues that hinder interoperability as well as possible directions towards their harmonization [52].

Paliouras et al. proposed an approach towards the automation of knowledge acquisition from multimedia content [53]. In particular, with reference to the BOEMIE project, they adopted a synergistic approach that combines multimedia extraction and ontology evolution in a bootstrapping process.

In relation to the state of the art we proposed the use of different domain ontologies in a specific context, and our approach, here presented, can open up innovative ways to categorize.

2.1.2 Standards

The wide diffusion of mobile devices integrating their normal functionalities with the possibility of consulting the Web and download, edit or enrich its contents led our work to the research of formats suitable for the managing and cataloguing of different contents which can be edited and enriched or data coming from infomobility and georeferencing information.

As for our work (which we will describe in the next sections) we referred to standards as domain reference, which fit with the management and categorization of different types of content and georeferenced data.

The structure and the semantics is accurately modeled to be broadly consistent with existing multimedia description standards such as MPEG-7 as shown in [54]. These standards will be described below.

XMP standard

The Adobe Extensible Metadata Platform (XMP)[6] is a standard, created by Adobe Systems Inc.[7], for processing and storing standardized and proprietary information relating to the contents of a file. XMP standardizes the definition, creation, and processing of extensible metadata. Serialized XMP can be embedded into a significant number of popular file formats, without breaking their readability by non-XMP-aware applications. Embedding metadata avoids many problems that occur when metadata is stored separately. XMP is used in PDF, photography and photo editing applications.

XMP encapsulates metadata inside the file, using RDF, a basic tool proposed by W3C to encode, exchange and reuse the structured metadata as proven by W3C. In addition, the standard allows interoperability among the different applications interacting on the Web. The reason for its use is that it is a common standard for a wide range of applications, which allows us to work efficiently and effectively on metadata. These properties have encouraged the rapid increase in popularity of XMP at many companies operating in the digital media, which integrate their applications with this technology.

Many applications and back-end systems got, through the use of XMP, a considerable help tool to share, import and protect precious metadata. XMP is completely customizable and developable; particularly, it allows work-teams and companies to adapt metadata in order to optimize productive and editorial workflows. It involves different, largely spread and already existing metadata schemes such as Dublin Core, Exif and IPTC. XMP has also been created and conceived as a standard for the definition, the creation and the elaboration of metadata and it is based on the following points [55]:

- data models, it describes the types of metadata supported by XMP and how they are catalogued inside the schemes;
- serialization model: it describes the way in which XMP metadata are converted into an XML flow in compliance with a document with the syntax defined in the XML standard. An XML document is considered “Well Formed“ if it does not have syntax mistakes, every tags are balanced and there is only one root node containing the other ones;
- definition of schemes: they are sets of metadata gathered adequately. XMP offers an accurate description for the so-called standard schema, used by many applications. They also offer guidelines to define new schemes, compatible with the standards;
- XMP packet: it describes how XMP metadata must be encapsulated inside the various file formats.

The following table shows the formats supported by the XMP standard:

[6] Extensible Metadata Platform, http://www.adobe.com/products/xmp

[7] Adobe Systems Inc, http://www.adobe.com

Table 2.1: *Formats supported by XMP*

Image Formats	**Dynamic Media Formats**	**Video Package Formats**	**Adobe Application Formats**	**Markup Formats**	**Document Formats**
DNG	ASF (WMA, WMV)	AVCHD	INDD, INDT	HTML	PDF
GIF	AVI	P2	XML		PS, EPS
JPEG	FLV	HDV			UCF
JPEG-2000	MOV	XDCAM EX			
PNG	MP3	XDCAM, FAM			
TIFF	MPEG-2	AI			
	MPEG-4	PSD			
	SWD	INDD, INDT			
	WAV				

Dublin Core standard

Dublin Core (DC)[8] is a metadata system consisting of a core of essential elements for the description of any digital material accessible via computer network.

Becker et al. proposed a set of 15 basic elements also extended to sub-elements or qualifiers: each element is defined by using a set of 10 properties obtained by a standard ISO 11179 [56]. The aim was to establish a base set of descriptive elements which could be provided by the author or the editor of the resource. Thus, a users consortium started to develop an architecture for metadata, with the aim of meeting the needs of the users and producers of information.

DC's elements can lead to ambiguities or be totally missing. For this reason, things were solved by defining qualifiers which are associated to DC's main elements; it is required that the programs which base their functioning on the usage of DC are able to interpret it. Qualifiers are divided into two categories:

1 element refinement: they make the meaning of an element more specific;

2 encoding scheme: they simplify the interpretation of an element and can include controlled vocabularies or rules of elaboration.

The main features of DC are the following [57]:

1 ease of use: the standard is aimed at both specialized cataloguers and to non-expert users to catalogue;

2 semantic interoperability, which gives rise to a complex and precise data system whose meaning has been agreed in advance, along with a value that allows the DC to be a standard for quality researches on the Internet;

[8] Dublin Core Metadata Initiative, http://dublincore.org

3 flexibility, as it allows to integrate and develop data structure with different semantic meanings and a congenial application environment.

Exif standard

Exif (Exchangeable image file format) is a standard created by Japan Electronics and Information Technology Industries[9] to specify the formats of digital systems handling image and sound files such as the ones used by digital cameras, scanners, and so on [58]. It is a standard supported by the main producers of digital cameras and it gives users the opportunity to supply photos with interchangeable information between imaging devices to improve processing and printing.

The rapid spread of digital cameras and related tools (e.g. smartphones) increased the need to exchange images directly from cameras or other instruments, or to display an image taken with a camera through either another, or a different device altogether. Everything must be in compliance with DCF (Design rule for Camera File system)[10] specifications, which want to create a setting where the user can freely combine products and simply change media.

Exif offers a set of specific tags in itself, concerning shooting parameters and settings of the device at the time of capture. These cover a wide spectrum, including:

- time and date information, memorizing the current date and time;
- camera settings, containing static information about the camera's model and producer, information about the orientation, aperture, shutter click speed, focal length, white balancing and ISO speed information for every image;
- information about shutter click's location, coming from a GPS receiver connected to the camera;
- information and descriptions about the copyrights.

IPTC standard

International Press Telecommunication Council (IPTC)[11] is a standard that offers an advantage to relations and exchanges among entities devoted to information creation and distribution [59]. It is sponsored and defined by a consortium based in London that encompasses the leading news companies in the information world, such as Reuters, Associated Press and France Press. IPTC does not hold, among its metadata, fields related to technical information on a digital object; in fact, the metadata of the digital object itself, existing in other standards like EXIF, are not defined. The focus of IPTC in defining the standard is on analyzing what surrounds the various situations of telecommunications, and on studying their production process. Therefore, a range of metadata was defined, which is useful to define and certificate all digital object production activities for printing or editing. The first IPTC standard treated those metadata involved in the transmission of texts to newspapers, information agencies, news agencies and so on. It has been constantly updated, and it is efficient to date.

[9] JEITA, http://www.jeita.or.jp

[10] DCF, http://www.exif.org/dcf.pdf

[11] International Press Telecommunication Council, http://www.iptc.org

During the ’80s it was extended to other areas as well, such as TV production and radio broadcasting, to protect the interests of telecommunication industry. From the cooperation between IPTC and NAA (Newspaper association of America) a new IPTC-IIM (Information Interchange Model)[12] standard was born. It defines a model which can be applied to whichever kind of data and represents the first multimedia format for the exchanging of data.

In 2001 the IPTC-core standard was released. It took inspiration from many fields of the old IPTC-IIM model, still used, such as Description, Keywords, Author/Creator, Headline, eliminating others such as Urgency, Categories, and introducing new fields such as Genre, Rights, UsageTermes, etc. After the agreement between IPTC Consortium and Adobe, IPTC-core was integrated in the XMP standard, mapping its fields on the defined schemes in XMP. Some fields could not be directly mapped on the defined schemes in XMP: for them, a specific group named IPTC4XMP[13] was created. The agreement between IPTC and XMP brought to a fast diffusion of XMP in the field of newspapers, publishing and mass-media.

2.2 Domain knowledge

UGC stands for User Generated Content referring to micro-contents produced by users for the Web sites: users create communities, share comments, opinions and above all their own knowledge and experience. In the Web 2.0 era, many Web sites include UGC, in fact UGC points out how the Web is evolving more and more towards being a product made by its very users, labeled with the new name of “prosumer“ (producers and consumers). Every publicly accessible content type, with an added share customized by the user, is part of the UGC universe. Nowadays a large number of Web sites contain user generated content, and they have become massive repositories in which users share the results of their use of Web resources.

The most powerful applications and the most common platforms usually have these features: easy and fast content search by keywords, link usage for an easy navigation through contents, content editing by users themselves either iteratively (Wikipedia) or cumulatively (blogs and forums), content classification through “tags“, possibility to direct users to offers (any kind) through “collaborative filtering“-type algorithms, real time notifications through RSS for content change or editing. The usage of all these new technologies encouraged the success of such systems for socializing, where a remarkable exchange of information of many types (text, video, audio) and from different sources takes place.

We considered two kinds of such content: we analyzed and compared metadata from YouTube and Flickr, which, despite handling customizable multimedia content by users, consider a different way to represent information. In YouTube’s case, it is often possible to create relation by direct mapping in general, and indirect mapping in some special cases. Regarding Flickr, instead, some information are natively represented, other are included in metadata. The differences can be immediately noticed. In the first case, there is mapping possibility, directly or not, through schemas and standard properties; in the second, a new cataloguing method is used, typical of the platform and coming from a new school of thought, with no compliance with any standards.

[12] IPTC-IIM, www.iptc.org/IIM

[13] IPTC4XMP, www.iptc.org/IPTC4XMP

2.2.1 YouTube

YouTube[14] is a Web site for video sharing. Founded in February 2005 by Chad Hurley, Steve Chen and Jawed Karim (all former PayPal employees), it is now property of Google Inc. It is the third most visited Web site in the world next to Facebook and Google itself.

In 2006, Sony bought for US 65 million US dollars its competitor Web site, Grouper. At that time, this event made people think that Youtube's value on the market could be of around one billion dollars, but that estimation happened to be undersized, because the 10th of October 2006, Google acquired Youtube for 1,65 billions. Since the 19th of June 2007 the Web site has been available in different languages, among which Italian.

YouTube uses the Adobe Flash technology to play its contents. It aims to host only videos created directly by the uploader, but it often contains materials belonging to third parts, uploaded without permission, such as TV shows and music videos. YouTube allows for the embedding of its video in other Web sites and also generates its code.

This platform makes an intensive usage of feeds containing objects, such as Web link to content sources. Such feeds are used to give users frequently updated contents.

The interaction between YouTube and clients is managed through a protocol named YouTube Data API Protocol, a program communication interface application. This protocol allows the client applications to carry out all functions and actions which are normally done on the YouTube's Web site, such as looking for videos, retrieve standard feeds for the videos, retrieve video comments and responses.

Youtube Data API obtains the search results from a special index optimized for the searches and created in order for it to promptly include new uploaded videos, ensuring, at the same time, high performances even under heavy working loads for the API server.

Data existing in the API are shown by the protocol as views or projections; these can edit the way in which a feed is presented, keeping is content unaltered. However, content is preserved as it is. In this way, two different projections from the same feed will identify the same objects, but using different XML tag sets [60].

2.2.2 Flickr

Flickr [15], developed by Ludicorp[16] (a Canadian company in Vancouver founded in 2002), is a multilingual Web site that allows users to share personal pictures with whoever has access to the Internet, in a Web 2.0 environment. The site, owned by the Yahoo! group, has an ever growing library and was one of the first to implement tag clouds, visual representations of user-generated tags.

Tag clouds allow access to images tagged with the most popular keywords. Thanks to this support for tags, Flickr was mentioned as the first example of actual folksonomy use, although Thomas Vander Wal suggested Flickr was not the best example [61].

This Web site allows to simply organize large amounts of pictures taken with different tools (smartphones, compact cameras, webcams, reflex, analog cameras) directly through on-line applications, in the Web browser, via MMS or e-mail, soon after the picture is taken.

Flickr supports standard metadata sets (it shows the entire Exif metadata set for every picture), keywords for searches and a group of tags belonging to folksonomies.

[14] YouTube, http://www.youtube.com

[15] Flickr, www.flickr.com

[16] Ludicorp, http://ludicorp.com

The step from standardized metadata to folksonomies was due to some choices made by considering the usage of the Web site made by users (or the use that should be made):

- the vast majority of Flickr's users is not composed by professionals or contents creators, but by simple 'users' of contents and documents, and folksonomies show to be the best tool directly reflecting their vocabulary, their choices in diction, terminology and precision;
- Browsing vs Finding: it is a philosophy in which it is preferred that users do not unequivocally find the pictures they are looking for. Although the use of controlled vocabularies helps the availability of contents, the fact of "browsing" the system and its related links to sets allows users to find unexpected contents which they might like. It is like finding answers to a specific question and exploring an area of issues to formulate questions with regards to it. Folksonomies are a tool which naturally fosters this philosophy.

As far as the georeferencing is concerned, Flickr allows users to organize their pictures in "sets", namely groups of images sharing the same gallery. Sets are more flexible than the traditional folder organization method for files: a picture can belong to one or more sets, or to none. Those sets represent a form of metadata category, instead of a physical hierarchy. The pictures in a set can be geotagged, and every set of geotagged pictures can be put in relation with a map using ImapFlickr[17]. Such a map can then be embedded in Web sites.

2.3 Proposed approach

The main aim of the work was to suggest an approach for the management and categorization of UGC that comes from different sources like popular digital platforms.

We proposed an approach organized into three steps:

1 implementing an ontology to represent information typically associated with such contents, among what is already available, using well known standards as references;

2 analyzing the data to define a formalization that allow the representation of information which come from sources such as UGCs not complying to standards. This technique can exploit different relations when possible, or create new ones whenever necessary. This is especially true for information which can be found in many contents;

3 integrating the information contained in the ontology with other in fields that can store non-mappable information with the above mentioned technique. In such fields typical tags for the platforms, as well as tags defined by users (folksonomies) can be stored.

With the proposed approach, we can categorize and catalogue all non-standard and unclassifiable information inside the ontology, using pre-made schemas. The purpose of our ontology is to associate semantic value to all non-standard, mappable tags as well as storing information found in non-mappable tags, not to represent all properties of multimedia content.

[17] ImapFlickr, http://imapflickr.com

The ontology does not need to be able to represent everything, but to use what is already available for representing known and classified information such as author, URL, etc. It also must use that mapping amongst infrastructures and information provided by the platform.

Folksonomies, in general, are used as an alternative for every information for which no standards based on schemas or tags exist, i.e. for everything non-standard, such as user comments and other new default information.

2.3.1 Top-down analysis

As KB for the starting domain, the approach we followed in building a multimedia content ontology assumes the XMP, Dublin Core, Exif and IPCT standards, as well as the related XML schemas and the integration with the semantics through RDF according to Lassila and Swick [62] and by Brickley and Guha [32]. In this way it is possible to have a complete modeling of the domain of multimedia content properties, together with a uniform representation of the variety of associated metadata that come from different sources.

We assume to use this approach because such standards allow to catalogue different aspects of multimedia content and natively possess the specification tools for georeferenced information. The ontology was then modeled on those standards, selecting the relevant elements. Once the basis ontology was decided, the next step was to analyze, catalogue and classify the metadata of contents that come from the main software platforms of the Net. Thanks to this we could acknowledge alternative standards and proprietary formats used.

We decided to narrow the scope and choose which ones should be considered, because, given the great number of available platforms on the Net, we considered the task to analyze all of them too onerous.

After a study of all the features related to metadata on the chosen UGC sites, we worked on a mapping mechanism that allows such data and associated metadata to be represented within the ontology.

Ontology modeling on standards

Modeling the semantics of metadata from various multimedia contents, providing for georeferencing and mapping of the different standards related to metadata, was the main purpose of our ontology. For this reason the representation of the metadata can thus comply with the reconciliation standard provided by the MWG and with Adobe XMP. Acquiring knowledge about the domain to be modeled is the first step to take into consideration when creating an ontology [34]. In fact, we started from the assumption that the reference domain is the one that includes every kind of multimedia content, both currently available on the Web or through modern digital technologies, equipped with sets of metadata belonging to the above mentioned standards.

The ontology must be able to receive a content coming from social networks or software platforms for content management without information loss or alteration. The ontology that we obtain can also be used as a KB supporting the geolocalized guide.

Due to its computational completeness, its decidability and the fact it guarantees maximum expressiveness [63], OWL DL is the chosen sublanguage in order to categorize the concepts related to the ontology [64][65]. Following a middle-out approach and by modeling the concepts mentioned above as classes or properties we created the structure of the ontology. First of all we proceeded with the definition of relations and main entities which were progressively generalized and specialized. These structures were integrated with RDF schemas. In particular, the entire set of metadata required by the Exif standard, together with the entire Dublin Core set (complete with its refinement terms), was imported. Both schemas allow the ontology to exploit their metadata, making them available as particular properties, datatype properties and object properties at the same time, probably so as to satisfy every kind of usage needs.

In this specific case they were used only as object properties, i.e. to link class instances with other class instances. The main classes involved in ontology building are explained as follows:

1 'MultimediaContent': this class is responsible to model the concept of multimedia content. It is a simple class, without subclasses, which formalizes its link with the class representing file formats ('MultimediaFormat') through a property called 'hasMultimediaFormat', which respectively has 'MultimediaContent' and 'MultimediaFormat' as domain and co-domain;

2 'MultimediaFormat': it represents the most common file format currently available on the Net. This class is structured into a two-level hierarchy. The first level represents format file categorizations depending on the content type they express. The second level is represented within each categorization, where classes representing the actual formats are located. Each format is identified by its own extension;

3 'Metadata': its subclasses represent every type of metadata considered in the study of reference standards and reference application context;

4 'XMPtype': it represents the co-domain of all properties concerning the 'Metadata' class and its subclasses. It includes a number of classes which represent the different data type which the XMP standard uses to describe information inside its tags.

Some applications avoid the complex operation of storing information inside files, finding it a problem. They opt instead for executing it in external files or databases, although such operation could lead to the loss of metadata as well, when the same file is used in different applications. XMP, for example, is one of the standards that requires writing of its own metadata set inside the file, but it is not the only standard enabling this action. Often, every file format has its own blocks, different from the ones used by XMP, to store certain metadata schemas. For example, a JPEG image has some containers to store the EXIF, IPTC-IIM, and Photoshop standards.

Metadata are stored in different semantic groups inside each block. For example, the following groups can be found inside the XMP APP1 block: Dublin Core, IPTC-Core, EXIF/TIFF; inside the Photoshop APP13 block is the IPTC-IIM group. This problem required a data reconciliation which was performed through the mapping technique. The creation of the mapping meant the execution, where feasible, of a set of non-automatable, strongly subjective operations. The search for XMP tags that could map the ones used on the analyzed platforms was an integral part of our work.

We searched for tags with the same semantics as the ones we needed, among those available in the standards within XMP. This search was performed with particular care so as to avoid mistakes due to unclear or poor descriptions and consequent semantic association mistakes.

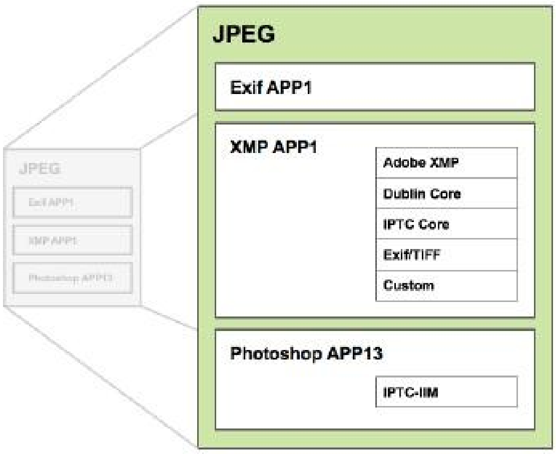


Figure 2.1: *XMP schema blocks*

2.3.2 Bottom-up analysis

In order to define the ontology we used a significant number of metadata compliant with our goals related to YouTube and Flickr, very important multimedia UGC repositories.

During the analysis phase we faced several problems: the most important one was to reduce the semantics of certain tags to a single representation. In fact, each tag was represented with all of its attributes and subtags within the ontology. This subset describes the information related to videos, which are the main content of YouTube, and has pieces of information that must be taken into account when complying to specifications, such as comments and georeferencing information.

YouTube has three categories of contents:

1 videos with their associated information (“video location“ with georeferenced information, etc.);

2 descriptive fields (title, description, category, video properties, etc.);

3 the ’Tags’ field contain ”keywords to help people discover your video” where the user freely insert words (folksonomies).

For type 1 and 2 data we could perform two kinds of mapping: direct or indirect, according to whether the semantic correspondence was direct (same meaning of information, and same format, same data type as well) or indirect, that is to say there was discordance in its form (same meaning but different representation). In the direct case, we exploited the feature by which it is possible to create property hierarchies: each property can have its own sub-properties, which specialize their super-properties just like a subclass specializes a super-class. This means that an implication relation among nested properties is in place: if the super-properties have a domain and a co-domain, those will necessarily be inherited by their sub-properties. Even on a visual level, mapped tags will appear under the mapping ones. Therefore direct mappings were performed by assigning to the mapped tag its mapping tag as super-property.

We have also some information in YouTube fields that we don't want to represent in the ontology and consequently don't have mapping, but we stored these data in a bulk class 'YouTubeOther'.

The words in the 'Tags' field which are non-standard information, but folksonomies, are stored in the bulk class 'YouTubeFolksonomies' to preserve every information associated to the content.

To make such operation clearer and the ontology more readily accessible by users, every direct mapping came together with an annotation of the rdf:comment type with information related to the 'mapping' tag.

In the indirect case the implication relation cannot be used, because the information must firstly be broken down in its elementary parts, and then those parts must be traced back to direct mode. These steps are described inside the rdf:comment associated to the mapped tag. In particular it explains how to split and convert the information, and where to store it.

On the other hand, Flickr has three important types of contents:

1 images with associated Exif, with Exif coordinates or user-defined coordinates;

2 descriptive fields (description, title, etc.);

3 free tags created by users (folksonomies).

Initially, Flickr used to equip its content with a simple set of pure Exif data, so they are natively mappable in the ontology because they are strictly compliant with the standard.

Exif metadata which are not fully compliant with standard are not mapped in the ontology. They were stored in the bulk class 'FlickOther', as we did for YouTube.

In Flickr the folksonomies do not belong to any kind of hierarchy. They were represented inside the ontology, with a bulk class called 'FlickrFolksonomies', where the free tags of Flickr are stored. For operational purpose, the class has, as a property, a set of tags which allow to generate a Feed Atom, that in turn includes all such information in bulk, non-standardized ones. The set of metadata belonging to folksonomies must be stored inside the Atom Syndication Format tag atom:content. *"This specification describes Atom's XML markup vocabulary. Markup from other vocabularies (foreign markup) can be used in an Atom Document. Note that the atom:content element is designed to support the inclusion of arbitrary foreign markup."* [66].

2.3.3 Example

In this section we consider a peculiar example of Web content acquisition for content related to the image "Castello di Arco" (Arco's Castle) stored in the Flickr platform, and of related metadata management [67].

Figure 2.2: *Arco's Castle*

The metadata associated to this picture by Flickr are listed in Table 2.2:

Flickr provides a tool, flickr.photos.getExif[18], that allows to read the metadata set associated to a given content. Entering the last number of the address into the tool, we obtain as output a list of tags including that information.

In order to enter such data in the ontology, it is necessary to create various instances to represent content, format, the Exif schema describing it, instances for each data type associated to each tag and related values. Not all of the retrieved data were created inside our ontology, and an Atom feed was to be associated to the content so that it could collect unknown metadata in bulk. These metadata are partly complying with the Exif standard and mapped with the typical rules of the standard as such.

As for the mapping, it was necessary to enter what was not provided for by the scheme of the ontology. We inserted the information related to all properties and created the link amongst them - and between them and the various metadata - so that they could be represented univocally and no information could be lost. In our example the first thing to be created was, with the aid of the tool, the 'MultimediaContent' class; the name "Castellodi-Arco" was then associated to it, exploiting the 'instance' browser. It could be noticed that, for the properties previously created, the 'hasMetadataLocation' and 'doesExpress' fields appear already compiled.

On the other hand, we had to define the elements to insert in the 'hasMetadataDescription' field and the 'ExifSchema', 'ExifSchemaCastellodiArco', 'UnknownMetadata' and 'UnknownmetadataCastellodiArco' instances. The latter belongs to the class devoted to the representation of unknown metadata belonging to a standard. At this stage, the 'ExifSchemaCastellodiArco' instance could be filled out with all the fields returned by the Flickr tool. In this way a univocal correspondence between information and metadata related to it was created.

[18] flickr.photos.getExif, http://www.flickr.com/services/api/flickr.photos.getExif.html

Table 2.2: *Metadata Arco's Castle*

Camera	Sony DSC-H3
Exposure	0,006 sec (1/160)
Aperture	f/9.0
Lens	29.6 mm
ISO	125
Exposure Bias	0 EV
Flash	Auto, Did not fire
Orientation	Horizontal (normal)
X-Resolution	72 dpi
Y-Resolution	72 dpi
Software	Picture Motion Browser
Date and Time (Modified)	2009:08:17 22:15:32
YCbCr Positioning	Co-sited
Exposure Program	Program AE
Date and Time (Original)	2009:07:19 17:50:00
Date and Time (Digitized)	2009:07:19 17:50:00
Compressed Bits Per Pixel	4
Max Aperture Value	3.5
Metering Mode	Multi-segment
Light Source	Unknown
Color Space	sRGB
Custom Rendered	Normal
Exposure Mode	Auto
White Balance	Auto
Scene Capture Type	Standard
Contrast	Normal
Saturation	Normal
Sharpness	Normal
Color Reproduction	Standard
Macro	Off
Exposure Mode	Program
Quality	Normal
Anti-Blur	On (Shooting)
Long Exposure Noise Reduction	Off
Compression	JPEG (old-style)
Orientation	Horizontal (normal)

In order to know which tags of the picture are existing or not, the entire Exif schema must be checked. Once the values were ready to be entered into the tags, we created a different data-type instance for each data. Since the data type belongs to the Exif schema, it requires some additional attributes for temporal information; thanks to the existing relations, the fields related to such attributes were displayed as well. The other information are stored in the bulk class ’FlickrOther’.

Since the data type belongs to the Exif schema, it requires some additional attributes for temporal information (exif:subSecTimeDigitized,exif:subSecTimeOriginalexif:subSecTime); thanks to the existing relations, the fields related to such attributes were displayed as well. The other information are stored in the bulk class ’FlickrOther’.

The final result of this example is that the ontology among the 60 tags found:

- mapped 3 tags of XMP metadata;
- 42 were natively mapped onto the ontology;
- 15 tags are stored in the bulk class ’FlickrOther’ because there is no corresponding tag in our ontology.

The Title: ”Il Castello di Arco - Arco’s Castle” and the Description ”The city is developed...“ are natively mapped in the ontology. The Tags (Europa, Europe, Italia, Italy, Trentino Alto Adige, Arco di Trento, Vacanze 2009, Holidays 2009, Castello, Castle) are stored in the bulk class ’FlickrFolksonomies’.

Chapter 3

The Issue of Linguistic Information in Audio Content

Pieces of information are more and more frequently published on the Internet in a well structured and organized manner through Knowledge Management Systems (KMSs), either created ad hoc, or within a structured database (e.g. management systems of images, like Flickr, or videos, like YouTube).

Organization and availability of content in KMSs basically depend on two factors: one is whether KMSs have effective tools for information indexing and retrieval; the other is how they do that. The use of metainformation, i.e. data used to describe and classify information, is a possible solution to this issue. Through organized schemas and relevant standardized metadata, or data tiers, it is possible to describe, classify, and organize basic information, allowing retrieval and use.

In this study we propose a way to associate linguistic and musicological annotations (information associated to specific text portions) to the corpus by treating them as metadata, so as to insert and manage them in the archive of choice after formalizing them in XML, the universally used markup language for representing metainformation.

The chapter is structured as follows: in the first and in the second section we recall some aspects about KM and the Linguistic Corpus; we also define our domain of interest. In the third, we present our proposed approach for knowledge formalization and management. Our approach was experimented and validated during a project that aimed to create the Analytic Sound Archive of Sardinia.

3.1 Related concepts

Computational linguistics originated in the second half of the Twentieth century, when information science met linguistics and languages in general. Their encounter was inevitable: where information science deals with data elaboration through computers, gathering and processing information, linguistics deals with the analysis and study of the most powerful tool human beings possess to express information, that is language. Computational approaches have been recently devised for music studies as well, with applications in both analytics and Music Information Retrieval (MIR).

In the latest thirty years, a progressive integration and assimilation of IT tools in linguistics and literature studies has been active. At the same time, languages have started to be considered as possible sources and objectives of countless software applications. In this way, theoretical contributions from information theory, language statistics, formal languages and AI entwined themselves with the astounding technical possibilities offered by the development of microprocessors for computing power, hard disks for storing vast amounts of multimedia data and telecommunications for transmitting information on the Internet.

Computational linguistics could thus be defined as the meeting place between theoretic (and applied) linguistics and information technology. Theoretic and applicative problems related to language merge with theoretic and applicative problems related to informatics and computers.

An electronic corpus is generally a homogeneous collection of written or oral texts in digital format, processed with coherent criteria in order to build an empirical basis for language analysis. Its advantage is that it can be annotated by adding linguistic information in a specific portion of text. In our case study, the electronic corpus is formed by a collection of audio recordings from poetry contests and singing performances in Sardinian language and from interviews in Sardinian and/or Italian, stored and annotated on different linguistic and musicological levels.

Linguists and musicologists, creators of the corpus, needed to study and research the documents in it, and they asked for the possibility to save their work in a readily available digital archive to store, index and manage for both access and communication inside the scientific community.

Corpus-based linguistics uses corpora to integrate linguistic theories with real and linguistic data; such natural data are taken from texts actually produced. Corpus-driven linguistics, instead, is based on corpus data for constructing theories and general linguistic hypotheses. From this point of view a corpus is essentially a collection of linguistic texts or utterances. “Corpus“ is a Latin term describing any “complete and orderly collection of writings, of one or more authors, concerning a certain matter”, or “in language proper a culture sample examined in the description of a language” [68].

Various kinds of works can be collected in a corpus, i.e. articles, interviews, court documents, private calls, texts and more. We can, also, put together texts, phrases, words, etc. In general, however, corpora contain real and authentic texts and not smaller or altered sections, which would be otherwise called samples of phrases, words, etc.

A linguistic corpus is only a sample of a given language, since it will never coincide with all its virtual textual expressions, which, with statistical term, define a “population“. However, the productions of a language, such as texts, form only a potential population, being continuously expanded with new contributions. A corpus can render a population as a sample only for the type of study for which it was designed. According to the guidelines of EAGLES, a corpus, in general, can be defined as: "a collection of parts of language that are selected and ordered according to linguistic criteria explicit to be used as a sample of the language" [69]. Instead, an electronic corpus is defined as: ”a corpus that is encoded in a standardized and homogeneous way for recovery purposes without limitation”. With the advent of computer technology, the creation and maintenance of electronic corpora have been considerably facilitated to the extent that electronic corpora are now totally identified with the general notion of corpus.

In our analysis we also consider another type of corpus, widespread and extremely useful in software applications, i.e. the linguistic parallel corpus. In this case, the texts are original in one language and translations into one or more foreign languages. Original and translated texts are then aligned, and equipped with a series of correspondences between the translated text portions. Parallel corpora are especially used in international bodies such as the European Union, that recognizes different official languages, but are also an excellent device for preparation of tools for machine translation of texts and for teaching foreign languages.

The first step in the construction of an electronic corpus is the choice of linguistic corpora to be set up: the first selection is made between corpora of written or spoken texts. Annotated corpora are another type of electronic corpus that has gained considerable importance. An annotated corpus contains information regarding portions of text. The annotation can cover a large amount of possible areas: phonetics, morphology, syntax, etc.

The preparation phase is one of the most important stages in building a corpus. The preparation of the format in which the text will be stored and processed is very important because it allows its circulation and sharing.

Among the linguistic representation formats, the XML standard has been standing out for a long time. It is a metalanguage describing markup languages, by defining the labels (tags) and the structure of the data and metadata. The use of XML has many advantages: it facilitates exchanging computational resources and primarily lends itself very well to multi-level annotations, especially to the tagging of levels with a good "structure" which are represented for "hierarchical" components and categories.

Any aspect of the linguistic analysis can be labeled: phonology, phonetics, morphology, syntax, semantics, pragmatics, text, etc.

The main aim of this annotation process on the corpus is an easy and fast extraction of data, linguistic or not, from the text or from collected texts. Different types of annotations can be created: some basic annotations are relevant to the syntactic structure of the utterances such as morph-syntactic and syntactic annotations.

The set of relations that are established between the basic parts of a sentence (words, monemes, morphemes, phrases) and among these and the functions (subject, predicate, object, etc.) or categories (of time, of space, of reference to interlocutors, etc.) about a language are shown in a record.

Phonetic annotations identify and label the phones corresponding to a class of linguistic sounds that are similar to each other for both way and place of articulation and specific acoustic features. Phonological annotations identify phonemes. A phoneme is a sound corresponding to a distinguished, indivisible unit which is capable to carry a precise semantic meaning (as opposed to phones being able to express the same meaning despite corresponding to distinct sounds). Pragmatic annotations identify the function that a segment of text has with regard to the context, understood as verbal context (the statements preceding and following) or situational context (what actions participants do in the dialogue, or beliefs that a given message induces in the listener, appearing as owned by the speaker). An annotated corpus should satisfy specific formal requirements.

The raw material of the corpus is made by all the digitized texts that have to be separated from the annotation language and all the other encodings. The coding of the corpus and the record must be in a standard format. On the other hand, an annotated electronic corpus allows, through simple tools, to consult and search within the corpus itself. Electronic corpora are by their nature dynamic and, due to that, much more flexible and richer than static corpora: electronic consultation makes it possible to make an infinite number of searches.

3.1.1 Metadata

Metadata play a key role in the organization and management of digital resources, especially when the amount of available information is high and must be indexed and catalogued for easy search and retrieval [70][71][72][73].

Metadata must make the resource accessible by adding tags to the content according to a consistent pattern. The term “metadata“ recalls the idea of "second-level data", i.e. data used to describe and categorize the primary data or "information level" that makes up the information resources themselves. Just as the choice of metadata, even the choice of a possible classification and organization depends on the type of information that is being handled and the objectives that are being pursued. Therefore, the choice of one classification, of course, may not be the only possible one, or the most complete and satisfactory one. The search for a standardization is, therefore, an essential requirement in the construction of metadata schemas.

We considered a metadata schema as a set of structured metadata, developed with the aim to establish a standard of metadata structure and terminology, and to associate different types of metadata. Every metadata schema includes a definite number of elements, called metadata elements, each with its own meaning and purpose, i.e. describing the information resource [74][75]. Metadata schemas typically have a complex nature, highly structured, often hierarchical, and include not only different metadata, but also metadata of different types or with different functions.

Metadata represent the means by which big quantities of information, that come from digital resources, must be indexed and catalogued to facilitate search and retrieval. Indeed, the aim of metadata is to manage the primary information in a functional and standardized fashion, not only by humans but also by software agents and programs. There is also a greater attention to accuracy.

Our idea was to use largely used metadata schemas rather than creating new ones. Considering metadata schemas and their metadata, we made various application profiles that were aimed to create tools for particular applications while keeping interoperability with the original base schema. This procedure and the application of common rules can make different systems interoperable, like those in libraries, museums and archives, making them able to share a part of common metadata.

3.1.2 Dublin Core metadata schema

A support to content management is offered by the DC metadata schema, which easily pairs up with other metadata schemas in the OAI architecture, improving granularity and refinement of their structures [76]. The rapid diffusion of DC as a metadata schema was fostered by its remarkable simplicity, thanks to which it could adapt to many kinds of resources and usage environments. It is important, for a semantic model used in resource discovery, not to be dependent on the format of the resource it needs to describe. DC is increasingly used in many fields in order to describe, organize, and manage resources belonging to institutions and international organizations, and also to support and provide a base format for the aggregation and exchange of metadata collections. The use of a standardized general classification system allows for metadata in such collections to be combined and for knowledge inside each collection to be shared.

3.1.3 Institutional Repositories

Starting in the 90s, the process of scientific communication and knowledge sharing was affected by the appearance and spread of digital repositories of scientific contributions in order to make the transfer of information more “agile”. In 1991, Paul Ginsparg, at Los Alamos National Laboratory (USA), proposed the arXiv1, a repository of works on Physics and Mathematics. In June 1994 Stevan Harnad introduced a new idea to the students of the Virginia Polytechnic Institute: in order to communicate their results more effectively, they would have to share ideas through the contributions of self-archiving on the Internet. The new trend for open archives were Institutional Repositories (IR), supported and managed by an institution, such as universities, which incorporated the contributions of its researchers.

At the end of the 90s, an important fact that marks the turning point occurred: a group of researchers and librarians in Santa Fe (USA) employed the OAI as operating tools and for indexing. In fact they considered the IR essential to the management of technical aspects such as protocols and data exchange standards, localization and subsequent retrieval of scientific contribution, and software. After ten years, when the archives were already open and operational, the expression “Open Access“ was used for the first time in a public document: Budapest Open Access Initiative Manifesto (2002) [77]. It suggested for the first time to adopt both strategies, called “complementary“, to encourage the open access system spread: the "self-archiving", i.e. archiving in institutional and disciplinary "open electronic archives", of articles by researchers and "open access journals", the new generation of scientific open access journals.

The main need was to have an efficient tool able to classify and store the knowledge regarding a specific electronic corpus and that could likewise allow a high usability in terms of ease of reference as well as ease of query and communication.

According to an industrial project concerning the use of the KB of the Analytic Sound Archive of Sardinia, the idea to create an IR to solve the problems of organization and availability of information came forth.

3.1.4 DSpace

DSpace[1], an open source software suite developed in 2000 during a joint project between Massachusetts Institute of Technology (MIT) and Hewlett-Packard, provides all the necessary tools for the creation and management of an IR based on the Open Access model. Such an IR can collect, store, index, preserve and make accessible the information output created by universities and research institutes in a digital format. DSpace originated as a natural addition to MIT Libraries, and from the actual need to archive and store all documents produced by MIT in an electronic format for a long time. The resources, usually stored in the hard drives of the personnel or in the servers of the departments, were at risk of being lost with time or with the constant updates in technology.

Systems for electronic document management were already being operational for a long time, but they usually were commercial and proprietary, so not very customizable; moreover, their survival was tied to the reliability of the company producing them. This feature in particular made guaranteeing a stable software support or a long-term storage not feasible. Therefore, MIT and HP developed the Open Source platform DSpace and made it publicly

[1] Dspace, http://www.dspace.org

available under BSD, with the apparent intention of fostering the formation of an active open source community that could participate to its improvement [78].

The development model was based on one of the most successful Open Source projects in the world, that is Apache Foundation, which is currently leader in Web server applications with a market share close to 70

The service was supported by DSpace Federation up to 2009. It was made by different research organizations and universities like Columbia University, Cornell University, Ohio State University, University of Rochester, University of Toronto, University of Washington in Seattle, Cambridge University, MIT.

The biggest efforts of this joint project were towards the fostering of interoperability among institutional archives, the development and management of the code of DSpace, the long-term preservation of scientific resources and the technical support to the institutions for programming and configuring DSpace.

From 2009 on, the DSpace project is managed by the no-profit organization DuraSpace, with the purpose to maintain it and spread it so that the access to digital resources and knowledge sources stored in the archives that use DSpace can be granted for a long time.

DSpace is designed as a central storage facility able to collect all kinds of content from the community relating to the institution through a user interface as simple and intuitive as possible. It can collect various types of digital resources including text, images, video, audio, articles and preprints, technical reports, working papers, datasets, and learning objects directly from the creators. DSpace can recognize and manage a great number of file formats and MIME types. Currently, some of the most common formats it can manage and automatically recognize are PDF, Microsoft Word, JPEG, MPEG, TIFF, but the support can be extended to every file format. Moreover, DSpace has a registry of the file formats in which not recognized formats can be registered in, in order to have them easily identified in the future.

DSpace is not only the basis for the creation of institutional archives using the Open Access model, but it also includes all the set of services that the institution may offer for storage, preservation, organization, and management of the digital resources created by the institution itself and its members. DSpace was chosen to create the Analytic Sound Archive of Sardinia as it fulfils all the requirements asked by linguists and musicologists. It is in fact completely customizable, supports natively the Qualified DC metadata schema and is compatible with OAI with the support of OAI-PMH [79].

3.2 Domain knowledge

The basic starting concept is the definition of a KB as a centralized repository for information: a public library, or a database of related information about a particular subject, could be considered to be examples of KBs. The most important aspect of a KB is the quality of the information it contains. The best KBs have carefully written articles that are kept up to date, an excellent information retrieval system (search engine), and a carefully designed content format and classification structure. In Information Technology, a KB is a machine-readable resource for the dissemination of information, generally online, or that can be put online. An integral component of KMSs is the KB used to optimize information collection, organization, and retrieval, for an organization or the general public. A KB can give customers an easy access to information that would otherwise require contact with an organization's staff;

as a rule, this opportunity should make the interaction simpler for both the customer and the organization. A KB is not a static collection of information, but a dynamic resource that may be able to learn, as part of an AI expert system.

In this case, the KB is used to collect and categorize the information belonging to musicological and linguistic fields. How and whether KMSs have effective tools for information indexing and retrieval are the main factors that characterized the organization and availability of contents in KMSs. Experience suggested to use metainformation, i.e. data used to describe and classify information, as a possible solution. A good example came by analyzing the case of the library and archive industry: in fact they faced issues related to organization and collection of information since way before the digital revolution. Metadata could be considered as a tool to enter and manage contents on the Internet and allow for entering and retrieving organized and relevant metainformation.

3.2.1 Formalization of linguistic annotations

Language production, either spoken or written or sung, is studied in linguistics corpora by observing its characteristics: lexicon, syntax, collocations, phonic chain, morphological structures, etc. Any complete and orderly collection of written texts, by one or more authors, on a certain topic, or, linguistically speaking, the sample of a language as examined in the description of the same language, could be considered as a corpus. We can take advantage of the wealth of information stored if we enhance it with additional linguistic or meta linguistic information i.e. the adding of linguistic or metalinguistic information to different portions of text [69].

The most efficient way to use the information in the corpus, organized as informal annotations, is formalizing those annotations through metadata schemas. In this way, not only annotations can be associated to their texts, but they can also be used as search parameters for finding texts. The formalization of annotations in a metadata schema can be achieved using a BU or inductive reasoning. Starting with the analysis of the structure of each annotation in the file and applying inductive logic, a "category" is abstracted from every linguistic level. The use of BU or inductive reasoning could help to formalize the annotations in a metadata schema. All annotations in the same linguistic level, e.g. phonetics, can be formalized in XML as different occurrences of the same metadata, called "phoneme", whose value can be made up of two terms: linguistic label and eventually time interval. For an easy representation and coding of annotations we can consider XML, a markup language: in fact, the use of tags or markers could help with the representation of metadata and their qualifiers, inside which a linguistic label is found.

3.2.2 Annotations through PRAAT

Linguists and musicologists working on the Analytic Sound Archive of Sardinia use the PRAAT [2] software to annotate the electronic corpus. This allows for multilevel segmentation and linguistic annotations of audio files and to perform spoken language analysis. The software provides a graphic interface in which we can see waveforms and voice spectrograms that make annotators' work easier and make acoustic phenomena which can be found by an accurate spectrum analysis visible, followed by annotation levels. It is also possible to choose

[2] PRAAT, http://www.fon.hum.uva.nl/praat

the list of possible annotation levels (syllable, tone, morpheme, syntagm, accents, etc.). This is geared towards linguistic and musical analysis of audio recordings.

An example of one TextGrid providing different types of annotation is displayed in Figure 3.1.

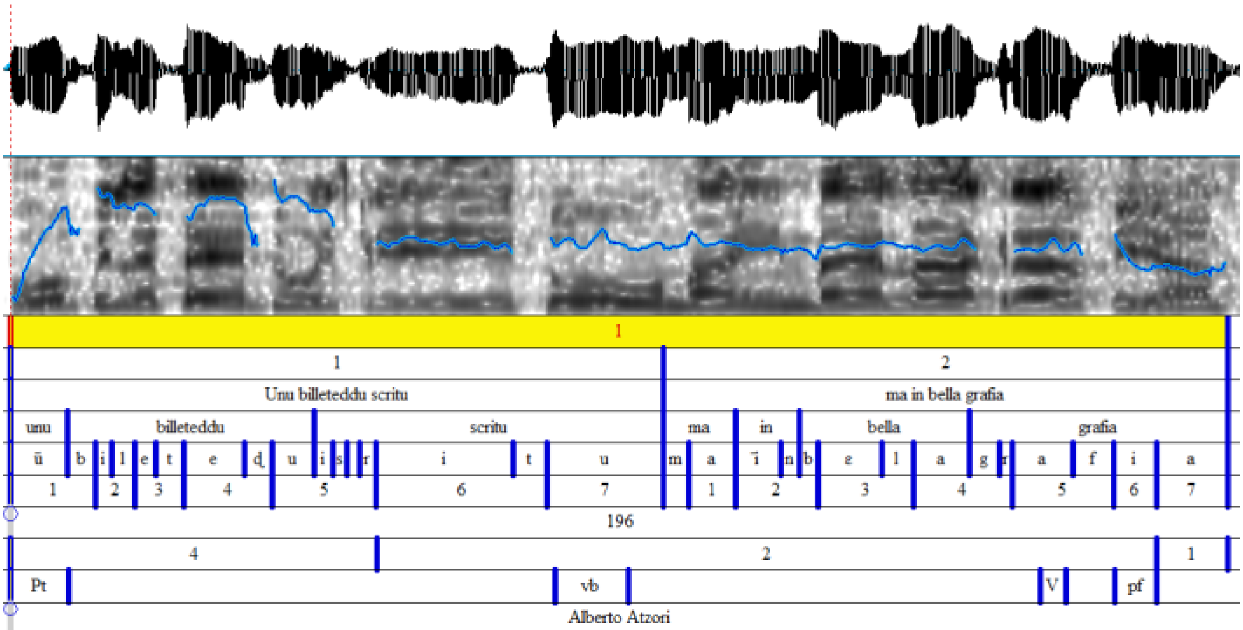


Figure 3.1: *Example of multilevel annotation of an audio file by means of a PRAAT TextGrid*

In this case, the audio file is a sung performance coming from a poetry contest in the style known as *a s'arrepentina* [80][81] and the annotations are relevant to some essential metrical, linguistic and musical features, that are (from the highest to the lowest tiers of the TextGrid):

1. Metrical unit (verse);
2. Metrical sub-unit (hemistich);
3. Transcription by unit (in this case, by hemistich);
4. Transcription by word;
5. Phonetic transcription (IPA code);
6. Metrical-Musical position (by hemistich);
7. Tonal centre level (in Hertz);
8. Numeric transcription;
9. Ornamentation;
10. Performer.

These kinds of annotations allow a detailed examination of the performance and an accurate definition of the expressive means and acoustic features defining each performer's singing style. Three examples of usage of the annotation for analytical purpose in the field of the ethno-musicological studies will be presented here. The first example concerns a typical feature in the *a s'arrepentina* singing style, i.e. the *accelerando* within each intervention sung by the poets. This characteristic can be observed and accurately measured, and the results are useful to differentiate their individual singing styles.

Figure 3.2 shows three different *accelerando* steps in three poets participating in one *a s'arrepentina* contest: the inter-beat durations undergo a contraction which is almost linear in one poet (see central panel) and more rapid in the first part of the intervention in the other two (see left and right panels).

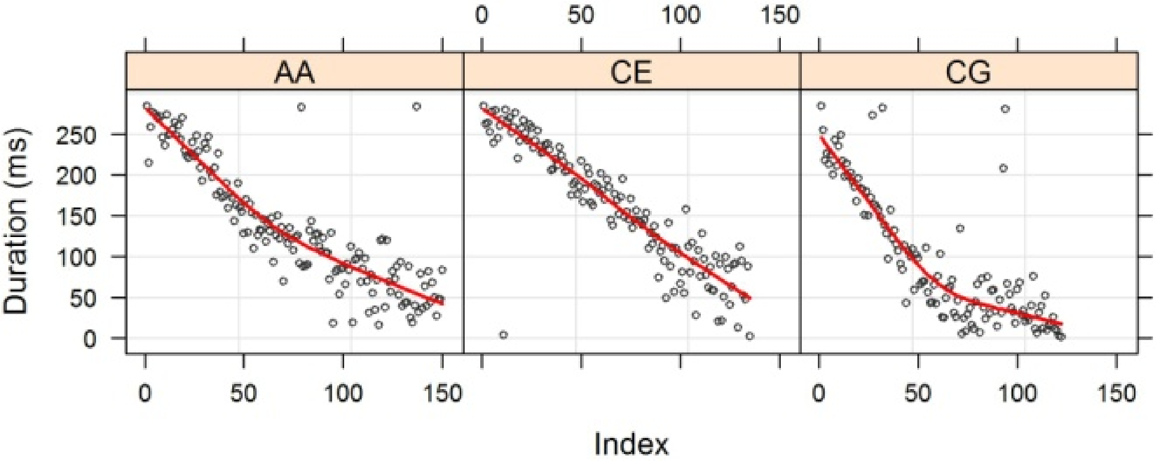


Figure 3.2: *Accelerando in three performances of a s'arrepentina improvised poetry, evaluated by means of the inter-beat durations.*

The second and third examples of analysis made possible by the annotation and presented here concern the scales used by the singing poets. By measuring the duration of the segments of the tier with the numeric transcription, it is possible to get an accurate evaluation of the presence of the scale degrees (in this case, within the range of six degrees) in the three performers' singing style (Figure 3.3).

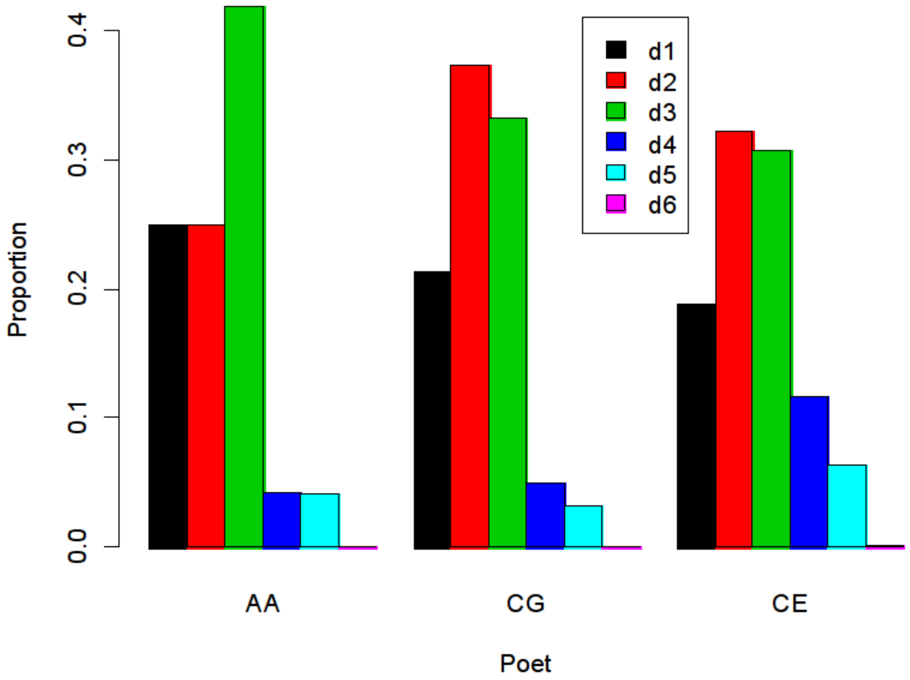


Figure 3.3: *Presence (in proportion) of the scale degrees in the performances of three poets*

By using the annotations regarding the performers along with the pitch extractor provided by PRAAT, one can accomplish a detailed examination of the musical intervals actually used by the singers. In Figure 3.4, a tonagram (i.e. a histogram representing the frequency distribution of sampled F0 values [82]) displays the interval structure of the performance of one poet, with a visible lowering of the third degree with respect to the tempered scale.

Linguists and musicologists, other than with annotations, would like to enhance every audio recording by expressing it with a set of information, chosen among the most relevant aspects of the recordings. Such metainformation concerned properties like author, title, object, recording date, etc., up to more technical information like the different singing types, speech types, accompaniment or instruments. The information could be used to manage recordings in the archive, because by describing them they allow for selection and organization, facilitating efficient retrieval and usage.

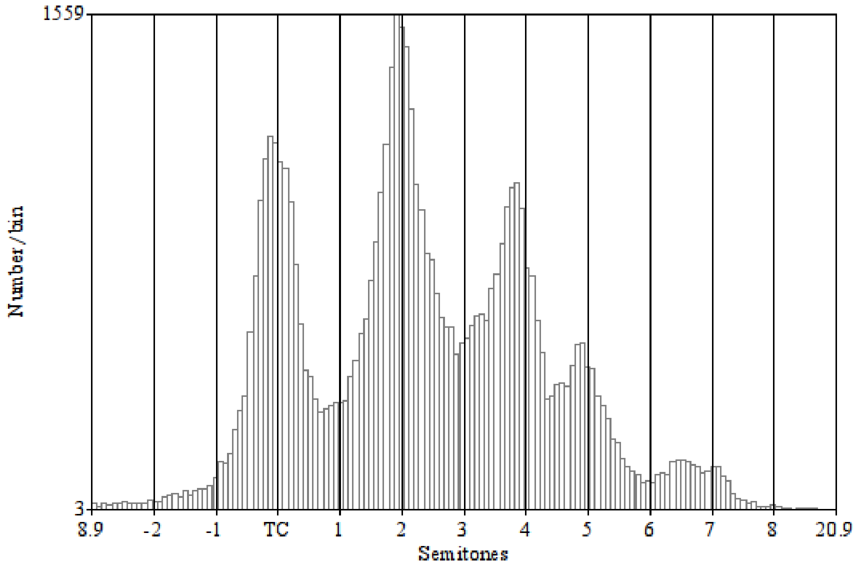


Figure 3.4: *Interval structure of one a s'arrepentina performance displayed through a tonagram*

3.3 Proposed approach

Our KB is represented not only by the texts of the corpus, but especially by the metalanguage and linguistic annotations that enhance them. For each audio clip, a set of metainformation describing the content is needed in order to enable the search and retrieval of data by local author, title, date of recording, to more particular features like linguistic variety or singing type. Each audio clip is also enhanced by a set of linguistic annotations. The management of the information associated to the audio clips required the formalization of all the associated metainformation in the form of a structured set of metadata.

The TD phase starts from the organization of information associated to each audio-clips like author, etc., using the DC standard as a reference and existing formalizations and definitions in the linguistic corpus.

To formalize this knowledge in the BU phase we chose an approach based on the classification of the annotations. In order to classify these annotations we identified a list of possible annotation levels (syllable, tone, morpheme, syntagm, accents, etc.), useful for both linguistic and musical analysis of audio recordings, able to represent the knowledge related to the Analytic Sound Archive of Sardinia. We used a KMS to manage this formalized knowledge with the annotated electronic corpus.

The formalization activities are described below. We defined a metadata schema to organize the information, using linguistic annotations. We defined the different kinds of metadata using two different orthogonal approaches: a TD approach for the information related to the files, and a BU approach to manage the linguistic annotations.

3.3.1 Top-down approach to define the general metadata

We used a TD or deductive approach to formalize the general semantic characteristics of the files in the corpus.

Twelve general metadata were found: title, author, publisher, object, contributor, date, place, occasion, document accessibility, language, description and format. Those metadata outline the necessary information to describe the spoken texts in the corpus, conveyed in a particular singing or speech type, the occasion in which the audio was recorded, and the linguistic variety it belongs to.

The TD approach further specializes the metadata. More specific, or qualified, metadata are represented by adding a qualifier to the name of the more general metadata and using the common syntax metadata.qualifier.

Lastly, "relational" metadata are defined as well, in order to define a certain relation among two or more different objects belonging to the corpus. An inclusion relation must be specified in order to describe the belonging of one or more objects to the same recording set, for example different songs in a singing contest.

We needed to manage and organize the information that made up the corpus, and we needed to associate organized and relevant information to a text when it was processed. Metadata schemas reflect the complex nature of data and are often strongly structured and hierarchical, including many kinds of metadata, with many different functions.

Building an effective system of structured metadata means creating a conceptual model to formalize and model the essential semantic characteristics of the knowledge domain. After designing the conceptual model of the knowledge domain, a TD approach was used to structure the metadata schema. The knowledge domain is an electronic corpus, and its objects are associated texts. We deducted and formalized the essential metadata (author, title, language, publishing date, etc.) using their semantic characteristics. Some of those metadata may be further specified according to a hierarchical structure: for example, the metadata “author“ may be further refined as “main author“, “illustrator“, and "curator".

We used both a deductive and an inductive approach to formalize the metainformation and linguistic annotations in a single structured metadata schema. We defined the metadata using the Qualified Dublin Core as main schema for interoperability reasons. There are four main criteria for choosing a metadata schema, with different approaches in metadata organization:

1 mapping of native metadata on existing DC elements;

2 mapping of native metadata on DC elements and creation of new customized qualifiers for DC elements;

3 creation of a customized metadata schema, identical to the native metadata set;

4 creation of DC metadata records as abstractions of native metadata records and entering the latter as attachments to the resource.

Out of the criteria mentioned above, the first one is the least satisfactory for preservation and reuse of descriptive metadata of resources, while the third one is the most preserving of the integrity and granularity of original metadata but needs greater efforts for the creation of a customized metadata schema, together with high maintenance costs for the archive. The second and fourth criteria combine preservation and granularity needs with archive management costs better than the other two. Choosing between them depends solely upon the particular requirements of the archive.

Once the decision on which criterion to use is settled, the archive must be configured so that it is compatible with the approach of choice for metadata management. In particular, if the second criterion is adopted, the DC schema must be updated with new, customized qualifiers; if the third criterion is chosen, the entire metadata schema created ad hoc must be entered into the system. In this way, customized metadata and qualifiers can be used to describe texts of the corpus inside the archive.

Linguists and musicologists selected 38 pieces of metainformation associated to audio recordings: title, author, object, description, format, etc.

Choosing a metadata schema

For legal reasons and interoperability needs, importing the metadata schema that was just created into the KMS may not have been appropriate or convenient. Most archives use Qualified DC as a main schema for indexing and displaying metadata and Simple DC to show them through the OAI-PMH standard. Therefore, the adoption of Dublin Core must be thoroughly evaluated when an archive must be compliant with the interoperability principles required by OAI.

The most suitable technique for the case study is a hybrid model between the mapping of native metadata on DC elements and the creation of new customized qualifiers for DC elements using the second criterion among the criteria listed above and the creation of a customized metadata schema, identical to the native metadata set using the third criterion. The latter criterion is more convenient to organize linguistic annotations that are compliant with the rules of linguistic corpus, so that a dedicated metadata schema can be created to preserve their granularity; while the second criterion is best suited for all other metadata, because it combines the advantages of granularity as provided by qualifiers to interoperability provided by DC metadata.

Application profile for the ASAS

In creating a specific application profile for the ASAS, we used a "conservative" approach towards the original Qualified DC elements and qualifiers in order to use as many of them as possible for the formalization of descriptive and relational metadata. We created a special schema instead for annotations. Its metadata were entered into the DC application profile as outlined in Table 3.1.

Table 3.1: *Application profile for the ASAS*

Metainformation or ASAS Annotation	**DC Application Profile Metadata**
Title	dc.title
Author	dc.creator
Publisher	dc.publisher
Object	dc.type
Description	dc.type.category
Contributor	dc.contributor
Annotator	dc.contributor.annotatore
Location	dc.coverage.spatial
Date	dc.date.created
Occasion	dc.subject
Source	dc.relation.isbasedon
Document Accessibility	dc.rights
Performer	dc.contributor.sperakerPerformer
Performer's Age	dc.description.speakerPerformer
Performer's Place of Origin	dc.description.speakerPerformer
Language	dc.language
Source Completeness	dc.description.integritá
Source No.	dc.relation.ispartofseries
Source Section No.	dc.relation.ispartofseries
Document Type	dc.format.audioVideo
Format	dc.format.medium
Acquisition Method	dc.format.modoAcquisizione
Reading Type	dc.type.lettura
Interview Type	dc.type.intervista
Monody Type	dc.type.monodia
Unison / Heterophony	dc.type.unisonoEterofonia
Accompaniment Type	dc.type.monodiaAccompagnamento
Polyphony Type	dc.type.polifonia
Instrumental	dc.type.strumentale
Instrument	dc.type.strumento
Singing Type	dc.type.tipoCanto
Other	dc.description
Syllable	asas.annotazione.sillaba
Tone	asas.annotazione.toni
Morpheme	asas.annotazione.morfema
Phone	asas.annotazione.fono
Word	asas.annotazione.parola
Part of Speech	asas.annotazione.pos
Syntagm	asas.annotazione.sintagma
Sentence	asas.annotazione.frase
Information Structure	asas.annotazione.strutturaInformativa
TurnPerf	asas.annotazione.turnPerf
Musical Syllable	asas.annotazione.sillabaMusicale
Metric Segment	asas.annotazione.segmentoMetrico
Musical Segment	asas.annotazione.segmentoMusicale
Tonal Centre	asas.annotazione.centroTonale
Notation	asas.annotazione.notazione
Ornamentation	asas.annotazione.ornamentazione
Accents	asas.annotazione.accenti
Melismatic Syllable	asas.annotazione.sillabaMelismatica
ADD1	asas.annotazione.annotazioneLibera

Subsequently, we inserted the metadata in the KMS: once metainformation were organized and structured, the KMS was configured so that it could be adapted to the selected metadata schema.

3.3.2 Bottom-up approach to define the linguistic annotations

The formalization of annotations in a metadata schema can be achieved using a BU or inductive reasoning, as explained in the previous section. The structure of annotations is defined by the PRAAT software according to the linguistic annotation previously defined. Annotations are organized with a precise structure: each annotation is made of a time interval and a text label or of an instant and a marker with its text. All annotations in the same linguistic category are collected in the same tier (or annotation level), which can be considered as the category they belong to, giving its name to the corresponding metadata. In this way, a repeatable metadata is found in each annotation level of the TextGrid (the text file where PRAAT stores all tiers with their own segmentations and annotations) and each annotation can be represented as multiple occurrences of that metadata. We defined a metadata for each annotation level, defined by the domain experts, used as tier in PRAAT. For each metadata, we defined, using the BU approach, the specific attributes and type.

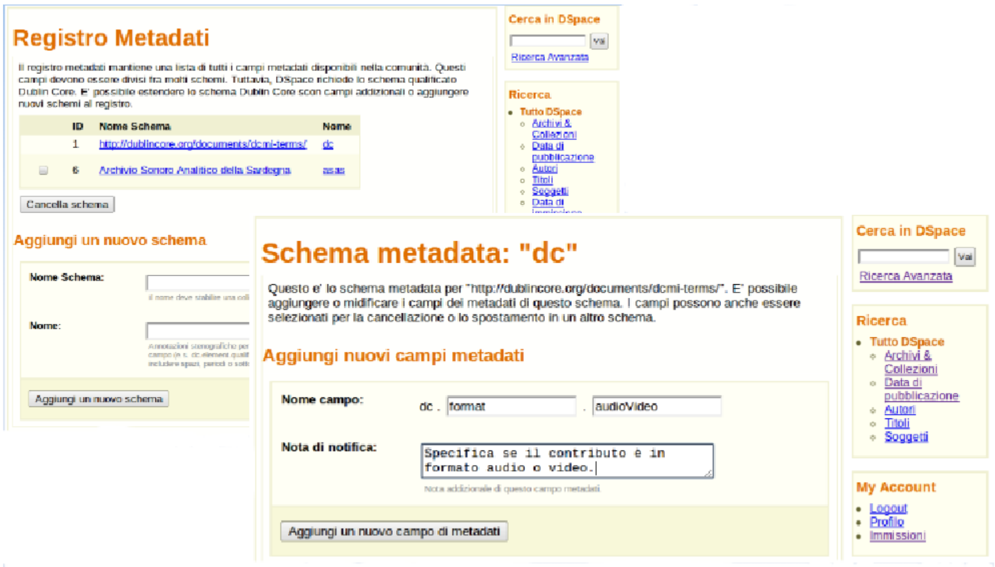
We organized the information in the corpus by formalizing those annotations through metadata schemas, using the informal annotations made by the domain experts. We used this approach to associate the annotations to their texts, using the selected linguistic level.

The linguistic annotations for audio files created using PRAAT are generally stored in a semi-structured manner. In fact, each annotation is distinctly represented inside the file, according to a defined, repetitive structure where the annotation texts is paired with the instant or the time interval it refers to.

Managing linguistic annotations

We propose an approach that allows to insert the corpus and the associated knowledge inside of DSpace, ensuring the maintenance of its structure and the ability to query and update it easily by adding or modifying its contents. Each text of the corpus is inserted into a DSpace item so that it can be uniquely associated with all of the metadata needed for the linguistic analysis. The audio file contains the recordings, and the original files with the annotations are loaded inside of the item as a bitstream, while metadata are stored in the system database.

Our first step was the customization of new qualifiers for the Dublin Core descriptive metadata representation and the creation of a new scheme called "asas" for the representation of the annotations. When inserting the corpus into DSpace it was decided to create a specific item for each audio clip. It was therefore necessary to set the release wizard offered by DSpace by changing the specific XML file responsible for entry forms ('input-forms.xml'). The descriptive metadata, identified by researchers, such as title, author, type of song, instrument, etc., and all metadata corresponding to linguistic annotations (phone, morpheme, word, etc.), were associated to each item, together with the original file containing the audio recording and the original annotation file.



Registro Metadati

Il registro metadati mantiene una lista di tutti i campi metadati disponibili nella comunità. Questi campi devono essere divisi fra molti schemi. Tuttavia, DSpace richiede lo schema qualificato Dublin Core. E' possibile estendere lo schema Dublin Core scon campi addizionali o aggiungere nuovi schemi al registro.

	ID	Nome Schema	Nome
	1	http://dublincore.org/documents/dcmi-terms/	dc
□	6	Archivio Sonoro Analitico della Sardegna	asas

Schema metadata: "dc"

Questo e' lo schema metadata per "http://dublincore.org/documents/dcmi-terms/". E' possibile aggiungere o midificare i campi dei metadati di questo schema. I campi possono anche essere selezionati per la cancellazione o lo spostamento in un altro schema.

Aggiungi nuovi campi metadati

Nome campo: dc . format . audioVideo

Nota di notifica: Specifica se il contributo è in formato audio o video.

Nota addizionale di questo campo metadati

Figure 3.5: *Customization of DSpace metadata register*

After the insertion of metadata, we customized the interface by replacing the standard forms provided by DSpace using modules specifically designed to allow the creation of items and the release of DC metadata according to the specific needs of the project. The metadata on the annotations were inserted instead using direct import because the high number of occurrences for each item made it difficult to enter them manually, as shown by Hillman and Westbrooks [70].

Finally, we proceeded to customize the search interface of DSpace in order to adapt it to new metadata and to the particular needs of the Analytic Sound Archive of Sardinia. In essence, all metadata corresponding to linguistic annotations needed to be indexed in DSpace's search engine so that we could find a certain audio clip even through the search of an associated record. Furthermore, some descriptive metadata such as location, type of performer and contribution were indexed to allow effective searching that exploited the granularity of the metadata.

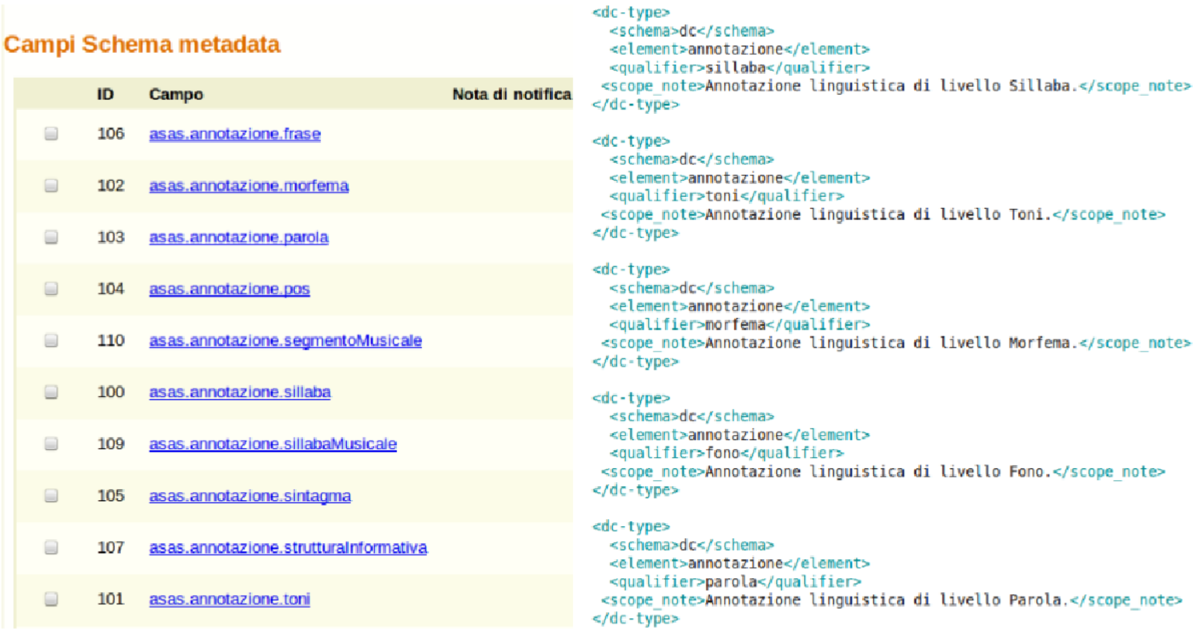
3.3.3 Implementing metadata schema and inserting in DSpace

The metadata are stored and managed by DSpace through a special tool, the Metadata Registry, where the Qualifed Dublin Core schema is configured by default. It can nevertheless be changed, and new customized schemas can be added.

Once a schema is created, the metadata can be added one at a time, with any qualifiers and related notes. This feature is crucial for the updating and maintenance of the system as it can make adjustments quickly and easily without the intervention of a computer expert. Likewise, one can choose the second solution, in which using a specific command makes it possible to import the metadata schema expressed in XML into the register, according to a specific syntax.

During the creation of the Sound Archive, metadata was to be gradually defined and refined by linguists and musicologists, so it was decided to insert and manage it through the DSpace Web interface.

We obtained a customized Qualified DC schema with new specific qualifiers and a new schema “asas” with the metadata to use for the insertion of linguistic annotations, as shown in Figure 3.6.



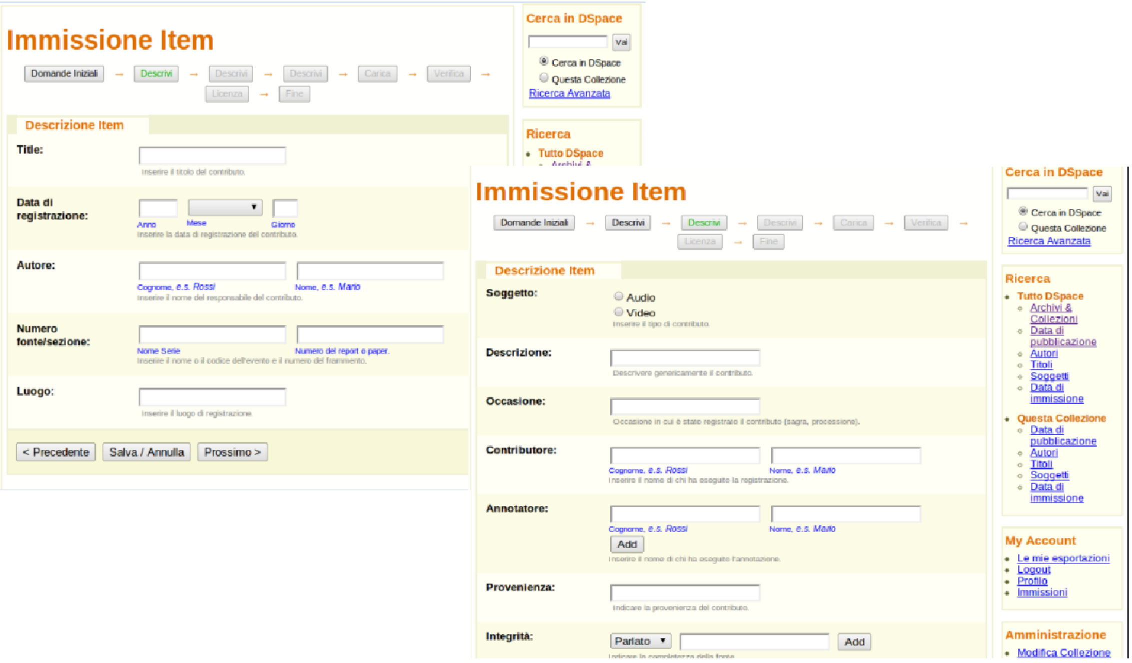
Campi Schema metadata

	ID	Campo	Nota di notifica
☐	106	asas.annotazione.frase	
☐	102	asas.annotazione.morfema	
☐	103	asas.annotazione.parola	
☐	104	asas.annotazione.pos	
☐	110	asas.annotazione.segmentoMusicale	
☐	100	asas.annotazione.sillaba	
☐	109	asas.annotazione.sillabaMusicale	
☐	105	asas.annotazione.sintagma	
☐	107	asas.annotazione.strutturaInformativa	
☐	101	asas.annotazione.toni	

```
<dc-type>
  <schema>dc</schema>
  <element>annotazione</element>
  <qualifier>sillaba</qualifier>
 <scope_note>Annotazione linguistica di livello Sillaba.</scope_note>
</dc-type>

<dc-type>
  <schema>dc</schema>
  <element>annotazione</element>
  <qualifier>toni</qualifier>
 <scope_note>Annotazione linguistica di livello Toni.</scope_note>
</dc-type>

<dc-type>
  <schema>dc</schema>
  <element>annotazione</element>
  <qualifier>morfema</qualifier>
 <scope_note>Annotazione linguistica di livello Morfema.</scope_note>
</dc-type>

<dc-type>
  <schema>dc</schema>
  <element>annotazione</element>
  <qualifier>fono</qualifier>
 <scope_note>Annotazione linguistica di livello Fono.</scope_note>
</dc-type>

<dc-type>
  <schema>dc</schema>
  <element>annotazione</element>
  <qualifier>parola</qualifier>
 <scope_note>Annotazione linguistica di livello Parola.</scope_note>
</dc-type>
```

Figure 3.6: *The new schema “asas”*

Once we reached the final version of the schema, however, we decided to formalize the metadata in XML so that, should the system be reinstalled or transferred, the register could be quickly configured via the “import metadata“ command. DSpace is preset to use, during the insertion phase of an item, the Qualified Dublin Core metadata schema but at the same time allows to customize it, or to create new metadata schemas according to the users and their requirements. The first step in order to submit the metadata in DSpace was to organize and structure them as metadata schemas. Our second step was instead to configure the KMS so that it could be adapted to the particular metadata schema we had chosen.



Descrizione Item

Title:
Inserire il titolo del contributo.

Data di registrazione:
Anno Mese Giorno
Inserire la data di registrazione del contributo.

Autore:
Cognome, e.s. Rossi Nome, e.s. Mario
Inserire il nome del responsabile del contributo.

Numero fonte/sezione:
Nome Serie Numero del report o paper.
Inserire il nome o il codice dell'evento e il numero del frammento.

Luogo:
Inserire il luogo di registrazione.

Descrizione Item

Soggetto:
○ Audio
○ Video
Inserire il tipo di contributo.

Descrizione:
Descrivere genericamente il contributo.

Occasione:
Occasione in cui è stato registrato il contributo (sagra, processione).

Contributore:
Cognome, e.s. Rossi Nome, e.s. Mario
Inserire il nome di chi ha eseguito la registrazione.

Annotatore:
Cognome, e.s. Rossi Nome, e.s. Mario
Inserire il nome di chi ha eseguito l'annotazione.

Provenienza:
Indicare la provenienza del contributo.

Integrità:
Indicare la completezza della fonte.

Figure 3.7: *Interface customization metadata input*

The graphic interface offers a tool to manage the Metadata Register that allows to customize the preconfigured Dublin Core schema in a quick and intuitive way.

Every new qualifier can be entered in the Metadata Register with its related comments to obtain the desired application profile, which coexists in the Qualified Dublin Core standard with its qualifiers, made for the specific project.

In DSpace this entity is called “item“ and is "built" with a wizard that, using the preconfigured modules, allows to specify the values of the meta-information to be formalized as metadata. These modules are ready for the insertion of the metadata belonging to the Qualified Dublin Core schema, but in the case study they were fully customized to fit the new specific qualifiers for the archive.

Research modules

DSpace offers a powerful search interface that is configured to use the main DC metadata (like title, author, language) or the free full-text research as parameters for the search.

The researchers asked for specific search criteria, besides the standard ones, so one can search for audio clips by place of recording, by the participation to a particular event, but also by the annotations they contain.

Search indexes were changed so that all needed metadata were selected as criteria in the search interface and information like place of recording, performers' information, the number indicating the event place and all metadata corresponding to the levels of annotation were specifically added.

Chapter 4

The Issue of Construction Processes

The amount of information needed by all the operatives involved in a construction process to work properly and successfully is always growing. For this reason, construction processes - while still largely relying on intuition and experience - need to be rationalized through new procedures and tools for a strict formulation and implementation of efficiency criteria.

The purpose of this study is to show how KM techniques may be one of those tools, supporting those activities through a rational organization of the large quantity of data/information and a capitalization of consolidated knowledge.

KM is described as follows: *"The Knowledge Management is the systematic, explicit and deliberate organization, application and renewal of a company internal knowledge, aiming at maximizing the effectiveness of the cognitive ground and of the related advantages"* [83]. This definition makes it easy to understand why including a KM policy in an organization means considering knowledge as a key resource to develop, capitalize, and share, that will determine the future of its operating strategy. *"Knowledge is the information that changes and modifies the organization, making the agent capable of new and/or effective actions"* [84].

Introducing a KM policy into a company means making knowledge into a key wealth, to develop, capitalize, share, and to use as a base for a companyâs operational strategy. The aim of KM is, in fact, to express, making it accessible to the entire company, all the knowledge that every operative has gained with their work, so that the company can gain an advantage both economically and from a service quality point of view. An increase in performance and competitive advantage are the main benefits of KM; this is the reason why more and more effort and resources are being spent to define and implement KM policies into companies [85][86]. Some of the application of this research may involve third-party inspection services (verification and validation of the project, technical control of the building) in the construction sector. In fact, tools that can manage elementary products as defined are the foundation of good quality in project validation for public works, thus being vital for a systematic approach by contracting authorities. One of the instruments of KM is its KB: developing a KB means rationalizing and clearly conveying the dynamics and know-how structure of a company [1][3][87].

This work sought a rational organization of large amounts of data using the knowledge that characterizes the various stages of a construction process.

The first step to implement a KMS is to define its base content, schemes and structures, in order to enter and offer the knowledge collected by all the participants to a project. We suggest the concept of elementary product, described further below, as the basic unit needed to create the KB of a construction project.

In this chapter, we present an overview about the construction process and the state of the art (in the first and in the second section), and we propose an approach to formalize the knowledge associated with construction processes (in the third section).

4.1 Related concepts

According to some researches, knowledge exchange in the construction industry is based on non-developed models [88], and studies for the application of KM techniques to the sector were developed only recently, as proven in [89] and [90]. An essential aspect of that is the development of tools to support management of variables in construction processes [91]. Tools are being defined that could make the flow of information pertaining a construction project more efficient and unequivocal, outlining a new model that includes both a qualitative description of the work and its production. It means structuring projects so that the information they contain can flow efficiently, without letting construction site the option of inferring things that could cause substantial changes.

The research starts from the development of preliminary concepts, described also in [92], functional to the innovative approach introduced above. Limiting the chances of inferring, in fact, is giving an objective value to the project, which now can register all those reasoning the designer does not report for brevity's sake but that would offer an unequivocal interpretation to all the other professionals (designers, commissioners, builders). It actually means borrowing the approach from the techniques of Project Management: it starts from the description of the building through a multi-level tree structure (i.e., creating a Project Breakdown Structure, PBS). This approach allows for a description where components are listed in detail, down to the most basic ones.

Currently, many international researches have been developed, using different approaches: the use of KM techniques and the theorization of virtual models suggested that knowledge sharing and the ability to manage the whole cycle of knowledge is indispensable for the process, so that no knowledge is lost.

A hierarchical knowledge structure is defined in [93], starting from information and applying it to a specific context. Contextualization of information is one of the pre-requisites of the construction sector, so approaches to safety during manufacturing [94], and timing and budgeting algorithms [95][97], were developed with that focus.

KM is based on information tools and cutting-edge technologies, defined and developed in the last 15 years, where knowledge has become the real added value, and as such, the real competitive advantage for those companies that choose to organize it [97].

4.2 Domain knowledge

The KB started from the experience in building projects of the DICAAR[1] team, which had information and objects from many real building projects. We selected for the analysis a representative subset of that project for different kinds of buildings: Hospitals, Primary Schools, Houses, University Departments, etc.

[1] DICAAR, http://idra.unica.it

A construction process is a very complex process, with many legal constraints and technical elements, like plans, design, construction site pictures, product data sheets, construction notes, etc. Each construction project has many associated objects: a simple house construction project could produce 100 different objects. A more complex construction project, like that of a hospital, could produce 1000-2000 different objects during its life. Many of those objects are multimedia objects.

Our KB contains several thousands of elementary objects and 80 building projects. For example, a product data sheet is a document summarizing the performance and other technical characteristics of a product, component, material, in sufficient detail to be used by a design engineer to integrate the component into a system. Depending on the specific purpose, a data sheet may offer typical values, tolerance, colours. Specific materials have technical data in specific sheets: thermodynamic properties, spectral data, vapour pressure, etc.

4.3 Proposed approach

We needed to formalize already well-defined knowledge to extract information embedded in the objects produced in the construction process [98][99].

The TD phase started by splitting this process into subprocesses in an iterative approach, in order to define the elementary components and objects involved in the process. The analysis can work orthogonally with a breakdown process of the building objects in sub-elements.

The BU phase analyzed the objects created in the construction process and the information associated to them. The objects are varied and with different kinds of information.

Using the mixed approach, we could group the objects analyzed during the BU phase in elements with a semantic meaning based on the Elementary Product (EP) concept defined in TD phase.

The formalized knowledge could be managed using a KMS, using defined metadata and the multimedia objects as defined with the analysis method introduced above. Our choice has fallen on DSpace, because we had to manage many multimedia objects and we wanted to promote availability of that information also for maintenance purposes. DSpace is designed as a central storage facility able to collect various types of digital resources: text, images, video, audio, articles, technical reports, working papers, datasets, etc.

4.3.1 Top-down analysis: defining the elementary product

The TD phase started from the theory to split the construction process in subprocesses in a logic that uses an iterative approach, to define the EP and objects that are involved in the different phases of the building process. We proceeded with an orthogonal breakdown of the building objects in sub-elements.

Describing the building object as a tree structure with several levels, following the TD technique [100], lead to a representation that defines all of its components down to the most elementary ones. The building object was resolved into three elements, called macro products. They were further subdivided into products and by-products, progressively less complex, to the level of desired breakdown. Such a procedure allowed us to work on smaller and smaller portions, more easily controllable and manageable, coordinated by a production simulation. The levels at the base of that hierarchical tree showed an in-depth and detailed definition of the work needed for the final product; the optimal breakdown level appears to be the one where the elements are:

1 flexible, interchangeable with other elementary products of different quality;

2 identifiable, and assigned to a manager;

3 manageable: of a determinable duration and cost;

4 measurable in their results;

5 significant and interface-able in their specific requirements.

Moreover, they had an identification code that highlighted their sequential order in the structure.

Should we make an example, it is easy to understand how de-structuring and performing a PBS (through production simulation) on a building leads to marking the elementary products. In particular, the figure shows that the product enclosure was split. This operation led, within the hierarchical structure, to the identification of the elementary products: brick wall and the PVC window. We show the related technical sheets in the next section. The elementary products are included into the structure as a group of vertical elements, placed in a given position, with a given dimension, made with form-work, etc. All these pieces of information, despite belonging to the same elementary product, are not to be conveyed to every person involved in the project, but are organized in a structure through which each person can access them differently.

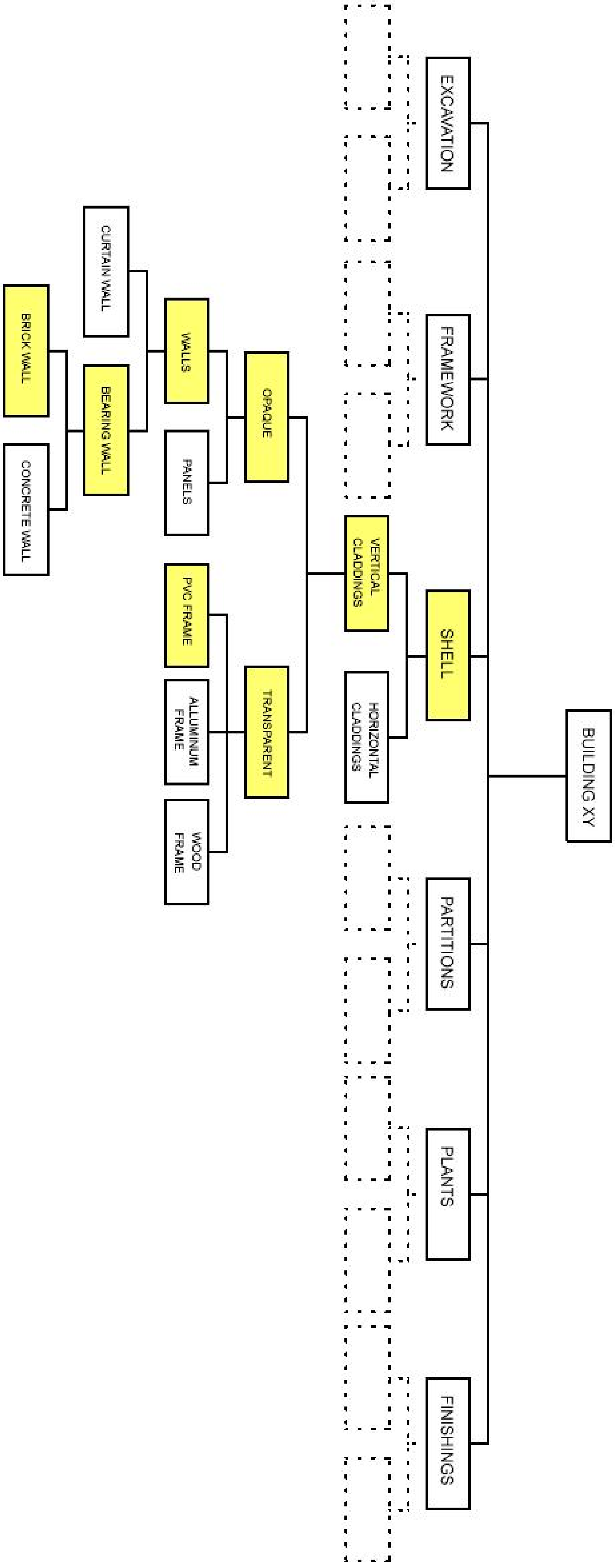


Figure 4.1: *An excerpt of a P.B.S.*

The creation of the PBS and its efficacy in a process are directly influenced by the level of accuracy used to identify all the parts of the building object. The breakdown process finishes when the required level of appropriate accuracy is reached. It is important to remember that the breakdown level varies according to the characteristics of the work to carry out. In fact it is correct to say that the PBS can be divided into any number of levels, according to the intervention complexity. Nevertheless, if the de-structuring is extreme, it is difficult to keep track of the general state of the work, particularly if it has a long-term planning. The products that belong to the lowest level of the breakdown are the Elementary Products [101].

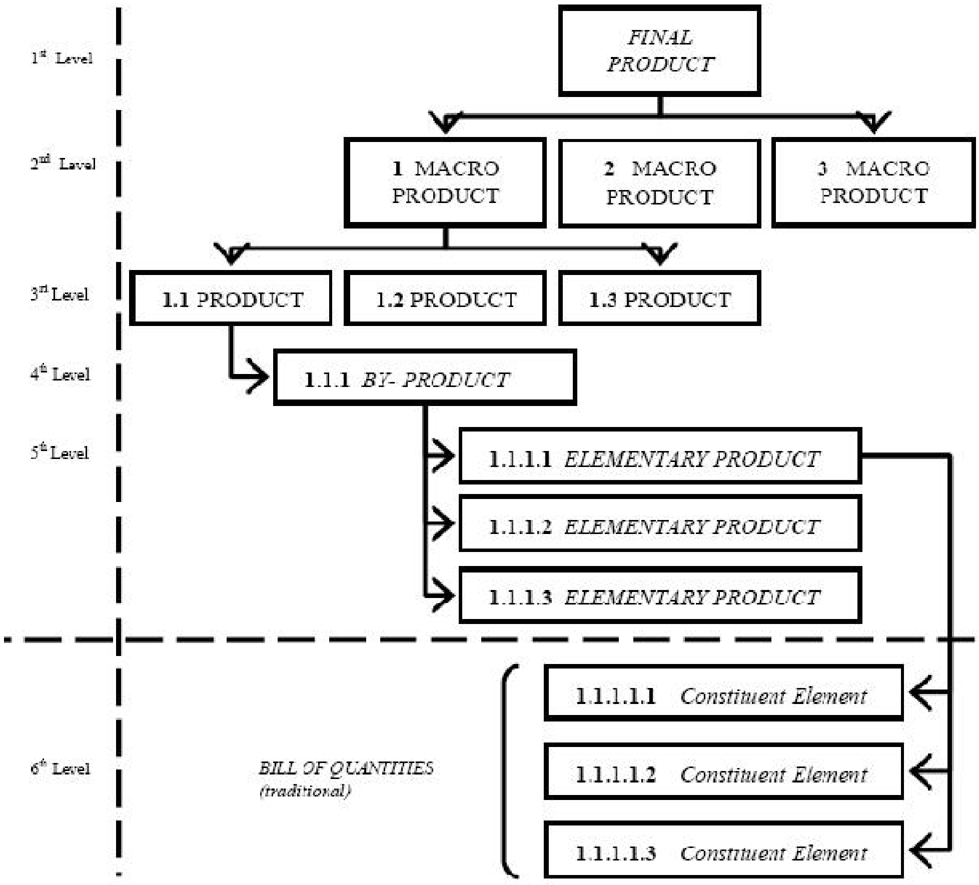

1st Level
2nd Level
3rd Level
4th Level
5th Level
6th Level
FINAL PRODUCT
1 MACRO PRODUCT
2 MACRO PRODUCT
3 MACRO PRODUCT
1.1 PRODUCT
1.2 PRODUCT
1.3 PRODUCT
1.1.1 BY- PRODUCT
1.1.1.1 ELEMENTARY PRODUCT
1.1.1.2 ELEMENTARY PRODUCT
1.1.1.3 ELEMENTARY PRODUCT
BILL OF QUANTITIES (traditional)
1.1.1.1.1 Constituent Element
1.1.1.1.2 Constituent Element
1.1.1.1.3 Constituent Element

Figure 4.2: *Hierarchical breakdown diagram (P.B.S.)*

The breakdown level, which the elementary product belongs to, allows for an effective management and control of the process in regard to the economic, time, and quality properties. So the project becomes the conception of a building object in relation to the production possibilities and methods, and to its employment and maintenance.

The elementary product, which represents the basic unit of the KB, is configured as the sum of four basic knowledge units, defined as follows:

- EPd: elementary design product;
- EPe: elementary executive product;
- EPc: elementary constructional product;

- EPm: elementary managerial product.

According to the four views of the EP, the building process is divided into four phases:

1 definition of architecture: based on a set of needs expressed by the customer, the designer defines the architecture of the building object, that is broken down and described as a set of elementary design products (EPd) related to each other. In order to meet both the constraints and the needs, the technical and performance characteristics are specified. Therefore, at this stage EPds are structured as a real storage of architectural design data, information and knowledge;

2 project engineering: after capitalizing on the information and the knowledge about the object in terms of EPd, each identified elementary product is defined, and as a consequence, the building itself is interpreted in terms of production techniques, technologies, resources, activities, etc. EPes are structured to contain all data, information and knowledge related to this stage;

3 construction: thanks to the capitalization of all the information on the specific products and materials selected and used to meet performance and requirements declared in EPe, EPe evolves in EPc during the accomplishment of the building process;

4 management and maintenance: EPcs are reliable and updated storages of information and knowledge, and a starting point to run and maintain the building object. Building deterioration, due to time, requires a planned ordinary and/or extraordinary maintenance, and consequently it is essential to record all information related to the life of the building and to its elementary products. The EPm is the basic unit to capitalize on the information and the knowledge concerning the building management and maintenance.

The building process gradually progresses, and EPd first becomes EPe, then EPc and finally EPm. Such a development is the integration of the information and the knowledge acquired during the Project Engineering and Construction stages. The EP is the outcome of the four structures defined above. Therefore, the EP has to keep track of all information and knowledge of a specific building process, including As Built documents and feedback on use. With respect to this aspect, in Italy, as in most European countries, authorities require drawings of the object to be built immediately after the design phase, while as-built drawings are not mandatory after construction. However, many changes occur during the construction phase, and a lack of information on such changes makes maintaining and/or renovating existing buildings particularly difficult and onerous. Moreover, the lack of usersâ feedback is an obstacle to innovate and develop new and more appropriate products and/or construction criteria for future building activities.

During the whole building process, EP is the basis for all parties involved. In fact, at any time they can dialogue and cooperate, and be kept up to date about the evolution of the process in terms of elementary products. Moreover, each involved actor can modify and/or add data, information and knowledge concerning each EP. Each EP is analyzed from different aspects (EPd, EPe, EPc, and EPm), that are complementary, since they represent different development stages of a specific building process.

4.3.2 Bottom-up analysis: the building objects

The BU analysis started from the objects produced in the construction process and the information associated to them. These objects are varied and rich in many kinds of information. We started to analyze these very different objects. We have many different kind of objects gathered during the different phases of the construction process, such as designs (Figure 4.3), pictures, technical sheets/specifications (Table 4.1 and Table 4.2), notes, etc.

The analysis shows that we have a KB with too much heterogeneous associated information, multimedia information, design information, product attributes (such as thermal resistance, insulating capability, etc.) and other information. All of those information can be represented using metadata, but their number changes depending on the object that we are analyzing, as shown in the data sheets in Table 4.1 and Table 4.2.

The main goal of the study is to make the knowledge available also for searching purposes in a smart mode. Making a system that manages all those information can be a solution for a database of all the elements involved in the construction process, but cannot be a solution to manage the knowledge using a KM approach.

We needed to manage the information at a higher level. We have to group the information in a single object and manage it as knowledge element. We used the semantic concept of EPs to aggregate these information and make available the informations using this level of abstraction.

Table 4.1: *Technical properties of brick wall*

Property	**Standard**	**Value**
Density	DIN 53420	Av 33 kg/m^3
Compressive Strength	ISO 3386	0.024 N/mm^2
Compression Set	ISO 1856	14%
Tensile Strength	ISO 1798	0.25 N/mm^2
Elongation at Break	ISO 1798	100%
Tear Resistance	DIN 53575	1.28 N/mm
Thermal Conductivity	ASTM C-177	0.038 W/mK
Water Absorption	DIN 53428	0.8 vol%
Water Vapour Transmission 230C	DIN 53429	23 $\mu g/(m^2$s)
Permeability	ISO 1663	10 ng/(Pa.sm^2)

Table 4.2: *Technical properties of PVC window*

Property	**Value**
Density (g/cm^3)	1.38
Tensile Strength (psi)	10,200
Tensile Modulus (psi)	425,000
Tensile Elongation at Break (%)	36
Flexural Strength (psi)	14,000
Flexural Modulus (psi)	425,000
Compressive Strength (psi)	12,000
IZOD Impact Notched (ft-lb/in)	0.52
Coefficient of Linear Thermal Expansion (x 10-5 in./in./°F)	7.0
Heat Deflection Temp (°F / °C) at 264 psi	138 / 59
Vicat Softening Temp (°F / °C)	52 / 67
Max Operating Temp (°F / °C)	130 / 54
Surface Resistivity (ohms/square) at 50% RH	10^6
3mm Transparent Clear Transmittance - Total (%)	69
Haze (%)	6

The Figure 4.4 shows Elementary Products (as part of WBS) PVC window and brick wall. Each EP keeps, together with its attributes, different data gathered during the different phases of the construction process, such as designs, pictures, technical sheets/specifications, notes.

An analysis on which kind of information is actually described is then necessary. Properties, considered as attributes, that could be searched in the context are stored in two bulk metadata fields called “General Description” and “Technical Description”, where the information are not managed as structured metadata, but with a free logic like in folksonomies (what is considered more interesting is tagged). The technical sheet becomes then the tool through which information is not transformed into structured metadata but left as information belonging to an object, so that it can be searched according to the most peculiar attributes of that same object. We selected only two important, according with the semantic of the Elementary Products, metadata for searching purposes: ’ProjectName’ and ’Phase’. These information qualify the Elementary Product as the Elementary Product associated with a Project Phase of a specific project, qualifying the single EP for a specific project. The other important information must be stored in the bulk metadata fields ’General Description’ and ’Technical Description’. The experts storing the objects decide which kinds of information have to be stored in these fields as folksonomies.

Where “thermal resistance” is important, it is marked with a proprietary tag (like in folksonomies) inside the general description, while most other attributes are stored inside the object. Naturally, important attributes vary depending on each case, and on each EP, so the description could show “designer name“, “planning supervisor name“, etc.

A simple management system is thus created, where knowledge elements are classified following a folksonomies logic, instead of structured information, but are available also as a full text search in these fields.

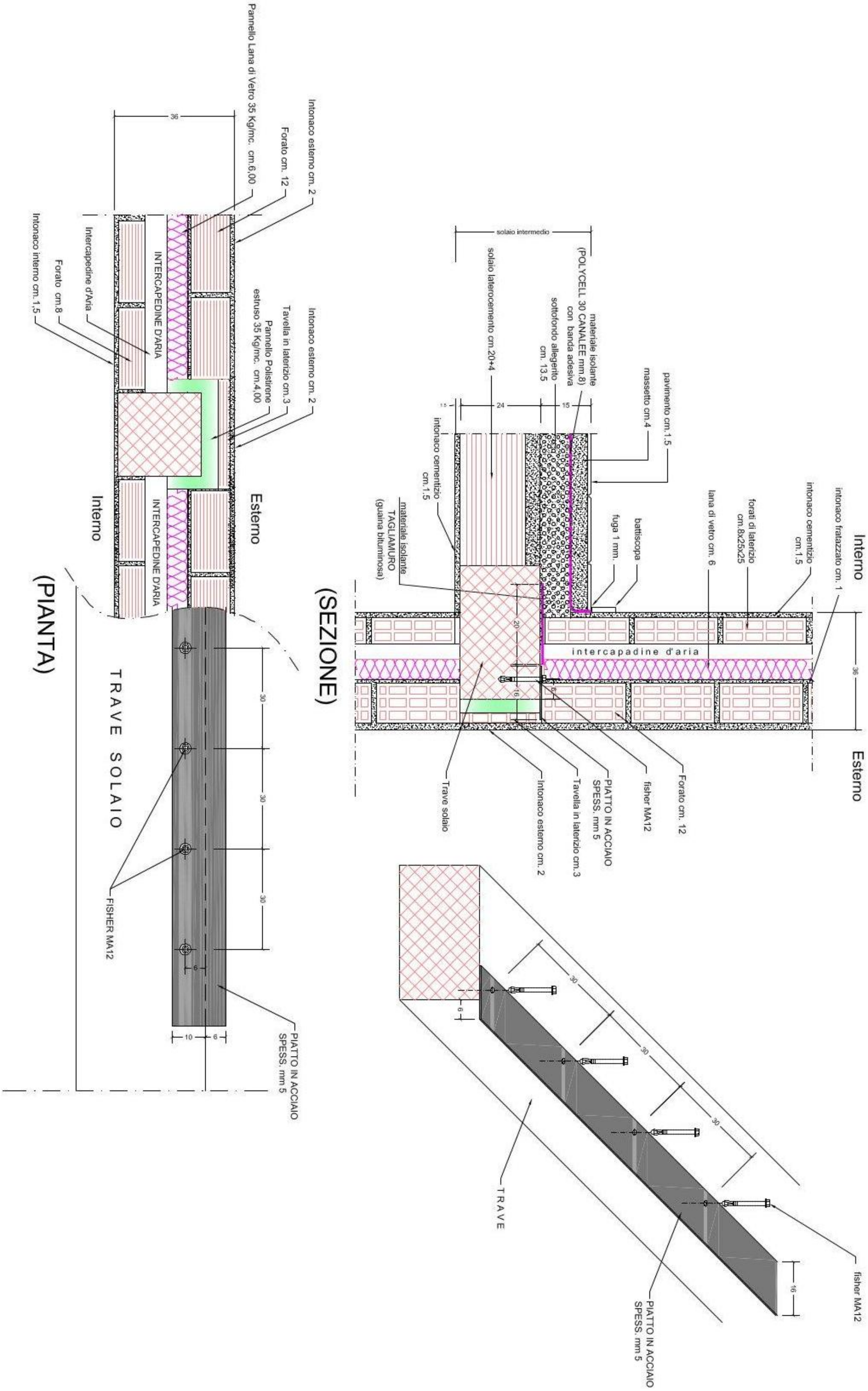

(PIANTA)
Pannello Lana di Vetro 35 Kg/mc. cm.6,00
Intonaco esterno cm. 2
Forato cm. 12
Intonaco esterno cm. 2
Tavella in laterizio cm.3
Pannello Polistirene estruso 35 Kg/mc. cm.4,00
Intercapedine d'Aria
Forato cm.8
Intonaco interno cm.1,5
INTERCAPEDINE D'ARIA
Esterno
Interno
36
TRAVE SOLAIO
FISHER MA12
PIATTO IN ACCIAIO SPESS. mm.5
30
6
10
(SEZIONE)
solaio intermedio
materiale isolante (POLYCELL 30 CANALEE mm.8) con banda adesiva
sottofondo alleggerito cm. 13,5
massetto cm.4
pavimento cm.1,5
solaio laterocemento cm.20+4
intonaco cementizio cm.1,5
materiale isolante TAGLIAMURO (guaina bituminosa)
fuga 1 mm.
battiscopa
lana di vetro cm. 6
forati di laterizio cm.8x25x25
intonaco cementizio cm.1,5
intonaco fratazzato cm. 1
Interno
Esterno
intercapedine d'aria
1.5
24
15
20
16
36
Forato cm. 12
fisher MA12
PIATTO IN ACCIAIO SPESS. mm.5
Tavella in laterizio cm.3
Intonaco esterno cm. 2
Trave solaio
TRAVE
30
6
16
fisher MA12
PIATTO IN ACCIAIO SPESS. mm.5

Figure 4.3: *Brick seal design*

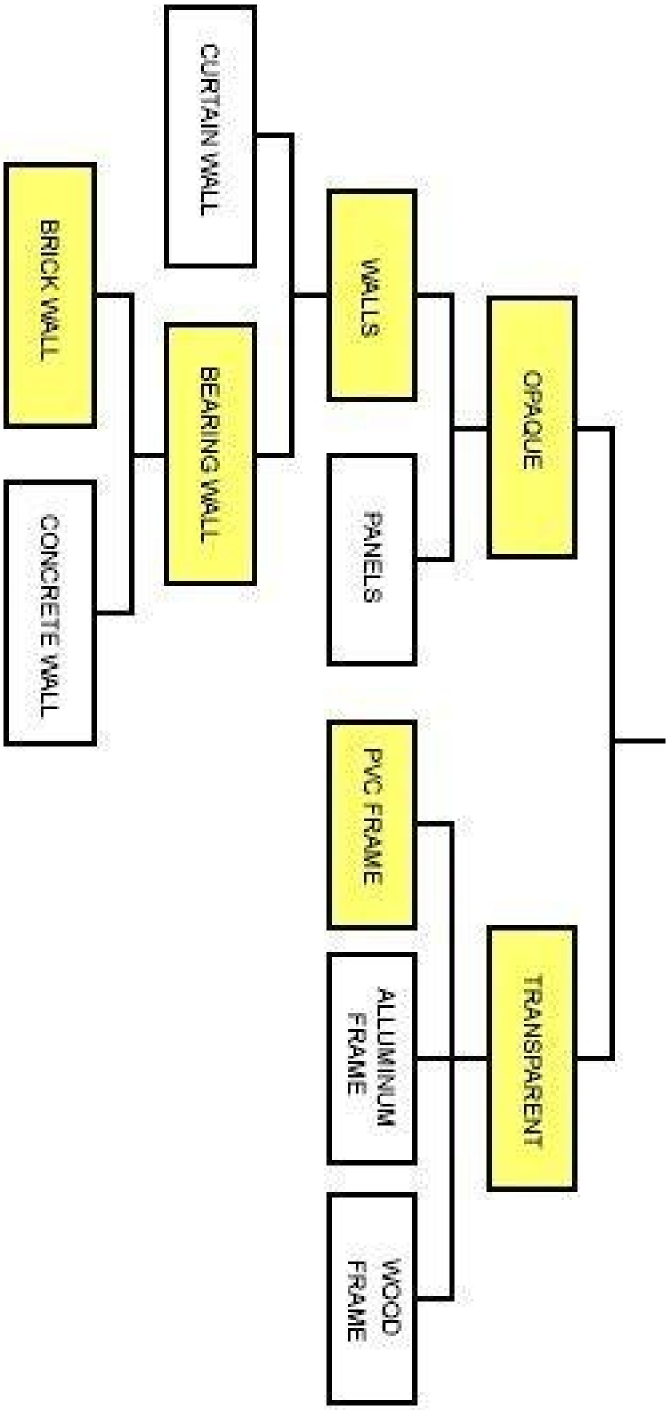


Figure 4.4: *The Elementary Products PVC window and brick wall*

Chapter 5

The Issue of Italian Wines Review

Over the last decade a broader knowledge of the Web has strengthened and fostered the developing of new applications: the Web has turned into a multifunctional platform where users no longer get the information passively; in fact, they become authors and makers. This has mainly been possible thanks to the developing of new applications which allow users to add contents without knowing any programming code. The social value which the Web has acquired recently is therefore unquestionable; the Web's structure grows and changes depending on the user's needs, becoming every day more complex.

The new frontier for the Internet is represented by the Web 3.0 [2]: with the evolution of the Web into its semantic version, a transition to a more efficient representation of knowledge is a necessary step. Particularly, data are no longer represented just by the description of their structure (syntax) but also by the definition of their meaning (semantics). In fact, a data can have a different meaning depending on the contexts; the use of tools like ontologies and taxonomies helps the classification of information, as shown also in [34], [44], [46], [50], [53], [102], [103], [104], [105] and [106].

This study aims at defining an approach for the problem of the contents on the Internet, especially semi-structured contents coming from heterogeneous sources referring to a common knowledge domain. Through a combined TD and BU approach, knowledge of a specific domain was extracted defining a common structure through a taxonomy, in order to classify and make the majority of such knowledge available. With the TD approach the knowledge of interest on the domain was defined, following the specifications and the analysis of the ontologies, in order to define a reference taxonomy. On the other hand, the BU approach started from the information and other classifications selected of some Web sites concerning the domain of interest, to pinpoint the knowledge in them. Then, these contents were classified with the taxonomy previously defined and the mapping rules between contents and taxonomy. This taxonomy allowed for the definition of a reference knowledge which may later be managed in terms of really usable and interesting knowledge, fostered by the whole knowledge of all the selected Web sites.

We chose to test this approach on the knowledge domain of Italian wines reviews. As for the validation, we verified how this KMS allowed such knowledge to become available on systems that were compliant with the Wines ontology as defined as an example of Semantic Web by W3C; then we checked other Web sites of Italian wines reviews, verifying how their contents of interest could be represented and managed on the KMS through some simple mapping rules.

This chapter is structured as follows: we present a brief overview about the context of our proposal and we describe the knowledge domain. Then, we present our approach to define a taxonomy able to represent knowledge through an iterative combined approach. Finally, we describe the analysis of results and verification.

5.1 Related concepts

The Web becomes clever and is conceived as a big database in which data are orderly classified. “Information“, therefore, is one of the keywords at the base of the success of both search engines (Google[1], Yahoo[2], Bing[3], etc.), which become more refined in data retrieval and presentation, and social networks (YouTube, Facebook[4], Twitter[5], Flickr, etc.), which allow exchange and sharing, creating an interconnection among users and content makers. However, such data, despite being formally available, are often unreachable as for their semantic meaning and cannot be used as real knowledge.

Various proposals to solve these problems can be found in literature, also to overcome the semantic heterogeneity problem [107] and to facilitate knowledge sharing and reuse [108][109]. In [45] an approach based on the use of an ontology to make annotating photos and searching for specific images more intelligent is described; and in [46] a data-driven approach to investigate semi-automatic construction of multimedia ontologies is used. With the emergence of the Semantic Web, a shared vocabulary is necessary to annotate the vast collection of heterogeneous media: in [51] an ontology is proposed to provide a meaningful set of relationships which may enable this process.

5.2 Domain knowledge

In this work we chose the domain of wines and, particularly, the one belonging to the technical files and/or descriptions of “Italian wines“: the choice was not made randomly as the world of wines is rich in contents and complete enough to give a good starting point for our study. In fact, there are thousands of contents which can be found on the Internet; also, there are different studies on the classification of wines from which we can draw on.

Contents on wine available on the Web are thousands, offering a significant KB. Our study takes into consideration a subdomain of wine, represented by all the most important reviews which can be found on the Internet.

From the analyses of the domain on the Web and the Google Ranking of these Web sites, we chose a list of suitable and representative Web sites, having considered the popularity and the reliability given by the Web.

[1] Google, http://www.google.it

[2] Yahoo, http://it.yahoo.com

[3] Bing, http://it.bing.com

[4] Facebook, https://it-it.facebook.com

[5] Twitter, https://twitter.com

The Web sites we took into consideration are the following:

1 Decanter.com

2 DiWineTaste.com

3 Lavinium.com

4 GamberoRosso.it

5 Vintrospective.com

6 Snooth.com

7 Vinix.com

These Web sites are considered as representative for our study also because of their own information structures, particularly various and differentiated ones. Each Web site has its own structure and a different representation of the information. To correctly define our domain it was therefore necessary to precisely analyze the contents in each of them and the layouts. The structure of the page showing the review is useful to understand if the same Web site always uses the same structure and the same items for every review. Unfortunately we saw that some of them show the same information differently depending on the review, using, for instance, different tags for the same information. This, obviously, is a limit in the process of classification of contents. It is thus necessary to align the different items for the same Web site, used to represent the same information.

5.3 Proposed approach

The knowledge to be represented is the most popular among users of a certain domain. To determine which is the users' knowledge of real interest we chose to select the most used Web sites by users, the most important and looked up ones. For this definition, Web sites with a higher ranking on Google among the domain of interest are typically chosen.

5.3.1 Top-down phase

In this phase we analyzed the existing formalizations for the representation of knowledge of this domain.

A very interesting formalization which we pinpointed was the one by the *Associazione Italiana Sommelier* (AIS, Italian Sommelier Association), providing a detailed description of all the terms associated with wine. Another important formalization was the one by the European law defining the reference features of a certain wine, such as type, colour, grape variety, etc. From these two, a reference taxonomy for those features was created.

As an additional formalization, we chose a reference scheme, represented by an ontology already existing on the Web and made by W3C: Wine ontology [110]. An ontology is surely more complex than a taxonomy. It has, apart from class hierarchies, property hierarchies with cardinality ties for the assignable values. It offers a general view of the world of wines, with a less detailed description for certain fields as stated on the reviews found on the Web. Moreover, from this ontology we took into consideration only the areas of interest existing in our classification, omitting those ones representing elements not of interest (such as, for instance, each winemakerâs property).

Starting from these reference formalizations, a first taxonomy was built in which we pinpointed the items to create the reference table. After choosing the items of interest in the reference ontology, we analyzed the direct correspondence among tags of the two representations, directly extracting the ontology ones from the OWL code. To standardize our taxonomy we decided to take into consideration the RDF standard indicating, just for the items with a correspondence, its URI.

Table 5.1: *Correspondence among tags*

Our taxonomy	**W3C ontology**	**URI**
wine	wine	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine
wine.winery	wine.winery	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineWinery
wine.color	Wine.wineDescriptor.WineColor	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineWineColor
wine.color.white	Wine.wineDescriptor.WineColor.White	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineWhite
wine.color.red	Wine.wineDescriptor.WineColor.Red	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineRed
wine.color.rose	Wine.wineDescriptor.WineColor.Rose	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineRose
wine.grape	Food.grape.winegrape	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineWineGrape
wine.state	Wine.Region	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineRegion
wine.tastingNotes.vintageNotes	vintage	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineVintage
wine.tastingNotes.OlfactoryAnalysis	Wine.wineDescriptor.WineTaste.WineFlavor	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineWineFlavor
wine.tastingNotes.TasteAnalysis	Wine.wineDescriptor.WineTaste.WineSugar	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineWineSugar
wine.tastingNotes.FinalConsiderations	Wine.wineDescriptor.WineTaste.WineBody	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineWineBody

The RDF standard allows to associate a URI also to the properties. The list of the common ones associated to their classes, which can be used to standardize our taxonomy, is shown below. In the first column the properties of the ontology which can be used for our taxonomy are indicated; in the second column the classes of the taxonomy for each feature; in the third one the associated URI.

Table 5.2: *Association properties*

Property	**Class**	**URI**
hasWinery	Wine - Winery	http://www.w3.org/TR/2003/PR-owl-guide-20031209/winehasMaker
hasWineDescriptor.hasColor	Wine - Color	http://www.w3.org/TR/2003/PR-owl-guide-20031209/winehasColor
madeFromFruit.madeFromGrape	Wine - Grape	http://www.w3.org/TR/2003/PR-owl-guide-20031209/winemadeFromGrape
locatedIn	Wine - Region	http://www.w3.org/TR/2003/PR-owl-guide-20031209/winelocatedIn
hasVintageYear	TastingNotes - VintageNotes	http://www.w3.org/TR/2003/PR-owl-guide-20031209/winehasVintageYear
hasWineDescriptor.hasFlavor	TastingNotes - WineFlavor	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineWineFlavor
hasWineDescriptor.hasSugar	TastingNotes - WineSugar	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineWineSugar
hasWineDescriptor.hasBody	TastingNotes - WineBody	http://www.w3.org/TR/2003/PR-owl-guide-20031209/wineWineBody

5.3.2 Bottom-up phase

The BU analysis required a detailed analysis of the contents of these Web sites, trying to pinpoint the information we considered as important; then we studied the structure of each single source, useful to see the existing data and their position in the layout of the page.

Once the KB for the domain of interest composed by the Web sites was defined, the next step was to classify all the chosen information. Such classification is made by creating a classification of the BU contents because it was built from the bottom: information on the Web sites are thus accurately analyzed.

We start from the analysis of the specific to reach a general classification of data. One of the initial steps of our project contemplated the study of the structure of each source, useful to see the existing data and their position on the page's layout. This procedure happens to be important also at this point of the study, because allows for the evaluation of the classification of information. Both the item *"maturazione"*, but also the organoleptic analysis (visual, olfactory and gustatory test) if existing, are systematically shown on the Web sites taken into consideration, into the area which we identified as "tasting notes". For this reason, to build the hierarchy we tried to respect the original, already existing one. The type of classification was also revealed during the data analysis phase, during the study of semantics and standardization.

With the creation of the tables we tried to represent the knowledge in the shape of fields as faithful as possible to those ones already existing in the samples taken into consideration.

The evaluation of this phase is subjective and left to the intuition of the analyst, which freely interprets the information at their disposal, intuitively obtaining the taxonomic tree. This step happens to be very tricky, because is susceptible to accidental mistakes. However, we could say that the various structures found in the domain which we considered, apart from the caption used to define each field, are not so different, thus the classification did not raise any big doubt, as for the representation.

Thus, the macrosystem made 7 tables, one for each Web site. Every table has the list of information of the Web site it represents.

5.3.3 Iteration of phases

During this phase, the items of the fields existing in the taxonomy defined in the TD phase were compared to the fields of the tables created in the BU phase. To do this, we built a mapping macro-table of knowledge containing, for each item of the taxonomy, the correspondence if and where that information exists on the various Web sites and also the information existing on the Web sites which were not represented by the taxonomy.

To each item we thus assigned a numerical value to represent this mapping:

- 1: existing and extractable information;
- 2: existing but not extractable information;
- 3: sometimes existing and extractable information;
- 4: sometimes existing but not extractable information;
- 5: always missing information.

For the fields with values 1 and 3, the corresponding field and the mapping rule to extrapolate the information are also indicated. The information with value 2 and 4 is embedded (hidden in the text) and, therefore, should be specifically looked for with tools of semantic analysis. Anyway, the field in which it exists is indicated.

With this analysis and classification of every single data, we managed to solve the heterogeneity of the information existing in the Web, as for the domain of interest. This allowed to study both its structure and the type of information existing, giving us the chance to examine how data are presented and the classification given for each Web site.

When creating the taxonomy, which wants to be a semantic classification tool, we also tried to represent the structure of data and the existing hierarchies of the sample Web sites.

This activity was iteratively repeated to best represent the knowledge and its connections described in the macro-system mentioned above. As expected, not all the fields were taken into consideration, neither among those existing in the initial taxonomy nor among the extrapolated ones, and those ones which appear just once in the whole macro-system were rejected (evaluation made considering the field value = 5), such as, for instance, *"Bicchiere consigliato"* or *"Temperatura di servizio consigliata"*.

The inhomogeneity among the information existing in the different Web sites was analyzed by looking for the semantic correspondences represented in the macrosystem with the column âfield detailsâ. The same principle was used to uniform fields with numerical values. The final range takes into account the classification used by the majority of Web sites.

A simplifying table summarizing the procedure of classification described above is shown below. The result of these phases was the KB formalized through the taxonomy. The table shows some items of it, with a field of value 1 or 3 and expressed in textual form (for instance, those ones directly extractable through tags or metadata). Other fields, represented by an icon, were rejected, though their presence was considered.

Table 5.3: *Classification*

Macrosystem item	**Final tag**
Wine's identification name	Wine
<Produttore> <Winery> <Producer>	Winery: address, telephone, fax, e-mail, web, map, other wine, other info winery
<classification> <denominazione> <tipologia>	Classification: Vino da tavola, IGT, DOC, DOCG
<tipologia> <type>	Colour: white, rose, red
<type> <tipologia>	specification
Qualification: embedded	Qualification: classic, reserve, superior
<typical grape composition> <Varietal> <vitigni> <uve>	Grape
<titolo alcolometrico> <alcohol> <alcol>	alcohol
Label	Label
<origin> <region> <zona>	State/Region
<tasting notes> <reviews> <overview>	Tasting notes
<prezzo enoteca> <prezzo> <starting at> <$> <average bottle price>	Price
<abbinamento> <suggested recipe pairings> <food pairing suggestions>	Food pairing suggestion
<posted by> <source>	author
<posted on> <inserito> <degustazione in data>	Date
<decanter rating>: max 5 stelle <rated>: max 5 bicchieri <valutazione>: max 5 chiocciole <punteggio>: max 5 diamanti <voto>: max 5 chiocciole Punteggio: max 3 bicchieri	Rate: 60-70; 71-75; 76-80; 81-85; 86-90; 91-100

5.4 Analysis of results and verification

During the validation phase we verified how our KMS made the acquired knowledge usable for the systems compliant with another ontology of wines and for other Web sites on Italian wines reviews. We went on verifying how the contents of interest of these Web sites could be represented and managed on the KMS through some simple mapping rules. Then, we tried to solve the clear inhomogeneity by paying more attention to the semantic meaning and not to the notation used to represent those contents. In fact, the purpose of the study was not to describe the whole world of wines, but just the part of it represented by the information which can be found on the Web.

After matching the two systems, ontology and taxonomy, the information were generalized and made coherent. This allowed us to verify that our system is able to represent and combine specific information, and at the same time understands the main variances between the two systems, namely the difference of some considered information.

This kind of study can also be used to enrich an already existing ontology with fields coming from a general classification, evaluating a possible integration of such information without damaging the existing hierarchy, so that we can have a broader and more accurate view over the analyzed domain.

To continue with the phase of verification of the created taxonomy, we decided to take into consideration another set of samples - again, wine reviews which can be found on the Web.

The choice of the Web sites for the testing phase followed the same criteria used during the analysis of the domain. The main obstacle we found was due to the popularity of the product and the large amount of followers who have a very subjective way of representing the information about wine and the acquired knowledge. Here comes the need of pinpointing sources with clear, easily extractable and objective information.

One of the main features which these sources needed to have was the presence of differentiated fields with a single notation rather than a broad textual field. So, also in this case, all the Web sites gathering a large quantity of information in just a macro-textual area were rejected. In fact, these kind of Web sites, though full of contents, were not suitable for the testing phase. The embedded information, though fostering the acquisition of a general knowledge, do not facilitate its own structured classification. Similarly, some apparently suitable sources happened to have very few contents, with a database so poor that it did not mention the most appreciated wines.

After these considerations, the Web sites we decided to take into consideration for the tests were the following:

- guida-vino.com
- vinogusto.com
- kenswineguide.com
- buyingguide.winemag.com

5.4.1 Testing phase

For each sample Web site, in this testing phase we verified whether the information in them could be found in the classification proposed by us, and whether our taxonomy could be able to represent them. For each Web site, therefore, the following table was built, representing the specific fields of information which was the same for every review that we analyzed.

Table 5.4: *Testing phase*

Existing information	**Field details**	**Taxonomy item**
Label	Label's image	Wine.label
Producer <Winery> <Producer>	About the producer	Wine.winery
Classification	IGT, DOC,DOCG	Wine.classification
Grape variety	Grape variety	Wine.grape
Range of prices	Price	Wine.Price
Other years	Other years	Wine.winery.infoWinery.otherWines
Presentation/comments	Wine tasting	Wine.tastingNotes
Rate: max 5 stars	Rate	Wine.rate

In the light of the results obtained in this testing phase, we are satisfied with the taxonomy which we created. In fact, with this testing phase, we saw that the classification defined in our study reflects the type of contents needed. Such classification, therefore, is usable, re-usable and possibly extendible to the domain of interest of wine.

Chapter 6

Conclusions and Future Work

We focused our research on some issues concerning Knowledge Management, namely the problem of how to make multimedia content-related knowledge available to people.

6.1 User Generated Content

We proposed an approach to solve the problem of managing the knowledge of UGCs. This approach is especially suited for all those instances when a multimedia content is considered, for which associated information do not comply with standards in categorizing metadata. Special attention has to be paid to widespread standards such as Adobe XMP, Dublin Core, Exif, IPTC.

The general goal was to study, design and create an ontology that could formalize the multimedia content semantics and geocoded data, starting from the already mentioned standards in representing that domain. In fact, in those cases, a synergistic integration of an ontology based on the standard with the usage of a clearly set mapping technique allows for representing a great number of contents and metadata as proven in the mapping example. This mapping technique was especially useful to sort out a vast and complex knowledge field such as multimedia content. Dealing with mapping arose the necessity of using shared standards rather than proprietary ones, now very widespread. The proposed approach may be used as support for a software platform that allows different actors to develop added-value services. Such services could be based on multimedia content insertion into a semantic organization context. It is clear that such an approach should rely on a powerful tool which could map all the information concerning entered contents in relation to the form decided as representation standard within itself. The purpose was to offer a structure enhanced with semantics, which could serve as base support for the creation of a Web content management software platform.

The platform, thanks to the modelled concepts, could give users the chance to collect and add contents originated from varied sources (Web sites, Web portals, local files) and to influence the value of the contents though ratings, comments and preferences. Thus contents could be gathered, aggregated and geocoded, and then distributed to each user. Such a platform should clearly be provided with a powerful tool capable to “conform“ every piece of information about the added contents to the form designated as a representation standard within itself. In other words, it must be able to map any kind of metadata present in contents.

Once again the ontology we created would be an impressive tool to fulfill that requirement. The system could be accessible through mobile devices such as PNAs (Personal Navigator Assistants), that would use a geolocalization system to know their location.

6.2 Linguistic information in audio content

We proposed an approach for formalization and management of knowledge. In this case, the knowledge was represented by a set of audio recordings in a corpus and linguistic information added to that corpus with annotations. We organized the information in the corpus formalizing those annotations through metadata schemas using the informal annotations made by the domain experts. We used this approach to associate the annotations to their texts, using the selected linguistic level. The proposed approach was experimented and validated during a project that aimed to create the Analytic Sound Archive of Sardinia. The ASAS is a joint project by linguists and musicologists at University of Cagliari that had the purpose to create an institutional archive with a linguistically and musically annotated electronic corpus. This archive has an electronic corpus of spoken texts, linguistically annotated at various levels.

DSpace was chosen to make the ASAS, since it fulfils all the requirements set by linguists and musicologists. This tool in fact is very efficient, easy to use, customizable and flexible to allow the management, the classification and the storage of a vast amount of knowledge contained in an electronic corpus of Sardinian spoken and sung language. At the same time, it can also allow a high usability in terms of ease of reference as well as ease of query and communication. It natively supports the Qualified DC metadata schema and is compatible with OAI with the support of OAI-PMH.

The formalization of a structured metadata schema was reached through the creation of an application profile for the Qualified Dublin Core metadata schema, where customized qualifiers were added to the standard elements and qualifiers. Metadata in non-standard schemas could then be better represented.

Linguistic annotations were formalized as well through a metadata schema. Corpus interrogation was thus made easier and quicker, since it used the knowledge management system's search tool. This work leaves space for future research on ways to improve the service. A dedicated Web site or the integration of this system in an institutional portal through an exploration interface would be particularly interesting. Another feature that could be implemented may be a virtual map where recordings could be explored by geographic location.

6.3 Construction process

We analyzed the objects and verified the structure, and the factors were all well defined and analyzed. The knowledge base that we used is very big and representative of the general knowledge. The analysis can be replicated on different data. It was assumed that breakdown of building products in the top-down phase and the analysis of objects in bottom-up phase had been applied in the case study. The results are compliant with the general theory of mixed approach to analyze knowledge.

The breakdown process of building components in Elementary Products defines the reference elements which can manage the multimedia objects.

The Elementary Product:

- is a classification that can be used to define formalized metadata;
- groups all the multimedia objects in a single semantic object;
- makes users select information in form of folksonomies tag;
- can be connected with other concepts like designer, project manager, etc.

The Elementary Product is the core concept of this knowledge; every instance of a single building project and the construction process can be managed using this semantic concept. The formalized information are the metadata defined for the Elementary Product. All other information, like technical data or data sheet (a PVC window), are present in the multimedia objects associated with the Elementary Product, and the interesting information regarding the project can be represented as a folksonomy tag.

The structural information of the Elementary Product is represented as metadata, as well as the 'Project Name, 'Phase', 'Technical description' and 'General Description' where the relevant information of the project selected by the user can be found. With this approach and formalization we can manage all the relevant and embedded information.

The management of very complex knowledge is a big problem in Knowledge Management research; the proposed approach reaches its main goal to find a rational organization of such large amounts of information. The technical and multimedia information are varied and contain interesting embedded information. The solution proposed is based on the very interesting concept of Elementary Product, which guides the organization of the knowledge. This knowledge formalization suggests its implementation in a KMS such as DSpace using metadata schemas. Further studies could analyze the results of the use of this system and the result of the experience could be used to define further interesting information that can be formalized as metadata associated with the Elementary Product.

6.4 Italian wines reviews

The spread of the Social Web is significantly influencing the evolution of Semantic Web: users themselves are creating rules for the representation of information. The structure of the Web grows and changes giving the user the chance to actively participate in the developing of the Web. For this reason, our study took into consideration this feature with the standardization of UGCs, trying to link the two worlds: Social Media and Semantic Web. Also the main search engines (Google, Yahoo, Bing, etc.) and the main Social Networks (YouTube, Facebook, Twitter, Flickr, etc.) are evolving, specializing and interconnecting themselves on data retrieval, presentation, exchange and sharing. That being so, the basic idea of our study was to propose a solution to the problem of the contents of the Web, present in different Web sites but belonging to the same domain of knowledge.

Our proposal is to define a taxonomy able to represent knowledge through a mixed iterative approach, articulated in a top-down and a bottom-up analysis to define a reference taxonomy. Then, the knowledge we considered as important (and as an element of common interest) was extracted from a selection of Web sites belonging to the domain of interest. These contents are to be classified in the taxonomy mentioned before, also using mapping rules made ad-hoc.

We apply this approach to the KB of Italian wine reviews from the analysis of the KB on the Web and the Google Ranking of many Web sites, we chose a list of some suitable and representative ones after considering popularity and reliability given from the Web.

The taxonomy created allowed for a definition of the reference knowledge which could then be managed as an actual usable knowledge, fostered by all the information existing on the selected Web sites.

We chose to validate the resulting taxonomy by verifying how the KMS allowed to make the acquired knowledge usable and accessible to the systems compliant with the Ontology of Wines. We validated the taxonomy by analyzing the content in other Web sites of Italian wine reviews, underlining how, also in this case, the collected information could be represented and managed on the KMS through some simple mapping rules.

A further, interesting development could be the creation of repositories able to collect the information previously classified and, through a system made ad-hoc, they would be presented to the final user in a structured and customized way, depending on the requests, and possibly developing a graphic interface which could be able to draw the curiosity and the interest of the user.

6.5 An interesting industrial exploitation of the studies

A very interesting example of industrial exploitation still under development regards the results of the research on ontologies in UGCs. The ontology contains all specifications to create a multimedia content management system able to manage information from UGCs as georeferenced multimedia contents.

The platform can manage UGCs, making them usable in an aggregated way. In fact, it is possible to use the ontology as a basis on which a system can be created, which will allow for searching and classifying multimedia content with a semantic reference given by the ontology, making data usable.

The ontology was particularly apt to make order in a wide and complex knowledge field such as the one pertaining to descriptive metadata of multimedia content. This context has a large number of different standards, some proprietary, some even with no regulation at all, which makes things difficult to people who want to work in that field. Tackling the issue of mapping made light on how working in this field would be much more efficient and convenient if one could refer to shared standards instead of proprietary ones, as it usually happens.

The project sets some specific extractors to be developed for each UGC source to power the platform. The extractors would follow the dates of the ontology and implement mapping rules defined at a semantic level, and so would be able to retrieve the contents from UGC repositories and transform the information associated to them into manageable information in the platform.

Thanks to the modelled concepts, the platform would thus offer to users the opportunity to use the contents coming from various sources, already gathered, aggregated and geocoded.

The use of such contents could happen through an application that could show aggregated data either by type and by location. Were the use of the contents to be performed with a smartphone or a tablet device, it could be extremely strategic to show them as Points of Interest (POI) located near the user, exploiting georeferenced information and the GPS function of the devices.

The results of this research are the basis of a software platform allowing different customers (content producers, public administration, communication companies, public service suppliers, etc.) to develop added-value services based on georeferenced multimedia contents.

The users of such services could interact with the platform using the data already there and also show their preferences and adding their own contents. The platform is an enabling technology which gives the proponents the opportunity to enter an emerging, highly innovative and not yet covered market, that is the one of UGC-based georeferenced contents. They would have a solid starting ground for a complete, articulated and definitely wider business solution offer.

The platform itself is the vital element on which a number of solutions can be defined depending on the contents the client has, which would be distributed according to their own business models.

The reasons behind this project are connected to a business opportunity born from many factors, among which the widespread mobile information devices such as smartphones and tablets that have mapping features (Google Maps). Users who are interested in receiving information on the places they are in, thanks to the UGC, could receive information that are much richer than the traditional POI present in the current systems.

6.6 Final remarks

Through this study, we could see how a mixed-iterative approach made of top-down and bottom-up analyses of a knowledge domain could be efficient when formalizing knowledge. Our approach to Knowledge Management is a simple process of applying a systematic analysis to capture, structure and manage knowledge. Our real goal was to make interesting knowledge available for sharing and reuse, and we focused our attention on interesting information on the knowledge domain which had to be represented.

In all the case studies we used, we studied a process to identify existing formalizations and knowledge sources within the domain, paying attention to multimedia objects. Valuable knowledge was represented into explicit form through formalization and codification of information, in order to facilitate the availability of knowledge.

At the end of these analyses we defined a formalization, in form of ontologies, taxonomies, metadata schemas, able to represent the knowledge of interest for the domain.

As a final remark, we can notice how such a process could lead to many different solutions for formalization. We used ontologies, taxonomies and metadata schemas to formalize knowledge. The concept of metadata is clearly vital, also because it is easy to represent for operational purposes. In fact, metadata are very suited to being represented through different standards (XML Schema, RDF, etc.), and managed with many tools (DataBase, KMS, CMS, etc.).

A mixed approach, already proposed in the literature, means surely a demanding manual analysis work, but some Knowledge Engineering activity is necessary to represent knowledge. As regards the top-down phase, the number of formalization coming from ontologies, taxonomies and existing standards allow an articulated structuring of knowledge. Such representation are unlikely to represent the same knowledge we wanted to represent for our purposes: for this reason, they are very useful in a top-down analysis but cannot be used as-is to represent our knowledge of interest.

At the same time, basic knowledge is seldom natively structured; it usually contains a number of pieces of information that can be extracted from it and formalized, then used and enclosed in a representation of knowledge. In particular, contents on the Internet can be a source of raw knowledge where some information can be acquired. Such information could then be formalized and acquire a remarkable value. An example of that may be the studies on UGCs and on the reviews of Italian wines that exist on the Web. In fact, even Semantic Web tools proved unsatisfactory in managing this kind of issue.

Another important example could be the analysis of the information contained in multimedia objects pertaining to the ASAS and the Construction Process, where the sheer quantity of available resources and rich information would have been very difficult to manage without a formalization. The formalizations made it possible for domain experts to locate the truly useful information in the operational context and classify them, with the purpose of managing and sharing them through a simple KMS like DSpace.

The most important result we achieved with this thesis was the opportunity to make this disparate knowledge available and manageable. In the current market, exploiting existing knowledge is a mainstream business, but in order to exploit it, one must be able to manage it first. As a token of this importance, not only about ten scientific publications (listed in the References), but most of all a number of industrial research projects, in partnership with ICT companies - one of which with a total value above one million Euros - stemmed from the studies discussed in this thesis.

Bibliography

[1] Maier, R., Knowledge Management Systems: Information and Communication Technologies for Knowledge Management, Springer London, 2010. ISBN: 3540714081, 9783540714088. [cited at p. 1, 45]

[2] Berners-Lee T., Hendler J., Lassila, O., The Semantic Web. In: Scientific American, pp. 29-37., 2001. [cited at p. 1, 57]

[3] Malhotra, Y., Knowledge Management for the new world of business. In: Journal for Quality and Participation, pp. 58-60, 1998. [cited at p. 1, 45]

[4] Gashaw, K., Knowledge management: An information science perspective. In: International Journal of Information Management, Vol. 30, Issue 5, pp. 416-424, 2010. [cited at p. 1]

[5] Jakubik, M., Exploring the knowledge landscape: Four emerging views of knowledge. In: Journal of Knowledge Management, 11(4), pp. 6-19, 2007. [cited at p. 1, 2]

[6] Dalkir, K., Knowledge management in theory and practice, Elsevier Butterworth-Heinemann, 2005. [cited at p. 1, 2]

[7] Rowley, J., The wisdom hierarchy: Representations of the DIKW hierarchy. In: Journal of Information Science, 33(2), pp. 163-180, 2007. [cited at p. 1, 2]

[8] Wild, R., Griggs, K., A model of information technology opportunities for facilitating the practice of knowledge management. In: VINE (Journal of information and knowledge management systems), 38(4), pp. 490-506, 2008. [cited at p. 1]

[9] Ajiferuke, I., Role of information professionals in knowledge management programs: Empirical evidence from Canada. In: Informing Science Journal, Vol. 6, pp. 147-157, 2003. [cited at p. 1]

[10] Blair, D. C., Knowledge management: Hype, hope, or help. In: Journal of the American Society for Information Science and Technology, 53(12), pp. 1019-1028, 2002. [cited at p. 1]

[11] Chua, A. Y. K., The dark side of knowledge management initiatives. In: Journal of Knowledge Management, 13(4), pp. 32-40, 2009. [cited at p. 1]

[12] Ekbia, H. R., Hara, N., The quality of evidence in knowledge management research: Practitioner versus scholarly literature. Journal of Information Science, 34(1), 1-17, 2007. [cited at p. 1]

[13] Jashapara, A., The emerging discourse of knowledge management: A new dawn for information science research? Journal of Information Science, 31(2), pp. 136-148., 2005. [cited at p. 1, 2]

[14] McInerney, C., Knowledge management and the dynamic nature of knowledge. Journal of the American Society for Information Science and Technology, 53(12), pp. 1009-1018, 2002. [cited at p. 1]

[15] Sarrafzadeh, M., Martin, B., Hazeri, A., LIS professionals and knowledge management: Some recent perspectives. Library Management, 27(9), 621-635, 2006. [cited at p. 2]

[16] Widen-Wulff, G., Allen, D., Maceviciute, E., Moring, C., Papik, R., Wilson, T., Knowledge management/information management. In Leif Kajberg, Leif Lorring (Eds.), European curriculum reflections on library and information science education (pp. 121-130). Copenhagen: The Royal School of Library and Information Science, 2005. [cited at p. 2]

[17] Martin, B., Knowledge management. In C. Blaise (Ed.), Annual review of information science and technology (ARIST), Vol. 42 pp. 371-424. Medford, NJ: Information Today, Inc., 2008. [cited at p. 2]

[18] Sinotte, M., Exploration of the field of knowledge management for the library and information profession. Libri, 54, 190-198, 2004. [cited at p. 2]

[19] Bouthillier, F., Shearer, K., Understanding knowledge management and information management: The need for an empirical perspective. Information Research, 8(1), paper no. 141, 2002. [cited at p. 2]

[20] Bouthillier, F., Shearer, K., Knowledge management and information management: Review of empirical evidence. In Elena Maceviciute, T. D. Wilson (Eds.), Introducing information management: An information research reader (pp. 139-150). London: Facet Publishing, 2005. [cited at p. 2]

[21] Corrall, S., Knowledge Management. In: Ariadne Issue 18, 1998. http://www.ariadne.ac.uk/issue18/knowledge-mgt [cited at p. 2]

[22] Hlupic, V., Pouloudi, A., Rzevski, G., Towards an integrated approach to knowledge management: ’Hard’, ’soft’ and ’abstract’ issues. Knowledge and Process Management, 9(2), pp. 90-102, 2002. [cited at p. 2]

[23] Maceviciute, E., Wilson, T., Part D: Knowledge management. In E. Maceviciute, T. D.Wilson (Eds.), Introducing information management: An information research reader (pp. 137-138). London: Facet Publishing, 2005. [cited at p. 2]

[24] Morrow, N. Mac., Knowledge management. In M. E. Williams (Ed.), An introduction annual review of information science and technology (ARIST), vol. 35 (pp. 381-422). Medford, NJ: Information Today, Inc., 2001. [cited at p. 2]

[25] Ponelis, S., Fairer-Wessels, F. A., Knowledge management: A literature review. South African Journal of Library and Information Science, 66(1), 1-9, 1998. [cited at p. 2]

[26] Wilson, T.D., The nonsense of knowledge management. Information Research, 8(1), paper no. 144, 2002. [cited at p. 2]

[27] Nonaka, I., and Takeuchi, H., The knowledge-creating company: how Japanese companies create the dynamics of innovation. New York: Oxford University Press, 1995. [cited at p. 2]

[28] Pasternack, B., and Viscio, A., The centerless corporation. New York: Simon Schuster, 1998. [cited at p. 2]

[29] Pfeiffer, J., Sutton, R., The knowing-doing gap: How smart companies turn knowledge into action. Boston: Harvard Business School Press, 1999. [cited at p. 2]

[30] Ruggles, R., Holtshouse, D., The knowledge advantage. Dover, N. H., Capstone Publishers, 1999. [cited at p. 2]

[31] Gruber, T., A Translation Approach to Portable Ontology Specification. In: Knowledge Acquisition, Vol. 5, pp. 199-220, 1993. [cited at p. 3, 4, 11]

[32] Brickley, D., Guha, R.V., Resource Description Framework (RDF) Schema Specification. Proposed Recommendation, World Wide Web Consortium, 1999. http://www.w3.org/TR/PR-rdf-schema [cited at p. 3, 19]

[33] Hendler, J., McGuinness, D.L., The DARPA Agent Markup Language. IEEE Intelligent Systems 16(6), pp. 67-73, 2000. [cited at p. 3]

[34] Noy, N. F., McGuinness, D. L., Ontology Development 101: A Guide to Creating Your First Ontology, Stanford Knowledge Systems, Laboratory Technical Report KSL-01-05, 2001. [cited at p. 3, 4, 19, 57]

[35] Musen, M.A., Dimensions of knowledge sharing and reuse. Computers and Biomedical Research 25: 435-467, 1992. [cited at p. 4]

[36] Lunesu, M. I., Pani F. E., Concas G., An Approach to manage semantic informations from UGC. In: Proceedings of the 3rd International Conference on Knowledge Engineering and Ontology Development (KEOD 2011), Paris, France, 2011. ISBN: 978-989-8425-80-5. [cited at p. 6]

[37] Lunesu, M. I., Pani F. E., Concas G., Using a standards-based approach for a multimedia knowledge-base. In: Proceedings of the 3rd International Conference on Knowledge Management and Information Sharing (KMIS 2011), Paris, France, 2011. ISBN: 978-989-8425-81-2. [cited at p. 6]

[38] Pani, F. E., Lunesu M. I., Concas G., Stara C., Tilocca M. P., Optimization of Knowledge Availability in an Institutional Repository. In: Proceedings of the 4th International Conference on Knowledge Engineering and Ontology Development, KEOD 2012, Barcelona, Spain, 2012. ISBN: 978-989-8565-30-3. [cited at p. 7]

[39] Pani, F. E., Lunesu M. I., Concas G., Stara C., Tilocca M. P., Knowledge Formalization and Management in KMS. In: Proceedings of the 4th International Conference on Knowledge Management and Information Sharing, KMIS 2012, Barcelona, Spain, 2012. ISBN: 978-989-8565-31-0. [cited at p. 7]

[40] Concas, G., Pani F. E., Lunesu M. I., Knowledge Management using Open Source Repository. In: Proceedings of the 6th European Computing Conference (ECC '12), Prague, Czech Republic, 2012. ISBN: 978-1-61804-129-6. [cited at p. 7]

[41] Guarino, N., Giaretta, P., Ontologies and Knowledge Bases, Towards a Terminological Clarification. In N. Mars (Ed.), Towards Very Large Knowledge Bases, Knowledge Building and Knowledge Sharing (pp. 25-32), Amsterdam, IOS Press, 1995. [cited at p. 10]

[42] Hepp, M.: Ontologies: State of the Art, Business Potential, and Grand Challenges, in: Hepp, M.; De Leenheer, P.; de Moor, A.; Sure,Y. (Eds.), Ontology Management: Semantic Web, Semantic Web Services, and Business Applications, ISBN 978-0-387-69899-1, Springer, pp. 3-22, 2007. [cited at p. 10]

[43] Swartout, B., Patil, R., Knight, K., Russ, T., Toward Distributed Use of Large-Scale Ontologies Ontological Engineering. In: AAAI-97 Spring Symposium Series, pp. 138-148, 1997. [cited at p. 11]

[44] Gruber, T., Ontology. In: Liu L., Ãzsu, M. T. (Eds.), Encyclopedia of Database Systems, Springer-Verlag, 2008. [cited at p. 11, 57]

[45] Schreiber, A. Th., Dubbeldam, B., Wielemaker, J., Wielinga, B., Ontology-Based Photo Annotation. In: IEEE Intelligent Systems, Vol. 16, pp. 66-74, 2001. [cited at p. 11, 58]

[46] Jaimes, A., Smith, J., Semi-automatic, data-driven construction of multimedia ontologies. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), Vol. 2, 2003. [cited at p. 11, 57, 58]

[47] Benitez, A., Chang, S., Automatic multimedia knowledge discovery, summarization and evaluation, IEEE Transactions on Multimedia, 2003. [cited at p. 12]

[48] Strintzis, J., Bloehdom, S., Handschuh, S., Staab, S., Simou, N., Tzouvatras, V., Petridis, K., Kompatsiaris, I., Avrithis, Y., Knowledge representation for semantic multimedia content analysis and reasoning. In: Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media technology, 2004. [cited at p. 12]

[49] Bertini, M., Cucchiara, R., Del Bimbo, A., Torniai, C., Video annotation with pictiorally enriched ontologies. In: Proceedings of IEEE Int'l Conference on Multimedia and Expo, 2005. [cited at p. 12]

[50] Bertini, M., Del Bimbo, A., Torniai, C., Cucchiara, R., Grana, C., MOM: Multimedia Ontology Manager, A Framework for Automatic Annotation and Semantic Retrieval of Video Sequences. ACM, Santa Barbara, California, USA, 2006. [cited at p. 12, 57]

[51] Jewell, M. O., Lawrence, K. F., Tuffield, M. M., Prugel-Bennett, A., Millard, D. E., Nixon, M. S., Schraefel, M. C., Shadbolt N. R., OntoMedia: An Ontology for the Representation of Heterogeneous Media. In: Multimedia Information Retrieval Workshop, ACM SIGIR, 2005. [cited at p. 12, 58]

[52] Dasiopoulou, S., Tzouvaras, V., Kompatsiaris, I., Strintzis, M. G., Enquiring MPEG-7 based Ontologies. In: Multimedia Tools and Applications, Vol. 46, issue 2, pp. 331-370, 2010. [cited at p. 12]

[53] Paliouras, G., Spyropoulos, C. D., Tsatsaronis, G. (Eds.), Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, Bridging the Semantic Gap, Lecture Notes in Computer Science, Vol. 6050, 1st Edition, IX, 245 p., ISBN 978-3-642-20794-5, 2011. [cited at p. 12, 57]

[54] Martinez, J. M., Koenen, R., Pereira, F., MPEG-7: the generic Multimedia Content Description Standard, part 1. In: IEEE Multimedia, Vol. 9, pp. 78-87, 2002. [cited at p. 12]

[55] Adobe Systems Incorporated, Adobe XMP Specifications, additional properties, 2010. http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart2.pdf [cited at p. 13]

[56] Becker, H., Chapman, A., Daviel, A., Kaye, K., Larsgaard, M., Miller, D., Nebert, D., Prout, A., Wolf, M.P., Dublin Core element: Coverage, 1997. http://www.alexandria.ucsb.edu/public-documents/metadata/dc_coverage.html [cited at p. 14]

[57] Hillmann, D., Using Dublin Core, Dublin Core Metadata Initiative, 2004. http://dublincore.org/documents/usageguide [cited at p. 14]

[58] Technical Standardization Committee on AV IT Storage Systems and Equipment. Exchangeable image file format for digital still cameras: Exif version 2.2. Published by Standard of Japan Electronics and Information Technology Industries Association, 2002. http://www.exif.org/Exif2-2.pdf [cited at p. 15]

[59] IPTC, Information Technology for news Standard Photo Metadata 2008 IPTC Core Specification 1.1 and IPTC Extension Specification 1.0, 2008. http://www.iptc.org/std/photo-metadata/2008/specification/IPTC-Photo-Metadata-2008.pdf [cited at p. 15]

[60] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Extensible Markup Language (XML) 1.0, W3C Recommendation, 1998. http://www.w3.org/TR/1998/REC-xml-19980210 [cited at p. 17]

[61] Vander Wal, T., Folksonomy Coinage and Definition, 2007. http://www.vanderwal.net/folksonomy.html [cited at p. 17]

[62] Lassila, O., Swick, R., Resource Description Framework (RDF): Model and Syntax Specification. Recommendation W3C, 1999. http://www.w3.org/TR/REC-rdf-syntax [cited at p. 19]

[63] Carroll, J. J., De Roo, J., OWL Web Ontology Language Test Cases. World Wide Web Consortium (W3C) Recommendation, 2004. http://www.w3.org/TR/2004/REC-owl-test-20040210 [cited at p. 19]

[64] McGuinness, D.L., Van Harmelen F., OWL Web Ontology Language Overview. World Wide Web Consortium (W3C) Recommendation, 2004. http://www.w3.org/TR/owl-features [cited at p. 19]

[65] Heflin, J., OWL Web Ontology Language Use Cases and Requirements. World Wide Web Consortium (W3C) Recommentation, 2004. http://www.w3.org/TR/webont-req [cited at p. 19]

[66] RFC 4287, The Atom Syndication Format, 2005. http://www.ietf.org/rfc/rfc4287 [cited at p. 22]

[67] “Castello di Arco” picture. Retrieved from: http://www.flickr.com/photos/cristina63/3830632607 [cited at p. 22]

[68] De Mauro, T.: GRADIT, The GRAnde Dizionario ITaliano dell’Uso, UTET, 1999. [cited at p. 28]

[69] Llisterri, J., Text Corpora Working Group Reading Guide, EAGLES (Expert Advisory Group on language Engineering Standards) Document EAG-TCWG-FR-2, CNR, Istituto di Linguistica computazionale, 1996. [cited at p. 28, 33]

[70] Hillman, D. I., Westbrooks, E. L., Metadata in practice, American Library Association, 2004. [cited at p. 30, 42]

[71] Chopey, M. A., Planning and Implementing a Metadata-Driven Digital Repository, Haworth Press Inc., 2005. [cited at p. 30]

[72] Dunsire, G., Collecting metadata from institutional repositories. In: OCLC Systems Services, Vol. 24, No. 1, pp. 51-58, 2008. [cited at p. 30]

[73] Solodovnik, I., Metadata issues in Digital Libraries: key concepts and perspectives. In: Italian Journal of Library and Information Science, Vol. 2, No. 2, 2011. [cited at p. 30]

[74] Heery, R., Patel, M., Application profiles: mixing and matching metadata schemas, Ariadne, 2000. [cited at p. 30]

[75] Lagoze, C., Van de Sompel, H., The making of the Open Archives Initiative protocol for metadata harvesting, Library Hi Tech, 2003. [cited at p. 30]

[76] Hutt, A., Riley, J., Semantics and Syntax of Dublin Core Usage in Open Archives Initiative Data. In: Joint Conference on Digital Libraries, ACM Press, 2005. [cited at p. 30]

[77] The Budapest Open Access Initiative, Manifesto of the The Budapest Open Access Initiative, 2002. http://www.opensocietyfoundations.org/openaccess/read [cited at p. 31]

[78] Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., Smith, M.: The DSpace Institutional Digital Repository System: Current Functionality. In: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries (2003). [cited at p. 32]

[79] Jackson, A. S., Han, M. J., Groetsch, K., Mustafoff, M.: Dublin Core Metadata Harvested Through OAI-PMH. In: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries (2008). [cited at p. 32]

[80] Lutzu, M., Forme e struttura della gara poetica a sa repentina. In: AA. VV., Su cantu de sei in Sardinnia, Alfa Editrice, pp. 7-25, 2007. [cited at p. 34]

[81] Bravi, P., La musica de s'arrepentina. In: P. Zedda, Enciclopedia della musica sarda, Vol. 14: Poesia improvvisata, pp. 160-163, Unione Sarda, 2012. [cited at p. 34]

[82] Fransson, F., Sundberg, J., Tjernlund, P., Grundfrequenzmessungen an schwedischen Kernspaltflöten. In: Studia Instrumentorum Musicae Popularis II (ed. E. Stockmann), Stockholm, pp. 77-96, 1972. [cited at p. 36]

[83] Wiig, K., Introducing Knowledge Management into the Enterprise. In Knowledge Management Handbook, J. Liebowitz (Ed.), CRC Press LCL, Boca Raton (FL), 1999. [cited at p. 45]

[84] Drucker, P. F., Il grande cambiamento, Sperling Kupfer, 1996. [cited at p. 45]

[85] Alavi, M., Leidner, D. E., Knowledge management systems: issues, challenges, and benefits. In Journal Communications of the AIS, Vol. 1, Issue 2es, Article No. 1, 1999. [cited at p. 45]

[86] Firestone, J. M., Key issues in knowledge management. In Knowledge and innovation, 1(3), pp. 8-17, 2001. [cited at p. 45]

[87] Stankosky, M., Creating the Discipline of Knowledge Management, Elsevier, Burlington, MA, 2005. [cited at p. 45]

[88] Egbu, C., Suresh, S., Knowledge Mapping Techniques Within The Construction Industry: An Exploratory Study, Joint CIB Conference W102 Information and Knowledge Management in Building W096 Architectural Management, 2008. ISBN: 978-951-758-492-0. [cited at p. 46]

[89] Alsakini, W., Kiiras, J., Huovinen P., An integrated information system of a virtual construction management service company, Joint CIB Conference W102 Information and Knowledge Management in Building W096 Architectural Management, 2008. ISBN: 978-951-758-492-0. [cited at p. 46]

[90] Loforte Ribeiro, F., Knowledge management in construction sites: a comparative case study, Joint CIB Conference W102 Information and Knowledge Management in Building W096 Architectural Management, 2008. ISBN: 978-951-758-492-0. [cited at p. 46]

[91] Argiolas, C., Quaquero, E., Lâelaborazione dellâofferta economicamente piú vantaggiosa. In Argiolas, C., L'edilizia disegnata dal Decreto Legislativo n.163 del 2006: Analisi dei punti critici per gestire la complessitÃ , Lithos Grafiche Editore, Cagliari, pp. 119-147, 2008. ISBN: 978-88-95398-03-7. [cited at p. 46]

[92] Argiolas, C., Raccontando il dettaglio: ricerca e sperimentazione nel processo di produzione e management edilizio, Lithos Grafiche Editore, Cagliari, 2008. ISBN: 978-88-95398-04-4. [cited at p. 46]

[93] Beckman, T., The current state of Knowledge Management, Knowledge Management Handbook, J. Liebowitz (Editor). CRC Press LCL, Boca Raton (FL), 1999. [cited at p. 46]

[94] Argiolas, C., Melis, F., Quaquero, E., Knowledge Management as a safety management strategy in building sites, Joint CIB Conference W102 Information and Knowledge Management in Building W096 Architectural Management, 2008. ISBN: 978-951-758-492-0. [cited at p. 46]

[95] Rigamonti, G., La gestione dei processi di intervento edilizio, UTET, 2001. [cited at p. 46]

[96] Bove, A., Project Management: la metodologia dei 12 step, Hoepli, 2008. [cited at p. -]

[97] Tronconi, O., Tecnologie informatiche e imprese di costruzioni, Il Sole 24 Ore, 2005. [cited at p. 46]

[98] Civi E., Knowledge management as a competitive asset: a review, Marketing Intelligence Planning, Vol. 18, Issue 4, pp.166-174, 2000. [cited at p. 47]

[99] McKeen, J. D., Zack, M. H., Knowledge Management and Organizational Performance: An Exploratory Survey. In Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS '06), Vol. 7, 2006. [cited at p. 47]

[100] Nepi, A., Introduzione al project management. Che cos'Ã¨, come si applica, tecniche e metodologie, Edizioni Angelo Guerini e Associati, 1997. [cited at p. 47]

[101] Argiolas, C., Melis, F., Quaquero E. (2011). Knowledge management in building process, Joint CIB Conference W078-W102 Computer, Knowledge, Building, Sophia Antipolis, France, ISBN: 978-90-6363-068-3. [cited at p. 50]

[102] Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M., Horrocks, I., The Semantic Web: the roles of XML and RDF. In: Internet Computing, IEEE, Vol. 4, Issue 5, pp. 63-73, Sep/Oct 2000. [cited at p. 57]

[103] Maedche, A., Staab, S., Ontology learning for the Semantic Web. In: Intelligent Systems, IEEE, Vol. 16, Issue 2, pp. 72 - 79, Mar-Apr 2001. [cited at p. 57]

[104] Jacob, E. K., Ontologies and the Semantic Web. In: Bulletin of the American Society for Information Science and Technology, Wiley Periodicals, Vol. 29, Issue 4, pp. 19-22, April/May 2003. [cited at p. 57]

[105] Davies, J., Fensel, D., van Harmelen, F., Towards the Semantic Web: Ontology-driven Knowledge Management, John Wiley Sons, 2003, ISBN: 9780470848678 [cited at p. 57]

[106] Simperi, E., Reusing ontologies on the Semantic Web: A feasibility study. In: Data Knowledge Engineering, Vol. 68, Issue 10, pp. 905-925, October 2009. [cited at p. 57]

[107] Euzenat, J., Shvaiko, P., Ontology matching, Springer-Verlag, Berlin Heidelberg (DE), 2007, ISBN: 978-3-540-49612-0 [cited at p. 58]

[108] Fensel, D., Van Harmelen, F., Horrocks, I., McGuinness, D. L.; Patel-Schneider, P. F., OIL: an ontology infrastructure for the Semantic Web. In: Intelligent Systems, IEEE, Vol. 16, Issue 2, pp. 38-45, Mar-Apr 2001. [cited at p. 58]

[109] Gómez-Pérez, A., Corcho, O., Ontology languages for the Semantic Web. In: Intelligent Systems, Vol. 17, Issue 1, pp. 54-60, Jan-Feb 2002. [cited at p. 58]

[110] W3C, Wine ontology, http://www.w3.org/TR/owl-guide/wine.rdf [cited at p. 59]

List of Publications Related to the Thesis

Published papers

- Pani, F. E., Lunesu, M. I., Concas, G., Stara, C., Tilocca, M. P., Optimization of Knowledge Availability in an Institutional Repository. In: Proceedings of the 4th International Conference on Knowledge Engineering and Ontology Development, KEOD 2012, Barcelona, Spain, 2012. ISBN: 978-989-8565-30-3.

- Pani, F. E., Lunesu, M. I., Concas, G., Stara, C., Tilocca, M. P., Knowledge Formalization and Management in KMS. In: Proceedings of the 4th International Conference on Knowledge Management and Information Sharing, KMIS 2012, Barcelona, Spain, 2012. ISBN: 978-989-8565-31-0.

- Concas, G., Pani, F. E., Lunesu, M. I., Knowledge Management using Open Source Repository. In: Proceedings of the 6th European Computing Conference (ECC '12), Prague, Czech Republic, 2012. ISBN: 978-1-61804-129-6.

- Lunesu, M. I., Pani, F. E., Concas, G., An Approach to manage semantic informations from UGC. In: Proceedings of the 3rd International Conference on Knowledge Engineering and Ontology Development (KEOD 2011), Paris, France, 2011. ISBN: 978-989-8425-80-5.

- Lunesu, M. I., Pani, F. E., Concas, G., Using a standards-based approach for a multimedia knowledge-base. In: Proceedings of the 3rd International Conference on Knowledge Management and Information Sharing (KMIS 2011), Paris, France, 2011. ISBN: 978-989-8425-81-2.

Book Chapters (in Press)

- Concas, G., Pani, F. E., Lunesu, M. I., Mannaro, K., Using an Ontology for Multimedia Content Semantics. In: Lai, C., Giuliani, A., Semeraro, G. (Eds.), New Challenges in Distributed Information Filtering and Retrieval, Studies in Computational Intelligence.

- Concas, G., Pani, F. E., Lunesu, M. I., Puddu, N., Macchiarella, I., Bravi, P., A New Approach for Knowledge Management on Linguistic Information. In: Lai, C., Giuliani, A., Semeraro, G. (Eds.), New Challenges in Distributed Information Filtering and Retrieval, Studies in Computational Intelligence.

Journal Article (in Press)

- Concas, G., Pani F. E., Lunesu M. I., A New Approach for Knowledge Management and Optimization using an Open Source Repository, Transactions on Information Science and Applications.

Currently submitted papers

- Argiolas, C., Concas, G., Di Francesco, M., Lunesu, M. I., Melis, F., Pani, F. E., Quaquero, E., Sanna, D., Knowledge in Construction Processes, International Conference on Knowledge Management and Information Sharing, KMIS 2013.

- Pani, F. E., Lunesu, M. I., Concas, G., Baralla, G., An Approach to Manage the Web Knowledge, International Conference on Knowledge Engineering and Ontology Development, KEOD 2013.